Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 20, No. 4, pp. 629–644, 2024

AN INEXACT REGULARIZED PROXIMAL NEWTON-TYPE METHOD FOR NONCONVEX COMPOSITE OPTIMIZATION PROBLEMS*

Danqi Zhu, Can Wu and Dong-Hui Li†

Abstract: In this paper, we consider a broad class of optimization problems with objective functions that are the sum of a smooth function and a possibly nonsmooth difference of convex (DC) functions. We develop an inexact regularized proximal DC Newton-type method that combines a DC algorithm with the proximal Newton-type method. We prove the global convergence of the method and demonstrate its effectiveness through numerical experiments on large-scale data sets.

Key words: *nonconvex composite optimization problem, DC algorithm, proximal Newton-type method, global convergence*

Mathematics Subject Classification: *15A69*

1 Introduction

Consider the following nonconvex composite optimization problem

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + g_1(x) - g_2(x), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuous differentiable function (and may be not convex) whose gradient ∇f is Lipschitz continuous on $\mathbb{R}^n$ with Lipschitz constant $L > 0$, $g_1 : \mathbb{R}^n \to (-\infty, +\infty]$ is a closed proper convex function, and $g_2 : \mathbb{R}^n \to \mathbb{R}$ is a real-valued convex function.

Problem (1.1) is a common optimization problem encountered in machine learning and statistics. In machine learning, f is typically a data-fidelity loss function such as the least squares loss or logistic regression loss, and $g_1 - g_2$ is a difference of convex (DC) regularizer that induces special structures in the solution, like sparsity. While convex regularizers such as $g_1(x) = \lambda\|x\|_1$ and $g_2(x) = 0$ for any $x \in \mathbb{R}^n$ [27] can yield a global optimal solution in polynomial time and allow for the characterization of statistical properties, it has been well known that they introduce nonnegligible estimation biases [8, 34]. To address this issue, several nonconvex regularizers have been proposed, including the smooth clipped absolute

*This work was supported in part by the National Natural Science Foundation of China under Grant 12271187, in part by the Hong Kong Research Grant Council under Grant 15304721, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515010368.

†Corresponding author

DOI: https://doi.org/10.61208/pjo-2023-052

deviation (SCAD) [8], the log-sum penalty function (LSP) [5], the minimax concave penalty (MCP) [31], and the capped ℓ_1-regularization (Capped ℓ_1)[32], among others. Compared to their convex counterparts, the nonconvex regularizers can eliminate estimation biases and achieve more refined statistical rates of convergence [9]. However, solving problem (1.1) becomes more challenging due to the nonconvex and nonsmooth nature of the DC regularizer.

For the problem (1.1) when f is convex and $g_2 = 0$, a large number of proximal Newton-type methods (PNMs) (also known as sequential quadratic approximal methods) has been proposed. In a generic proximal Newton-type method, the kth iteration involves first computing an approximate solution $\hat{x}_k$ to the subproblem obtained by approximating the smooth part f quadratically in (1.1):

$$\min_{x\in\mathbb{R}^n} f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T H_k(x - x_k) + g_1(x), \tag{1.2}$$

where H_k is a symmetric positive definite matrix as an approximation of the Hessian $\nabla^2 f(x_k)$. Subsequently, the iterate x_{k+1} is obtained by performing a backtracking line search along the direction $\hat{x}_k - x_k$. In fact, several PNMs and their variants (see, e.g., [10, 29, 12, 22, 3, 25, 33]) for solving special instances of problem (1.1) with f being convex and $g_2 = 0$ have been proposed. Although they exhibit good numerical performance, achieving global convergence, the methods need to solve subproblem (1.2) exactly, which is time consuming in practice. To address this issue, several families of the PNMs along with inexact stopping criterion for the subproblem (1.2) were proposed. Those inexact methods reserve global convergence and local superlinear convergence rates [13, 4] under the condition that f is strongly convex. Unfortunately, such strong convexity is often absent in many interesting applications, such as the ℓ_1-regularized least squares regression problem, especially when the number of features is much greater than the sample size. To overcome these challenges, Yue et al. [30] and Mordukhovich et al. [19] proposed similar frameworks of inexact PNMs, respectively. Both of them allow inexact solutions to the subproblem (1.2). Without the requirement of strong convexity on f, both methods still possess global convergence and local superlinear convergence rates.

