JMLR: Workshop and Conference Proceedings 63:97–109, 2016 ACML 2016

Linearized Alternating Direction Method of Multipliers for Constrained Nonconvex Regularized Optimization

Linbo Qiao[†‡] QIAO.LINBO@NUDT.EDU.CN
Bofeng Zhang[†] BFZHANG@NUDT.EDU.CN
Jinshu Su[†‡] SJS@NUDT.EDU.CN
Xicheng Lu[†‡] XCLU@NUDT.EDU.CN
[†] *College of Computer, National University of Defense Technology, ChangSha 410073, China*
[‡] *National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, ChangSha 410073, China*

Editors: Robert J. Durrant and Kee-Eung Kim

Abstract

In this paper, we consider a wide class of constrained nonconvex regularized minimization problems, where the constraints are linearly constraints. It was reported in the literature that nonconvex regularization usually yields a solution with more desirable sparse structural properties beyond convex ones. However, it is not easy to obtain the proximal mapping associated with nonconvex regularization, due to the imposed linearly constraints. In this paper, the optimization problem with linear constraints is solved by the Linearized Alternating Direction Method of Multipliers (LADMM). Moreover, we present a detailed convergence analysis of the LADMM algorithm for solving nonconvex compositely regularized optimization with a large class of nonconvex penalties. Experimental results on several real-world datasets validate the efficacy of the proposed algorithm.

Keywords: LADMM, Constrainded Nonconvex Regularized Optimization

1. Introduction

In this paper, we are going to focus on solving constrained nonconvex regularized optimization problems:

$$\begin{aligned} \min_{x,z} \quad & l(x) + r(z), \\ \text{s.t.} \quad & Ax - Bz = 0 \end{aligned} \tag{1}$$

where $l : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex function associated with the prediction rule x, $r : \mathbb{R}^l \to \mathbb{R}$ is a nonconvex regularization function, $A \in \mathbb{R}^{m\times d}$, and $B \in \mathbb{R}^{m\times l}$.

When r is a convex function, problem (1) can cover graph-guided regularized minimization (Hastie et al., 2009) and generalized Lasso (Tibshirani and Taylor, 2011). However, it was reported in the literature that nonconvex regularization usually yields a solution with more desirable sparse structural properties. And people manage to impose some nonconvex regularizations on Eq. (1), which have been proven to be better approximations of ℓ_0-norm theoretically and computationally beyond ℓ_1-norm, for example, the compressive sensing (Xiao et al., 2011). The existing nonconvex regularizations include ℓ_p-norm $(0 < p < 1)$ (Foucart and Lai, 2009), Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001),

Log-Sum Penalty (LSP) (Candes et al., 2008), Minimax Concave Penalty (MCP) (Zhang, 2010a), and Capped-ℓ_1 penalty (Zhang, 2010b, 2013).

Another challenge of problem (1) comes from the linear constraints. Specifically speaking, it is very likely that the proximal mapping associated with $r(z)$ is not easy to be obtained. Fortunately, since $l(x)$ is smooth and the solution of the proximal mapping associated with $r(z)$ can be explicitly given for many commonly used nonconvex regularizers, the Linearized Alternating Direction Method of Multipliers (LADMM) (Yang and Yuan, 2013) can be applied regardless of the availability of the proximal mapping on $l(x)$. However, it remains unclear whether the LADMM algorithm converges when applied to the nonconvex problem in Eq. (1), although its global convergence is established for convex objectives (He and Yuan, 2012; Hong and Luo, 2012). This issue is addressed in this paper affirmatively. The detailed convergence analysis is presented. The efficacy of the proposed algorithm is demonstrated by encouraging empirical evaluations of nonconvex graph-guided regularized minimization on several real-world datasets.

2. Related Work

In this section, we review some existing algorithms and discuss their connections to our work. When $A = I$ and $B = I$, the commonly used approaches for solving problem (1) include the multi-stage (MS) convex relaxation (or CCCP, DC programming) algorithm (Zhang, 2010b), the sequential convex programming (SCP) algorithm (Lu, 2012), the general iterative shrinkage and thresholding (GIST) algorithm (Gong et al., 2013), and the recent hybrid optimization algorithm (HONOR) which combines the quasi-Newton method and the gradient descent method (Gong and Ye, 2015). However, the MS algorithm does not admit a closed-form solution for graph-guided regularized optimization problems and hence leads to an expensive per-iteration computational cost. When A or B is non-diagonal, neither the SCP algorithm nor the GIST algorithm is efficient for solving problem (1) since the proximal mapping of $r(\cdot)$ is typically not available.

