TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/09 pp328–341
Volume 22, Number 3, June 2017

Adaptive Linearized Alternating Direction Method of Multipliers for Non-Convex Compositely Regularized Optimization Problems

Linbo Qiao, Bofeng Zhang, Xicheng Lu, and Jinshu Su*

Abstract: We consider a wide range of non-convex regularized minimization problems, where the non-convex regularization term is composite with a linear function engaged in sparse learning. Recent theoretical investigations have demonstrated their superiority over their convex counterparts. The computational challenge lies in the fact that the proximal mapping associated with non-convex regularization is not easily obtained due to the imposed linear composition. Fortunately, the problem structure allows one to introduce an auxiliary variable and reformulate it as an optimization problem with linear constraints, which can be solved using the Linearized Alternating Direction Method of Multipliers (LADMM). Despite the success of LADMM in practice, it remains unknown whether LADMM is convergent in solving such non-convex compositely regularized optimizations. In this research, we first present a detailed convergence analysis of the LADMM algorithm for solving a non-convex compositely regularized optimization problem with a large class of non-convex penalties. Furthermore, we propose an Adaptive LADMM (AdaLADMM) algorithm with a line-search criterion. Experimental results on different genres of datasets validate the efficacy of the proposed algorithm.

Key words: adaptive linearized alternating direction method of multipliers; non-convex compositely regularized optimization; capped-l_1 regularized logistic regression

1 Introduction

In this research, we focus on solving a class of non-convex compositely regularized learning problems:

$$\min_{x\in\mathbb{R}^d} \; l(x) + r(Fx) \tag{1}$$

where $r : \mathbb{R}^l \to \mathbb{R}$ is a non-convex regularization function and $l : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex function associated with the prediction rule x. Furthermore, we denote $F \in \mathbb{R}^{l\times d}$ as a penalty matrix (not necessarily diagonal) specifying the desired structural sparsity pattern in x.

When r is a convex function, Formula (1) can cover graph-guided regularized minimization[1] and generalized Lasso[2]. However, non-convex regularization usually yields a solution with more desirable structural properties. Let us take the ℓ_0-norm regularized least-squares problem (i.e., l is a least-squares function) as an example. It is well known that such a problem is NP-hard because of its combinatorial nature. To this end, the ℓ_1-norm regularized model was proposed to pursue the computational tractability and has been widely used in signal and image processing[3,4], biomedical informatics[5], and computer vision[6]. And there is potential advances in high performance computing[7], tracking[8], energy saving[9], vehicular ad hoc network[10], and hierarchical

• Linbo Qiao, Xicheng Lu, and Jinshu Su are with College of Computer, National University of Defense Technology, and National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha 410073, China. E-mail: sjs@nudt.edu.cn.
• Bofeng Zhang is with College of Computer, National University of Defense Technology, Changsha 410073, China.
* To whom correspondence should be addressed.
Manuscript received: 2016-07-19; revised: 2016-10-04; accepted: 2016-10-20

reinforcement learning[11]. Despite of computational advantages and successful applications, the ℓ_1 model has some limits under certain scenarios[12], because the ℓ_1-norm comes at the price of shifting the resulting estimator by a constant[13] and hence leads to an over penalized problem. To circumvent the issues pertaining to the ℓ_1-norm, researchers impose some non-convex regularizations on problem (1), which have been proven to provide better approximations of the ℓ_0-norm theoretically and computationally. The existing non-convex regularizations include the ℓ_p-norm ($0 < p < 1$)[14], Smoothly Clipped Absolute Deviation (SCAD)[13], Log-Sum Penalty (LSP)[12], Minimax Concave Penalty (MCP)[15], and Capped-ℓ_1 penalty[16,17].

In addition to the non-convex structures, another challenge of problem (1) derives from the linear composition. Specifically, when F is not diagonal, it is very likely that the proximal mapping associated with $r(Fx)$ is not easily obtained. A standard technique which is useful in this case is to introduce an auxiliary variable z and reformulate problem (1) with linear constraints as follows:

$$\begin{aligned} \min_{x,\, z} \quad & l(x) + r(z), \\ \text{s.t.} \quad & Fx - z = 0 \end{aligned} \tag{2}$$

Moreover, since $l(x)$ is smooth and the solution of the proximal mapping associated with $r(z)$ can be explicitly given for many commonly used non-convex regularizers, the Linearized Alternating Direction Method of Multipliers (LADMM)[18] can be applied regardless of the availability of the proximal mapping on $l(x)$. However, it is unclear whether the LADMM algorithm converges when applied to the non-convex problem in problem (2), although its global convergence is established for convex objectives[19,20]. This issue is successfully addressed in this research affirmatively. Moreover, we propose a novel Adaptive LADMM (AdaLADMM) algorithm for solving problem (2), which achieves faster convergence by incorporating a line-search criterion into determining an appropriate penalty parameter at each iteration compared to the LADMM algorithm. The detailed convergence analysis is presented for both the LADMM and AdaLADMM algorithms. The efficacy of the proposed AdaLADMM algorithm is demonstrated by encouraging empirical evaluations of the non-convex graph-guided regularized minimization on several real-world datasets.

This research provides the first convergence analysis of the LADMM algorithm attemping to solve non-convex compositely regularized optimization problems, which is an extension to the prior research[21]. In addition, this research shows that the convergence can still be guaranteed if the penalty parameter is adaptively adjusted. In particular, this penalty parameter is determined according to a line-search criterion at each step of the AdaLADMM algorithm. Numerically, this adaptive strategy for the penalty parameter can lead to faster convergence, as observed in Ref. [22]. Therefore, our results can be viewed as a partial justification of this phenomenon from the theoretical perspective.

2 Related Work

In this section, we review some existing algorithms and discuss their connections to our research. When $F = I$, the commonly used approaches for solving problem (1) include the Multi-Stage (MS) convex relaxation algorithm[16], the Sequential Convex Programming (SCP) algorithm[23], the General Iterative Shrinkage and Thresholding (GIST) algorithm[24], and the recent Hybrid Optimization for NOn-convex Regularized problems (HONOR) combining the quasi-Newton and gradient descent methods[25].

To be specific, the MS algorithm reformulates problem (1) as follows:

$$\min_{x \in \mathbb{R}^d} \; f_1(x) - f_2(x),$$

where $f_1(x)$ and $f_2(x)$ are both convex functions. It solves problem (1) by generating a sequence $\{x^k\}$ via

$$x^{k+1} := \operatorname*{argmin}_{x \in \mathbb{R}^d} \; f_1(x) - f_2(x^k) - \langle g, x - x^k \rangle \tag{3}$$

where $g \in \partial f_2(x^k)$ and $\partial f_2(x^k)$ is the sub-differential of the function $f_2(x)$ at $x = x^k$. Although problem (3) is a convex optimization problem, it does not admit a closed-form solution in general and hence leads to an expensive computational cost per iteration. In contrast, the SCP and GIST algorithms solve problem (1) by generating a sequence $\{x^k\}$ as follows:

$$x^{k+1} := \operatorname*{argmin}_{x \in \mathbb{R}^d} \Big[\langle \nabla l(x^k), x - x^k \rangle + \frac{t^k}{2} \|x - x^k\|^2 + r_1(x) - r_2(x^k) - \langle s_2, x - x^k \rangle \Big],$$

and

$$x^{k+1} := \operatorname*{argmin}_{x \in \mathbb{R}^d} \langle \nabla l(x^k), x - x^k \rangle + \frac{t^k}{2} \|x - x^k\|^2 + r(x),$$

respectively, where $s_2 \in \partial r_2(x^k)$ and r_1 and r_2 are defined in Assumption 3 shown in Section 3. The HONOR algorithm improves the GIST algorithm by using the second-order information of $l(x)$. Specifically, it designs a framework to determine whether we can perform a quasi-Newton step at each iteration, which greatly speeds up the convergence. However, when F is non-diagonal, neither the SCP algorithm nor the GIST algorithm is efficient for solving problem (1) as the proximal mapping of $r(Fx)$ is typically not available.