Our purpose in this paper is to solve the nonconvex problem (1.1). Several variants of proximal gradient methods (combined with a majorization technique) for solving (1.1) have been proposed [18, 11, 15]. Those methods are first-order methods. We are interested in second-order methods. Specifically, we approximate the smooth part of f quadratically while approximate the concave term $-g_2$ at x_k linearly, which yields the subproblem

$$\min_{x\in\mathbb{R}^n} f(x_k) - g_2(x_k) + (\nabla f(x_k) - \xi_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T H_k(x - x_k) + g_1(x), \tag{1.3}$$

where $H_k \in \mathcal{S}^n_{++}$ (the set of all symmetric and positive definite matrices) is an approximate Hessian of f at x_k, and ξ_k is a subgradient of the subdifferential of g_2 at x_k. Liu et al. [17] proposed an inexact successive quadratic approximation method (sSQA$_{\text{major}}$) and derived the iteration complexity for obtaining an approximate ε-stationary point. Nakayama et al. [20] proposed an inexact proximal DC Newton-type method (mLBGS), in which H_k was updated by some memoryless BFGS formula, and also discussed the numerical method for solving the subproblem (1.3). To ensure the convergence, both above methods require the assumption that the smallest eigenvalue $\lambda_{\min}(H_k)$ of the matrix H_k has a positive lower bound. We notice that in a recent work, Mordukhovich et al. [19] proposed a PNM method for solving the convex problem (1.1) with $g_2 = 0$. The global convergence and locally superlinear convergence rate were achieved without the assumption that $\lambda_{\min}(H_k)$ has a

positive lower bound. Our purpose of this paper is to develop an efficient PNM method called the inexact regularized proximal Newton-type method (IRP), for solving (1.1).

The contributions of the paper can be summarized as follows:

(1) We propose the IRP (Algorithm 1) for solving nonconvex composite optimization problem (1.1). Particularly, we develop a novel inexact rule (2.7) to assess the inexactness of solution to subproblem (1.3).

(2) We establish the global convergence (Theorem 2.5) of the IRP without the requirement of the positive lower bound of $\{\lambda_{\min}(H_k)\}_{k\geq 0}$.

(3) We conduct numerical experiments to solve the problem (1.1) with convex and nonconvex loss functions, along with four distinct DC regularizers on large-scale data sets, and demonstrate the superiority of IRP over three existing methods in terms of both objective function values achieved and the number of iterations required to reach the stopping criteria.

The rest of the paper is organized as follows. In Section 2, we propose the IRP and analyze its global convergence. In Section 3, we present numerical results of four methods for solving the problem (1.1) using a convex loss function or a nonconvex loss function under four different DC regularizers on large-scale data sets. Finally, Section 4 concludes this paper.

2 The Algorithm and Its Convergence

In this section, we will design an inexact proximal Newton-type method for solving the nonconvex composite optimization problem (1.1) and also show its global convergence. We first introduce the concept of the critical point, which will be useful for characterizing the convergence result of our algorithm.

Definition 2.1 ([28]). A point $x^* \in \mathbb{R}^n$ is said to be a critical point of the problem (1.1) if

$$0 \in \nabla f\left(x^*\right) + \partial g_1\left(x^*\right) - \partial g_2\left(x^*\right),$$

where $\partial g_i(x^*)$ is the subdifferential of g_i at x^* for $i = 1, 2$, i.e.,

$$\partial g_i(x^*) = \{\xi_i \mid g_i(x) \geq g_i(x^*) + \xi_i^T(x - x^*), \forall x \in \mathbb{R}^n\}, \quad i = 1, 2.$$

It is well-known that if f is convex and $g_2 \equiv 0$, then x is a solution to (1.1) if and only if it is a critical point of (1.1).

