On the Global Convergence of Majorization Minimization Algorithms for Nonconvex Optimization Problems

Yangyang Kang
Department of Computer Science and Engineering
Shanghai Jiao Tong University
kcany27@gmail.com

Zhihua Zhang
Department of Computer Science and Engineering
Shanghai Jiao Tong University
zhang-zh@cs.sjtu.edu.cn

Wu-Jun Li
National Key Laboratory for Novel Software Technology
Department of Computer Science and Technology
Nanjing University
liwujun@nju.edu.cn

Abstract

In this paper, we study the global convergence of majorization minimization (MM) algorithms for solving nonconvex regularized optimization problems. MM algorithms have received great attention in machine learning. However, when applied to nonconvex optimization problems, the convergence of MM algorithms is a challenging issue. We introduce theory of the Kurdyka-Łojasiewicz inequality to address this issue. In particular, we show that many nonconvex problems enjoy the Kurdyka-Łojasiewicz property and establish the global convergence result of the corresponding MM procedure. We also extend our result to a well known method that called CCCP (concave-convex procedure).

Keywords: nonconvex optimization, majorization minimization, Kurdyka-Łojasiewicz inequality, global convergence

1. Introduction

Majorization minimization (MM) algorithms have wide applications in machine learning and statistical inference (Lange et al., 2000, Lange, 2004). The MM algorithm can be regarded as a generalization of expectation-maximization (EM) algorithms, and it aims to turn an otherwise hard or complicated optimization problem into a tractable one by alternatively iterating an *Majorization* step and an *Minimization* step.

More specifically, the majorization step constructs a tractable surrogate function to substitute the original objective function and the minimization step minimizes this surrogate function to obtain a new estimate of parameters in question. In the conventional MM al-

gorithm, convexity plays a key role in the construction of surrogate functions. Moveover, convexity arguments make the conventional MM algorithm have the same convergence properties as EM algorithms (Lange, 2004).

Alternatively, we are interested in use of MM algorithms in solving nonconvex (nonsmooth) optimization problems. For example, nonconvex penalization has been demonstrated to have attractive properties in sparse estimation. In particular, there exist many nonconvex penalties, including the ℓ_q ($q \in (0,1)$) penalty, the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), the minimax concave plus penalty (MCP) (Zhang, 2010a), the capped-ℓ_1 function (Zhang, 2010b, Zhang et al., 2012, Gong et al., 2013), the LOG penalty (Mazumder et al., 2011, Armagan et al., 2013), etc. However, they might yield computational challenges due to nondifferentiability and nonconvexity that they have. An MM algorithm would be a desirable choice (Lange, 2004).

In this paper we would like to address the global convergence property of MM algorithms for nonconvex optimization problems. Our motivation comes from the novel Kurdyka-Łojasiewicz inequality. In the pioneer work (Lojasiewicz, 1963, Łojasiewicz, 1993), the author provided the "Łojasiewicz inequality" to derive finite trajectories. Later on, Kurdyka (1998) extended the Łojasiewicz inequality to definable functions and applications. Bolte et al. (2007) then extended to nonsmooth subanalytic functions. Recently, the Kurdyka-Łojasiewicz property has been used to establish convergence analysis of proximal alternating minimization or coordinate descent algorithms (Attouch et al., 2010, Xu and Yin, 2013, Bolte et al., 2013).

We revisit a generic MM procedure of solving nonconvex optimization problems. We observe that many nonconvex penalty functions satisfy the Kurdyka-Łojasiewicz inequality and such a property is shared by a number of machine learning problems arising in a wide variety of applications. Specifically, we demonstrate several examples, which admit the Kurdyka-Łojasiewicz property. Thus, we conduct the convergence analysis of the MM procedure based on theory of the Kurdyka-Łojasiewicz inequality. More specifically, our work offers the following major contributions.

- We discuss a family of nonconvex optimization problems in which the objective function consists of a smooth function and a non-smooth function. We give the constructive criteria of surrogates that approximate the original functions well. Additionally, we also illustrate that many existing methods for solving the nonconvex optimization problem can be regarded as an MM procedure.

- We establish the global convergence results of a generic MM framework for the nonconvex problem which are obtained by exploiting the geometrical property of the objective function around its critical point. To the best of our knowledge, our work is the first study to address the convergence property of MM algorithms for nonconvex optimization using the Kurdyka-Łojasiewicz inequality.

- We also show that our global convergence results can be successfully extended to many popular and powerful methods such as iteratively re-weighted ℓ_1 minimization method Candes et al. (2008), Chartrand and Yin (2008), local linear approximation (LLA) Zou and Li (2008), Zhang (2010b), concave-convex procedure (CCCP) Yuille and Rangarajan (2003), Lanckriet and Sriperumbudur (2009), etc.

1.1 Related Work and Organization

We discuss some related work about the convergence analysis of nonconvex optimization. Vaida (2005) established the global convergence of EM algorithms and extended it to the global convergence of MM algorithms under some conditions. However, they considered the differentiable objective function, whereas the objective function in our paper can be nonsmooth (also nonconvex). This implies that the problem we are considering is more challenging. Additionally, Vaida (2005) assumed that all the stationary points of objective function are isolated. In our paper, we don't require this assumption. The isolation assumption does not always hold, or holds but is difficult to verify, for many objective functions in practice. This motivates us to employ the Kurdyka-Lojasiewicz inequality to establish the convergence. Moreover, it is usually easily verified that the objective function admits the Kurdyka-Lojasiewicz inequality. Gong et al. (2013) proposed an efficient iterative shrinkage and thresholding algorithm to solve nonconvex regularized problems. The key assumption is that the computation of proximal operator of the regularizer has a closed form. We note that this method falls into our MM framework. However, the authors only showed that the subsequence converges to a critical point. Mairal (2013) studied instead asymptotic stationary point conditions with first-order surrogate functions, but he did not propose the convergent sequence which converges to the solution point.

Attouch et al. (2010), Xu and Yin (2013), Bolte et al. (2013) employed the Kurdyka-Łojasiewicz inequality to analyze the convergence of nonconvex optimization problems. They are mainly concerned with the convergence analysis of the block coordinate approaches. In this paper, we pay attention to the global convergence analysis of the MM framework for solving nonconvex regularization problems. Specifically, we construct surrogates both on the smooth and nonsmooth terms. To achieve the global convergence, we exploit the geometry property of the objective function around its critical point.

The remainder of the paper is organized as follows. Section 2 provides preliminaries about the nonsmooth and nonconvex analysis and introduces the Kurdyka-Łojasiewicz property. We also give some examples which enjoy the Kurdyka-Łojasiewicz inequality. In Section 3, we formulate the problem we are interested in and make some common assumptions. A generic majorization minimization algorithm is revisited in Section 4. Section 5 is the key part of our paper which gives the global convergence results. In Section 6 we extend our work to CCCP. In Section 7 we conduct numerical examples to verify our theoretical results. Finally, we conclude our work in Section 8.

2. Preliminaries

In this section we introduce the notion of Fréchet's subdifferential and a limiting-subdifferential. Then we present the novel Kurdyka-Łojasiewicz inequality. First of all, for any $\mathbf{u} = (u_1, \ldots, u_p)^T \in \mathbb{R}^p$ and $\mathbf{v} = (v_1, \ldots, v_p)^T \in \mathbb{R}^p$, we denote $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^p u_i v_i$ and $\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ here and later.

Definition 1 (Subdifferentials) *(Rockafellar et al., 1998) Consider a proper and lower semi-continuous function $f : \mathbb{R}^p \to (-\infty, +\infty]$ and a point $\mathbf{x} \in \mathrm{dom}(f)$.*

(i) *The Fréchet subdifferential of* f *at* $\mathbf{x}$*, denoted* $\hat{\partial} f(\mathbf{x})$*, is the set of all vectors* $\mathbf{u} \in \mathbb{R}^p$ *which satisfy*

$$\liminf_{\substack{\mathbf{y} \neq \mathbf{x} \\ \mathbf{y} \to \mathbf{x}}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \mathbf{u}^T(\mathbf{y} - \mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} \geq 0.$$

(ii) *The limiting-subdifferential of* f *at* $\mathbf{x}$*, denoted* $\partial f(\mathbf{x})$*, is defined as*

$$\partial f(\mathbf{x}) \equiv \left\{ \mathbf{u} \in \mathbb{R}^p : \exists \mathbf{x}_k \to \mathbf{x}, f(\mathbf{x}_k) \to f(\mathbf{x}) \ \ and \ \ \mathbf{u}_k \in \hat{\partial} f(\mathbf{x}_k) \to \mathbf{u} \ \ as \ \ k \to \infty \right\}.$$

Remark 2 *Here* $\mathrm{dom} f \triangleq \{\mathbf{x} : f(\mathbf{x}) < +\infty\}$*. If* $\mathbf{x} \notin \mathrm{dom} f$*, one sets* $\hat{\partial} f(\mathbf{x}) = \varnothing$*. It is worth pointing out that* $\hat{\partial} f(\mathbf{x})$ *for each* $\mathbf{x}$ *is closed and convex while* $\partial f(\mathbf{x})$ *is closed. If* f *is differentiable at* $\mathbf{x}_0$*, then* $\hat{\partial} f(\mathbf{x}_0) = \{\nabla f(\mathbf{x}_0)\}$ *and* $\nabla f(\mathbf{x}_0) \in \partial f(\mathbf{x}_0)$*. More details are referred to Rockafellar et al. (1998). As we see, both the Fréchet subdifferential and limiting-subdifferential are applicable for nonconvex functions.*

Corollary 3 (Rockafellar and Wets, 1998) *Suppose* $F = f + r : \mathbb{R}^p \to \mathbb{R}$*. Moreover,* f *is smooth in the neighborhood of* $\mathbf{x}_0$ *and* r *is finite at* $\mathbf{x}_0$*. Then, we have*

$$\hat{\partial} F(\mathbf{x}_0) = \nabla f(\mathbf{x}_0) + \hat{\partial} r(\mathbf{x}_0) \ \ and \ \ \partial F(\mathbf{x}_0) = \nabla f(\mathbf{x}_0) + \partial r(\mathbf{x}_0).$$

Definition 4 *It is said that* $\mathbf{x}^* \in \mathbb{R}^p$ *is a critical point of a lower semi-continuous function* $F : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$*, if the following condition holds*

$$\mathbf{0} \in \partial F(\mathbf{x}^*).$$

Remark 5 *If* $\mathbf{x}^*$ *is a minimizer (not necessarily global) of function* F*, we can conclude that* $\mathbf{0} \in \partial F(\mathbf{x}^*)$*. The set of critical points of* F *is denoted by* $\mathrm{crit} F$*.*

2.1 Kurdyka-Łojasiewicz properties

With the notion of subdifferentials, we now briefly recall the Kurdyka-Łojasiewicz inequality, which plays a central role in our globally convergence analysis.