Another related stream of works are the ADMM-type algorithms which are suitable to solve problem (1) when A or B is not diagonal (Zhong and Kwok, 2013; Zhang and Kwok, 2014; Wang et al., 2014; Zhao et al., 2015). Such a kind of algorithms have recently been shown effective to handle some nonconvex optimization problems (Magnsson et al., 2014; Jiang et al., 2014; Hong et al., 2015; Yang et al., 2015; Wang et al., 2015a,b; Li and Pong, 2015). However, the results of (Magnsson et al., 2014; Jiang et al., 2014) require a not well-justified assumption about the generated iterates, while some other works focus on certain specific problems such as the consensus and sharing problems (Hong et al., 2015) and the background/foreground extraction problems (Yang et al., 2015). The rest works (Wang et al., 2015a,b; Li and Pong, 2015) consider proximal ADMM applied to the linear constrained problems with convergence established under some mild conditions. However, they all assume that the proximal mapping of l is easily obtained, which is not the case for many objective functions encountered in machine learning, such as the logistic function.

3. Preliminaries

To proceed, we make the following assumptions (Assumptions 1-6) throughout this paper.

Assumption 1 *B is column full rank.*

Assumption 2 *$l(x)$ is continuously differentiable with Lipschitz continuous gradient, i.e., there exists a constant $L > 0$ such that*

$$\|\nabla l(x_1) - \nabla l(x_2)\| \leq L\|x_1 - x_2\|, \ \forall x_1, x_2 \in \mathbb{R}^d.$$

Assumption 3 *$l(x)$ is lower-bounded, i.e., $\inf_x l(x) \geq l^* > -\infty$. In addition, there exists $\beta_0 > 0$ such that $\bar{l}(x) = l(x) - \beta_0\|\nabla l(x)\|^2$ is lower-bounded and coercive, i.e., $\inf_x \bar{l}(x) \geq \bar{l}^* > -\infty$ and $\bar{l}(x) \to +\infty$ as $\|x\| \to +\infty$.*

We remark that Assumption 2 and Assumption 3 are not restrictive. In fact, they are easily satisfied by many popular functions in machine learning, such as least squares and logistic functions:

$$l(x) = \frac{1}{2n}\|Ax - \boldsymbol{b}\|^2 \text{ or } \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(b_i \cdot \boldsymbol{a}_i^\top x\right)\right),$$

where $A = \left[\boldsymbol{a}_1^\top; \cdots; \boldsymbol{a}_n^\top\right] \in \mathbb{R}^{n\times d}$ is a data matrix and $\boldsymbol{b} = [b_1, \cdots, b_n]^\top \in \mathbb{R}^n$. Specifically, when $l(x)$ is the least squares function, we have

$$\bar{l}(x) = \frac{1}{2n}\|Ax - \boldsymbol{b}\|^2 - \frac{\beta_0}{n^2}\|A^\top(Ax - \boldsymbol{b})\|^2.$$

Therefore, $\bar{l}(x)$ is lower-bounded and coercive when $\beta_0 \leq \frac{n}{2\lambda_{\max}(AA^\top)}$, where $\lambda_{\max}(AA^\top)$ is the largest eigenvalue of $AA^\top$. When $l(x)$ is the logistic function, $\|\nabla l(x)\|^2$ is bounded. Consequently, $\bar{l}(x)$ is lower-bounded and coercive for any $\beta_0 > 0$.

Assumption 4 *$r(x)$ is a continuous function, which is possibly nonconvex and non-smooth, can be rewritten as the difference between two convex functions, i.e.,*

$$r(x) = r_1(x) - r_2(x),$$

where $r_1(x)$ and $r_2(x)$ are convex functions. Moreover, $r(x)$ is lower-bounded, i.e., $\inf_x r(x) \geq r^* > -\infty$.

In Table 1, we present some nonconvex regularizers widely used in sparse learning, which satisfy Assumption 4.[1] It should be noted that $r(x)$ is not necessarily assumed to be coercive in our paper, which however is required in (Wang et al., 2015a; Li and Pong, 2015; Wang et al., 2015b). Indeed, this property does not hold true for some nonconvex penalty functions such as Capped-ℓ_1 regularization.

Assumption 5 *The smallest eigenvalue of $((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top$ is positive, i.e., $\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top) > 0$.*

1. We refer interested readers to (Gong et al., 2013) for the detailed decomposition of each nonconvex regularizer presented in Table 1.

Table 1: Examples of the penalty function $r(x)$ satisfying Assumption 4. $\gamma > 0$ is the regularization parameter. $[x]_+ = \max(0, x)$ and $r(x) = \sum_i r_i(x_i)$.