Other related studies include the ADMM-type algorithms that are suitable to solve problem (2) when F is not diagonal[26–29]. Such algorithms have recently been shown to effectively manage some non-convex optimization problems[30–36]. However, the results of Refs. [30, 31] require a not-well-justified assumption about the generated iterations, while some other studies focus on certain specific problems, such as the consensus and sharing problems[32] and the background/foreground extraction problems[33]. Several studies[34–36] consider proximal ADMM applied to the following problem:

$$\begin{aligned} \min_{x,z} \quad & l(x) + r(z), \\ \text{s.t.} \quad & Ax + Bz = b \end{aligned} \tag{4}$$

where $x \in \mathbb{R}^d$, $z \in \mathbb{R}^l$, $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{p\times l}$, and $b \in \mathbb{R}^p$. The convergence is established under some mild conditions. Obviously, the above problem includes problem (2) as a special case. However, all these algorithms 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 throughout this research.

Assumption 1 $l(x)$ is continuously differentiable with the Lipschitz continuous gradient, i.e., there exists a constant $L > 0$ such that

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

Assumption 2 $l(x)$ is lower-bounded, i.e., $\inf_x l(x) \geqslant l^* > -\infty$. 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., $\inf_x \bar{l}(x) \geqslant \bar{l}^* > -\infty$ and $\bar{l}(x) \to +\infty$ as $\|x\| \to +\infty$.

We remark that Assumptions 1 and 2 are not restrictive. In fact, they are easily satisfied by many popular functions in machine learning, such as the least-squares and logistic functions:

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

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 $\beta_0 \leqslant \dfrac{n}{2\lambda_{\max}(AA^\top)}$, 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 3 $r(x)$ is a continuous function, which is possibly *non-convex* and *non-smooth*, and can be rewritten as the difference between two convex functions, namely,

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

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

The merit of the above decomposition over $r(x)$ is that the sub-differential is non-empty for any convex function. Thus, the optimality condition of the sub-problem associated with $r(x)$ is readily obtained. In Table 1, we list some non-convex regularizers widely used in sparse learning, which satisfy Assumption 3. We refer interested readers to Ref. [24] for the detailed decomposition of each non-convex regularizer presented in Table 1. It should be noted that $r(x)$ is not necessarily assumed to be coercive in our research, which however is required in Refs. [34–36]. Indeed, this property does not hold true for some non-convex penalty functions, such as the Capped-ℓ_1 regularization.

Assumption 4 The smallest eigenvalue of $FF^\top$ is positive, i.e., $\lambda_{\min}(FF^\top) > 0$.

Assumption 5 The critical point set of problem (1) is non-empty, i.e., there exist x^*, $g_1^* \in \partial r_1(Fx^*)$, and $g_2^* \in \partial r_2(Fx^*)$ such that

$$\nabla l(x^*) + F^\top(g_1^* - g_2^*) = 0 \tag{5}$$

Recall that x^* is called a critical point of problem (1)[37] when Eq. (5) holds. Moreover, the Lagrangian function of problem (2) is expressed as follows:

$$\mathcal{L}(y, x, \lambda) = l(x) + r(y) - \langle \lambda, Fx - y \rangle,$$

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

Table 1 Examples of the penalty function $r(x)$ and the corresponding convex functions $r_1(x)$ and $r_2(x)$. $\gamma>0$ is the regularization parameter. $[w]_+ = \max(0, w)$, $r(x) = \sum_i r_i(x_i)$, $r_1(x) = \sum_i r_{1,i}(x_i)$, and $r_2(x) = \sum_i r_{2,i}(x_i)$.

	$r_i(x_i)$	$r_{1,i}(x_i)$	$r_{2,i}(x_i)$
LSP	$\gamma \log(1+\lvert x_i\rvert/\theta)$ $(\theta>0)$	$\gamma\lvert x_i\rvert$	$\gamma(\lvert x_i\rvert - \log(1+\lvert x_i\rvert/\theta))$
SCAD	$\gamma\int_0^{\lvert x_i\rvert}\min\left(1,\frac{[\theta\gamma-y]_+}{(\theta-1)\gamma}\right)\mathrm{d}y(\theta>2)=$ $\begin{cases}\gamma\lvert x_i\rvert, & \text{if } \lvert x_i\rvert\leqslant\gamma;\\ \frac{-x_i^2+2\theta\gamma\lvert x_i\rvert-\gamma^2}{2(\theta-1)}, & \text{if } \gamma<\lvert x_i\rvert\leqslant\theta\gamma;\\ \frac{(\theta+1)\gamma^2}{2}, & \text{if } \lvert x_i\rvert>\theta\gamma\end{cases}$	$\gamma\lvert x_i\rvert$	$\gamma\int_0^{\lvert x_i\rvert}\frac{[\min(\theta\gamma,y)-\gamma]_+}{(\theta-1)\gamma}\mathrm{d}y=$ $\begin{cases}0, & \text{if } \lvert x_i\rvert\leqslant\gamma;\\ \frac{x_i^2-2\gamma\lvert x_i\rvert+\gamma^2}{2(\theta-1)}, & \text{if } \gamma<\lvert x_i\rvert\leqslant\theta\gamma;\\ \gamma\lvert x_i\rvert-\frac{(\theta+1)\gamma^2}{2}, & \text{if } \lvert x_i\rvert>\theta\gamma\end{cases}$
MCP	$\gamma\int_0^{\lvert x_i\rvert}\left[1-\frac{y}{\theta\gamma}\right]_+\mathrm{d}y$ $(\theta>0)=$ $\begin{cases}\gamma\lvert x_i\rvert-x_i^2/(2\theta), & \text{if } \lvert x_i\rvert\leqslant\theta\gamma;\\ \theta\gamma^2/2, & \text{if } \lvert x_i\rvert>\theta\gamma\end{cases}$	$\gamma\lvert x_i\rvert$	$\gamma\int_0^{\lvert x_i\rvert}\min\left(1,\frac{y}{\theta\gamma}\right)\mathrm{d}y$ $(\theta>0)=$ $\begin{cases}x_i^2/(2\theta), & \text{if } \lvert x_i\rvert\leqslant\theta\gamma;\\ \gamma\lvert x_i\rvert-\theta\gamma^2/2, & \text{if } \lvert x_i\rVert\leqslant\theta\gamma\end{cases}$
Capped-ℓ_1	$\gamma\min(\lvert x_i\rvert,\theta)$ $(\theta>0)$	$\gamma\lvert x_i\rvert$	$\gamma[\lvert x_i\rvert-\theta]_+$

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

where $g_1^* \in \partial r_1(Fx^*)$ and $g_2^* \in \partial r_2(Fx^*)$. Hence, x^* is a critical point of problem (1) as well.

4 LADMM

In this section, we first review the LADMM[18] and discuss how it can be applied to solve problem (2). Then, we present a detailed convergence analysis of LADMM.

4.1 Algorithm

It is well known that problem (2) can be solved by the standard ADMM[38] when the proximal mappings of $l(x)$ and $r(z)$ are both easily obtained. Its typical iteration can be written as follows:

$$\begin{aligned}x^{k+1} &:= \operatorname*{argmin}_x \mathcal{L}_\beta\left(x, z^k, \lambda^k\right),\\ \lambda^{k+1} &:= \lambda^k - \beta\left(Fx^{k+1} - z^k\right),\\ z^{k+1} &:= \operatorname*{argmin}_z \mathcal{L}_\beta\left(x^{k+1}, z, \lambda^{k+1}\right),\end{aligned}$$

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

$$\mathcal{L}_\beta(x,z,\lambda) = l(x)+r(z)-\langle\lambda, Fx-z\rangle+\frac{\beta}{2}\|Fx-z\|^2.$$

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, thereby making ADMM inefficient in computing the proximal mapping of the function $l(x)$. This inspires a linearized ADMM algorithm[18] by linearizing $l(x)$ in the x-subproblem. In particular, this algorithm considers a modified augmented Lagrangian function:

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

Then, the LADMM algorithm solves problem (2) by generating a sequence $\left\{x^{k+1}, \lambda^{k+1}, z^{k+1}\right\}$ as follows:

$$\begin{aligned}x^{k+1} &:= \operatorname*{argmin}_x \bar{\mathcal{L}}_\beta\left(x, x^k, z^k, \lambda^k\right),\\ \lambda^{k+1} &:= \lambda^k - \beta\left(Fx^{k+1} - z^k\right),\\ z^{k+1} &:= \operatorname*{argmin}_z \bar{\mathcal{L}}_\beta\left(x^{k+1}, x^k, z, \lambda^{k+1}\right)\end{aligned} \quad (6)$$