Definition 6 *Let the function* $F : \mathbb{R}^p \to (-\infty, +\infty]$ *be proper and lower semi-continuous. Then* F *is said to have the Kurdyka-Łojasiewicz property at* $\bar{\mathbf{u}} \in \mathrm{dom}\, \partial F$ *if there exist* $\eta \in (0, +\infty]$*, a neighborhood* $\mathcal{U}$ *of* $\bar{\mathbf{u}}$*, and a continuous concave function* $\phi : [0, \eta) \to \mathbb{R}_+$ *with the following properties:*

(a) $\phi(0) = 0$,

(b) ϕ *is* C^1 *on* $(0, \eta)$,

(c) *for all* $t \in (0, \eta)$, $\phi'(t) > 0$,

such that for all $\mathbf{u}$ *in* $\mathcal{U} \cap [F(\bar{\mathbf{u}}) < F(\mathbf{u}) < F(\bar{\mathbf{u}}) + \eta]$*, the following* **Kurdyka-Łojasiewicz inequality** *holds true:*

$$\phi'(F(\mathbf{u}) - F(\bar{\mathbf{u}}))\mathrm{dist}(\mathbf{0}, \partial F(\mathbf{u})) \geq 1.$$

Here $\mathrm{dist}(\mathbf{u}, \mathcal{A}) = \inf_{\mathbf{v}} \{\|\mathbf{u} - \mathbf{v}\|, \mathbf{v} \in \mathcal{A}\}$.

It is well established that real analytic and sub-analytic functions satisfy the Kurdyka-Łojasiewicz property (Bolte et al., 2007). Moreover, the sum of a real analytic function and a subanalytic function is subanalytic (Bochnak et al., 1998). Thus, the sum admits the Kurdyka-Łojasiewicz property. Many functions involved in machine learning satisfy the Kurdyka-Łojasiewicz property. For example, both the logistic loss and the least squares loss are real analytic.

We also find that many nonconvex penalty functions, such as MCP, LOG, SCAD, and Capped ℓ_1, enjoy the Kurdyka-Łojasiewicz property. Here we give two examples. First, the MCP function is defined as

$$\zeta(t;\lambda,\gamma) = \begin{cases} \lambda(|t| - \frac{t^2}{2\lambda\gamma}) & \text{if } |t| < \lambda\gamma, \\ \frac{\lambda^2\gamma}{2} & \text{if } |t| \geq \lambda\gamma, \end{cases}$$

where $\lambda, \gamma > 0$ are constants. The graph of ζ is the closure of the following set

$$\left\{(t,s) : s = \frac{\lambda^2\gamma}{2}, t < -\lambda\gamma\right\} \cup \left\{(t,s) : s = \frac{\lambda^2\gamma}{2}, t > \lambda\gamma\right\}$$
$$\cup \left\{(t,s) : s = -\lambda t - \frac{t^2}{2\gamma}, -\lambda\gamma < t < 0\right\} \cup \left\{(t,s) : s = \lambda t - \frac{t^2}{2\gamma}, 0 < t < \lambda\gamma\right\}.$$

This implies that MCP is a semi-algebraic function (Bochnak et al., 1998), which is sub-analytic (Bolte et al., 2007). Thus, MCP satisfies the Kurdyka-Łojasiewicz property. Similarly, we can obtain that the SCAD and capped ℓ_1 penalties have the Kurdyka-Łojasiewicz property.

Second, the LOG penalty is defined as

$$\zeta(t;\lambda,\alpha) = \lambda \log(1 + \alpha|t|), \text{ for } \alpha > 0.$$

The graph of ζ is the closure of the following set

$$\left\{(t,s) : s = \lambda \log(1+\alpha t), t > 0\right\} \cup \left\{(t,s) : s = \lambda \log(1-\alpha t), t < 0\right\}.$$

Note that the graph is sub-analytic (Bolte et al., 2007), so the LOG penalty is sub-analytic, which enjoys the Kurdyka-Łojasiewicz property.

3. Problem and Assumptions

In this paper we are mainly concerned with the following optimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^p} \Big\{F(\mathbf{w}) = f(\mathbf{w}) + r(\mathbf{w})\Big\}. \tag{1}$$

Many machine learning problems can be cast into this formulation. Typically, $f(\mathbf{w})$ is defined as a loss function and $r(\mathbf{w})$ is defined as a regularization (or penalization) term. Specifically, given a training dataset $\mathcal{D} = \big\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\big\}$, one defines $f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{w}; \mathbf{x}_i, y_i)$. A very common setting for the penalty function $r(\mathbf{w})$ is $\sum_{i=1}^p r_i(w_i;\lambda)$, where λ is the tuning parameter controlling the trade-off between the loss function and the regularization.

Recently, many nonconvex penalty functions, such as LOG (Mazumder et al., 2011, Armagan et al., 2013), SCAD (Fan and Li, 2001), MCP (Zhang, 2010a), and the capped-ℓ_1 function (Zhang, 2010b), have been proposed to model sparsity. These penalty functions have been demonstrated to have attractive properties theoretically and practically.

Meanwhile, iteratively reweighted methods haven widely used to solve the optimization problem in (1). Usually, the iteratively reweighted method enjoys a majorization minimization (MM) procedure. In this paper we attempt to conduct convergence analysis of the MM procedure. For our purpose, we make some assumptions about the objective function.

Assumption 7 *Suppose $f : \mathbb{R}^p \to \mathbb{R}_+$ is a smooth function of the type $C^{1,1}$. Moreover, the gradient of f is L_f-Lipschitz continuous; that is,*

$$\|\nabla f(\mathbf{u}) - \nabla f(\mathbf{v})\| \leq L_f \|\mathbf{u} - \mathbf{v}\| \tag{2}$$

for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, where $L_f > 0$ is called a Lipschitz constant of ∇f.

Corollary 8 *Let $h(\mathbf{w}) = \sum_{i=1}^n \alpha_i f_i(\mathbf{w})$. Suppose $f_i(\mathbf{w})$ is differentiable for any $i \in [n] \triangleq \{1, 2, \cdots, n\}$. If each $\nabla f_i(\mathbf{w})$ is L_i-Lipschitz continuous ($L_i > 0$), then $h(\mathbf{w})$ is differentiable and $\nabla h(\mathbf{w})$ is $\sum_{i=1}^n |\alpha_i| L_i$-Lipschitz continuous.*

Lemma 9 *If $f : \mathbb{R}^p \to \mathbb{R}$ is differentiable and ∇f is L_f-Lipschitz continuous. Then*

$$f(\mathbf{u}) \leq f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \frac{L_f}{2} \|\mathbf{u} - \mathbf{v}\|^2 \tag{3}$$

for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$.

This is a classical result whose proof can be seen from Nesterov and Nesterov (2004).

Assumption 10 *$F : \mathbb{R}^p \to \mathbb{R}$ is lower semi-continuous and coercive [1], and it satisfies $\inf_{\mathbf{w} \in \mathbb{R}^p} F(\mathbf{w}) > -\infty$.*

We give several examples to show that the assumptions hold in many machine learning problems. For the linear regression, $f(\mathbf{w}) = \frac{1}{2n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, where $\mathbf{w} \in \mathbb{R}^p, \mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$ is the input matrix and $\mathbf{y} = [y_1, \cdots, y_n]^T \in \mathbb{R}^n$ is the output vector. In this example, the Lipschitz constant of $\nabla f(\mathbf{w})$ is lower-bounded by the maximum eigenvalue of $\frac{1}{n}\mathbf{X}^T\mathbf{X}$. In binary classification problems in which $y_i \in \{-1, 1\}$, we consider the logistic regression loss function. Specifically, $f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{x}_i^T \mathbf{w}))$. The Lipschitz constant of $\nabla f(\mathbf{w})$ is lower-bounded by $\frac{1}{4n}\sum_{i=1}^n \mathbf{x}_i^T \mathbf{x}_i$.

4. Majorization Minimization Algorithms

We consider a minimization problem with the objective function $F(\mathbf{w})$. Given an estimate $\mathbf{w}^{(k)}$ at the kth iteration, a typical MM algorithm consists of the following two steps:

1. A function $g(\mathbf{u})$ on $\mathbb{R}^p$ is said to be coercive if $\lim_{\|\mathbf{u}\| \to \infty} g(\mathbf{u}) = \infty$

Table 1: Examples of nonconvex penalties for one dimension

FUNCTION	$\zeta(t)$
LOG	$\frac{\lambda}{\log(\theta+1)}\log(1+\theta\|t\|),\ (\theta>0)$
SCAD	$\begin{cases}\lambda\|t\| & \text{IF } \|t\|\le\lambda,\\ -\frac{\|t\|^2-2\theta\lambda\|t\|+\lambda^2}{2(\theta-1)} & \text{IF } \lambda<\|t\|\le\theta\lambda,\quad(\theta>2)\\ \frac{(\theta+1)\lambda^2}{2} & \text{IF } \theta\lambda<\|t\|,\end{cases}$
MCP	$\begin{cases}\lambda(\|t\|-\frac{t^2}{2\lambda\gamma}) & \text{IF } \|t\|<\lambda\gamma,\\ \frac{\lambda^2\gamma}{2} & \text{IF } \|t\|\ge\lambda\gamma.\end{cases}$
CAPPED ℓ_1-PENALTY	$\lambda\min(\|t\|,\theta),\ (\theta>0)$

Majorization Step: Substitute $F(\mathbf{w})$ by a tractable surrogate function $Q(\mathbf{w}|\mathbf{w}^{(k)})$, such that

$$Q(\mathbf{w}|\mathbf{w}^{(k)}) \geq F(\mathbf{w})$$

for any $\mathbf{w} \in \mathrm{dom}F$, with equality holding at $\mathbf{w} = \mathbf{w}^{(k)}$.

Minimization Step: Obtain the next parameter estimate $\mathbf{w}^{(k+1)}$ by minimizing $Q(\mathbf{w}|\mathbf{w}^k)$ with respect to $\mathbf{w}$. That is,

$$\mathbf{w}^{(k+1)} = \operatorname*{argmin}_{\mathbf{w}}\ Q(\mathbf{w}|\mathbf{w}^{(k)}).$$

In order to address the global convergence of MM for solving the problem in (1), we propose a generic MM framework under the assumptions given in the previous section. We particularly present criteria to devise the majorant functions of the loss function and penalty function, respectively.

4.1 Majorization of Loss Function

We first consider the majorization of the loss function $f(\mathbf{w})$. Recall that $\nabla f(\mathbf{w})$ is assumed to be Lipschitz continuous (Assumption 7). Given the estimate $\mathbf{w}^{(k)}$ of $\mathbf{w}$ at the kth iteration, one would derive the majorization of $f(\mathbf{w})$ to obtain $\mathbf{w}^{(k+1)}$. For the sake of simplicity, we denote the corresponding surrogate function as $Q_f(\mathbf{w}|\mathbf{w}^{(k)})$. In our work, we claim that $Q_f(\mathbf{w}|\mathbf{w}^{(k)})$ should have the following two properties so that the surrogate can approximate the objective f well and also lead to efficient computations.

