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, Constrained Nonconvex Regularized Optimization

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

\[
\min_{x, z} \quad l(x) + r(z), \\
\text{s.t.} \quad Ax - Bz = 0
\] (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 \( \ell_0 \)-norm theoretically and computationally beyond \( \ell_1 \)-norm, for example, the compressive sensing (Xiao et al., 2011). The existing nonconvex regularizations include \( \ell_p \)-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-\(\ell_1\) 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 (Magnusson 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 (Magnusson 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:
\[
l(x) = \frac{1}{2n}\|Ax - b\|^2 \text{ or } \frac{1}{n}\sum_{i=1}^n \log \left(1 + \exp \left(b_i \cdot a_i^\top x\right)\right),
\]
where $A = [a_1^\top; \cdots; a_n^\top] \in \mathbb{R}^{n \times d}$ is a data matrix and $b = [b_1, \cdots, b_n]^\top \in \mathbb{R}^n$. Specifically, when $l(x)$ is the least squares function, we have
\[
\bar{l}(x) = \frac{1}{2n}\|Ax - b\|^2 - \frac{\beta_0}{n^2}\|A^\top (Ax - b)\|^2.
\]

Therefore, $\bar{l}(x)$ is lower-bounded and coercive when $\beta_0 \leq \frac{\lambda_{\max}(AA^\top)}{2\lambda_{\max}(AA^\top)}$, 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-$\ell_1$ 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 $r(x) = \sum_i r_i(x_i)$.

<table>
<thead>
<tr>
<th>Name</th>
<th>$r_i(x_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSP</td>
<td>$\gamma \log (1 +</td>
</tr>
<tr>
<td>SCAD</td>
<td>$\gamma \int_0^{</td>
</tr>
<tr>
<td>MCP</td>
<td>$\gamma \int_0^{</td>
</tr>
<tr>
<td>Capped-$\ell_1$</td>
<td>$\gamma \min (</td>
</tr>
</tbody>
</table>

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

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

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

$$L(y, x, \lambda) = l(x) + r(y) - \left< \lambda, ((B^T B)^{-1}(B^T A))x - y \right>,$$

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

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

where $g_1^* \in \partial r_1((B^T B)^{-1}(B^T A)x^*)$ and $g_2^* \in \partial r_2(((B^T B)^{-1}(B^T A))x^*)$. 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

\[ x^{k+1} := \text{argmin}_x \mathcal{L}_\beta \left(x, z^k, \lambda^k\right), \]

\[ \lambda^{k+1} := \lambda^k - \beta \left( ((B^\top B)^{-1}(B^\top A)) x^{k+1} - z^k \right), \]

\[ z^{k+1} := \text{argmin}_z \mathcal{L}_\beta \left(x^{k+1}, z, \lambda^{k+1}\right), \]

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

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

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 \( l(x) \) cannot 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:

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

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

\[ x^{k+1} := \text{argmin}_x \tilde{\mathcal{L}}_\beta \left(x, x^k, z^k, \lambda^k\right), \]

\[ \lambda^{k+1} := \lambda^k - \beta \left( ((B^\top B)^{-1}(B^\top A)) x^{k+1} - z^k \right), \]

\[ z^{k+1} := \text{argmin}_z \tilde{\mathcal{L}}_\beta \left(x^{k+1}, x^k, z, \lambda^{k+1}\right). \] (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

\[ x^{k+1} := \text{argmin}_x \tilde{\mathcal{L}}_\beta \left(x, x^k, z^k, \lambda^k\right) + \frac{\delta}{2} \|x - x^k\|^2, \]

which leads to a closed-form solution

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

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

\[ z^{k+1} := \text{argmin}_z \left[ \frac{1}{2} \|z - x^k\|^2 + \frac{1}{\beta} r(z) \right], \] (5)
Algorithm 1  LADMM