In this research, we modify the above LADMM algorithm by imposing a proximal term on the subproblem of x and update x^{k+1} using the following equation:

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

which leads to a closed-form solution:

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

The updating rule of z^{k+1} is the same as that of Eq. (6) and is equivalent to the proximal operator problem:

$$z^{k+1} := \operatorname*{argmin}_z\left[\frac{1}{2}\|z-u^k\|^2 + \frac{1}{\beta}r(z)\right] \quad (7)$$

where $u^k = Fx^{k+1} - \dfrac{\lambda^{k+1}}{\beta}$. For all the regularized functions listed in Table 1, Eq. (7) has a closed-form solution even though $r(z)$ is non-convex and non-smooth (details are provided in Ref. [24]). Considering the capped-ℓ_1 regularized function, for example, its closed-form expression can be expressed as follows:

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

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

4.2 Convergence analysis

This subsection is dedicated to the convergence analysis for the LADMM algorithm. 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, namely:

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

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

$$\|\lambda^{k+1} - \lambda^k\|^2 \leqslant \frac{3L^2 + 3\delta^2}{\lambda_{\min}(FF^\top)}\|x^k - x^{k-1}\|^2 + \frac{3\delta^2}{\lambda_{\min}(FF^\top)}\|x^{k+1} - x^k\|^2.$$

To proceed, we define a potential function Φ_1 as follows:

Algorithm 1 LADMM

Choose the parameter β such that Formula (9) is satisfied;
Initialize an iteration counter $k \leftarrow 0$ and a bounded starting point (x^0, λ^0, z^0);
repeat
  Update x^{k+1} according to Eq. (7);
  $\lambda^{k+1} \leftarrow \lambda^k - \beta\left(Fx^{k+1} - z^k\right)$;
  Update z^{k+1} according to Eq. (7);
  if some stopping criterion is satisfied; **then**
    Break;
  else
    $k \leftarrow k + 1$;
  end if
until exceed the maximum number of outer loop.

$$\Phi_1(x, \hat{x}, z, \lambda) = l(x) + r(z) - \langle \lambda, Fx - z\rangle + \frac{\beta}{2}\|Fx - z\|^2 + \frac{3L^2 + 3\delta^2}{\beta\lambda_{\min}(FF^\top)}\|x - \hat{x}\|^2.$$

This function is built to measure the violation of the optimality of the current iteration. Some key properties of $\Phi_1(x^{k+1}, x^k, z^{k+1}, \lambda^{k+1})$ are stated below.

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

$$\beta \geqslant \max\left\{\left(3L^2 + 6\delta^2\right)/\lambda_{\min}(FF^\top)\left(\delta - \frac{L}{2}\right), 3/(2\beta_0\lambda_{\min}(FF^\top))\right\} \quad (8)$$

where β_0 is defined in Assumption 2.

Then, $\Phi_1(x^{k+1}, x^k, z^{k+1}, \lambda^{k+1})$ is monotonously decreasing and uniformly lower-bounded.

Note that when $\delta = \dfrac{L}{2} + \beta_0$ in the LADMM algorithm, Formula (8) implies that $\beta \geqslant (3L^2 + 6\delta^2)/(\lambda_{\min}(FF^\top)\left(\delta - \dfrac{L}{2}\right))$ as

$$\frac{3L^2 + 6\delta^2}{\lambda_{\min}(FF^\top)\left(\delta - \frac{L}{2}\right)} = \frac{3L^2 + 6\delta^2}{\beta_0\lambda_{\min}(FF^\top)} \geqslant \frac{3}{2\beta_0\lambda_{\min}(FF^\top)}.$$

Now, we present the convergence result of LADMM in the following theorem.

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

$$\begin{aligned} \|x^{k+1} - x^k\| &\to 0, \\ \|z^{k+1} - z^k\| &\to 0, \\ \|Fx^{k+1} - z^{k+1}\| &\to 0, \end{aligned}$$

and any limit point of the sequence $\{x^{k+1}, z^{k+1}, \lambda^{k+1}\}$ is a critical point of problem (2). Finally, we have

$$\min_{0\leqslant k\leqslant n}\|x^k - x^{k+1}\|^2 \leqslant \frac{\Phi_1(x^1, x^0, z^1, \lambda^1) - \Phi^*}{n\delta_{\min}} \quad (9)$$

where Φ^* is the uniform lower-bound of $\Phi_1(x^{k+1}, x^k, z^{k+1}, \lambda^{k+1})$, and $\delta_{\min}$ is defined as

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

We remark that $\|x^{k+1} - x^k\|^2 \to 0$ is the key condition for the convergence of Algorithm 1. The proof of Theorem 1 is presented in Appendix and it shows that both $\|Fx^{k+1} - z^{k+1}\|$ 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 Algorithm 1 (i.e., LADMM).

5 AdaLADMM

Note that Formula (8) is a sufficient condition to guarantee the convergence of LADMM. In practice, the value of β in Formula (8) could be very large, and it consequently slows down the convergence. In this section, we propose a novel AdaLADMM for solving problem (2), wherein a line-search criterion is adopted to determine an appropriate penalty parameter β^k at each iteration. The detailed convergence analysis of the AdaLADMM algorithm is provided next.

5.1 Algorithm

We first introduce the adaptive augmented Lagrangian function

$$\bar{\mathcal{L}}(x,\hat{x},z,\lambda,\beta) = l(\hat{x}) + \langle \nabla l(\hat{x}), x-\hat{x}\rangle + r(z) - \langle \lambda, Fx-z\rangle + \frac{\beta}{2}\|Fx-z\|^2$$

by allowing β in $\bar{\mathcal{L}}_\beta(x,\hat{x},z,\lambda)$ to be a variable rather than a constant. In fact, the adaptive proximal LADMM algorithm is based on the framework of Algorithm 1. In particular, we generate x^{k+1} from (x^k,λ^k,z^k) via

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

which has a solution in the form of Eq. (7) with β replaced by β^k, and z^{k+1} from (x^{k+1},λ^{k+1}) via

$$z^{k+1} := \operatorname*{argmin}_z \bar{\mathcal{L}}(x^{k+1},x^k,z,\lambda^{k+1},\beta^k).$$

Equivalently, z^k is obtained by Eq. (7) with β replaced by β^k. The detailed procedure of the AdaLADMM algorithm is presented in Algorithm 2. There are two issues that persist: how to initialize β^k, and how to select a line-search criterion at each outer iteration.

5.2 Initialization of the penalty parameter β^k

It is known that a good initialization strategy for step-size of outer iterations can greatly reduce the line-search cost and hence speed up the overall convergence of the respective algorithm. In this paper, we adopt the so-called last rule to initialize the penalty parameter. In the last rule, β^{k+1} is initialized by using the finally accepted value of β at the last iteration, i.e., the value of β^k identified by the line-search. In the experiment, we compare this strategy with the constant initialization strategy; the experimental results show the advantage of the former strategy over the latter one.

5.3 Line-search criterion

We use the monotone line-search criterion, which requires that the value of the potential function should decrease after updating x and λ. In particular, we propose to accept the penalty parameter β^k if the following monotone line-search criterion is satisfied:

$$\Phi_2(x^k,x^{k-1},z^k,\lambda^k,\beta^k) - \sigma\left(\delta-\frac{L}{2}\right)\|x^{k+1}-x^k\|^2 \geqslant \Phi_2(x^{k+1},x^k,z^{k+1},\lambda^{k+1},\beta^k),$$

where σ is a constant in the interval $(0,1)$, and the potential function Φ_2 is defined as

$$\Phi_2(x,\hat{x},z,\lambda,\beta) = l(x) - \langle \lambda, Fx-z\rangle + \frac{\beta}{2}\|Fx-z\|^2 + r(z) + \frac{3L^2+3\delta^2}{\beta\lambda_{\min}(FF^\top)}\|x-\hat{x}\|^2.$$

Algorithm 2 AdaLADMM

Choose parameters $\eta > 1$ and $\beta_{\min}, \beta_{\max}$ with $0 < \beta_{\min} < \beta_{\max} < +\infty$;
Initialize an iteration counter $k \leftarrow 0$ and a bounded starting point (x^0,λ^0,z^0);
repeat
  Initialize $\beta^k \in [\beta_{\min},\beta_{\max}]$;
  repeat
    Update x^{k+1} according to Eq. (7) with β replaced by β^k;
    $\lambda^{k+1} \leftarrow \lambda^k - \beta^k(Fx^{k+1}-z^k)$;
    if some line-search criterion is satisfied; **then**
      Break;
    else
      $\beta^k \leftarrow \eta\beta^k$;
    end if
  until exceed the maximum number of inner loop;
  Update z^{k+1} according to Eq. (7) with β replaced by β^k;
  if some stopping criterion is satisfied; **then**
    Break;
  else
    $k \leftarrow k+1$;
  end if
until exceed the maximum number of outer loop.