Assumption 11 *Let $Q_f(\mathbf{w}|\mathbf{w}^{(k)})$ be the majorization of $f(\mathbf{w})$ such that $Q_f(\mathbf{w}|\mathbf{w}^{(k)}) \geq f(\mathbf{w})$ and $Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) = f(\mathbf{w}^{(k)})$. Additionally, the following properties also hold:*

(i) *$Q_f(\mathbf{w}|\mathbf{w}^{(k)}) - f(\mathbf{w})$ is γ-strongly convex, where $\gamma > 0$;*

(ii) *$\nabla Q_f(\mathbf{w}|\mathbf{w}^{(k)})$ is Lipschitz continuous, and $\nabla Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) = \nabla f(\mathbf{w}^{(k)})$.*

Let us see several extant popular algorithms which meet Assumption 11. Proximal algorithms (Rockafellar, 1976, Lemaire, 1989, Iusem, 1999, Combettes and Pesquet, 2011,

Parikh and Boyd, 2013) solve optimization problems by using a so-called proximal operator of the objective function. Suppose we have an objective function $f(\mathbf{w})$ at hand. Given the kth estimate $\mathbf{w}^{(k)}$, the proximal algorithm aims to solve the following problem

$$\mathbf{w}^{(k+1)} = \operatorname*{argmin}_{\mathbf{w}} \left\{ f(\mathbf{w}) + \frac{1}{2\alpha_k} \|\mathbf{w} - \mathbf{w}^{(k)}\|^2 \right\},$$

where α_k is the step size at each iteration. Typically, $\mathbf{w}^{(k+1)}$ is written as:

$$\mathbf{w}^{(k+1)} \triangleq \mathrm{Prox}_{\alpha_k f}(\mathbf{w}^{(k)}).$$

The majorization function is defined as $Q_f(\mathbf{w}|\mathbf{w}^{(k)}) \triangleq f(\mathbf{w}) + \frac{1}{2\alpha_k}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2$. When $f(\mathbf{w})$ is convex and $\nabla f(\mathbf{w})$ is Lipschitz continuous, it is easy to check that $Q_f(\mathbf{w}|\mathbf{w}^{(k)})$ satisfies Assumption 11.

Another powerful algorithm is the proximal gradient algorithm. The algorithm is more efficient when dealing with the following problem

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \left\{ f(\mathbf{w}) + r(\mathbf{w}) \right\}, \tag{4}$$

where f is differentiable and convex and r is nonsmooth. The proximal gradient algorithm first approximates $f(\mathbf{w})$ based on a local linear expansion plus a proximal term, both at the current estimate $\mathbf{w}^{(k)}$. That is,

$$f(\mathbf{w}) \approx f(\mathbf{w}^{(k)}) + \langle \nabla f(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle + \frac{1}{2\alpha_k}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2,$$

where $\alpha_k > 0$ is the step size. Then the $(k+1)$th estimate of $\mathbf{w}$ is given as

$$\begin{aligned} \mathbf{w}^{(k+1)} = \operatorname*{argmin}_{\mathbf{w}} \Big\{ & f(\mathbf{w}^{(k)}) + \langle \nabla f(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle \\ & + \frac{1}{2\alpha_k}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2 + r(\mathbf{w}) \Big\}. \end{aligned} \tag{5}$$

Equivalently,

$$\mathbf{w}^{(k+1)} = \mathrm{Prox}_{\alpha_k r}(\mathbf{w}^{(k)} - \alpha_k \nabla f(\mathbf{w}^{(k)})).$$

Intuitively, the proximal gradient algorithm would take the gradient descent step first and then does the proximal minimization step. In this algorithm, $Q_f(\mathbf{w}|\mathbf{w}^{(k)}) \triangleq f(\mathbf{w}^{(k)}) + \langle \nabla f(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle + \frac{1}{2\alpha_k}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2$. It is also immediately verified that $Q_f(\mathbf{w}|\mathbf{w}^{(k)})$ satisfies Assumption 11 when $\alpha_k < L_f^{-1}$, where L_f is the Lipschitz constant of $\nabla f(\mathbf{w})$.

In fact, Lemma 9 implies that there always exists a quadratic surrogate of f only if (2) holds. In particular, we can define Q_f as

$$Q_f(\mathbf{w}|\mathbf{w}^{(k)}) = f(\mathbf{w}^{(k)}) + \langle \nabla f(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)} \rangle + \frac{\mu^{(k)}}{2}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2, \tag{6}$$

where we require that $\mu^{(k)} > L_f$.

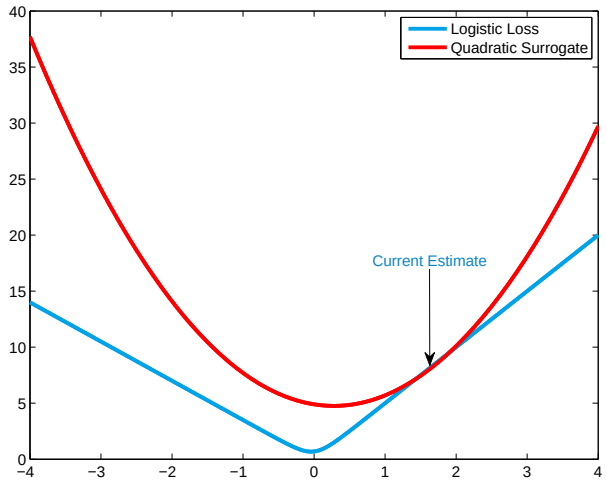


Figure 1: Surrogate for logistic loss

4.2 Majorization for Nonconvex Penalty Functions

We assume that the penalty function $r(\mathbf{w}) = \sum_{i=1}^{p} \zeta(|w_i|)$. Thus we can construct the surrogates for ζ separately. It should be emphasized that the majorization of the penalty function is not always necessary. For instance, when one can easily obtain

$$\mathbf{w}^{(k+1)} = \operatorname*{argmin}_{\mathbf{w}} \left\{ Q_f(\mathbf{w}|\mathbf{w}^{(k)}) + r(\mathbf{w}) \right\}, \tag{7}$$

the surrogate for $r(\mathbf{w})$ may not be considered. That is to say, this procedure is optional. However, the surrogate for $r(\mathbf{w})$ can result in efficient computations sometimes, especially when handling the proximal operator of $r(\mathbf{w})$ suffers a large computation burden.

We consider a more general case and give some assumptions.

Assumption 12 *Let $r(\mathbf{w}) = \sum_{i=1}^{p} \zeta(|w_i|)$, where the map ζ: $\mathbb{R}_+ \to \mathbb{R}_+$ is concave and differentiable. Moreover, $\zeta'(t)$ is Lipschitz continuous on $[0, +\infty)$. That is,*

$$|\zeta'(t_1) - \zeta'(t_2)| \leq L_\zeta |t_1 - t_2|,$$

for any $t_1, t_2 \geq 0$.

Many nonconvex penalties admit such properties, such as nonconvex LOG penalty, MCP, SCAD, etc. Although the ℓ_q-norm ($q \in (0, 1)$) may not satisfy the gradient Lipschitz continuous condition, we alternatively consider $\zeta(|w|) = \lambda(1 + \alpha|w|)^q$, with $\alpha > 0$, which is gradient Lipschitz continuous on $[0, +\infty)$.

Thanks to concavity, we have

$$\zeta(|w_i|) \leq \zeta(|w_i^{(k)}|) + \zeta'(|w_i^{(k)}|)(|w_i| - |w_i^{(k)}|), \tag{8}$$

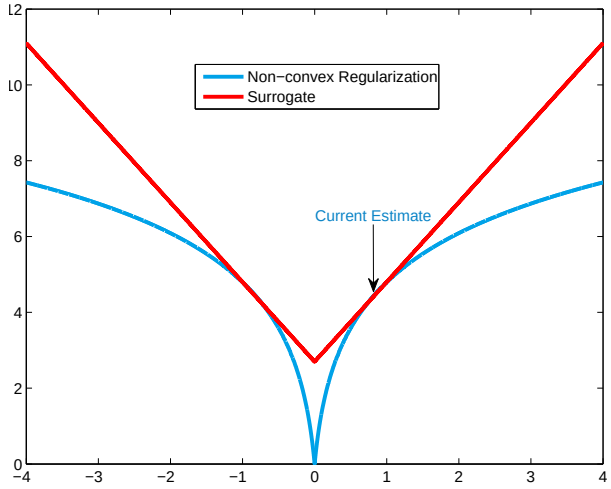


Figure 2: Surrogate for nonconvex regularization

for any $i \in \{1, \cdots, p\}$. Thus, the majorant function for $r(\mathbf{w})$, denoted by $Q_r(\mathbf{w}|\mathbf{w}^{(k)})$, is

$$Q_r(\mathbf{w}|\mathbf{w}^{(k)}) = \sum_{i=1}^{p} \Big[\zeta(|w_i^{(k)}|) + \zeta'(|w_i^{(k)}|)(|w_i| - |w_i^{(k)}|)\Big]. \tag{9}$$

It is easy to see that $Q_r(\mathbf{w}|\mathbf{w}^{(k)}) \geq r(\mathbf{w})$ and $Q_r(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) = r(\mathbf{w}^{(k)})$. Moreover, the corresponding surrogates transfer nonconvex objectives into convex ones, which brings efficient and stable computations. As illustrated in Figure 2, at each iteration, we optimize the tangent above the nonconvex penalty which is tight at the current estimate.

The key idea was early studied in DC programming (Gasso et al., 2009), which linearizes iteratively concave functions to obtain convex surrogates. The idea has been also revisited by Zou and Li (2008), Candes et al. (2008), Chartrand and Yin (2008).

Specifically, Zou and Li (2008) developed the local linear approximation (LLA) algorithm and pointed that the LLA algorithm can be cast as an EM algorithm under certain condition. The LLA algorithm uses the same majorant function as in (9) for the nonconvex and nonsmooth penalty function.

Candes et al. (2008) studied a so-called iteratively re-weighted ℓ_1 minimization, which also falls into the MM procedure. For example, when

$$r(\mathbf{w}) = \lambda \sum_{i=1}^{p} \log(1 + \frac{1}{\epsilon}|w_i|), \text{ where } \epsilon > 0,$$

the re-weighted ℓ_1 minimization scheme is given as

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ f(\mathbf{w}) + \lambda \sum_{i=1}^{p} \frac{|w_i|}{|w_i^{(k)}| + \epsilon} \right\}, \tag{10}$$

which can be also derived from (9).

Algorithm 1 Majorization Minimization Algorithm for Nonconvex Penalization

1: Initialize $\mathbf{w}^{(0)}$ and T (the maximum number of iterations), set $k = 0$.
2: **repeat**
3: Compute $Q_f(\mathbf{w}|\mathbf{w}^{(k)})$, which is the surrogate of $f(\mathbf{w})$;
4: Compute $Q_r(\mathbf{w}|\mathbf{w}^{(k)})$, which is the surrogate of $r(\mathbf{w})$;
5: Update $\mathbf{w}^{(k+1)}$ by (7) or (12);
6: $k = k + 1$;
7: **until** Some stopping criterion is satisfied.

To be the best of our knowledge, there is few complete convergence results for these algorithms. In particular, it is hard to address the convergence of the sequence $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$. In most of traditional treatments, the asymptotic stationary point is studied. These treatments usually follow the general convergence results for MM algorithms (Lange, 2004) without exploiting the property of the objective function. Our global convergence results given in Section 5 are based on theory of the Kurdyka-Łojasiewicz inequality. Our results directly apply to the LLA and iteratively re-weighted ℓ_1 minimization algorithms.