Name	$r_i(x_i)$
LSP	$\gamma \log\left(1 + |x_i|/\theta\right)\ (\theta > 0)$
SCAD	$\gamma \int_0^{|x_i|} \min\left(1, \frac{[\theta\gamma - y]_+}{(\theta-1)\gamma}\right) dy\ (\theta > 2) = \begin{cases} \gamma|x_i|, & \text{if } |x_i| \le \gamma, \\ \frac{-x_i^2 + 2\theta\gamma|x_i| - \gamma^2}{2(\theta-1)}, & \text{if } \gamma < |x_i| \le \theta\gamma, \\ \frac{(\theta+1)\gamma^2}{2}, & \text{if } |x_i| > \theta\gamma, \end{cases}$
MCP	$\gamma \int_0^{|x_i|} \left[1 - \frac{y}{\theta\gamma}\right]_+ dy\ (\theta > 0) = \begin{cases} \gamma|x_i| - x_i^2/(2\theta), & \text{if } |x_i| \le \theta\gamma, \\ \theta\gamma^2/2, & \text{if } |x_i| > \theta\gamma, \end{cases}$
Capped-ℓ_1	$\gamma \min\left(|x_i|, \theta\right)\ (\theta > 0)$

Assumption 6 *The critical point set of problem (1) is nonempty, i.e., there exist* x^*, $g_1^* \in \partial r_1(((B^\top B)^{-1}(B^\top A))x^*)$ *and* $g_2^* \in \partial r_2(((B^\top B)^{-1}(B^\top A))x^*)$ *such that*

$$\nabla l(x^*) + ((B^\top B)^{-1}(B^\top A))^\top (g_1^* - g_2^*) = \mathbf{0}. \tag{2}$$

Recall that x^* is called a critical point of problem (1) (Toland, 1979) when Eq. (2) holds. Moreover, the Lagrangian function of problem (1) is given by

$$\mathcal{L}(y, x, \lambda) = l(x) + r(y) - \left\langle \lambda, ((B^\top B)^{-1}(B^\top A))x - y \right\rangle,$$

and it can be easily verified that a critical point (y^*, x^*, λ^*) of the Lagrangian function satisfies:

$$\begin{aligned} \mathbf{0} &= \nabla l(x^*) - ((B^\top B)^{-1}(B^\top A))^\top \lambda^*, \\ \mathbf{0} &= g_1^* - g_2^* + \lambda^*, \\ \mathbf{0} &= ((B^\top B)^{-1}(B^\top A))x^* - y^*, \end{aligned}$$

where $g_1^* \in \partial r_1(((B^\top B)^{-1}(B^\top A))x^*)$ and $g_2^* \in \partial r_2(((B^\top B)^{-1}(B^\top A))x^*)$. Hence, x^* is a critical point of problem (1) as well.

4. Linearized Alternating Direction Method of Multipliers (LADMM)

In this section, we first review the Linearized Alternating Direction Method of Multipliers (LADMM) (Yang and Yuan, 2013), and discuss how it can be applied to solve problem (1). Then we present detailed convergence analysis of LADMM.

4.1. Algorithm

It is well known that problem (1) can be solved by the standard ADMM (Gabay and Mercier, 1976) when the proximal mappings of $l(x)$ and $r(z)$ are both easily obtained. Its typical

iteration can be written as

$$\begin{aligned} x^{k+1} &:= \underset{x}{\operatorname{argmin}}\ \mathcal{L}_\beta\left(x, z^k, \lambda^k\right), \\ \lambda^{k+1} &:= \lambda^k - \beta\left(((B^\top B)^{-1}(B^\top A))x^{k+1} - z^k\right), \\ z^{k+1} &:= \underset{z}{\operatorname{argmin}}\ \mathcal{L}_\beta\left(x^{k+1}, z, \lambda^{k+1}\right), \end{aligned}$$

where the augmented Lagrangian function $\mathcal{L}_\beta(x, z, \lambda)$ is defined as

$$\mathcal{L}_\beta(x, z, \lambda) = l(x) + r(z) - \left\langle \lambda, ((B^\top B)^{-1}(B^\top A))x - z \right\rangle + \frac{\beta}{2}\|((B^\top B)^{-1}(B^\top A))x - z\|^2.$$