5.4 Convergence analysis

Here, we present the detailed convergence analysis for the AdaLADMM algorithm (Algorithm 2) using the last rule strategy. We first present a key lemma which guarantees that the monotone line-search criterion is satisfied in Algorithm 2.

Lemma 3 Let the constant $\sigma \in (0,1)$ be given and $\delta = \frac{L}{2}+\beta_0$. Then, for any integer $k \geqslant 0$, the monotone line-search criterion is satisfied whenever

$$\beta^k \geqslant \frac{3L^2+6\delta^2}{\lambda_{\min}(FF^\top)(1-\sigma)\left(\delta-\frac{L}{2}\right)}.$$

Furthermore, $\Phi_2(x^{k+1},x^k,z^{k+1},\lambda^{k+1},\beta^k) \geqslant \Phi^*$.

In the following lemma, we show that when k is

sufficiently large, a constant β^k suffices.

Lemma 4 There exists $K_0 \geqslant 0$ such that when β^k remains as a constant $\bar{\beta}$, the monotone line-search criterion is satisfied for any $k \geqslant K_0$.

Based on Lemmas 3 and 4, we present the convergence result of Algorithm 2 in the following theorem.

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

$$\begin{aligned} \|x^{k+1} - x^k\| &\to 0, \\ \|z^{k+1} - z^k\| &\to 0, \\ \|Fx^{k+1} - z^{k+1}\| &\to 0, \end{aligned}$$

and any limit point of $\{x^{k+1}, z^{k+1}, \lambda^{k+1}\}$ is a critical point of problem (2). Finally, we have

$$\min_{K_0 \leqslant k \leqslant n} \|x^k - x^{k+1}\|^2 \leqslant \frac{\Phi_2(x^{K_0}, x^{K_0-1}, z^{K_0}, \lambda^{K_0}, \bar{\beta}) - \Phi^*}{(n - K_0)\sigma_{\min}} \tag{10}$$

where $\sigma_{\min}$ is defined by $\sigma_{\min} = \sigma\left(\delta - \frac{L}{2}\right) > 0$.

6 Experiments

6.1 Capped-ℓ_1 regularized logistic regression

In this section, we conduct an experiment to evaluate the performance of our method. The first task considered is to evaluate the capped-ℓ_1 regularized logistic regression problem used in Ref. [24]:

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

where l is the logistic function and γ is the regularization parameter. We formulate problem (11) by introducing $z = x$ as follows:

$$\begin{aligned} \min_{x,z} \quad & l(x) + \gamma \min\{\|z\|_1, \theta\}, \\ \text{s.t.} \quad & x - z = 0 \end{aligned} \tag{12}$$

For problem (11) with a simple structure, it is not necessary to formulate it as a two-variable equality constrained optimization. Instead, we can directly solve problem (11) without any constraint by using several popular algorithms, which are discussed in Section 2. We select the GIST algorithm as the baseline as it has been proven more effective than other competitive algorithms[24]. The Barzilai-Borwein (BB) initialization and the non-monotone line-search criterion are not used so as to provide an unbiased comparison. Furthermore, it is unfair to compare our method with the HONOR algorithm as the HONOR algorithm is a combination of the quasi-Newton method and the GIST algorithm, whereas our method is purely a first-order method.

Experiments are conducted on eight datasets, downloaded from https://www.shi-zhong.com/software/docdata.zip. These datasets are summarized in Table 2. They are sparse and highly dimensional. We transform the multi-classes datasets into two-classes 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 $\hat{L} = 0.25 \max_{1 \leqslant i \leqslant n} \|a_i\|^2$. All algorithms are implemented in Matlab and executed on an Intel(R) Core(TM) CPU (i7-4710MQ@2.50 GHz) with 16 GB memory, and we use the code of the GIST algorithm available online from http://www.public.asu.edu/~pgong5/. We set $\sigma = 1 \times 10^{-5}$, $\eta = 1.1$, $1/\beta_{\min} = \beta_{\max} = 10^{20}$ and choose the starting point of all algorithms as zero vectors. We terminate all algorithms when the relative change of the two consecutive objective function values is lower than 1×10^{-5} or the number of iterations exceeds 1000.

Figure 1 shows the objective values as a function of time with different parameter settings. Our observations are summarized as follows:

(1) We attempted different initializations of the very first $\beta \in \{0.1, 1, 10, 100\}$, the results show that setting β to 0.1 achieves better convergence speed on these datasets.

(2) LADMM-Monotone-Last-0.1 rapidly decreased the objective function value and achieved the fastest convergence speed, indicating that adopting the monotone line-search criterion greatly accelerates the convergence speed. Moreover, LADMM-Monotone-Last-0.1 consistently achieves the smallest objective function values

(3) LADMM-Monotone-Last and LADMM-Monotone-Constant may give rise to an increasing objective function at the beginning, however, it finally converges and has a faster overall convergence speed

Table 2 Statistics of the datasets: *n* is the number of samples; *d* is the dimensionality of the data.

Dataset	*n*	*d*	Dataset	*n*	*d*
classic	7094	41 681	sports	8580	14 866
hitech	2301	10 080	a9a	32 561	123
k1b	2340	21 839	20news	16 242	100
la12	2301	31 472	mushrooms	8124	112
la1	3204	31 472	w8a	64 700	300
la2	3075	31 472	lfcrc	84 776	234
reviews	4069	18 482			