4.3 A Generic MM Algorithm

We are now ready to summarize the whole MM procedure. Recall that the original problem (1) includes two parts. We first consider the simple case. The "simple" means the following problem can be handled easily:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \frac{1}{2\lambda}\|\mathbf{w} - \mathbf{u}\|^2 + r(\mathbf{w}) \right\}. \tag{11}$$

This implies that (7) can be efficiently solved. This leads to nonconvex proximal-gradient methods (Fukushima and Mine, 1981, Lewis and Wright, 2008). We thus generate a sequence $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ by (7).

However, when (11) is intractable, we substitute $r(\mathbf{w})$ with (9). Then in each iteration, the problem reduces to

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ Q_f(\mathbf{w}|\mathbf{w}^{(k)}) + Q_r(\mathbf{w}|\mathbf{w}^{(k)}) \right\}. \tag{12}$$

The whole procedure is briefly presented in Algorithm 1.

5. Convergence Analysis

We now study the convergence analysis of Algorithm 1. It should be claimed that the global convergence, which is our focus, means that for any $\mathbf{w}^{(0)} \in \mathbb{R}^p$, the sequence $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ generated by (7) or (12) converges to the critical point of $F(\mathbf{w})$.

Lemma 13 *Suppose* **Assumptions 10, 11** *hold or* **Assumptions 10, 11, and 12** *hold. Then the sequence* $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ *generated by (7) or generated by (12) satisfies the following properties.*

(i) *The generated sequence* $\left\{F(\mathbf{w}^{(k)})\right\}_{k\in\mathbb{N}}$ *is non-increasing, specifically,*

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^{(k+1)}) \geq \frac{\gamma}{2}\|\mathbf{w}^{(k)} - \mathbf{w}^{(k+1)}\|^2, \quad \forall k \geq 0.$$

(ii)

$$\sum_{k=0}^{\infty} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 < +\infty,$$

which implies $\lim_{k\to\infty} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| = 0$.

Lemma 13 enjoys the descent property of the MM approach which always makes the objective term decrease after each iteration. Moreover, the objective function value decreases at least $\frac{\gamma}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2$ for the kth step. By the fact that $\inf_{\mathbf{w}} F(\mathbf{w}) > -\infty$, we can draw the conclusion that the sequence $\{F(\mathbf{w}^{(k)})\}_{k\in\mathbb{N}}$ converges.

Because of the coerciveness of function $F(\mathbf{w})$, there exists a convergent subsequence $\left\{\mathbf{w}^{(n_k)}\right\}_{k\in\mathbb{N}}$ that converges to $\bar{\mathbf{w}}$. The set of all cluster or limit points which are started with $\mathbf{w}^{(0)}$ is denoted by $\mathcal{M}(\mathbf{w}^{(0)})$. That is,

$$\mathcal{M}(\mathbf{w}^{(0)}) \triangleq \left\{\bar{\mathbf{w}} \in \mathbb{R}^p : \exists\ n_k,\ \ \{n_k\}_{k\in\mathbb{N}},\ \text{such that}\ \ \mathbf{w}^{(n_k)} \to \bar{\mathbf{w}}\ \text{as}\ k\ \to\ \infty\right\}.$$

It is also easy to see that $F(\mathbf{w})$ is constant and finite on $\mathcal{M}(\mathbf{w}^{(0)})$. In the following lemma we attempt to demonstrate that all points which belong to $\mathcal{M}(\mathbf{w}^{(0)})$ are stationary or critical points of $F(\mathbf{w})$.

Lemma 14 *Suppose* **Assumptions 7, 10, 11** *hold, and the sequence* $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ *is generated by (7). Let* $A^{(k+1)} = \nabla f(\mathbf{w}^{(k+1)}) - \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)})$. *Then*

(i) $A^{(k+1)} \in \partial F(\mathbf{w}^{(k+1)})$;

(ii) $\|A^{(k+1)}\| \leq (L_{Q_f} + L_f)\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|$.

Lemma 15 *Suppose* $r(\mathbf{w}) = \sum_{i=1}^{p} \zeta(w_i)$, *where* ζ*:* $\mathbb{R}_+ \to \mathbb{R}_+$ *is concave and continuous differentiable on* $[0, +\infty)$. *Let* $Q_r(\mathbf{w}|\mathbf{w}^{(k)}) = \sum_{i=1}^{p} \zeta(|w_i^{(k)}|) + \zeta'(|w_i^{(k)}|)(|w_i| - |w_i^{(k)}|)$. *Then*

(i) $Q_r(\mathbf{w}|\mathbf{w}^{(k)}) \geq r(\mathbf{w})$ *and* $Q_r(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) = r(\mathbf{w}^{(k)})$;

(ii) $\partial Q_r(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) = \partial r(\mathbf{w}^{(k)})$.

Lemma 15 shows the relationship between the nonconvex (nonsmooth) penalty function and the corresponding surrogate. This also implies that the surrogate approximates the penalty function well.

We introduce the notion of $\mathrm{sgn}(u)$, which is defined as

$$\mathrm{sgn}(u) \triangleq \begin{cases} 1 & \text{if } u > 0, \\ c & \text{if } u = 0, \\ -1 & \text{if } u < 0. \end{cases} \tag{13}$$

Here c is some real number in $[-1, 1]$. We emphasize that $\mathrm{sgn}(u)$ is a scalar rather than a set.

Lemma 16 (Main Lemma) *Suppose* **Assumptions 7, 10, 11, 12** *hold, and the sequence* $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ *is generated by (12). Let* $b_i^{(k)} = \operatorname{sgn}(w_i^{(k+1)})[\zeta'(|w_i^{(k)}|) - \zeta'(|w_i^{(k+1)}|)]$ *for* $i \in [p]$, *and* $B^{(k+1)} = \nabla f(\mathbf{w}^{(k+1)}) - \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - (b_1^{(k)}, b_2^{(k)}, \cdots, b_p^{(k)})^T$. *Then*

(i) $B^{(k+1)} \in \partial F(\mathbf{w}^{(k+1)})$;

(ii) $\|B^{(k+1)}\| \leq (L_{Q_f} + L_f + L_\zeta)\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|$.

Both Lemma 14 and Lemma 16 suggest a subgradient lower bound for the iterate gap. Due to the majorization of the nonconvex and nonsmooth penalty functions, it is more challenging to bound the subgradient. The ingredient is to observe that the majorant function and the original one share the same subgradient at the current estimate. With Lemmas 14 and 16, we are now ready to state the following lemma.

Lemma 17 *Suppose* **Assumptions 7, 10, 11, 12** *hold. Let the sequence* $\left\{\mathbf{w}^{(k)}\right\}_{k\in\mathbb{N}}$ *be generated by (7) or (12). Then*

(i) $\mathcal{M}(\mathbf{w}^{(0)})$ *is not empty and* $\mathcal{M}(\mathbf{w}^{(0)}) \subset \operatorname{crit} F$;

(ii)
$$\lim_{k\to\infty} \operatorname{dist}\left(\mathbf{w}^{(k)}, \mathcal{M}(\mathbf{w}^{(0)})\right) = 0. \tag{14}$$

Lemma 17 implies that $\mathcal{M}(\mathbf{w}^{(0)})$ is the subset of stationary or critical points of $F(\mathbf{w})$ and $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ are approaching to one point of $\mathcal{M}(\mathbf{w}^{(0)})$. Our current concern is to prove $\lim_{k\to\infty} \mathbf{w}^{(k)} = \mathbf{w}^*$. From Lange (2004), we know that $\mathcal{M}(\mathbf{w}^{(0)})$ is connected. Additionally, if $\mathcal{M}(\mathbf{w}^{(0)})$ is finite, $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ converges .

We can obtain the global convergence based on the assumption that $\mathcal{M}(\mathbf{w}^{(0)})$ is finite. However, the assumption is not practical, because it is usually unknown. Moreover, it is hard to check this assumption. To avoid this issue, the Kurdyka-Łojasiewicz property of the objective function enters in action, because it is often a very easy task to verify the Kurdyka-Łojasiewicz property of a function.

Theorem 18 *Suppose that* F *has Kurdyka-Łojasiewicz property at each point of* $\operatorname{dom}\partial F$, *and* **Assumptions 7, 10, 11, 12** *hold. Let the sequence* $\left\{\mathbf{w}^{(k)}\right\}_{k\in\mathbb{N}}$ *be generated by scheme (7) or (12). Then the following assertions hold.*

(i) *The sequence* $\left\{\mathbf{w}^{(k)}\right\}_{k\in\mathbb{N}}$ *has finite length.*

$$\sum_{k=0}^{\infty} \left\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right\| < \infty \tag{15}$$

(ii) *The sequence* $\left\{\mathbf{w}^{(k)}\right\}_{k\in\mathbb{N}}$ *converges to a critical point* $\mathbf{w}^*$ *of* F.

Theorem 18 shows the global convergence of Algorithm 1. As we have stated, many methods for solving a nonconvex and nonsmooth problem, such as the re-weighted ℓ_1 (Candes et al., 2008) and LLA (Zou and Li, 2008), share the same convergence property as in

Theorem 18. Attouch et al. (2010), Bolte et al. (2013) have well established the global convergence for nonconvex and nonsmooth problems based on the Kurdyka-Łojasiewicz inequality. It is also interesting to point out that their procedures fall into (7). However, they focused on a coordinate descent procedure. The work of Attouch et al. (2010), Bolte et al. (2013) cannot be trivially extended to our more general case.

Theorem 19 (Convergence Rate) *Suppose* **Assumptions 7, 10, 11, 12** *hold, and* $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ *is generated by (7) or (12) which converges to a critical point* $\mathbf{w}^*$ *of* F*, which satisfies the Kurdyka-Łojasiewicz property at each point of* $\operatorname{dom}\partial F$ *with* $\phi(t) = ct^{1-\theta}$ *for* $c > 0$ *and* $\theta \in [0, 1)$*. We have*

(i) *if* $\theta = 0$, $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ *converges to* $\mathbf{w}^*$ *in finite iterations;*

(ii) *if* $\theta \in (0, \frac{1}{2}]$, $\|\mathbf{w}^{(k)} - \mathbf{w}^*\| \le C\rho^k$, $\forall k \ge K_0$, *for some* $K_0 > 0, C > 0, \rho \in (0, 1)$*;*

(iii) *if* $\theta \in (\frac{1}{2}, 1)$, $\|\mathbf{w}^{(k)} - \mathbf{w}^*\| \le Ck^{-\frac{1-\theta}{2\theta-1}}, \forall k \ge K_0$, *for some* $K_0 > 0, C > 0$.

Theorem 19 tells us the convergence rate of our MM procedure for solving the nonconvex regularized problem, which is based on the geometrical property of the function F around its critical point. We see that the convergence rate is at least sublinear.

6. Extension to Concave-Convex Procedure

In this section we show that our work can be extended to the concave-convex procedure (CCCP) (Yuille and Rangarajan, 2003). It is worth noting that CCCP can be also unified into the MM framework.