The penalty parameter $\beta > 0$ is a constant, and can be seen as a dual step-size. Unfortunately, in many machine learning problems, the proximal mapping of the function $l(x)$ can not be explicitly computed, thus making ADMM inefficient. This inspires a linearized ADMM algorithm (Yang and Yuan, 2013) by linearizing $l(x)$ in the x-subproblem. Specifically, this algorithm considers a modified augmented Lagrangian function:

$$\begin{aligned} \bar{\mathcal{L}}_\beta(x, \hat{x}, z, \lambda) &= l(\hat{x}) + \langle \nabla l(\hat{x}), x - \hat{x}\rangle + r(z) - \left\langle \lambda, ((B^\top B)^{-1}(B^\top A))x - z \right\rangle \\ &\quad + \frac{\beta}{2}\|((B^\top B)^{-1}(B^\top A))x - z\|^2. \end{aligned}$$

Then the LADMM algorithm solves problem (1) by generating a sequence $\{x^{k+1}, \lambda^{k+1}, z^{k+1}\}$ as follows:

$$\begin{aligned} x^{k+1} &:= \underset{x}{\operatorname{argmin}}\ \bar{\mathcal{L}}_\beta\left(x, x^k, z^k, \lambda^k\right), \\ \lambda^{k+1} &:= \lambda^k - \beta\left(((B^\top B)^{-1}(B^\top A))x^{k+1} - z^k\right), \\ z^{k+1} &:= \underset{z}{\operatorname{argmin}}\ \bar{\mathcal{L}}_\beta\left(x^{k+1}, x^k, z, \lambda^{k+1}\right). \end{aligned} \tag{3}$$

In this paper, we slightly modify the above LADMM algorithm by imposing a proximal term on the subproblem of x and update x^{k+1} via

$$x^{k+1} := \underset{x}{\operatorname{argmin}}\ \bar{\mathcal{L}}_\beta\left(x, x^k, z^k, \lambda^k\right) + \frac{\delta}{2}\|x - x^k\|^2,$$

which leads to a closed-form solution

$$\begin{aligned} x^{k+1} &:= \left[\delta I + \beta((B^\top B)^{-1}(B^\top A))^\top((B^\top B)^{-1}(B^\top A))\right]^{-1} \\ &\quad \cdot\left[((B^\top B)^{-1}(B^\top A))^\top\lambda^k + \beta((B^\top B)^{-1}(B^\top A))^\top z^k + \delta x^k - \nabla l(x^k)\right]. \end{aligned} \tag{4}$$

The updating rule of z^{k+1} is the same as Eq. (3), and is equivalent to the proximal operator problem:

$$z^{k+1} := \underset{z}{\operatorname{argmin}}\left[\frac{1}{2}\|z - \boldsymbol{u}^k\|^2 + \frac{1}{\beta}r(z)\right], \tag{5}$$

Algorithm 1 LADMM

Choose the parameter β such that Eq. (6) is satisfied;
Initialize an iteration counter $k \leftarrow 0$ and a bounded starting point $\left(x^0, \lambda^0, z^0\right)$;
repeat
  Update x^{k+1} according to Eq. (4);
  $\lambda^{k+1} \leftarrow \lambda^k - \beta\left(((B^\top B)^{-1}(B^\top A))x^{k+1} - z^k\right)$;
  Update z^{k+1} according to Eq. (5);
  if some stopping criterion is satisfied; **then**
    Break;
  else
    $k \leftarrow k + 1$;
  end if
until exceed the maximum number of outer loop.

where $\boldsymbol{u}^k = ((B^\top B)^{-1}(B^\top A))x^{k+1} - \frac{\lambda^{k+1}}{\beta}$. For all the regularized functions listed in Table 1, the above problem has a closed-form solution even though $r(z)$ is nonconvex and nonsmooth (details are provided in (Gong et al., 2013)). Taking the Capped-ℓ_1 regularized function for example, its closed-form expression is given by

$$z_i^{k+1} := \begin{cases} x_1, & \text{if } h_i(x_1) \leq h_i(x_2), \\ x_2, & \text{otherwise,} \end{cases}$$

where $h_i(x) = \frac{1}{2}\left(x - u_i^k\right)^2 + \gamma \min(|x|, \theta)/\beta$, $x_1 = \text{sign}(u_i^k)\max\left(|u_i^k|, \theta\right)$, and $x_2 = \text{sign}(u_i^k)\min\left(\theta, \left[|u_i^k| - \gamma/\beta\right]_+\right)$. We describe the details of the LADMM algorithm in Algorithm 1.

4.2. Convergence Analysis

This subsection is dedicated to the convergence analysis of the LADMM algorithm for nonconvex regularized optimization. We first present a couple of technical lemmas as preparation.