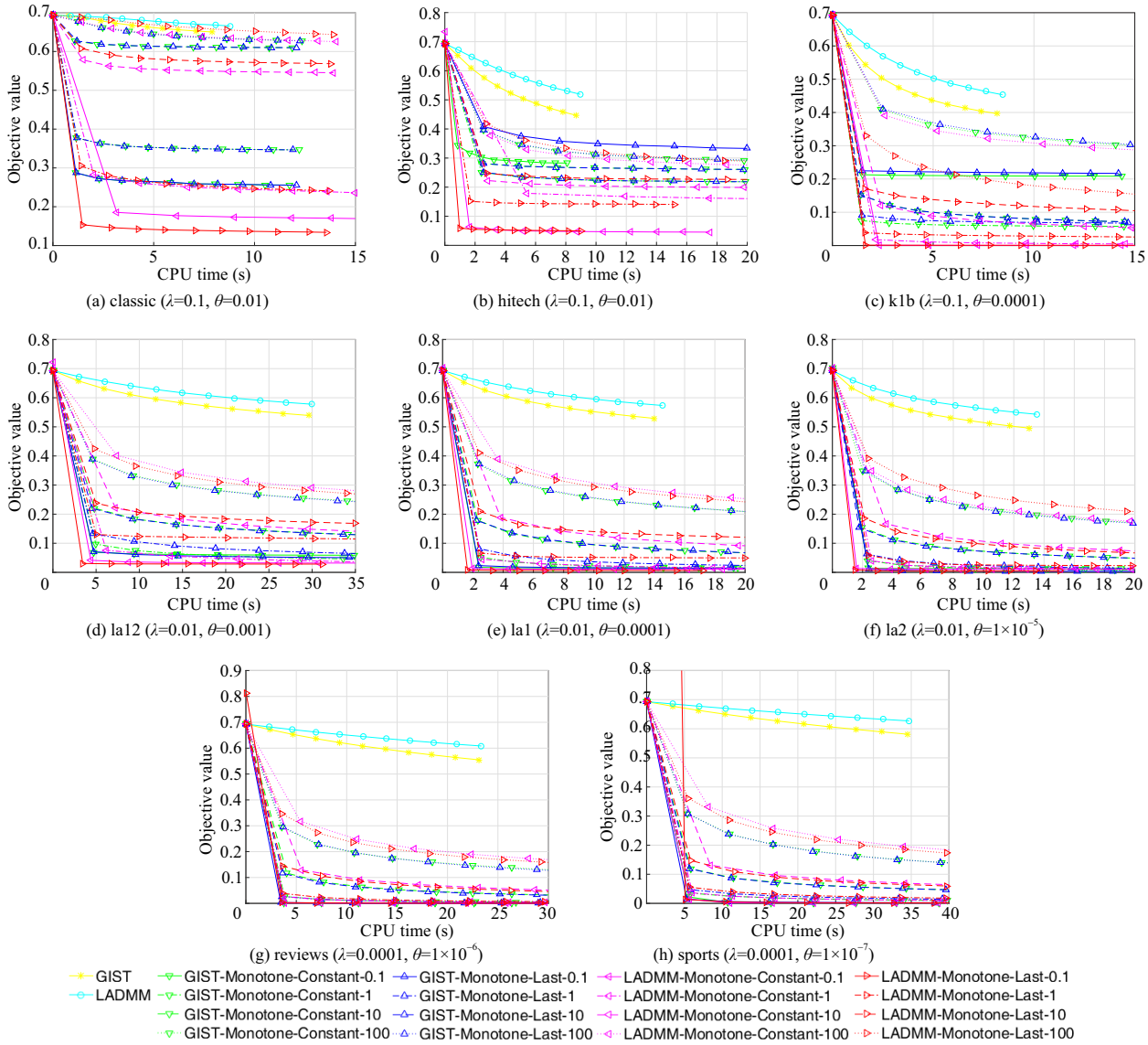


Fig. 1 Objective value as a function of CPU time on capped-ℓ_1 regularized logistic regression problem. LADMM-Monotone-Last/GIST-Monotone-Last refers to the adaptive linearized alternating direction method of multipliers (AdaLADMM)/general iterative shrinkage and thresholding (GIST) algorithm using the monotone line-search criterion and the last rule to initialize β. LADMM-Monotone-β_0/GIST-Monotone-β_0 refers to the AdaLADMM/GIST algorithm using the monotone line-search criterion and β_0 to initialize β, and $\beta_0 \in \{0.1, 1, 10, 100\}$ are compared. LADMM/GIST refers to the LADMM/GIST algorithm using the sufficiently large constant β.

than GIST-Monotone-Last and GIST-Monotone-Constant, indicating the superiority of LADMM-type algorithms for solving problem (1).

(4) LADMM performs worse than GIST as β^k in LADMM needs to be set much larger than t^k to guarantee convergence, demonstrating the significance of using the AdaLADMM algorithm.

6.2 Generalized capped-ℓ_1 regularized logistic regression

The LADMM and AdaLADMM algorithms are more powerful for solving 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[2]:

$$\min_x \ l(x) + \gamma \|Fx\|_1 \tag{13}$$

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 explore the sparse structure of the graph, we replace the ℓ_1-norm by the

non-convex capped-ℓ_1 norm and obtain the generalized capped-ℓ_1 regularized logistic regression expressed as follows:

$$\min_x \ l(x) + \gamma \min\{\|Fx\|_1, \theta\} \tag{14}$$

By introducing $z = Fx$, problem (14) is reformulated as follows:

$$\begin{aligned} \min_{x,y} \quad & l(x) + \gamma \min\{\|y\|_1, \theta\}, \\ \text{s.t.} \quad & Fx - y = 0 \end{aligned} \tag{15}$$

Experiments are conducted on five binary classification datasets: 20news, a9a, mushrooms, w8a, and lfcrc. 20news is downloaded from http://www.cs.nyu.edu/~roweis/data.html. a9a, mushrooms, and w8a are downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. lfcrc is the London financial credit risk control (lfcrc) dataset, provided by Data Scientist Yichi Zhang. We use 80% sample data for training and 20% of the data for testing and the regularization parameter $\lambda = 1 \times 10^{-5}$ for all datasets. We generate F using sparse inverse covariance selection[39]. In addition, we use the metrics in Ref. [26], and the test loss to verify the quality of the solution obtained using the AdaLADMM algorithm for solving problem (15).

Experimental results of solving the generalized capped-ℓ_1 regularized logistic regression are presented in Fig. 2. We observe that the AdaLADMM algorithm solves both problem (13) and problem (15) efficiently. Compared with ℓ_1 regularization, we observe that capped-ℓ_1 regularization term recovers a better sparse solution, which results in the smaller test loss. This coincides with the results about statistical learning[16,17] and further demonstrates the efficacy of the AdaLADMM algorithm for solving non-convex compositely regularized optimization problems.

7 Conclusion

We presented the first detailed convergence analysis of the LADMM algorithm in solving the non-convex compositely regularized optimization problem with a large number of non-convex penalties. Furthermore, we proposed an efficient adaptive LADMM algorithm with a monotone line-search criterion, which greatly accelerates the convergence speed. The results indicate that the proposed AdaLADMM algorithm achieves the same rate of convergence as that of the LADMM algorithm. Experimental results on eight datasets demonstrated that the AdaLADMM algorithm outperforms the LADMM algorithm and the GIST algorithm.

Both LADMM and AdaLADMM algorithms are well-suited for addressing compositely regularized loss minimization when the penalty matrix F is non-diagonal. In fact, the proximal splitting methods like GIST and HONOR are no longer applicable to these types problems. Experimental results for the other four datasets demonstrated that the AdaLADMM algorithm for solving a non-convex compositely regularized optimization problem can attain better solutions than those obtained through solving its convex counterpart, which again validates the efficacy of the proposed algorithm.

Appendix

A.1 Proof of Lemma 1

It follows from the updates of x^{k+1} that

$$\nabla l(x^k) - F^\top \lambda^k + \delta(x^{k+1} - x^k) +$$

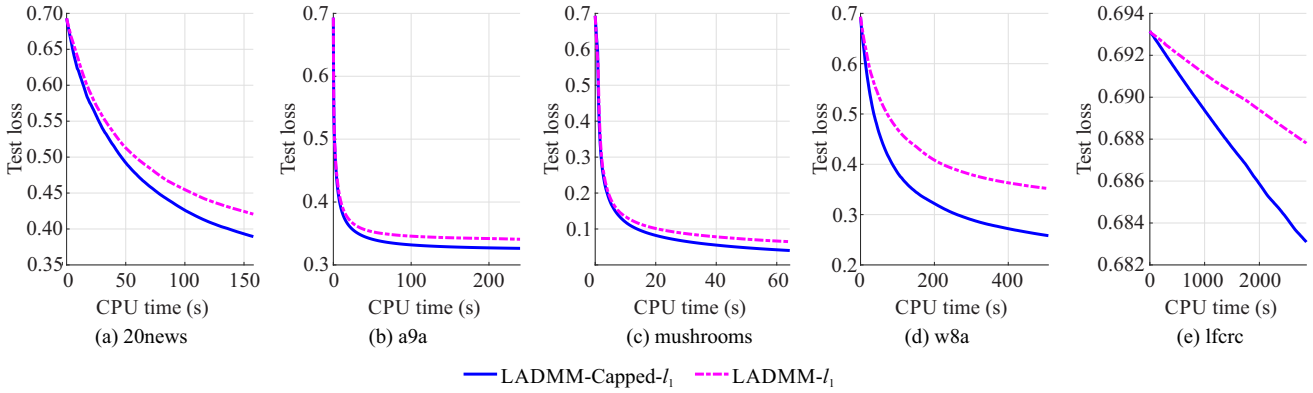


Fig. 2 Test loss as a function of time for the generalized capped-ℓ_1 regularized logistic regression and generalized Lasso problems.

$$\beta F^\top(Fx^{k+1}-z^k)=0 \tag{16}$$

Combining $\lambda^{k+1}=\lambda^k-\beta\left(Fx^{k+1}-z^k\right)$ and Eq. (16) yields that

$$\nabla l(x^k)+\delta(x^{k+1}-x^k)=F^\top\lambda^{k+1}.$$

Therefore, we conclude that

$$\begin{aligned}\|\lambda^{k+1}\|^2 \leqslant\ & \frac{1}{\lambda_{\min}(FF^\top)}\|F^\top\lambda^{k+1}\|^2= \\ & \frac{1}{\lambda_{\min}(FF^\top)}\|(\nabla l(x^k)-\nabla l(x^{k+1}))+ \\ & \delta(x^{k+1}-x^k)+\nabla l(x^{k+1})\|^2\leqslant \\ & \frac{3}{\lambda_{\min}(FF^\top)}\|\nabla l(x^{k+1})\|^2+ \\ & \frac{3L^2+3\delta^2}{\lambda_{\min}(FF^\top)}\|x^{k+1}-x^k\|^2\end{aligned} \tag{17}$$

and

$$\begin{aligned}\|\lambda^{k+1}-\lambda^k\|^2 \leqslant\ & \frac{1}{\lambda_{\min}(FF^\top)}\|F^\top\lambda^{k+1}-F^\top\lambda^k\|^2= \\ & \frac{1}{\lambda_{\min}(FF^\top)}\|(\nabla l(x^k)-\nabla l(x^{k-1}))+ \\ & \delta(x^{k+1}-x^k)-\delta(x^k-x^{k-1})\|^2\leqslant \\ & \frac{3L^2+3\delta^2}{\lambda_{\min}(FF^\top)}\|x^k-x^{k-1}\|^2+ \\ & \frac{3\delta^2}{\lambda_{\min}(FF^\top)}\|x^{k+1}-x^k\|^2.\end{aligned}$$