Notice that in the case where f is convex, the problem (1.1) is a DC programming. In a typical DC algorithm (see, e.g, [26]), the subproblem is the the following convex subproblem

$$\min_{x\in\mathbb{R}^n} f(x) + g_1(x) - g_2\left(x_k\right) - \xi_k^T\left(x - x_k\right), \tag{2.1}$$

where $\xi_k \in \partial g_2(x_k)$. The optimality condition for the above subproblem (2.1) can be expressed as

$$0 \in \nabla f\left(x\right) - \xi_k + \partial g_1(x).$$

Then the proximal residual mapping associated with (2.1) is given by

$$\mathcal{G}_k(x) := x - \mathrm{Prox}_{g_1}(x - \nabla f(x) + \xi_k),$$

where for a closed proper convex mapping $g : \mathbb{R}^n \to (-\infty, +\infty]$, its proximal mapping is defined as

$$\mathrm{Prox}_g(u) := \mathrm{argmin}\left\{ g(x) + \frac{1}{2}\|x-u\|^2 \;\Big|\; x \in \mathbb{R}^n \right\}, \quad u \in \mathbb{R}^n. \tag{2.2}$$

In what follows, we extend this idea to solve the nonconvex problem (1.1). In stead of (2.1), in our method, the subproblem is the following quadratic approximation to it

$$\min_{x\in\mathbb{R}^n} q_k(x) := f(x_k) - g_2(x_k) + (\nabla f(x_k) - \xi_k)^T (x - x_k) + \frac{1}{2}(x-x_k)^T H_k (x-x_k) + g_1(x), \tag{2.3}$$

where the matrix H_k is an approximate to the Hessian $\nabla^2 f(x_k)$ that takes the form

$$H_k := B_k + \alpha_k I \text{ with } \alpha_k = \min\{\bar{\alpha}, c\|\mathcal{G}_k(x_k)\|^\rho\} \tag{2.4}$$

for some given positive constants $\bar{\alpha}, c$ and $\rho \in (0,1]$. In the case where f is convex and $g_2 \equiv 0$, the method reduces to the method in [19]. If f is convex and $g_1 = g_2 \equiv 0$, the method with $B_k = \nabla^2 f(x_k)$ reduces to the regularized Newton method for solving convex unconstrained optimization problem by Li et. al. [14].

We will let matrix B_k satisfy the following condition

$$B_k \in \mathcal{S}^n_+ \text{ and there exists a number } M \geq 0 \text{ such that } \|B_k\| \leq M \text{ for all } k = 0, 1, \ldots. \tag{2.5}$$

In the case that f is convex, if we take $B_k = \nabla^2 f(x_k)$, then the above condition is satisfied [1, Theorem 5.12].

We define the proximal residual mapping associated with (2.3) by

$$r_k(x) := x - \mathrm{Prox}_{g_1}\left(x - (\nabla f(x_k) - \xi_k) - H_k(x - x_k)\right). \tag{2.6}$$

It is single-valued and continuous [24, Theorem 2.26].

Observing that $\|r_k(\hat{x}_k)\| = 0$ if and only if $\hat{x}_k$ is an exact solution to subproblem (2.3). Thus it is reasonable to use $\|r_k(\hat{x}_k)\|$ for measuring the quality of the approximate solution $\hat{x}_k$ of (2.3). Taking this into account, we use the following two inequalities as the inexact stopping criteria for the subproblem (2.3)

$$\|r_k(\hat{x}_k)\| \leq \eta_k \|\mathcal{G}_k(x_k)\| \quad \text{and} \quad q_k(\hat{x}_k) \leq q_k(x_k), \tag{2.7}$$

where

$$\mathcal{G}_k(x) := x - \mathrm{Prox}_{g_1}(x - \nabla f(x) + \xi_k), \quad \forall\, x \in \mathbb{R}^n, \tag{2.8}$$

and $\eta_k := \nu \min\{1, \|\mathcal{G}_k(x_k)\|^\varrho\}$ with constants $\nu \in [0,1)$ and $\varrho > 0$.

In the latter part of the paper, we require the following assumption.

Assumption 2.1. Assume that the objective function F in problem (1.1) is bounded below on $\mathbb{R}^n$, i.e., there exists a constant $\widehat{C} \in \mathbb{R}$ such that $\inf_{x\in\mathbb{R}^n} F(x) \geq \widehat{C}$.

Now, we propose an inexact regularized proximal Newton-type method for solving (1.1) whose steps are given below.