Lemma 1 *The norm of the dual variable can be bounded by the norm of the gradient of the objective function and the iterative gap of primal variables*

$$\begin{aligned} \|\lambda^{k+1}\|^2 \quad \leq \quad & \frac{1}{\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)}\|\nabla l(x^{k+1})\|^2 \\ & + \frac{3L^2 + 3\delta^2}{\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)}\|x^{k+1} - x^k\|^2. \end{aligned}$$

Similarly, the iterative gap of dual variables can be bounded as follows:

$$\begin{aligned} \|\lambda^{k+1} - \lambda^k\|^2 \quad \leq \quad & \frac{3L^2 + 3\delta^2}{\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)}\|x^k - x^{k-1}\|^2 \\ & + \frac{3\delta^2}{\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)}\|x^{k+1} - x^k\|^2. \end{aligned}$$

To proceed, we define a potential function Φ as

$$\Phi(x,\hat{x},z,\lambda) = l(x)+r(z)-\left\langle \lambda,((B^\top B)^{-1}(B^\top A))x-z\right\rangle+\frac{\beta}{2}\|((B^\top B)^{-1}(B^\top A))x-z\|^2 + \frac{3L^2+3\delta^2}{\beta\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)}\|x-\hat{x}\|^2.$$

This function is built to measure the violation of the optimality of the current iterate. Some key properties of $\Phi(x^{k+1},x^k,z^{k+1},\lambda^{k+1})$ are stated below.

Lemma 2 *Let the sequence $\{x^{k+1},\lambda^{k+1},z^{k+1}\}$ be generated by LADMM, and δ and β satisfy that $\delta>\frac{L}{2}$ and*

$$\beta \geq \max\left\{\left(3L^2+6\delta^2\right)/\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)\left(\delta-\frac{L}{2}\right), 3/(2\beta_0\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top))\right\}, \tag{6}$$

where β_0 is defined in Assumption 3. Then $\Phi(x^{k+1},x^k,z^{k+1},\lambda^{k+1})$ is monotonously decreasing and uniformly lower-bounded.

Note that when $\delta=\frac{L}{2}+\beta_0$ in the LADMM algorithm, Eq. (6) implies that $\beta\geq(3L^2+6\delta^2)/(\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)\left(\delta-\frac{L}{2}\right))$ since

$$\begin{aligned}\frac{3L^2+6\delta^2}{\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)\left(\delta-\frac{L}{2}\right)} &= \frac{3L^2+6\delta^2}{\beta_0\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)}\\ &\geq \frac{3}{2\beta_0\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)}\end{aligned} \tag{7}$$

Theorem 3 *Let $\{x^{k+1},z^{k+1},\lambda^{k+1}\}$ be generated by LADMM, and β and δ be specified in Lemma 2. Then the sequence is bounded and has at least one limit point. Furthermore, we have*

$$\begin{aligned}\|x^{k+1}-x^k\| &\to 0,\\ \|z^{k+1}-z^k\| &\to 0,\\ \|((B^\top B)^{-1}(B^\top A))x^{k+1}-z^{k+1}\| &\to 0,\end{aligned}$$

and that any limit point of the sequence $\{x^{k+1},z^{k+1},\lambda^{k+1}\}$ is a critical point of problem (1). Finally, we have

$$\min_{0\leq k\leq n}\|x^k-x^{k+1}\|^2\leq\frac{\Phi(x^1,x^0,z^1,\lambda^1)-\Phi^*}{n\delta_{\min}}, \tag{8}$$

where Φ^* is the uniformly lower bound of $\Phi(x^{k+1},x^k,z^{k+1},\lambda^{k+1})$, and $\delta_{\min}$ is defined as

$$\delta_{\min}=\delta-\frac{L}{2}-\frac{3L^2+6\delta^2}{\beta\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)}>0.$$

We remark that $\|x^{k+1}-x^k\|^2\to 0$ is the key condition for the convergence of LADMM. The proof of Theorem 3 presented in Appendix shows that both $\|((B^\top B)^{-1}(B^\top A))x^{k+1}-z^{k+1}\|$ and $\|z^{k+1}-z^k\|$ can be bounded above by $\|x^{k+1}-x^k\|$. Therefore, $\|x^{k+1}-x^k\|^2$ can be used as a quantity to measure the convergence of the sequence generated by LADMM.