Choose the parameter $\beta$ such that Eq. (6) is satisfied;
Initialize an iteration counter $k \leftarrow 0$ and a bounded starting point $(x^0, \lambda^0, z^0)$;

repeat
  Update $x^{k+1}$ according to Eq. (4);
  $\lambda^{k+1} \leftarrow \lambda^k - \beta ((B^\top B)^{-1}(B^\top A)) x^{k+1} - z^k$;
  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 $u^k = ((B^\top B)^{-1}(B^\top A)) x^{k+1} - \frac{\lambda^{k+1}}{\beta}$. For all the regularized functions listed in Table 1, the above problem has a closed-form solution even though $r(z)$ is nonconvex and non-smooth (details are provided in (Gong et al., 2013)). Taking the Capped-$\ell_1$ regularized function for example, its closed-form expression is given by

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

where $h_i(x) = \frac{1}{2} (x - u^k_i)^2 + \gamma \min(|x|, \theta) / \beta$, $x_1 = \text{sign}(u^k_i) \max(|u^k_i|, \theta)$, and $x_2 = \text{sign}(u^k_i) \min(\theta, [\frac{|u^k_i| - \gamma / \beta}{+}])$. 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 non-convex 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

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

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

$$\|\lambda^{k+1} - \lambda^k\|^2 \leq \frac{3L^2 + 3\delta^2}{\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)} \|x^k - x^{k-1}\|^2 + \frac{3\delta^2}{\lambda_{\min}(((B^\top B)^{-1}(B^\top A))((B^\top B)^{-1}(B^\top A))^\top)} \|x^{k+1} - x^k\|^2.$$
To proceed, we define a potential function \( \Phi \) as
\[
\Phi(x, \hat{x}, z, \lambda) = l(x) + r(z) - \left\langle \lambda, (B^T B)^{-1}(B^T A)x - z \right\rangle + \frac{\beta}{2}\|((B^T B)^{-1}(B^T A))x - z\|^2 \\
+ \frac{3L^2 + 3\delta^2}{\lambda_{\min}(((B^T B)^{-1}(B^T A))(B^T B)^{-1}(B^T A))^\top)}\|x - \hat{x}\|^2.
\]
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 \( \delta \) and \( \beta \) satisfy that \( \delta > \frac{L}{2} \) and
\[
\beta \geq \max \left\{ \left(3L^2 + 6\delta^2\right)/\lambda_{\min}(((B^T B)^{-1}(B^T A))(B^T B)^{-1}(B^T A))^\top) \left(\delta - \frac{L}{2}\right), \right. \\
\left. 3/(2\beta_0\lambda_{\min}(((B^T B)^{-1}(B^T A))(B^T B)^{-1}(B^T A))^\top) \right\},
\]
where \( \beta_0 \) 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 \( \beta \geq (3L^2 + 6\delta^2)/(\lambda_{\min}(((B^T B)^{-1}(B^T A))(B^T B)^{-1}(B^T A))^\top) (\delta - \frac{L}{2}) \) since
\[
\frac{3L^2 + 6\delta^2}{\lambda_{\min}(((B^T B)^{-1}(B^T A))(B^T B)^{-1}(B^T A))^\top)} = \frac{3L^2 + 6\delta^2}{\beta_0\lambda_{\min}(((B^T B)^{-1}(B^T A))(B^T B)^{-1}(B^T A))^\top)} \geq \frac{3L^2 + 6\delta^2}{2\beta_0\lambda_{\min}(((B^T B)^{-1}(B^T A))(B^T B)^{-1}(B^T A))^\top)}.
\]

**Theorem 3** Let \( \{x^{k+1}, z^{k+1}, \lambda^{k+1}\} \) be generated by LADMM, and \( \beta \) and \( \delta \) be specified in Lemma 2. Then the sequence is bounded and has at least one limit point. Furthermore, we have
\[
\|x^{k+1} - x^k\| \to 0, \\
\|z^{k+1} - z^k\| \to 0, \\
\|((B^T B)^{-1}(B^T A))x^{k+1} - z^{k+1}\| \to 0,
\]
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
\[
\min_{0 \leq k \leq n} \|x^k - x^{k+1}\|^2 \leq \frac{\Phi(x^1, x^0, z^1, \lambda^1) - \Phi^*}{n\delta_{\min}},
\]
where \( \Phi^* \) is the uniformly lower bound of \( \Phi(x^{k+1}, x^k, z^{k+1}, \lambda^{k+1}) \), and \( \delta_{\min} \) is defined as
\[
\delta_{\min} = \delta - \frac{L}{2} - \frac{3L^2 + 6\delta^2}{\beta_0\lambda_{\min}(((B^T B)^{-1}(B^T A))(B^T B)^{-1}(B^T A))^\top)} > 0.
\]