Algorithm 1 Inexact regularized proximal Newton-type method for solving (1.1)

Initialization: Choose an initial point $x_0 \in \mathbb{R}^n$, parameters $0 < \mu < 1/2$, $0 < \sigma, \gamma < 1, C > F(x_0), \bar{\alpha}, c > 0$, and $\rho \in (0, 1]$. Set $k = 0$.

repeat

1: Update matrix B_k in (2.4) satisfying conditions (2.5) and choose $\xi_k \in \partial g_2(x_k)$.

2: Form the quadratic model (2.3) with H_k defined in (2.4).

3: Obtain an inexact solution $\hat{x}_k$ of (2.3) satisfying the conditions (2.7).

4: If $k = 0$, let $\vartheta_1 := \mathcal{G}_0(x_0)$ and go to Step 5. For $k \geq 1$, if $\|\mathcal{G}_k(\hat{x}_k)\| \leq \sigma\vartheta_k$ and $F(\hat{x}_k) \leq C$, let $t_k := 1, \vartheta_{k+1} := \|\mathcal{G}_k(\hat{x}_k)\|$, and go to Step 6. Otherwise, let $\vartheta_{k+1} := \vartheta_k$ and go to Step 5.

5: Perform a backtracking line search along the direction $d_k := \hat{x}_k - x_k$ by setting $t_k := \gamma^{m_k}$, where m_k is the smallest nonnegative integer m such that

$$F(x_k + \gamma^m d_k) \leq F(x_k) - \mu\alpha_k\gamma^m \|d_k\|^2. \tag{2.9}$$

6: Set $x_{k+1} := x_k + t_k d_k$.

until Some proper stopping criterion is satisfied.

Remark 2.2. There exist fruitful effective methods to solve the convex subproblems (2.3) inexactly, e.g., the block coordinate descent method [29], the fast iterative soft-shrinkage algorithm [2], and the semismooth Newton augmented Lagrangian method [16], among others. For our experiment, we opt to utilize the block coordinate descent method to solve the subproblems (2.3).

The remainder of this section is devoted to the global convergence of the above algorithm. We first prove the following two lemmas.

Lemma 2.3. *Given an approximate solution $\hat{x}_k$ to the subproblem (2.3), there exists a vector $e_k \in \mathbb{R}^n$ such that*

$$\begin{cases} e_k \in \nabla f(x_k) - \xi_k + H_k(\hat{x}_k - x_k) + \partial g_1(\hat{x}_k - e_k) \\ \|e_k\| \leq \nu \min\left\{\|\mathcal{G}_k(x_k)\|, \|\mathcal{G}_k(x_k)\|^{1+\varrho}\right\}. \end{cases} \tag{2.10}$$

Proof. Let $e_k := r_k(\hat{x}_k) = \hat{x}_k - \mathrm{Prox}_{g_1}(\hat{x}_k - (\nabla f(x_k) - \xi_k) - H_k(\hat{x}_k - x_k))$. It follows from (2.2) that

$$e_k \in \nabla f(x_k) - \xi_k + H_k(\hat{x}_k - x_k) + \partial g_1(\hat{x}_k - e_k).$$

By the use of the inexact conditions (2.7), it is easy to get (2.10). □

The next lemma describes the lower bound of the step size obtained by a backtracking line search in Algorithm 1 and the corresponding decrease in the objective function values in (1.1).

Lemma 2.4. *The steplength t_k obtained by Step 5 of Algorithm 1 satisfies*

$$t_k \geq \gamma \min\left\{1, \frac{(1-2\mu)\alpha_k}{L}\right\}. \tag{2.11}$$

The values of the objective function at x_k and x_{k+1} satisfy

$$F(x_{k+1}) - F(x_k) \leq -\mu\,\alpha_k\,\gamma \min\left\{1, \frac{(1-2\mu)\alpha_k}{L}\right\}\left(\frac{1-\nu}{1+M+\alpha_k}\right)^2 \|\mathcal{G}_k(x_k)\|^2. \quad (2.12)$$

Proof. Since $\hat{x}_k$ is an inexact solution to (2.3) obeying the conditions in (2.7), one has

$$0 \geq q_k(\hat{x}_k) - q_k(x_k) = l_k(\hat{x}_k) - l_k(x_k) + \frac{1}{2}(\hat{x}_k - x_k)^T H_k(\hat{x}_k - x_k),$$

where l_k is defined as

$$l_k(x) := f(x_k) - g_2(x_k) + (\nabla f(x_k) - \xi_k)^T(x - x_k) + g_1(x). \quad (2.13)$$