The CCCP is usually used to solve the following problem:

$$\begin{array}{ll} \min\limits_{\mathbf{w}} & u(\mathbf{w}) - v(\mathbf{w}), \\ \text{s.t.} & c_i(\mathbf{w}) \le \mathbf{0},\ i \in [n], \\ & d_j(\mathbf{w}) = \mathbf{0},\ j \in [m], \end{array} \tag{16}$$

where u, v and c_i are real-valued convex functions and d_j are affine functions. The CCCP algorithm aims to solve the following sequence of convex optimization problems:

$$\begin{array}{rll} \mathbf{w}^{(k+1)} = \operatorname*{argmin}\limits_{\mathbf{w}} & u(\mathbf{w}) - \nabla v(\mathbf{w}^{(k)})^T\mathbf{w}, \\ \text{s.t.} & c_i(\mathbf{w}) \le \mathbf{0},\ i \in [n], \\ & d_j(\mathbf{w}) = \mathbf{0},\ j \in [m]. \end{array} \tag{17}$$

Denote $\mathcal{C} = \Big\{\mathbf{w} : c_i(\mathbf{w}) \le \mathbf{0}, d_j(\mathbf{w}) = \mathbf{0}, i \in [n], j \in [m]\Big\}$, and let $\delta_{\mathcal{C}}(\mathbf{w})$ be the indicator function of the feasible set $\mathcal{C}$; that is,

$$\delta_{\mathcal{C}}(\mathbf{w}) = \begin{cases} 0, & \mathbf{w} \in \mathcal{C}, \\ +\infty, & \mathbf{w} \notin \mathcal{C}. \end{cases}$$

It is directly proved that $\delta_{\mathcal{C}}$ is a convex function. Now the original problem can be reformulated as

$$\min_{\mathbf{w}} \; F(\mathbf{w}) \triangleq \Big\{\delta_{\mathcal{C}}(\mathbf{w}) + u(\mathbf{w}) - v(\mathbf{w})\Big\}. \tag{18}$$

Thus, the CCCP approach would solve the following convex problem at each iteration:

$$\mathbf{w}^{(k+1)} = \operatorname*{argmin}_{\mathbf{w}} \quad \Big\{\delta_{\mathcal{C}}(\mathbf{w}) + u(\mathbf{w}) - \nabla v(\mathbf{w}^{(k)})^T\mathbf{w}\Big\}. \tag{19}$$

In fact, the CCCP approach can be viewed as an MM algorithm. In particular, since $v(\mathbf{w})$ is convex, $-v(\mathbf{w})$ is concave. As a result, we have

$$-v(\mathbf{w}) \leq -v(\mathbf{w}^{(k)}) - \nabla v(\mathbf{w}^{(k)})^T(\mathbf{w} - \mathbf{w}^{(k)}).$$

This leads us to the linear majorization of $-v(\mathbf{w})$. When the constant part is omitted, (17) or (19) are recovered. In summary, CCCP linearizes the concave part of the objective function. Next, we make some assumptions to address the convergence of CCCP.

Assumption 20 *Consider the problem in (18) where $\delta_{\mathcal{C}}$, u, and v are convex functions. Suppose the following three asserts hold.*

(i) *$u(\mathbf{w})$ and $v(\mathbf{w})$ are C^1 functions;*

(ii) *$u(\mathbf{w})$ is γ-strongly convex;*

(iii) *$\nabla v(\mathbf{w})$ is Lipschitz continuous.*

With the above assumption, the following theorem shows that the sequence $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ generated by CCCP converges to the critical point of $F(\mathbf{w})$.

Theorem 21 (Global Convergence of CCCP) *Suppose* **Assumption 10, 20** *hold. And F satisfy the the Kurdyka-Łojasiewicz property at each point of* $\operatorname{dom}\partial F$. *Let the sequence $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ be generated by (19). Then the conclusions of Theorems 18 and 19 hold.*

It is worth pointing out that the global convergence analysis for CCCP has also been studied by Lanckriet and Sriperumbudur (2009). Their analysis is based on the novel Zangwill's theory. Zangwill's theory is a very important tool to deal with the convergence issue of iterative algorithms. But it typically requires that $\mathcal{M}(\mathbf{w}^{(0)})$ is finite or discrete to achieve the convergent sequence $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ (Wu, 1983, Lanckriet and Sriperumbudur, 2009). In contrast, our analysis based on the Kurdyka-Łojasiewicz inequality does not need this requirement.

7. Numerical Analysis

In this paper our principal focus has been to explore the convergence properties of majorization minimization (MM) algorithms for nonconvex optimization problems. However, we have also developed two special MM algorithms based on (7) and (12), respectively. Thus, it is interesting to conduct empirical analysis of convergence of the algorithms. We particularly employ the logistic loss and LOG penalty for the classification problem. We refer to the algorithms as MM-(a) and MM-(b) for discussion simplicity.

We evaluate both MM-(a) and MM-(b) on binary datasets[2]. Descriptions of the datasets are reported in Table 2. For each dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$,

$$F(\mathbf{w}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i\mathbf{x}_i^T\mathbf{w}))}_{f(\mathbf{w})} + \underbrace{\lambda\sum_{i=1}^{p}\log(1+\alpha|w_i|)}_{r(\mathbf{w})}, \tag{20}$$

where $\lambda > 0$ and $\alpha > 0$ are hyperparameters. We adopt the corresponding majorizantion function

$$Q_f(\mathbf{w}|\mathbf{w}^{(k)}) = f(\mathbf{w}^{(k)}) + \langle \nabla f(\mathbf{w}^{(k)}), \mathbf{w} - \mathbf{w}^{(k)}\rangle + \frac{\mu^{(k)}}{2}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2$$

and

$$Q_r(\mathbf{w}|\mathbf{w}^{(k)}) = \lambda\sum_{i=1}^{p}\left\{\log(1+\alpha|w_i^{(k)}|) + \frac{\alpha}{1+\alpha|w_i^{(k)}|}(|w_i| - |w_i^{(k)}|)\right\}.$$

As mentioned in the previous section, the Lipschitz constant of $\nabla f(\mathbf{w})$ is bounded by $\frac{1}{4n}\sum_{i=1}^n \mathbf{x}_i^T\mathbf{x}_i$. Typically, to set the value $\mu^{(k)}$, one often uses the line-search method (Beck and Teboulle, 2009) to achieve better performance. However, since we are only concerned with the convergence behavior of MM, we just set $\mu^{(k)} = \frac{\rho}{4n}\sum_{i=1}^n \mathbf{x}_i^T\mathbf{x}_i$ where $\rho \geq 1$.

We plot the error between objective function values and the $F(\mathbf{w}^*)$(log scaled) vs. CPU times with respect to different hyperparameters settings in Figure 3. We observe that both MM-(a) and MM-(b) generate the monotone decreasing sequence $\{F(\mathbf{w}^{(k)})\}_{k\in\mathbb{N}}$ and achieve nearly the same optimal objective value. We also find that MM-(b) runs faster than MM-(a). This implies that it is efficient to construct the majorization function of the LOG penalty. In fact, MM-(a) will cost more computations when one directly calculates the proximal operator of the LOG penalty. In contrast, MM-(b) only needs to do the soft-thresholding (shrinkage) operator on the current estimate. In summary, numerical experiments show that both MM-(a) and MM-(b) make the objective function value decrease and converge.

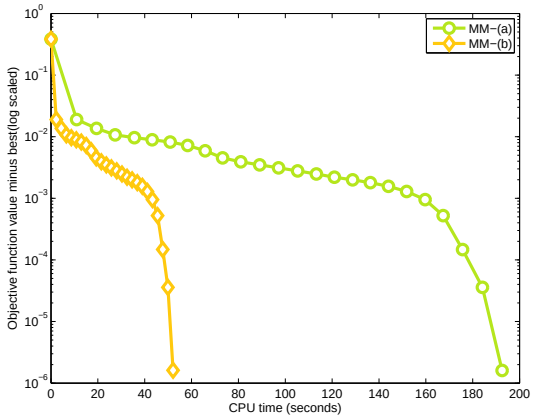
Table 2: Description of the datasets

Data sets	n	p	storage
leukemia	72	7129	sparse
news20	19996	1355191	sparse
covtype	581012	54	dense

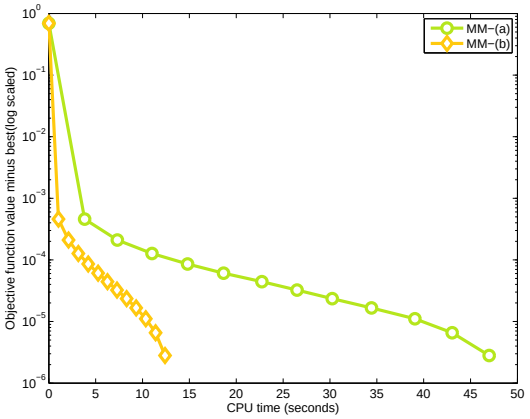
8. Conclusions

Majorization minimization (MM) algorithms are very popular in machine learning and statistical inference. In this paper, we have employed MM algorithms to solve the nonconvex regularized problems. However, the convergence analysis of MM for nonconvex and nonsmooth problems is a challenging issue. We have established the global convergence results of

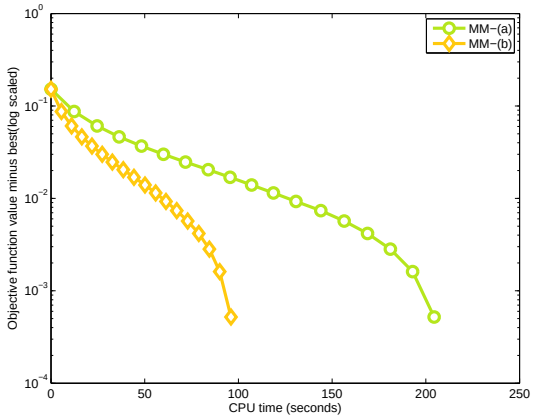
2. http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html



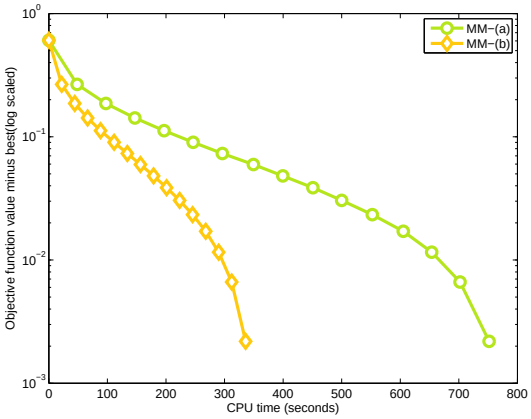
(a) leukemia ($\lambda = 0.3, \alpha = 0.3$)



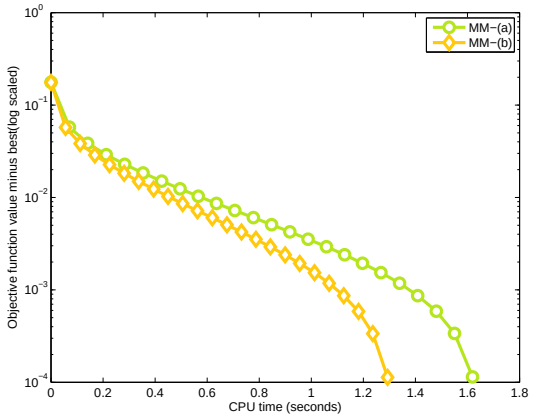
(b) leukemia ($\lambda = 0.0001, \alpha = 0.0003$)



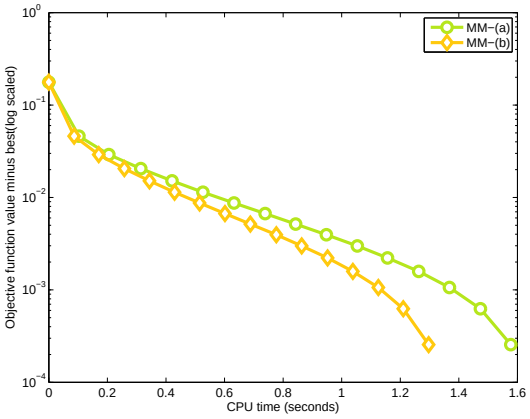
(c) news20 ($\lambda = 0.01, \alpha = 0.03$)



(d) news20 ($\lambda = 0.0001, \alpha = 0.0003$)



(e) covtype ($\lambda = 0.01, \alpha = 0.03$)



(f) covtype ($\lambda = 0.0001, \alpha = 0.0003$)

Figure 3: performance of MM with different parameter settings

the MM procedure using the geometrical property of the objective function. In particular, our results are built on the Kurdyka-Łojasiewicz inequality. We have shown that our results also apply to the iteratively re-weighted ℓ_1 minimization method, local linear approximation (LLA), and concave-convex procedure (CCCP).