A.2 Proof of Lemma 2

Combining Eq. (16) and the following inequality,

$$(x^k-x^{k+1})^\top\nabla l(x^k)-l(x^k)+l(x^{k+1})\leqslant\frac{L}{2}\|x^k-x^{k+1}\|^2,$$

we have

$$\begin{aligned}0 =\ & (x^k-x^{k+1})^\top[\nabla l(x^k)-F^\top\lambda^k+ \\ & \delta(x^{k+1}-x^k)+\beta F^\top(Fx^{k+1}-z^k)]\leqslant \\ & l(x^k)-l(x^{k+1})+(\frac{L}{2}-\delta)\|x^k-x^{k+1}\|^2- \\ & \langle\lambda^k,Fx^k-z^k\rangle+\langle\lambda^k,Fx^{k+1}-z^k\rangle+ \\ & \frac{\beta}{2}\|Fx^k-z^k\|^2-\frac{\beta}{2}\|Fx^{k+1}-z^k\|^2- \\ & \frac{\beta}{2}\|Fx^{k+1}-Fx^k\|^2\end{aligned} \tag{18}$$

Then, it follows from the update of z^{k+1} that,

$$\begin{aligned}& r(z^{k+1})-\langle\lambda^{k+1},Fx^{k+1}-z^{k+1}\rangle+ \\ & \frac{\beta}{2}\|Fx^{k+1}-z^{k+1}\|^2\leqslant r(z^k)- \\ & \langle\lambda^{k+1},Fx^{k+1}-z^k\rangle+\frac{\beta}{2}\|Fx^{k+1}-z^k\|^2\end{aligned} \tag{19}$$

Combining Formulas (18) and (19) and Lemma 1 yields that:

$$\begin{aligned}& r(z^{k+1})+l(x^{k+1})-\langle\lambda^{k+1},Fx^{k+1}-z^{k+1}\rangle+ \\ & \frac{\beta}{2}\|Fx^{k+1}-z^{k+1}\|^2+(\delta-\frac{L}{2})\|x^k-x^{k+1}\|^2\leqslant \\ & r(z^k)+l(x^k)-\langle\lambda^k,Fx^k-z^k\rangle+ \\ & \frac{\beta}{2}\|Fx^k-z^k\|^2+\frac{1}{\beta}\|\lambda^{k+1}-\lambda^k\|^2\leqslant\end{aligned} \tag{20}$$

$$\begin{aligned}& r(z^k)+l(x^k)-\langle\lambda^k,Fx^k-z^k\rangle+\frac{\beta}{2}\|Fx^k-z^k\|^2+ \\ & \frac{3L^2+3\delta^2}{\beta\lambda_{\min}(FF^\top)}\|x^k-x^{k-1}\|^2+ \\ & \frac{3\delta^2}{\beta\lambda_{\min}(FF^\top)}\|x^{k+1}-x^k\|^2,\end{aligned} \tag{21}$$

which implies that

$$\begin{aligned}& \left(\delta-\frac{L}{2}-\frac{3L^2+6\delta^2}{\beta\lambda_{\min}(FF^\top)}\right)\|x^k-x^{k+1}\|^2\leqslant \\ & \Phi_1(x^k,x^{k-1},z^k,\lambda^k)-\Phi_1(x^{k+1},x^k,z^{k+1},\lambda^{k+1})\end{aligned} \tag{22}$$

where

$$\begin{aligned}\Phi_1(x,\hat{x},z,\lambda)=\ & l(x)+r(z)-\langle\lambda,Fx-z\rangle+ \\ & \frac{\beta}{2}\|Fx-z\|^2+\frac{3L^2+3\delta^2}{\beta\lambda_{\min}(FF^\top)}\|x-\hat{x}\|^2\end{aligned} \tag{23}$$

Since $\delta>\frac{L}{2}$ and $\beta>0$ satisfy that

$$\beta>(3L^2+6\delta^2)/\lambda_{\min}(FF^\top)\left(\delta-\frac{L}{2}\right),$$

we conclude that $\Phi_1(x^{k+1},x^k,z^{k+1},\lambda^{k+1})$ is monotonically decreases as k increases.

On the other hand, we have

$$\begin{aligned}& \Phi_1\left(x^{k+1},x^k,z^{k+1},\lambda^{k+1}\right)= \\ & \quad l(x^{k+1})+r(z^{k+1})-\langle\lambda^{k+1},Fx^{k+1}-z^{k+1}\rangle+ \\ & \frac{\beta}{2}\|Fx^{k+1}-z^{k+1}\|^2+ \\ & \frac{3L^2+3\delta^2}{\beta\lambda_{\min}(FF^\top)}\|x^{k+1}-x^k\|^2\geqslant \\ & l(x^{k+1})+r(z^{k+1})-\frac{1}{2\beta}\|\lambda^{k+1}\|^2- \\ & \frac{\beta}{2}\|Fx^{k+1}-z^{k+1}\|^2+\frac{\beta}{2}\|Fx^{k+1}-z^{k+1}\|^2+ \\ & \frac{3L^2+3\delta^2}{\beta\lambda_{\min}(FF^\top)}\|x^{k+1}-x^k\|^2\geqslant \\ & l(x^{k+1})+r(z^{k+1})-\frac{3}{2\beta\lambda_{\min}(FF^\top)}\|\nabla l(x^{k+1})\|^2- \\ & \frac{3L^2+3\delta^2}{2\beta\lambda_{\min}(FF^\top)}\|x^{k+1}-x^k\|^2+ \\ & \frac{3L^2+3\delta^2}{\beta\lambda_{\min}(FF^\top)}\|x^{k+1}-x^k\|^2\geqslant\end{aligned}$$

$$\begin{aligned}& l(x^{k+1})+r(z^{k+1})-\beta_0\|\nabla l(x^{k+1})\|^2= \\ & \bar{l}(x^{k+1})+r(z^{k+1})\geqslant \\ & \bar{l}^*+r^*=\Phi^*\end{aligned} \tag{24}$$

where the second inequality holds due to Formula (17) and

the third inequality holds since

$$\beta \geqslant \frac{3}{2\beta_0\lambda_{\min}(FF^\top)}.$$

Therefore, we conclude that $\Phi_1(x^{k+1}, x^k, z^{k+1}, \lambda^{k+1})$ is uniformly lowe-bounded.