It yields from the definition of H_k that

$$l_k(x_k) - l_k(\hat{x}_k) \geq \frac{1}{2}(\hat{x}_k - x_k)^T H_k(\hat{x}_k - x_k) \geq \frac{1}{2}\alpha_k\|\hat{x}_k - x_k\|^2. \quad (2.14)$$

By the definition of the proximal residual mapping $\mathcal{G}_k$ in (2.8), we can deduce that

$$\mathcal{G}_k(x_k) \in \nabla f(x_k) - \xi_k + \partial g_1(x_k - \mathcal{G}_k(x_k)).$$

Based on the fact that the subdifferential ∂g_1 is monotone, associating with the inequality in (2.10), one has $\|e_k\| \leq \nu\|\mathcal{G}_k(x_k)\|$ and

$$(\mathcal{G}_k(x_k) + H_k(\hat{x}_k - x_k) - e_k)^T(x_k - \mathcal{G}_k(x_k) - \hat{x}_k + e_k) \geq 0,$$

which implies that

$$\begin{aligned}\|\mathcal{G}_k(x_k) - e_k\|^2 &\leq \|\mathcal{G}_k(x_k) - e_k\|^2 + (\hat{x}_k - x_k)^T H_k(\hat{x}_k - x_k)\\ &\leq (\mathcal{G}_k(x_k) - e_k)^T(x_k - \hat{x}_k + H_k(x_k - \hat{x}_k))\\ &\leq \|\mathcal{G}_k(x_k) - e_k\| \cdot \|x_k - \hat{x}_k + H_k(x_k - \hat{x}_k)\|.\end{aligned}$$

Combing the conditions $\|e_k\| \leq \nu\|\mathcal{G}_k(x_k)\|$ and $\|B_k\| \leq M$ from (2.5), we further obtain

$$\|\mathcal{G}_k(x_k)\| \leq \|\mathcal{G}_k(x_k) - e_k\| + \|e_k\| \leq (1 + M + \alpha_k)\|\hat{x}_k - x_k\| + \nu\|\mathcal{G}_k(x_k)\|.$$

Since $\nu \in [0, 1)$, the proximal residual mapping $\mathcal{G}_k(\cdot)$ at x_k can be estimated by

$$\|\mathcal{G}_k(x_k)\| \leq \frac{1+M+\alpha_k}{1-\nu}\|\hat{x}_k - x_k\|. \quad (2.15)$$

Next, we show that the backtracking line search along the direction $d_k = \hat{x}_k - x_k$ in Step 5 is well-defined and the proposed step size ensures a sufficient decrease in the cost function F. Based on the Lipschitz continuity of ∇f and $\xi_k \in \partial g_2(x_k)$, one has that for any $\tau \geq 0$,

$$f(x_k + \tau d_k) \leq f(x_k) + \tau\nabla f(x_k)^T d_k + \frac{L}{2}\tau^2\|d_k\|^2 \text{ and } g_2(x_k + \tau d_k) \geq g_2(x_k) + \tau\xi_k^\top d_k,$$

which further implies that

$$F(x_k) - F(x_k + \tau d_k) \geq l_k(x_k) - l_k(x_k + \tau d_k) - \frac{L}{2}\tau^2\|d_k\|^2. \quad (2.16)$$

Furthermore, by the convexity of l_k, we get

$$l_k(x_k) - l_k(x_k + \tau d_k) \geq \tau\left(l_k(x_k) - l_k(x_k + d_k)\right), \quad \forall \tau \in [0,1]. \tag{2.17}$$

Combing (2.16), (2.17), (2.14) and $\mu \in (0, 1/2)$, one has for any $\tau \in [0,1]$,

$$\begin{aligned} F(x_k) - F(x_k + \tau d_k) - \mu\alpha_k\tau\|d_k\|^2 &\geq l_k(x_k) - l_k(x_k + \tau d_k) - (L/2)\tau^2\|d_k\|^2 \\ &\quad - \mu\alpha_k\tau\|d_k\|^2 \\ &\geq \tau(l_k(x_k) - l_k(x_k + d_k)) - (L/2)\tau^2\|d_k\|^2 \\ &\quad - \mu\alpha_k\tau\|d_k\|^2 \\ &\geq (\tau/2)\alpha_k\|d_k\|^2 - (L/2)\tau^2\|d_k\|^2 \\ &\quad - \mu\alpha_k\tau\|d_k\|^2 \\ &= \tau[(1-2\mu)\alpha_k - L\tau]\|d_k\|^2/2. \end{aligned} \tag{2.18}$$