Appendix A. Proofs

A.1 The proof of Corollary 8

Proof Since $f_i(\mathbf{w})$ is differentiable for $i \in [n]$ and each $\nabla f_i(\mathbf{w})$ is L_i-Lipschitz continuous, we have

$$\|\nabla f_i(\mathbf{u}) - \nabla f_i(\mathbf{v})\| \leq L_i \|\mathbf{u} - \mathbf{v}\|,$$

for $i \in [n]$. Then

$$\begin{aligned}
\|\nabla h(\mathbf{u}) - \nabla h(\mathbf{v})\| &= \|\sum_{i=1}^{n} \alpha_i \nabla f_i(\mathbf{u}) - \sum_{i=1}^{n} \alpha_i \nabla f_i(\mathbf{v})\| \\
&\leq \sum_{i=1}^{n} |\alpha_i| \|\nabla f_i(\mathbf{u}) - \nabla f_i(\mathbf{v})\| \\
&\leq (\sum_{i=1}^{n} |\alpha_i| L_i) \|\mathbf{u} - \mathbf{v}\|
\end{aligned}$$

So, $\nabla h(\mathbf{w})$ is $\sum_{i=1}^{n} |\alpha_i| L_i$ Lipschitz continuous. ■

A.2 The proof of Lemma 13

Proof We first consider (7) procedure.

(i) Recall that

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ Q_f(\mathbf{w}|\mathbf{w}^{(k)}) + r(\mathbf{w}) \right\}.$$

We have

$$Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) + r(\mathbf{w}^{(k+1)}) - Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) - r(\mathbf{w}^{(k)}) \leq 0. \tag{21}$$

By the strongly-convex property of $Q_f(\mathbf{w}|\mathbf{w}^{(k)}) - f(\mathbf{w})$,

$$\begin{aligned}
&Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - f(\mathbf{w}^{(k+1)}) - Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) + f(\mathbf{w}^{(k)}) \geq \\
&\langle \nabla Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) - \nabla f(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)} \rangle + \frac{\gamma}{2} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \\
&= \frac{\gamma}{2} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2
\end{aligned}$$

The last equality complies with $\nabla Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) = \nabla f(\mathbf{w}^{(k)})$.

Combining with (21), we have

$$f(\mathbf{w}^{(k)}) + r(\mathbf{w}^{(k)}) - f(\mathbf{w}^{(k+1)}) - r(\mathbf{w}^{(k+1)}) \geq \frac{\gamma}{2} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2.$$

So

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^{(k+1)}) \geq \frac{\gamma}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2.$$

(ii) We summary the above inequality from $k = 0$ to $+\infty$. Then

$$\sum_{k=0}^{+\infty} F(\mathbf{w}^{(k)}) - F(\mathbf{w}^{(k+1)}) \geq \sum_{k=0}^{+\infty} \frac{\gamma}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2$$

Notice $\inf_{\mathbf{w}} F(\mathbf{w}) > -\infty$, so

$$\sum_{k=0}^{+\infty} \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \leq \frac{2}{\gamma}(F(\mathbf{w}^{(0)}) - F(\mathbf{w}^{(\infty)})) < +\infty,$$

which completes the proof.

Let's come to the sequence $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ generated by (12).
Similarly,

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ Q_f(\mathbf{w}|\mathbf{w}^{(k)}) + Q_r(\mathbf{w}|\mathbf{w}^{(k)}) \right\}.$$

We obtain

$$Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) + Q_r(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) - Q_r(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) \leq 0. \tag{22}$$

Since we have(similar to proof of Lemma 13)

$$Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - f(\mathbf{w}^{(k+1)}) - Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) + f(\mathbf{w}^{(k)}) \geq \frac{\gamma}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2.$$

On the other hand,

$$\begin{aligned}
Q_r(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - Q_r(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) &= \sum_{i=1}^{p} \zeta(|w_i^{(k)}|) + \zeta'(|w_i^{(k)}|)(|w_i^{(k+1)}| - |w_i^{(k)}|) \\
&\quad - \sum_{i=1}^{p} \zeta(|w_i^{(k)}|) + \zeta'(|w_i^{(k)}|)(|w_i^{(k)}| - |w_i^{(k)}|) \\
&= \sum_{i=1}^{p} \langle \zeta'(|w_i^{(k)}|)(|w_i^{(k+1)}| - |w_i^{(k)}|) \\
&\geq \sum_{i=1}^{p} \zeta(|w_i^{(k+1)}|) - \zeta(|w_i^{(k)}|) \\
&= r(\mathbf{w}^{(k+1)}) - r(\mathbf{w}^{(k)})
\end{aligned}$$

Combining the above three inequalities, we have

$$f(\mathbf{w}^{(k)}) + r(\mathbf{w}^{(k)}) - f(\mathbf{w}^{(k+1)}) - r(\mathbf{w}^{(k+1)}) \geq \frac{\gamma}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2,$$

which implies that

$$F(\mathbf{w}^{(k)}) - F(\mathbf{w}^{(k+1)}) \geq \frac{\gamma}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2.$$

■

A.3 The proof of Lemma 14

Proof

(i) Recall that

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ Q_f(\mathbf{w}|\mathbf{w}^{(k)}) + r(\mathbf{w}) \right\}.$$

Writing down the optimality condition, we have

$$\mathbf{0} = \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) + \mathbf{u}^{(k+1)}, \tag{23}$$

where $\mathbf{u}^{(k+1)} \in \partial r(\mathbf{w}^{(k+1)})$. Let's rewrite it as follow

$$\nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - \nabla f(\mathbf{w}^{(k+1)}) + \nabla f(\mathbf{w}^{(k+1)}) + \mathbf{u}^{(k+1)} = \mathbf{0}.$$

Because $A^{(k+1)} = \nabla f(\mathbf{w}^{(k+1)}) - \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)})$, we immediately have

$$A^{(k+1)} = \nabla f(\mathbf{w}^{(k+1)}) + \mathbf{u}^{(k+1)} \in \partial F(\mathbf{w}^{(k+1)}).$$

(ii) With the Lipschitz continuous of $\nabla Q_f(\mathbf{w}|\mathbf{w}^{(k)})$ and $\nabla f(\mathbf{w})$, we have

$$\begin{cases} \|\nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - \nabla Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)})\| \leq L_{Q_f}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|, \\ \|\nabla f(\mathbf{w}^{(k+1)}) - \nabla f(\mathbf{w}^{(k)})\| \leq L_f\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|. \end{cases} \tag{24}$$

Hence,

$$\begin{aligned} \|\mathbf{A}^{(k+1)}\| &= \|\nabla f(\mathbf{w}^{(k+1)}) - \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)})\| \\ &= \|\nabla f(\mathbf{w}^{(k+1)}) - \nabla f(\mathbf{w}^{(k)}) + \nabla Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) - \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)})\| \\ &\leq \|\nabla f(\mathbf{w}^{(k+1)}) - \nabla f(\mathbf{w}^{(k)})\| + \|\nabla Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) - \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)})\| \\ &\leq L_f\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| + L_{Q_f}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| \\ &= (L_{Q_f} + L_f)\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| \end{aligned} \tag{25}$$

■

A.4 The proof of Lemma 15

Proof

(i) By the concavity of ζ, we have

$$\zeta(|w_i|) \leq \zeta(|w_i^{(k)}|) + \zeta'(|w_i^{(k)}|)(|w_i| - |w_i^{(k)}|),$$

for any $i \in [p]$. We immediately obtain

$$Q_r(\mathbf{w}|\mathbf{w}^{(k)}) \geq r(\mathbf{w}) \text{ and } Q_r(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) = r(\mathbf{w}^{(k)}).$$

(ii) Notice the fact that the subdifferential calculus for separable functions yields the follows (Rockafellar et al., 1998).

$$\partial\Big(\sum_{i=1}^{p}\zeta(w_i)\Big) = \partial\zeta(w_1)\times\partial\zeta(w_2)\cdots\times\partial\zeta(w_n).$$

Since $r(\mathbf{w})$ and $Q_r(\mathbf{w})$ are separable, we consider each dimension independently. For any $i\in[p]$, if $w_i>0$, we have

$$\partial_i r(\mathbf{w}^{(k)}) = \{\zeta'(w_i)\},\ \partial_i Q_r(\mathbf{w}^{(k)}) = \{\zeta'(w_i)\}.$$

Similarly, if $w_i<0$, we have

$$\partial_i r(\mathbf{w}^{(k)}) = \{-\zeta'(-w_i)\},\ \partial_i Q_r(\mathbf{w}^{(k)}) = \{-\zeta'(-w_i)\}.$$

For the case $w_i=0$, we have

$$\partial_i r(\mathbf{w}^{(k)}) = [-\zeta'(0),\zeta'(0)],\ \partial_i Q_r(\mathbf{w}^{(k)}) = [-\zeta'(0),\zeta'(0)].$$

So

$$\partial Q_r(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) = \partial r(\mathbf{w}^{(k)}).$$

■

A.5 The proof of Lemma 16

Proof

(i) By using

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\left\{Q_f(\mathbf{w}|\mathbf{w}^{(k)}) + Q_r(\mathbf{w}|\mathbf{w}^{(k)})\right\}.$$

Alternatively

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\operatorname{argmin}}\left\{Q_f(\mathbf{w}|\mathbf{w}^{(k)}) + \sum_{i=1}^{p}\zeta(|w_i^{(k)}|) + \zeta'(|w_i^{(k)}|)(|w_i|-|w_i^{(k)}|)\right\}.$$

For the notation, we let $\mu_i^{(k)} = \zeta'(|w_i^{(k)}|)\mathrm{sgn}(w_i^{(k+1)})$. Then the optimal condition yields

$$\mathbf{0} = \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) + (\mu_1^{(k)},\mu_2^{(k)},\cdots,\mu_p^{(k)})^T. \tag{26}$$