Table 2: Statistics of datasets: n is the number of samples and d is the dimensionality of the data

dataset	*classic*	*hitech*	*k1b*	*la12*	*la1*	*la2*	*reviews*	*sports*	*a9a*	*20news*	*mrms*	*w8a*	*lfcrc*
n	7094	2301	2340	2301	3204	3075	4069	8580	32561	16242	8124	64700	84776
d	41681	10080	21839	31472	31472	31472	18482	14866	123	100	112	300	234

5. Experiments

5.1. Capped-ℓ_1 Regularized Logistic Regression

In this section, we conduct the experiment to evaluate the performance of our method. We propose a novel LADMM method as shown in Algorithm 2 for tackling problem (1). The first task considered is Capped-ℓ_1 regularized logistic regression problem:

$$\begin{aligned} \min_{x,z} \quad & l(x) + \gamma \min\{\|z\|_1, \theta\} \\ \text{s.t.} \quad & x - z = 0. \end{aligned} \tag{9}$$

where l is the logistic function, which is widely used in various application fields (Chen et al., 2016), and γ is the regularization parameter. We formulate problem (9) by eliminating z as used in (Gong et al., 2013):

$$\min_{x} \; l(x) + \gamma \min\{\|x\|_1, \theta\}. \tag{10}$$

For problem (10) with a simple structure, it is not necessary to formulate it as a two-variable equality constrained optimization. Instead, we can directly solve problem (10) without any constraint by using several popular algorithms discussed in Section 2. We select the GIST algorithm as the baseline since it has been proven more effective than other competitive algorithms (Gong et al., 2013). The Barzilai-Borwein (BB) initialization and the non-monotone line search criterion are not used for a fair comparison. Furthermore, it is unfair to compare our method with the HONOR algorithm since the HONOR algorithm is the combination of quasi-Newton method and the GIST algorithm while our method is purely a first-order method.

Experiments are conducted on eight datasets[2] summarized in Table 2. They are sparse and high dimensional. We transform the multi-class datasets into two-class by labeling the first half of all classes as the positive class. For each dataset, we calculate the lipschitz constant L as its classical upper bound $\hat{L} = 0.25 \max_{1 \le i \le n} \|\boldsymbol{a_i}\|^2$. All algorithms are implemented in Matlab and executed on an Intel(R) Core(TM) CPU (i7-4710MQ@2.50GHZ) with 16GB memory, and we use the code of the GIST algorithm available online[3]. We choose the staring point of all algorithms as zero vectors. We terminate all algorithms if the relative change of the two consecutive objective function values is lower than 10^{-5} or the number of iterations exceeds 1000.

2. https://www.shi-zhong.com/software/docdata.zip
3. http://www.public.asu.edu/~pgong5/

Algorithm 2 LADMM with line search

Initialize starting point $\left(x^0, \lambda^0, z^0\right)$;
repeat
  Initialize β^k;
  repeat
    Update x^{k+1} according to Eq. (4) with β replaced by β^k;
    $\lambda^{k+1} \leftarrow \lambda^k - \beta^k \left(((B^\top B)^{-1}(B^\top A))x^{k+1} - z^k\right)$;
    if line search criterion is satisfied **then**
      Break;
    end if
    update β^k;
  until exceed the maximum number of inner loop;
  Update z^{k+1} according to Eq. (5) with β replaced by β^k;
  if stopping criterion is satisfied **then**
    Break;
  end if
  update counter state;
until exceed the maximum number of outer loop.

Figure 1 shows the objective value as the function of time with different parameter settings. We have the following observations: (1) Both LADMM-Monotone-Last and LADMM-Monotone decrease the objective function value rapidly and achieve the fastest convergence speed. Moreover, LADMM-Monotone achieves the smallest objective function values consistently. (2) LADMM-Monotone-Last and LADMM-Monotone may give rise to an increasing the objective function at the beginning but finally converges and has a faster overall convergence speed than GIST, which indicates the superiority of LADMM-type algorithms for solving (1).

5.2. Generalized Capped-ℓ_1 Regularized Logistic Regression

The proposed algorithm is more powerful for problems with complex equality constraints, for which proximal splitting methods such as GIST and HONOR are no longer applicable. An important class of these problems is called the generalized lasso (Tibshirani and Taylor, 2011):

$$\min_x \; l(x) + \gamma \|Fx\|_1, \tag{11}$$

where l is the logistic function. γ is the regularization parameter and F is a penalty matrix promoting the desired sparse structure of x. To meet the goal of exploring the sparse structure of the graph, we replace the ℓ_1-norm by the nonconvex Capped-ℓ_1 norm, and obtain the Generalized Capped-ℓ_1 regularized logistic regression:

$$\min_x \; l(x) + \gamma \min\left\{\|Fx\|_1, \theta\right\}. \tag{12}$$

By introducing $z = Fx$, problem (12) is formulated as

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{y}} \quad & l(x) + \gamma \min\left\{\|y\|_1, \theta\right\} \\ \text{s.t.} \quad & Fx - y = 0. \end{aligned} \tag{13}$$