It means that the backtracking line search criterion (2.9) is satisfied when $0 < \tau \leq \min\{1, (1-2\mu)\alpha_k/L\}$, and thus the step size t_k satisfies the claimed condition (2.11). Substituting $\tau := t_k \geq \gamma\min\{1, (1-2\mu)\alpha_k/L\}$ into (2.18) and employing the estimate of $\|\mathcal{G}_k(x_k)\|$ in (2.15), we obtain

$$\begin{aligned} F(x_k) - F(x_k + t_k d_k) &\geq \mu\alpha_k t_k\|d_k\|^2 \\ &\geq \mu\alpha_k\gamma\min\left\{1, \frac{(1-2\mu)\alpha_k}{L}\right\}\left(\frac{1-\nu}{1+M+\alpha_k}\right)^2\|\mathcal{G}_k(x_k)\|^2, \end{aligned}$$

which verifies the decreasing condition (2.12) and thus completes the proof of the lemma. □

Now we are ready to prove the global convergence of Algorithm 1. Define the sets $K := \{0, 1, \ldots\}$ and

$$K_0 := \{0\} \cup \{k+1 \in K \mid \text{ Step 5 is not applied at iteration } k\}.$$

Theorem 2.5. *Let Assumption 2.1 hold and $\{x_k\}$ be generated by Algorithm 1. Then we have*

$$\liminf_{k\to\infty} \|\mathcal{G}_k(x_k)\| = 0. \tag{2.19}$$

Furthermore, if $\{x_k\}$ is bounded, then any accumulation point of $\{x_k\}$ is a critical point of (1.1).

Proof. Firstly, we consider the case that the index set K_0 is infinite. We can reorganize K_0 in such a way that $0 = k_0 < k_1 < k_2 < \ldots$. It follows from Step 4 of Algorithm 1 that the inequalities

$$\left\|\mathcal{G}_{k_{\ell+1}}\left(x_{k_{\ell+1}}\right)\right\| \leq \sigma\left\|\mathcal{G}_{k_\ell}\left(x_{k_\ell}\right)\right\| \quad \text{for } \ell = 0, 1, \ldots$$

hold, which follows from $\sigma \in (0,1)$ that

$$0 \leq \liminf_{k\to\infty}\|\mathcal{G}_k(x_k)\| \leq \limsup_{\ell\to\infty}\|\mathcal{G}_{k_\ell}(x_{k_\ell})\| \leq \lim_{\ell\to\infty}\sigma^\ell\|\mathcal{G}_{k_0}(x_{k_0})\| = 0.$$

Thus, the equality (2.19) holds if K_0 is infinite.

Secondly, consider the case that the index set K_0 is finite. Denote $\bar{k} = \max_{k\in K_0} k$. It follows from Lemma 2.4 that for any $k > \bar{k}$, we get

$$F(x_{k+1}) - F(x_k) \leq -\mu\,\alpha_k\,\gamma\min\left\{1, \frac{(1-2\mu)\alpha_k}{L}\right\}\left(\frac{1-\nu}{1+M+\alpha_k}\right)^2\|\mathcal{G}_k(x_k)\|^2.$$

It follows from Assumption 2.1 that

$$\sum_{k=\bar{k}}^{\infty} \mu\alpha_k\gamma \min\left\{1, \frac{(1-2\mu)\alpha_k}{L}\right\}\left(\frac{1-\nu}{1+M+\alpha_k}\right)^2 \|\mathcal{G}_k(x_k)\|^2 \leq F(x_{\bar{k}}) - \widehat{C} < +\infty,$$

which implies that

$$\lim_{k\to\infty} \alpha_k \min\left\{1, \frac{(1-2\mu)\alpha_k}{L}\right\}\left(\frac{1-\nu}{1+M+\alpha_k}\right)^2 \|\mathcal{G}_k(x_k)\|^2 = 0.$$

Based on the expression $\alpha_k := \min\{\bar{\alpha}, c\|\mathcal{G}_k(x_k)\|^\rho\}$ with positive numbers $\bar{\alpha}$, c, ρ and $\nu \in [0,1)$, one has

$$\lim_{k\to\infty} \|\mathcal{G}_k(x_k)\| = 0.$$

Thus, the equality (2.19) also holds when K_0 is finite.