We rewrite it as

$$\begin{aligned}\mathbf{0} &= \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - \nabla f(\mathbf{w}^{(k+1)}) + \nabla f(\mathbf{w}^{(k+1)})\\ &+ (\zeta'(|w_1^{(k)}|)\mathrm{sgn}(w_1^{(k+1)}),\zeta'(|w_2^{(k)}|)\mathrm{sgn}(w_2^{(k+1)}),\cdots,\zeta'(|w_p^{(k)}|)\mathrm{sgn}(w_p^{(k+1)}))^T\end{aligned}$$

On the other hand

$$b_i^{(k)} = \mathrm{sgn}(w_i^{(k+1)})(\zeta'(|w_i^{(k)}|) - \zeta'(|w_i^{(k+1)}|),$$

for $i \in [p]$. So

$$\mathbf{0} = \nabla f(\mathbf{w}^{(k+1)}) + (\zeta'(|w_1^{(k+1)}|)\text{sgn}(w_1^{(k+1)}), \zeta'(|w_2^{(k+1)}|)\text{sgn}(w_2^{(k+1)}), \cdots, \zeta'(|w_p^{(k+1)}|)\text{sgn}(w_p^{(k+1)}))^T$$
$$+ \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - \nabla f(\mathbf{w}^{(k+1)}) + (b_1^{(k)}, b_2^{(k)}, \cdots, b_p^{(k)})^T.$$

Then, we have

$$\begin{aligned}\nabla f(\mathbf{w}^{(k+1)}) &+ (\zeta'(|w_1^{(k+1)}|)\text{sgn}(w_1^{(k+1)}), \zeta'(|w_2^{(k+1)}|)\text{sgn}(w_2^{(k+1)}), \cdots, \zeta'(|w_p^{(k+1)}|)\text{sgn}(w_p^{(k+1)}))^T \\ &= \nabla f(\mathbf{w}^{(k+1)}) - \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - (b_1^{(k)}, b_2^{(k)}, \cdots, b_p^{(k)})^T \\ &= B^{(k+1)}\end{aligned}$$

Notice that

$$\begin{aligned}\nabla f(\mathbf{w}^{(k+1)}) &+ (\zeta'(|w_1^{(k+1)}|)\text{sgn}(w_1^{(k+1)}), \zeta'(|w_2^{(k+1)}|)\text{sgn}(w_2^{(k+1)}), \cdots, \zeta'(|w_p^{(k+1)}|)\text{sgn}(w_p^{(k+1)}))^T \\ &\in \partial F(\mathbf{w}^{(k+1)})\end{aligned}$$

So we have

$$B^{(k+1)} \in \partial F(\mathbf{w}^{(k+1)}).$$

(ii) Similarly, with the Lipschitz continuous of $\nabla Q_f(\mathbf{w}|\mathbf{w}^{(k)})$, $\nabla f(\mathbf{w})$ and $\zeta'(t)(t \geq 0)$ we have

$$\begin{cases} \|\nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - \nabla Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)})\| \leq L_{Q_f}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|, \\ \|\nabla f(\mathbf{w}^{(k+1)}) - \nabla f(\mathbf{w}^{(k)})\| \leq L_f\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|, \\ |\zeta'(t_1) - \zeta'(t_2)| \leq L_\zeta|t_1 - t_2|. \end{cases} \tag{27}$$

Now, we are ready to bound the subgradient $B^{(k+1)}$

$$\begin{aligned}\|B^{(k+1)}\| &= \|\nabla f(\mathbf{w}^{(k+1)}) - \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - (b_1^{(k)}, b_2^{(k)}, \cdots, b_p^{(k)})^T\| \\ &= \|\nabla f(\mathbf{w}^{(k+1)}) - \nabla f(\mathbf{w}^{(k)}) + \nabla Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) \\ &- \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)}) - (b_1^{(k)}, b_2^{(k)}, \cdots, b_p^{(k)})^T\| \\ &\leq \|\nabla f(\mathbf{w}^{(k+1)}) - \nabla f(\mathbf{w}^{(k)})\| + \|\nabla Q_f(\mathbf{w}^{(k)}|\mathbf{w}^{(k)}) - \nabla Q_f(\mathbf{w}^{(k+1)}|\mathbf{w}^{(k)})\| \\ &+ \|(b_1^{(k)}, b_2^{(k)}, \cdots, b_p^{(k)})^T\| \\ &\leq L_f\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| + L_{Q_f}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| + \|(b_1^{(k)}, b_2^{(k)}, \cdots, b_p^{(k)})^T\|\end{aligned} \tag{28}$$

Then let's bound $\|(b_1^{(k)}, b_2^{(k)}, \cdots, b_p^{(k)})^T\|$. For each $i \in [p]$, we have

$$b_i^{(k)} = \text{sgn}(w_i^{(k+1)})(\zeta'(|w_i^{(k)}|) - \zeta'(|w_i^{(k+1)}|)).$$

By using $|\text{sgn}(w_i^{(k+1)})| \leq 1$, we have

$$\begin{aligned}|b_i^{(k)}| &= |\text{sgn}(w_i^{(k+1)})(\zeta'(|w_i^{(k)}|) - \zeta'(|w_i^{(k+1)}|))| \\ &\leq |(\zeta'(|w_i^{(k)}|) - \zeta'(|w_i^{(k+1)}|))| \\ &\leq L_\zeta||w_i^{(k+1)}| - |w_i^{(k)}||\end{aligned}$$

Then

$$\begin{aligned}\|(b_1^{(k)}, b_2^{(k)}, \cdots, b_p^{(k)})^T\| &\leq L_\zeta \|(|w_1^{(k+1)}| - |w_1^{(k)}|, |w_2^{(k+1)}| - |w_2^{(k)}|, \cdots, |w_p^{(k+1)}| - |w_p^{(k)}|)^T\| \\ &\leq L_\zeta \|(|w_1^{(k+1)} - w_1^{(k)}|, |w_2^{(k+1)} - w_2^{(k)}|, \cdots, |w_p^{(k+1)} - w_p^{(k)}|)^T\| \\ &= L_\zeta \|(w_1^{(k+1)} - w_1^{(k)}, w_2^{(k+1)} - w_2^{(k)}, \cdots, w_p^{(k+1)} - w_p^{(k)})^T\| \\ &= L_\zeta \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| \end{aligned} \tag{29}$$

Combining (28) and (29),

$$\|B^{(k+1)}\| \leq (L_{Q_f} + L_f + L_\zeta)\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|.$$

■

A.6 The proof of Lemma 17

Proof Since the F($\mathbf{w}$) is coercive, the sequence $\left\{\mathbf{w}^{(k)}\right\}_{k\in\mathbb{N}}$ is bounded. Therefore there exists an increasing sequence $\{n_k\}_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} \mathbf{w}^{(n_k)} = \mathbf{w}^*.$$

Recall that $F(\mathbf{w}) = f(\mathbf{w}) + \lambda \sum_{i=1}^p \zeta(w_i)$ is continuous. We have

$$\lim_{k\to\infty} F(\mathbf{w}^{(n_k)}) = F(\mathbf{w}^*).$$

On the other hand, we know $A^{(k)} \in \partial F(\mathbf{w}^{(k)}), B^{(k)} \in \partial F(\mathbf{w}^{(k)})$. Moreover, from Lemma (14) and Lemma (16) it can be seen that as $k \to \infty$, $A^{(k)} \to \mathbf{0}$ and $B^{(k)} \to \mathbf{0}$. Remember that ∂F is close. So $\mathbf{0} \in \partial F(\mathbf{w}^*)$, which contributes to that $\mathbf{w}^*$ is a critical point of F. ■

A.7 Uniformized KŁ property

Before providing the global convergence result, we first introduce a class of concave and continuous functions. Let $\eta \in (0, +\infty]$. We are concerned with Φ_η which contain the class of all concave and continuous functions $\phi : [0, \eta) \to \mathbb{R}_+$ satisfying the following properties:

(a) $\phi(0) = 0$ and continuous at 0;

(b) ϕ is C^1 on $(0, \eta)$;

(c) $\phi'(t) > 0$, $\forall$ t $\in$ (0,η).

Lemma 22 (Bolte et al. (2013)) *Suppose Ω is a compact set and let $F : \mathbb{R}^p \to (-\infty, \infty]$ be a lower semi-continuous function. Moreover, F is constant on Ω and satisfy KL property*

at each point of Ω. *Then there exist* $\epsilon > 0, \eta > 0$ *and* $\phi \in \Phi_\eta$ *such that for all* $\bar{\mathbf{u}}$ *in* Ω *and all* $\mathbf{u}$ *in*

$$\Big\{\mathbf{u} \in \mathbb{R}^p : \ dist(\mathbf{u}, \Omega) < \epsilon\Big\} \bigcap \Big\{\mathbf{u} : F(\bar{\mathbf{u}}) < F(\mathbf{u}) < F(\bar{\mathbf{u}}) + \eta\Big\}$$

one has,

$$\phi'(F(\mathbf{u}) - F(\bar{\mathbf{u}}))dist(\mathbf{0}, \partial F(\mathbf{u})) \geq 1. \tag{30}$$

A.8 The proof of Theorem 18

Proof As is known, there exists an increasing sequence $\{n_k\}_{k\in\mathbb{N}}$ such that $\{\mathbf{w}^{(n_k)}\}$ converges to $\mathbf{w}^*$. Suppose that there exists an integer n_0 satisfy that $F(\mathbf{w}^{(n_0)}) = F(\mathbf{w}^*)$. Then it is clear that for any integer $N > n_0$, $F(\mathbf{w}^{(N)}) = F(\mathbf{w}^*)$ holds. Then it is trivial to achieve the convergent sequence. Otherwise, we consider the case that $F(\mathbf{w}^{(k)}) > F(\mathbf{w}^*)$, $\forall k \in \mathbb{N}$. Because the sequence $\{F(\mathbf{w}^{(k)})\}_{k\in\mathbb{N}}$ is convergent, it is clear that for any $\eta > 0$, there exist one integer m such that $F(\mathbf{w}^{(k)}) < F(\mathbf{w}^*) + \eta$ for all $k > m$. By using lemma 17, we have $\lim_{k\to\infty} dist(\mathbf{w}^{(k)}, \mathcal{M}(\mathbf{w}^{(0)})) = 0$ which implies that for any $\epsilon > 0$ there exists a positive integer n such that $dist(\mathbf{w}^{(k)}, \mathcal{M}(\mathbf{w}^{(0)})) < \epsilon$ for all $k > n$. Let $l = \max\{m, n\}$. Then for any $k > l$, we have

$$\phi'\Big(F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*)\Big)dist(\mathbf{0}, \partial F(\mathbf{w}^{(k)})) \geq 1.$$

By the Lemma 14 and Lemma 16, we have

$$\phi'\Big(F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*)\Big) \geq \rho\Big\|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\Big\|^{-1}, \tag{31}$$

where $\rho = \frac{1}{L_{Q_f} + L_f + L_\zeta}$.