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 \( \|((B^T B)^{-1}(B^T A))x^{k+1} - z^{k+1}\| \) 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

<table>
<thead>
<tr>
<th>dataset</th>
<th>classic</th>
<th>hitech</th>
<th>k1b</th>
<th>la12</th>
<th>la1</th>
<th>la2</th>
<th>reviews</th>
<th>sports</th>
<th>a9a</th>
<th>20news</th>
<th>mrms</th>
<th>w8a</th>
<th>lfcrc</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>7094</td>
<td>2301</td>
<td>2340</td>
<td>2301</td>
<td>3204</td>
<td>3075</td>
<td>4069</td>
<td>8580</td>
<td>32561</td>
<td>16242</td>
<td>8124</td>
<td>64700</td>
<td>84776</td>
</tr>
<tr>
<td>$d$</td>
<td>41681</td>
<td>10080</td>
<td>31839</td>
<td>31472</td>
<td>18482</td>
<td>14866</td>
<td>123</td>
<td>100</td>
<td>112</td>
<td>300</td>
<td>234</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Experiments

5.1. Capped-$\ell_1$ 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-$\ell_1$ regularized logistic regression problem:

$$\min_{x,z} l(x) + \gamma \min \{ \|z\|_1, \theta \}$$

s.t. $x - z = 0$.

(9)

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

$$\min_x l(x) + \gamma \min \{ \|x\|_1, \theta \}.$$  

(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 $\hat{L} = 0.25 \max_{1 \leq i \leq n} \|a_i\|^2$. 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 \((x^0, \lambda^0, z^0)\);

repeat

  Initialize \(\beta^k\);

  repeat

    Update \(x^{k+1}\) according to Eq. (4) with \(\beta\) replaced by \(\beta^k\);

    \(\lambda^{k+1} \leftarrow \lambda^k - \beta^k \left( ((B^\top B)^{-1}(B^\top A) \right) x^{k+1} - z^k \);

    if line search criterion is satisfied then
      Break;
    end if

    update \(\beta^k\);

  until exceed the maximum number of inner loop;

  Update \(z^{k+1}\) according to Eq. (5) with \(\beta\) replaced by \(\beta^k\);

  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-\(\ell_1\) 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):

\[
\min_x l(x) + \gamma \|Fx\|_1, \quad (11)
\]

where \(l\) is the logistic function, \(\gamma\) 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 \(\ell_1\)-norm by the nonconvex Capped-\(\ell_1\) norm, and obtain the Generalized Capped-\(\ell_1\) regularized logistic regression:

\[
\min_x l(x) + \gamma \min \{\|Fx\|_1, \theta\}. \quad (12)
\]

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

\[
\begin{align*}
\min_{x, y} & \quad l(x) + \gamma \min \{\|y\|_1, \theta\} \\
\text{s.t.} & \quad Fx - y = 0. \quad (13)
\end{align*}
\]
Figure 1: Objective Value vs Time of GIST and LADMM on Capped-\(\ell_1\) 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 AdaLADMM/GIST algorithm using the monotone line search criterion and last rule to initialize parameters. LADMM-Monotone/GIST-Monotone refer to the AdaLADMM/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 proposed algorithm solves both problem (11) and problem (13) efficiently. Compared with \(\ell_1\) regularization, we observe that Capped-\(\ell_1\) 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/~cjilin/libsvm/.
\(^6\) London financial credit risk control (lfcrc) dataset, provided by Data Scientist Yichi Zhang.
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