Next, we show that the boundedness of the sequence $\{x_k\}$ implies that any cluster point of $\{x_k\}$ is a critical point of problem (1.1). Assume that the sequence $\{x_k\}$ is bounded, and $\bar{x}$ is an arbitrary cluster point of $\{x_k\}$. Then there exists a subsequence $\{x_{k_i}\}$ of $\{x_k\}$ such that $\lim_{i\to\infty} x_{k_i} = \bar{x}$. Since g_2 is a real-valued convex function on $\mathbb{R}^n$, we know that g_2 is locally Lipschitz continuous on $\mathbb{R}^n$ [23, Theorem 10.4]. It further from [7, Proposition (a) and (c)] that there exists a positive integer N such that $\{\xi_{k_i}\}_{i\geq N}$ is bounded. Taking a subsequence of $\{\xi_{k_i}\}_{i\geq N}$ if necessary, assume that $\lim_{i\to\infty} \xi_{k_i} = \bar{\xi}$. From the closedness of $\partial g_1(\cdot)$, one has $\bar{\xi} \in \partial g_2(\bar{x})$. Then by the definition of $\mathcal{G}_{k_i}(\cdot)$ in (2.8) and continuities of $\mathrm{Prox}_{g_1}(\cdot)$ and $\nabla f(\cdot)$, we obtain

$$\begin{aligned} 0 =& \liminf_{i\to\infty} \|\mathcal{G}_{k_i}(x_{k_i})\| = \lim_{i\to\infty} \|x_{k_i} - \mathrm{Prox}_{g_1}(x_{k_i} - \nabla f(x_{k_i}) + \xi_{k_i})\| \\ =& \|\bar{x} - \mathrm{Prox}_{g_1}(\bar{x} - \nabla f(\bar{x}) + \bar{\xi})\|, \end{aligned}$$

which means that $\bar{x} = \mathrm{Prox}_{g_1}(\bar{x} - \nabla f(\bar{x}) + \bar{\xi})$, i.e.,

$$0 \in \nabla f(\bar{x}) + \partial g_1(\bar{x}) - \bar{\xi} \subseteq \nabla f(\bar{x}) + \partial g_1(\bar{x}) - \partial g_2(\bar{x}).$$

Thus $\bar{x}$ is a critical point of the problem (1.1) according to Definition 2.1. □

3 Numerical Experiments

In this section, we evaluate the numerical performance of our IRP for solving the problems (1.1) using a convex loss function or a nonconvex loss function under four different DC terms on large-scale data sets. To be specific, we consider four different DC terms $g_1(x) - g_2(x)$, with

$$g_1(x) = \sum_{i=1}^{n} g_{1,i}(x_i) \quad \text{and} \quad g_2(x) = \sum_{i=1}^{n} g_{2,i}(x_i).$$

Detailed $g_{1,i}$ and $g_{2,i}$ are given in Table 1.

All the numerical experiments were performed in MATLAB 2021a on a laptop with Apple M1 and 16GB memory running macOS Monterey.

We compare our IRP with other three algorithms: the general iterative shrinkage and thresholding (GIST) algorithm [11], the nonmonotone accelerated proximal gradient (nonAPG) algorithm [15], and the inexact proximal DC Newton-type method (mLBFGS) algorithm [20]. All the implementation details of above four algorithms are listed below.