we let $d_{k,k+1} \triangleq \phi(F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*)) - \phi(F(\mathbf{w}^{(k+1)}) - F(\mathbf{w}^*))$. With the property of concave functions, we have

$$\begin{aligned} d_{k,k+1} &\geq \phi'(F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*))(F(\mathbf{w}^k) - F(\mathbf{w}^{(k+1)})) \\ &\geq \frac{\gamma}{2}\phi'(F(\mathbf{w}^{(k)}) - F(\mathbf{w}^*))\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \\ &\geq \frac{\gamma}{2(L_f + L_{Q_f} + L_\zeta)}\|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|^{-1}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \end{aligned} \tag{32}$$

That is

$$Md_{k,k+1}\|\mathbf{w}^{(k)} - \mathbf{w}^{k-1}\| \geq \|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2$$

where $M = \frac{2(L_f + L_{Q_f} + L_\zeta)}{\gamma}$. Notice that

$$Md_{k,k+1}\|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\| \leq \Big(\frac{Md_{k,k+1} + \|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|}{2}\Big)^2.$$

So we have

$$2\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| \leq Md_{k,k+1} + \|\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}\|.$$

Then

$$\begin{aligned}\sum_{k=l+1}^{\infty}\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\| &\leq M\sum_{k=l+1}^{\infty}d_{k,k+1}+\sum_{k=l+1}^{\infty}\Big(\|\mathbf{w}^{(k)}-\mathbf{w}^{(k-1)}\|-\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\|\Big)\\ &\leq Md_{l+1,\infty}+\|\mathbf{w}^{(l+1)}-\mathbf{w}^{(l)}\|\\ &\leq M\phi\Big(F(\mathbf{w}^{(l+1)})-F(\mathbf{w}^{*})\Big)+\|\mathbf{w}^{(l+1)}-\mathbf{w}^{(l)}\|\end{aligned}$$

Let $l\to\infty$. Since $\lim_{l\to\infty}\|\mathbf{w}^{(l+1)}-\mathbf{w}^{(l)}\|=0$ and $\lim_{l\to\infty}F(\mathbf{w}^{(l+1)})=F(\mathbf{w}^{*})$, it is clear that

$$\lim_{l\to\infty}\sum_{k=l+1}^{\infty}\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\|=0$$

So

$$\sum_{k=0}^{\infty}\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\|<\infty.$$

Then, we have

$$\lim_{m\to\infty}\sum_{k=m}^{l}\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\|=0,$$

for any $m<l$. This suggests that $\{\mathbf{w}^{(k)}\}_{k\in\mathbb{N}}$ is Cauchy sequence. As a result, it is a convergent sequence that converges to $\mathbf{w}^{*}$. ■

A.9 The proof of Theorem 19

This is a classical result of KŁ function. Since the corresponding function

$$\phi(t)=ct^{1-\theta},\theta\in[0,1).$$

As Attouch and Bolte (2009), the conclusions of Theorem 19 hold.

A.10 The proof of Theorem 21

Proof Recall that

$$\mathbf{w}^{(k+1)}=\underset{\mathbf{w}}{\operatorname{argmin}}\quad\Big\{\delta_{\mathcal{C}}(\mathbf{w})+u(\mathbf{w})-\nabla v(\mathbf{w}^{(k)})^{T}\mathbf{w}\Big\}.$$

We immediately have

$$\mathbf{0}=\mathbf{g}^{(k+1)}+\nabla u(\mathbf{w}^{(k+1)})-\nabla v(\mathbf{w}^{(k)}),$$

where $\mathbf{g}^{(k+1)}\in\partial\delta_{\mathcal{C}}(\mathbf{w}^{(k+1)})$.

Because $\delta_{\mathcal{C}}(\mathbf{w})+u(\mathbf{w})-\nabla v(\mathbf{w}^{(k)})^{T}\mathbf{w}$ is γ-strongly convex, we have

$$\begin{aligned}&\delta_{\mathcal{C}}(\mathbf{w}^{(k)})+u(\mathbf{w}^{(k)})-\nabla v(\mathbf{w}^{(k)})^{T}\mathbf{w}^{(k)}-\delta_{\mathcal{C}}(\mathbf{w}^{(k+1)})-u(\mathbf{w}^{(k+1)})+\nabla v(\mathbf{w}^{(k)})^{T}\mathbf{w}^{(k+1)}\\ &\qquad\geq\langle\mathbf{0},\mathbf{w}^{(k)}-\mathbf{w}^{(k+1)}\rangle+\frac{\gamma}{2}\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\|^{2}.\end{aligned}\tag{33}$$

By the convexity of $v(\mathbf{w})$, we obtain

$$\nabla v(\mathbf{w}^{(k)})^T\mathbf{w}^{(k+1)} - \nabla v(\mathbf{w}^{(k)})^T\mathbf{w}^{(k)} \leq v(\mathbf{w}^{(k+1)}) - v(\mathbf{w}^{(k)}).$$

Thus,

$$\Big(\underbrace{\delta_{\mathcal{C}}(\mathbf{w}^{(k)}) + u(\mathbf{w}^{(k)}) - v(\mathbf{w}^{(k)})}_{F(\mathbf{w}^{(k)})}\Big) - \Big(\underbrace{\delta_{\mathcal{C}}(\mathbf{w}^{(k+1)}) + u(\mathbf{w}^{(k+1)}) - v(\mathbf{w}^{(k+1)})}_{F(\mathbf{w}^{(k+1)})}\Big) \geq \frac{\gamma}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2. \tag{34}$$

On the other hand, we known that

$$\mathbf{0} = \mathbf{g}^{(k+1)} + \nabla u(\mathbf{w}^{(k+1)}) - \nabla v(\mathbf{w}^{(k+1)}) + \nabla v(\mathbf{w}^{(k+1)}) - \nabla v(\mathbf{w}^{(k)}).$$

Let's denote $\nabla v(\mathbf{w}^{(k)}) - \nabla v(\mathbf{w}^{(k+1)}$ as $C^{(k+1)}$. Then $C^{(k+1)} \in \partial F(\mathbf{w}^{(k+1)})$.

$$\|C^{(k+1)}\| = \|\nabla v(\mathbf{w}^{(k)}) - \nabla v(\mathbf{w}^{(k+1)})\| \leq L_v\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| \tag{35}$$

Next we prove $\mathcal{M}(\mathbf{w}^{(0)})$ are the subset of the crit $F(\mathbf{w})$. Because of the coerciveness of the Function $F(\mathbf{w})$, there exists a bounded sequence $\{\mathbf{w}^{n_k}\}_{k\in\mathbb{N}}$, which satisfies that $\lim_{k\to\infty}\mathbf{w}^{n_k} = \bar{\mathbf{w}}$. Since $\delta_{\mathcal{C}}(\mathbf{w})$ is lower semicontinuous, we have

$$\liminf_{k\to\infty} \delta_{\mathcal{C}}(\mathbf{w}^{n_k}) \geq \delta_{\mathcal{C}}(\bar{\mathbf{w}}) \tag{36}$$

On the other hand,

$$\delta_{\mathcal{C}}(\mathbf{w}^{(k+1)}) + u(\mathbf{w}^{(k+1)}) - \nabla v(\mathbf{w}^{(k)})^T\mathbf{w}^{(k+1)} \leq \delta_{\mathcal{C}}(\bar{\mathbf{w}}) + u(\bar{\mathbf{w}}) - \nabla v(\mathbf{w}^{(k)})^T\bar{\mathbf{w}}$$

Rewrite the above formulation, we obtain

$$\delta_{\mathcal{C}}(\mathbf{w}^{(k+1)}) \leq \delta_{\mathcal{C}}(\bar{\mathbf{w}}) - (u(\mathbf{w}^{(k+1)}) - u(\bar{\mathbf{w}})) + \nabla v(\mathbf{w}^{(k)})^T(\mathbf{w}^{(k+1)} - \bar{\mathbf{w}}).$$

Substitute k with $n_k - 1$. By the fact u, v are C^1 functions, we have

$$\limsup_{k\to\infty} \delta(\mathbf{w}^{n_k}) \leq \delta_{\mathcal{C}}(\bar{\mathbf{w}}). \tag{37}$$

Combing (36) and (37), we immediately have

$$\lim_{k\to\infty} \delta_{\mathcal{C}}(\mathbf{w}^{n_k}) = \delta_{\mathcal{C}}(\bar{\mathbf{w}}).$$

Notice that $u(\mathbf{w}), v(\mathbf{w})$ are continuous, we have

$$\lim_{k\to\infty} F(\mathbf{w}^{n_k}) = F(\bar{\mathbf{w}}).$$

(35) implies that $C^{(k)} \to \mathbf{0}$ as $k \to \infty$. Moreover, $C^{(k)} \in \partial F(\mathbf{w}^{(k)})$. Remember the closeness of $\partial F(\mathbf{w})$, we have $\mathbf{0} \in \partial F(\bar{\mathbf{w}})$. So $\mathcal{M}(\mathbf{w}^{(0)})$ are the subset of the crit $F(\mathbf{w})$. With (34) and (35) ready, the next proof is the same as that of Theorem 18. ■

References

Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119, 2013. 2, 6

Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009. 25

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010. 2, 3, 14

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 16

Jacek Bochnak, Michel Coste, and Marie-Fran**c**coise Roy. *Real algebraic geometry*. Springer, 1998. 5

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007. 2, 5

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, pages 1–36, 2013. 2, 3, 14, 23

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008. 2, 10, 13

Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on*, pages 3869–3872. IEEE, 2008. 2, 10

Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011. 7

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. 2, 6

Masao Fukushima and Hisashi Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, 12(8): 989–1000, 1981. 11

Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *Signal Processing, IEEE Transactions on*, 57(12):4686–4698, 2009. 10

Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of The 30th International Conference on Machine Learning*, pages 37–45, 2013. 2, 3

AN Iusem. Augmented lagrangian methods and proximal point methods for convex optimization. *Investigación Operativa*, 8:11–49, 1999. 7

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783. Institut Fourier, 1998. 2

Gert R Lanckriet and Bharath K Sriperumbudur. On the convergence of the concave-convex procedure. In *Advances in neural information processing systems*, pages 1759–1767, 2009. 2, 15

Kenneth Lange. Optimization. springer texts in statistics. 2004. 1, 2, 11, 13

Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000. 1

Bernard Lemaire. The proximal algorithm. *International series of numerical mathematics*, 87:73–87, 1989. 7

Adrian S Lewis and Stephen J Wright. A proximal method for composite minimization. *arXiv preprint arXiv:0812.0423*, 2008. 11

Stanislas Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pages 1575–1595. Institut Fourier, 1993. 2

Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Colloques du CNRS, Les équations aux dérivés partielles*, 117, 1963. 2

Julien Mairal. Optimization with first-order surrogate functions. In *ICML 2013-International Conference on Machine Learning*, volume 28, pages 783–791, 2013. 3

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495), 2011. 2, 6

Yurii Nesterov and IU E Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004. 6

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013. 8

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976. 7

R Tyrrell Rockafellar, Roger J-B Wets, and Maria Wets. *Variational analysis*, volume 317. Springer, 1998. 3, 4, 21

Florin Vaida. Parameter convergence for em and mm algorithms. *Statistica Sinica*, 15(3): 831, 2005. 3

CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983. 15

Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013. 2, 3

Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003. 2, 14

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942, 2010a. 2, 6

Cun-Hui Zhang, Tong Zhang, et al. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012. 2

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:1081–1107, 2010b. 2, 6

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008. 2, 10, 13