A.3 Proof of Theorem 1

Combining Formula (24) and the fact that $\bar{l}(x)$ is coercive, we conclude that $\{x^{k+1}\}$ is bounded. Then, it directly follows from Formula (17) that $\{\lambda^{k+1}\}$ is bounded. Furthermore, we obtain the following from Formulas (20) and (24):

$$\left(\delta - \frac{L}{2} - \frac{3L^2+6\delta^2}{\beta\lambda_{\min}(FF^\top)}\right)\sum_{k=1}^{\infty}\|x^k - x^{k+1}\|^2 \leqslant \Phi_1(x^1, x^0, z^1, \lambda^1) - \Phi^* < +\infty \tag{25}$$

which implies that $\|x^k - x^{k+1}\| \to 0$, and hence, $\|\lambda^k - \lambda^{k+1}\| \to 0$ as $k \to +\infty$. Since $Fx^{k+1} - z^{k+1} = \frac{1}{\beta}(\lambda^k - \lambda^{k+1})$, we have $\|Fx^{k+1} - z^{k+1}\| \to 0$ which implies that $\{z^{k+1}\}$ is bounded and $\|z^k - z^{k+1}\| \to 0$ as $k \to +\infty$. In summary, we obtain that $\{x^{k+1}, z^{k+1}, \lambda^{k+1}\}$ is a bounded sequence, and

$$\|x^k - x^{k+1}\| \to 0, \quad \|z^k - z^{k+1}\| \to 0,$$
$$\|Fx^{k+1} - z^{k+1}\| \to 0.$$

Since $\{x^{k+1}, z^{k+1}, \lambda^{k+1}\}$ is bounded, this sequence must have at least one limit point. Let $\{x^*, z^*, \lambda^*\}$ be a limit point, that is, there exists a subsequence $\{k_q\}_{q=1}^{\infty}$ such that

$$\lim_{q\to+\infty}\left(x^{k_q}, z^{k_q}, \lambda^{k_q}\right) = \left(x^*, z^*, \lambda^*\right)$$

and it holds true that

$$\|x^{k_q} - x^{k_q+1}\| \to 0,$$
$$\|z^{k_q} - z^{k_q+1}\| \to 0, \quad \|Fx^{k_q+1} - z^{k_q+1}\| \to 0.$$

We consider the first-order optimality condition of updating x^{k_q+1}, z^{k_q+1}, and $r(z) = r_1(z) - r_2(z)$, i.e.,

$$0 = \nabla l(x^{k_q}) - F^\top\lambda^{k_q} + \delta\left(x^{k_q+1} - x^{k_q}\right) + \beta F^\top\left(Fx^{k_q+1} - z^{k_q}\right),$$

$$0 \in \partial r_1(z^{k_q+1}) - \partial r_2(z^{k_q+1}) + \lambda^{k_q} - \beta\left(Fx^{k_q+1} - z^{k_q+1}\right).$$

Letting $q \to +\infty$, by considering the semi-continuity of $\partial r_1(\cdot)$ and $\partial r_2(\cdot)$, we obtain that

$$0 = \nabla l(x^*) - F^\top\lambda^*,$$
$$0 \in \partial r_1(z^*) - \partial r_2(z^*) + \lambda^*.$$

Therefore, (x^*, z^*, λ^*) is a critical point.

Moreover, it follows from Formula (25) that

$$\min_{0\leqslant k\leqslant n}\|x^k - x^{k+1}\|^2 \leqslant \frac{\Phi_1(x^1, x^0, z^1, \lambda^1) - \Phi^*}{n\delta_{\min}},$$

where $\delta_{\min}$ is defined as

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

This completes the proof of Theorem 1.

A.4 Proof of Lemma 3

By the same argument as Lemma 2, we conclude that

$$\left(\delta - \frac{L}{2} - \frac{3L^2+6\delta^2}{\beta^k\lambda_{\min}(FF^\top)}\right)\|x^k - x^{k+1}\|^2 \leqslant \Phi_2(x^k, x^{k-1}, z^k, \lambda^k, \beta^k) - \Phi_2(x^{k+1}, x^k, z^{k+1}, \lambda^{k+1}, \beta^k),$$

where

$$\Phi_2(x, \hat{x}, z, \lambda, \beta) = l(x) + r(z) - \langle\lambda, Fx - z\rangle + \frac{\beta}{2}\|Fx - z\|^2 + \frac{3L^2+3\delta^2}{\beta\lambda_{\min}(FF^\top)}\|x - \hat{x}\|^2.$$

Therefore, the monotone line-search criterion is satisfied whenever

$$\left(\delta - \frac{L}{2} - \frac{3L^2+6\delta^2}{\beta^k\lambda_{\min}(FF^\top)}\right)\|x^k - x^{k+1}\|^2 \geqslant \sigma\left(\delta - \frac{L}{2}\right)\|x^k - x^{k+1}\|^2,$$

which implies that

$$\beta^k \geqslant \frac{3L^2+6\delta^2}{\lambda_{\min}(FF^\top)(1-\sigma)\left(\delta - \frac{L}{2}\right)}.$$

Since $\delta = \frac{L}{2} + \beta_0$, it also holds true that

$$\beta \geqslant \frac{3}{2\beta_0\lambda_{\min}(FF^\top)},$$

which implies that $\Phi_2(x^{k+1}, x^k, z^{k+1}, \lambda^{k+1}, \beta^k) \geqslant \Phi^*$. This completes the proof of the lemma.

A.5 Proof of Lemma 4

It is trivial to show that β^k is bounded from below, since $\beta^k \geqslant \beta_{\min}$ ($\beta_{\min}$ is defined as the AdaLADMM algorithm). Since β^k is a non-decreasing sequence, it is sufficient for us to prove that β^k is bounded from above. We prove this by contradiction. Without loss of generality, we assume that β^k increases to $+\infty$ and

$$\beta^k \geqslant \frac{\eta\left(3L^2+6\delta^2\right)}{\lambda_{\min}(FF^\top)(1-\sigma)\left(\delta - \frac{L}{2}\right)}.$$

Thus, we must try the following value of t in the previous iterations, i.e.,

$$t = \frac{\beta^k}{\eta} \geqslant \frac{3L^2+6\delta^2}{\lambda_{\min}(FF^\top)(1-\sigma)\left(\delta - \frac{L}{2}\right)}.$$

$t = \dfrac{\beta^k}{\eta}$ does not satisfy the line-search criterion. However, Lemma 3 states that the value of t is guaranteed to satisfy the monotone line-search criterion. This leads to a contradiction, and the conclusion β^k is bounded from below. Therefore, we state that there exists a $K_0 > 0$ such that β^k remains a constant $\bar{\beta} > 0$ for any $k \geqslant K_0$.

A.6 Proof of Theorem 2

From Lemma 4, we know that β^k remains as a

constant as $k \geqslant K_0$. Furthermore, the potential function $\Phi_2(x^{k+1}, x^k, z^{k+1}, \lambda^{k+1}, \beta^k)$ is uniformly lower bounded, and

$$\Phi_2(x^k, x^{k-1}, z^k, \lambda^k, \bar{\beta}) - \frac{\sigma(\delta - L/2)}{2}\|x^{k+1} - x^k\|^2 \geqslant \Phi_2(x^{k+1}, x^k, z^{k+1}, \lambda^{k+1}, \bar{\beta}) \tag{26}$$

By the same argument as Theorem 1, we obtain that $\{x^{k+1}, z^{k+1}, \lambda^{k+1}\}$ is a bounded sequence, and

$$\|x^k - x^{k+1}\| \to 0, \|z^k - z^{k+1}\| \to 0, \|Fx^{k+1} - z^{k+1}\| \to 0.$$

Since $\{x^{k+1}, z^{k+1}, \lambda^{k+1}\}$ is bounded, this sequence must have at least one limit point. Let $\{x^*, z^*, \lambda^*\}$ be a limit point, that is, there exists a subsequence $\{k_q\}_{q=1}^{\infty}$ such that

$$\lim_{q \to +\infty} (x^{k_q}, z^{k_q}, \lambda^{k_q}) = (x^*, z^*, \lambda^*)$$

and it holds true that

$$\|x^{k_q} - x^{k_q+1}\| \to 0, \|z^{k_q} - z^{k_q+1}\| \to 0, \|Fx^{k_q+1} - z^{k_q+1}\| \to 0.$$

We consider the first-order optimality condition of updating x^{k_q+1} and z^{k_q+1} and $r(z) = r_1(z) - r_2(z)$, i.e.,

$$0 = \nabla l(x^{k_q}) - F^\top \lambda^{k_q} + \delta\left(x^{k_q+1} - x^{k_q}\right) + \beta^k F^\top \left(Fx^{k_q+1} - z^{k_q}\right),$$

$$0 \in \partial r_1(z^{k_q+1}) - \partial r_2(z^{k_q+1}) + \lambda^{k_q} - \beta^k \left(Fx^{k_q+1} - z^{k_q+1}\right).$$

Letting $q \to +\infty$, by considering the semi-continuity of $\partial r_1(\cdot)$ and $\partial r_2(\cdot)$ and the boundedness of β^k, one obtains

$$\begin{aligned} 0 &= \nabla l(x^*) - F^\top \lambda^*, \\ 0 &\in \partial r_1(z^*) - \partial r_2(z^*) + \lambda^*. \end{aligned}$$

Therefore, (x^*, z^*, λ^*) is a critical point.