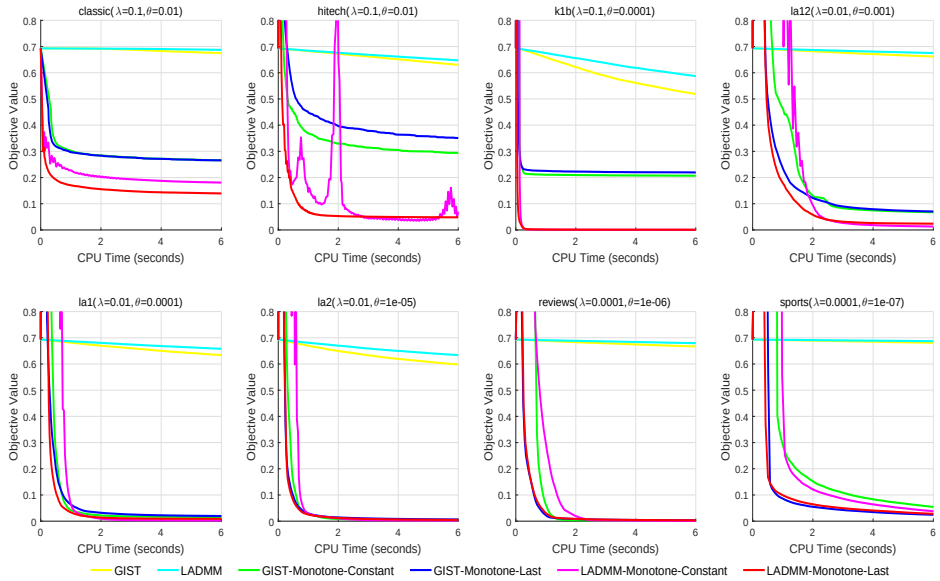

Figure 1: Objective Value vs Time of GIST and LADMM on Capped-ℓ_1 regularized logistic regression problem. The parameters of proposed method are setted exactly according to the theory analysis, while the parameters of GIST are setted by default. LADMM-Monotone-Last/GIST-Monotone-Last refer to the AdaL-ADMM/GIST algorithm using the monotone line search criterion and last rule to initialize parameters. LADMM-Monotone/GIST-Monotone refer to the AdaL-ADMM/GIST algorithm using the monotone line search criterion to initialize parameters. LADMM/GIST refer to the LADMM/GIST algorithm using the sufficiently large constant.

Experiments are conducted on five binary classification datasets: *20news*[4], *a9a*, *mushrooms*, *w8a*[5], and *lfcrc*[6]. We use 80% samples for training and 20% for testing and the regularization parameter $\lambda = 10^{-5}$ for all datasets. We generate F by sparse inverse covariance selection (Scheinberg et al., 2010).

Experimental results are presented in Figure 2. We observe that the propossed algorithm solves both problem (11) and problem (13) efficiently. Compared with ℓ_1 regularization, we observe that Capped-ℓ_1 regularization term recover the better sparse solution, which results in the smaller test loss. This coincides with some results about statistical learning (Zhang, 2010b, 2013), and further demonstrates the efficacy of the proposed algorithm for solving nonconvex compositely regularized optimization.

6. Conclusions

We presented the first detailed convergence analysis of the linearized alternating direction method of multipliers (LADMM) algorithm in solving constrained nonconvex regularized

4. www.cs.nyu.edu/~roweis/data.html.
5. https://www.csie.ntu.edu.tw/~cjlin/libsvm/.
6. London financial credit risk control (lfcrc) dataset, provided by Data Scientist Yichi Zhang.

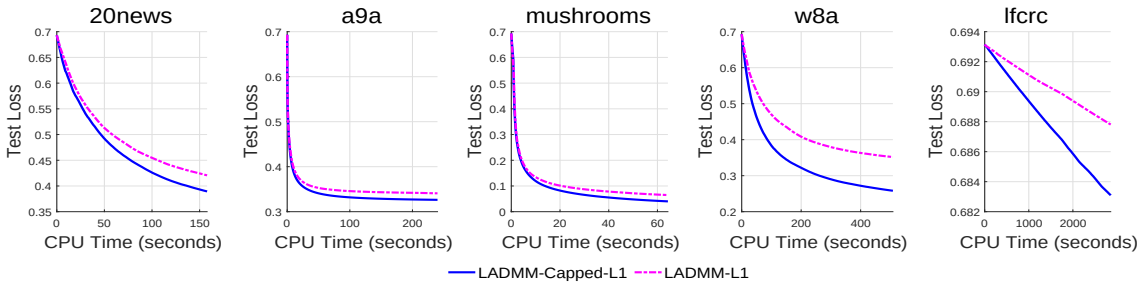

Figure 2: Test Loss vs Time on Generalized Capped-ℓ_1 regularized logistic regression and Generalized Lasso problems. For problems with complex equality constraints, for which *proximal splitting methods such as GIST and HONOR are no longer applicable.*

optimization with a large class of nonconvex penalties. It turns out that the proposed algorithm achieves the same rate of convergence as analysed. Experimental results on eight datasets demonstrated that the proposed algorithm outperforms the GIST algorithm. The proposed algorithm is well-suited for addressing constrained compositely regularized loss minimization when graph-guided regularization. In fact, the proximal splitting methods like GIST and HONOR are no longer applicable to this kind of problems. Experimental results on the other four datasets demonstrated that the proposed algorithm for solving a constrained nonconvex regularized optimization problem can attain better solutions than those obtained through solving its convex counterpart, which again validates the efficacy of the proposed algorithm.

Acknowledgments

The work was partially supported by the National Natural Science Foundation of China under Grant No. 61303264. We would like to thank Data Scientist Yichi Zhang for her kindly providing lfcrc dataset to validate our proposed algorithm.

References

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008. ISSN 1069-5869.

R Chen, M Hawes, L Mihaylova, J Xiao, and W Liu. Vehicle logo recognition by spatial-sift combined with logistic regression. *Proceedings of Fusion 2016*, 2016.

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. ISSN 0162-1459.

Simon Foucart and Ming-Jun Lai. Sparsest solutions of underdetermined linear systems via lq-minimization. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009. ISSN 1063-5203.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1): 17–40, 1976. ISSN 0898-1221.

Pinghua Gong and Jieping Ye. Honor: Hybrid optimization for non-convex regularized problems. In *Proceedings of NIPS*, pages 415–423, 2015.

Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Z Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of ICML*, volume 28, page 37. NIH Public Access, 2013.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, volume 2. Springer, 2009.

Bingsheng He and Xiaoming Yuan. On the o(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012. ISSN 0036-1429.

Mingyi Hong and Zhi-Quan Luo. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012.

Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. In *Proceedings of ICASSP*, pages 3836–3840. IEEE, 2015.

Bo Jiang, Shiqian Ma, and Shuzhong Zhang. Alternating direction method of multipliers for real and complex polynomial optimization models. *Optimization*, 63(6):883–898, 2014. ISSN 0233-1934.

Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015. ISSN 1052-6234.

Zhaosong Lu. Sequential convex programming methods for a class of structured nonlinear programming. *arXiv preprint arXiv:1210.3039*, 2012.

Sindri Magnsson, Pradeep Chathuranga, Michael Rabbat, and Carlo Fischione. On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. *arXiv preprint arXiv:1409.8033*, 2014. ISSN 2325-5870.

Katya Scheinberg, Shiqian Ma, and Donald Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *Proceedings of NIPS*, pages 2101–2109, 2010.

Ryan J Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011. ISSN 0090-5364.

JF Toland. A duality principle for non-convex optimisation and the calculus of variations. *Archive for Rational Mechanics and Analysis*, 71(1):41–61, 1979. ISSN 0003-9527.

Fenghui Wang, Wenfei Cao, and Zongben Xu. Convergence of multi-block bregman admm for nonconvex composite problems. *arXiv preprint arXiv:1505.03063*, 2015a.

Huahua Wang, Arindam Banerjee, and Zhi-Quan Luo. Parallel direction method of multipliers. In *Proceedings of NIPS*, pages 181–189, 2014.

Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*, 2015b.

Jingjing Xiao, Haoxing Yang, and Wusheng Luo. Compressed sensing by wavelet-based contourlet transform. In *International Conference on Information Technology, Computer Engineering and Management Sciences*, pages 75–78, 2011.

J. F. Yang and X. M. Yuan. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, 82(281):301–329, Jan 2013. ISSN 0025-5718.

Lei Yang, Ting Kei Pong, and Xiaojun Chen. Alternating direction method of multipliers for nonconvex background/foreground extraction. *arXiv preprint arXiv:1506.07029*, 2015.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010a. ISSN 0090-5364.

Ruiliang Zhang and James Kwok. Asynchronous distributed admm for consensus optimization. In *Proceedings of ICML*, pages 1701–1709, 2014.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:1081–1107, 2010b. ISSN 1532-4435.

Tong Zhang. Multi-stage convex relaxation for feature selection. *Bernoulli*, 19(5B):2277–2293, 2013. ISSN 1350-7265.

Peilin Zhao, Jinwei Yang, Tong Zhang, and Ping Li. Adaptive stochastic alternating direction method of multipliers. In *Proceedings of ICML*, pages 69–77, 2015.

Leon Wenliang Zhong and James T Kwok. Fast stochastic alternating direction method of multipliers. In *Proceedings of ICML*, 2013.