Moreover, it follows from Formula (26) that

$$\min_{K_0 \leqslant k \leqslant n} \|x^k - x^{k+1}\|^2 \leqslant \frac{\Phi_2(x^{K_0}, x^{K_0-1}, z^{K_0}, \lambda^{K_0}) - \Phi^*}{(n - K_0)\sigma_{\min}},$$

where $\sigma_{\min}$ is defined as

$$\sigma_{\min} = \sigma\left(\delta - \frac{L}{2}\right) > 0.$$

This completes the proof of Theorem 2.

Acknowledgment

The work was partially supported by the National Natural Science Foundation of China (Nos. 61303264, 61202482, and 61202488); Guangxi Cooperative Innovation Center of Cloud Computing and Big Data (No. YD16505); Distinguished Young Scientist Promotion of National University of Defense Technology. We would also like to thank Data Scientist Yichi Zhang for her kindly providing lfcrc dataset to validate the efficacy of the proposed algorithm.

References

[1] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer, 2001.

[2] R. J. Tibshirani and J. Taylor, The solution path of the generalized lasso, *Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, 2011.

[3] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, Sparse reconstruction by separable approximation, *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, 2009.

[4] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[5] S. K. Shevade and S. S. Keerthi, A simple and efficient algorithm for gene selection using sparse logistic regression, *Bioinformatics*, vol. 19, no. 17, pp. 2246–2253, 2003.

[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, Robust face recognition via sparse representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[7] Q. Lan, C. Xun, M. Wen, H. Su, L. Liu, and C. Zhang, Improving performance of gpu specific opencl program on cpus, in *Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2012 13th International Conference on*, 2012, pp. 356–360.

[8] J. Xiao, Single-target tracking of arbitrary objects using multi-layered features and contextual information, PhD Dissertation, University of Birmingham, Birmingham, UK, 2016.

[9] Z. Lai, K. T. Lam, C. L. Wang, and J. Su, Latency aware dvfs for efficient power state transitions on many-core architectures, *Journal of Supercomputing*, vol. 71, no. 7, pp. 1–28, 2015.

[10] Y. Sun, B. Zhang, B. Zhao, X. Su, and J. Su, Mixzones optimal deployment for protecting location privacy in vanet, *Peer-to-Peer Networking and Applications*, vol. 8, no. 6, pp. 1108–1121, 2015.

[11] X. Xu, C. M. Liu, S. X. Yang, and D. W. Hu, Hierarchical approximate policy iteration with binary-tree state space decomposition, *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 1863–1877, 2011.

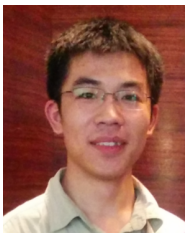
[12] E. J. Candes, M. B. Wakin, and S. P. Boyd, Enhancing sparsity by reweighted l1 minimization, *Journal of Fourier Analysis and Applications*, vol. 14, nos. 5&6, pp. 877–905, 2008.

[13] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[14] S. Foucart and M. J. Lai, Sparsest solutions of underdetermined linear systems via lq-minimization, *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 395–407, 2009.

[15] C. H. Zhang, Nearly unbiased variable selection under minimax concave penalty, *Annals of Statistics*, vol. 2, no. 1, pp. 894–942, 2010.
[16] T. Zhang, Analysis of multi-stage convex relaxation for sparse regularization, *Journal of Machine Learning Research*, vol. 11, pp. 1081–1107, 2010.
[17] T. Zhang, Multi-stage convex relaxation for feature selection, *Bernoulli*, vol. 19, no. 5B, pp. 2277–2293, 2013.
[18] J. F. Yang and X. M. Yuan, Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization, *Mathematics of Computation*, vol. 82, no. 281, pp. 301–329, 2013.
[19] B. He and X. Yuan, On the $o(1/n)$ convergence rate of the Douglas-Rachford alternating direction method, *SIAM Journal on Numerical Analysis*, vol. 50, no. 2, pp. 700–709, 2012.
[20] M. Hong and Z. Q. Luo, On the linear convergence of the alternating direction method of multipliers, arXiv preprint arXiv:1208.3922, 2012.
[21] L. Qiao, B. Zhang, J. Su, and X. Lu, Linearized alternating direction method of multipliers for constrained nonconvex regularized optimization, in *Asian Conference on Machine Learning (ACML)*, 2016, pp. 97–109.
[22] Z. Wen, D. Goldfarb, and W. Yin, Alternating direction augmented Lagrangian methods for semidefinite programming, *Mathematical Programming Computation*, vol. 2, nos. 3&4, pp. 203–230, 2010.
[23] Z. Lu, Sequential convex programming methods for a class of structured nonlinear programming, arXiv preprint arXiv:1210.3039, 2012.
[24] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye, A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems, in *Proceedings of ICML*, 2013, p. 37.
[25] P. Gong and J. Ye, Honor: Hybrid optimization for non-convex regularized problems, in *Proceedings of NIPS*, 2015, pp. 415–423.
[26] L. W. Zhong and J. T. Kwok, Fast stochastic alternating direction method of multipliers, in *Proceedings of ICML*, 2013.
[27] R. Zhang and J. Kwok, Asynchronous distributed admm for consensus optimization, in *Proceedings of ICML*, 2014, pp. 1701–1709.
[28] H. Wang, A. Banerjee, and Z. Q. Luo, Parallel direction method of multipliers, in *Proceedings of NIPS*, 2014, pp. 181–189.
[29] P. Zhao, J. Yang, T. Zhang, and P. Li. Adaptive stochastic alternating direction method of multipliers, in *Proceedings of ICML*, 2013, pp. 69–77.
[30] S. Magnsson, P. Chathuranga, M. Rabbat, and C. Fischione, On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems, arXiv preprint arXiv:1409.8033, 2014.
[31] B. Jiang, S. Ma, and S. Zhang, Alternating direction method of multipliers for real and complex polynomial optimization models, *Optimization*, vol. 63, no. 6, pp. 883–898, 2014.
[32] M. Hong, Z. Q. Luo, and M. Razaviyayn, Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems, *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.
[33] L. Yang, T. K. Pong, and X. Chen, Alternating direction method of multipliers for nonconvex background/foreground extraction, arXiv preprint arXiv:1506.07029, 2015.
[34] F. Wang, W. Cao, and Z. Xu, Convergence of multiblock bregman admm for nonconvex composite problems, arXiv preprint arXiv:1505.03063, 2015.
[35] Y. Wang, W. Yin, and J. Zeng, Global convergence of admm in nonconvex nonsmooth optimization, arXiv preprint arXiv:1511.06324, 2015.
[36] G. Li and T. K. Pong, Global convergence of splitting methods for nonconvex composite optimization, *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2434–2460, 2015.
[37] J. Toland, A duality principle for non-convex optimisation and the calculus of variations, *Archive for Rational Mechanics and Analysis*, vol. 71, no. 1, pp. 41–61, 1979.
[38] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.
[39] K. Scheinberg, S. Ma, and D. Goldfarb, Sparse inverse covariance selection via alternating linearization methods, in *Proceedings of NIPS*, 2010, pp. 2101–2109.

Xicheng Lu received the BS degree in computer science from Harbin Military Engineering Institute, China, in 1970. He was a visiting scholar at the University of Massachusetts from 1982 to 1984. He is now a professor at the College of Computer, National University of Defense Technology (NUDT), China. He has served as a member of editorial boards of several journals and co-chaired many professional conferences. He is an academician of the Chinese Academy of Engineering since 1999. His research interests include distributed computing, computer networks, parallel computing, etc.



Linbo Qiao received the MS and BS degrees in computer science and technology from the National University of Defense Technology (NUDT), China, in 2012 and 2010, respectively. He is currently pursuing the PhD degree in computer science and technology in NUDT. He worked as a research assistant with the SEEM, Chinese University of Hong Kong, from May 2014 to October 2014. His research interests include structured sparse learning, and online and distributed optimization.

Jinshu Su received the BS degree in mathematics from Nankai University, China, in 1985 and the MS and PhD degrees in computer science from the National University of Defense Technology, China, in 1988 and 2000, respectively. He is currently a professor with the College of Computer, NUDT. His research interests include network architecture and Internet routing and security.

Bofeng Zhang received the PhD degree in computer science from the National University of Defense Technology (NUDT), China, in 2007. He is currently an associate professor with College of Computer, National University of Defense Technology. His current research interests include computer networks, information security, data mining, and machine learning algorithms.