Table 1: Expressions of $g_{1,i}$ and $g_{2,i}$ with $[x]_+ = \max\{0, x\}$ and $\lambda > 0$

Name	$g_{1,i}(x_i)$	$g_{2,i}(x_i)$
LSP [5]	$\lambda\lvert x_i\rvert$	$\lambda(\lvert x_i\rvert - \log(1 + \lvert x_i\rvert/\theta))$
SCAD [8]	$\lambda\lvert x_i\rvert$	$\begin{cases} 0, & \text{if } \lvert x_i\rvert \le \lambda, \\ \frac{x_i^2 - 2\lambda\lvert x_i\rvert + \lambda^2}{2(\theta-1)}, & \text{if } \lambda \le \lvert x_i\rvert \le \theta\lambda, \\ \lambda\lvert x_i\rvert - \frac{(\theta+1)\lambda^2}{2}, & \text{if } \lvert x_i\rvert > \theta\lambda. \end{cases}$
MCP [31]	$\lambda\lvert x_i\rvert$	$\begin{cases} x_i^2/(2\theta), & \text{if } \lvert x_i\rvert \le \theta\lambda, \\ \lambda\lvert x_i\rvert - \theta\lambda^2/2, & \text{if } \lvert x_i\rvert > \theta\lambda. \end{cases}$
Capped ℓ_1 [32]	$\lambda\lvert x_i\rvert$	$\lambda[\lvert x_i\rvert - \theta]_+$

IRP We set $\nu = 0.9$ and $\varrho = 0.1$ in the inexact conditions (2.7). We also set $\mu = 0.1$, $\sigma = 0.25$, $\gamma = 0.5$, $C = 2F(x_0)$, $\bar{\alpha} = 10^{-4}$, and $\rho = 0.1$. The subproblem can be solved by the coordinate gradient descent method [29, 30], which is implemented in MATLAB as a C source MEX-file.

GIST It uses the nonmonotone line search criterion with Barzilai-Borwein rule as the step size initialization. And we set $\sigma = 10^{-5}$, $m = 5$, $\eta = 2$, and $1/t_{min} = t_{max} = 10^{30}$ as suggested in [11].

nonAPG We follow [15] to set $\sigma = 10^{-5}, \eta = 0.8, 1/t_{min} = t_{max} = 10^{20}$.

mLBFGS It chose the approximation of the Hessian $\nabla^2 f$ by the memoryless BFGS formula and applied semismooth Newton method to compute scaled proximal mappings. In our implementation, we set $\phi_k = 0$, $\tau_k = 1$, $\gamma_k = (s_{k-1}^T z_{k-1})/(z_{k-1}^T z_{k-1})$, and

$$\nu_k = \begin{cases} 0, & \text{if } s_{k-1}^T y_{k-1} \ge 10^{-6}\|s_{k-1}\|^2, \\ \max\left\{0, -\frac{s_{k-1}^T y_{k-1}}{s_{k-1}^T s_{k-1}}\right\} + 10^{-6}, & \text{otherwise.} \end{cases}$$

We also set $\delta = 0.5$ and $\beta_k = 0.5$ for its line search scheme, and set $\sigma = 10^{-4}$, $\rho = 0.5$, $\alpha_0 = (0, 0)^T$, and $\theta_k = 0.98$ for the semismooth Newton method.

3.1 Solving problem (1.1) with convex loss function f

In this subsection, we evaluate the numerical performance of our IRP for solving the logistic regression problem under four distinct DC regularizers on four large-scale real datasets. Specifically, the function $f(x) = \sum_{i=1}^s \log\left(1 + \exp\left(-b_i a_i^T x\right)\right)$ is chosen and is convex.

For all four methods, we set the initial point x_0 as the original point in $\mathbb{R}^n$, and the stopping criteria as $|F(x_{k+1}) - F(x_k)| < 10^{-5}\max(1, |F(x_k)|)$ or $\|u_k\| \le 10^{-2}$, where $u_k = \nabla f(x_k) + \xi_{1,k} - \xi_{2,k}$, $\xi_{i,k} \in \partial g_i(x_k)$ for $i = 1, 2$. In IRP, we choose $c = 10^{-8}$ and $B_k := \nabla^2 f(x_k)$.

We tested four real datasets "w2a.t","rcv1.binary" , "news20.binary" and "real-sim" downloaded from the SVMLib repository [6]. Their sizes are given in Table 2.

The code is downloaded from https://github.com/ZiruiZhou/IRPN.
The code is downloaded from https://github.com/iamtu/OPE_GIST.git
The code is downloaded from https://zhouchenlin.github.io/NIPS2015_code.zip
http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/.

Table 2: The size of data sets

Data sets	w2a.t	rcv1.binary	news20.binary	real-sim
s (sample size)	46279	20242	19996	72309
n (dimension of features)	300	47236	1355191	20958

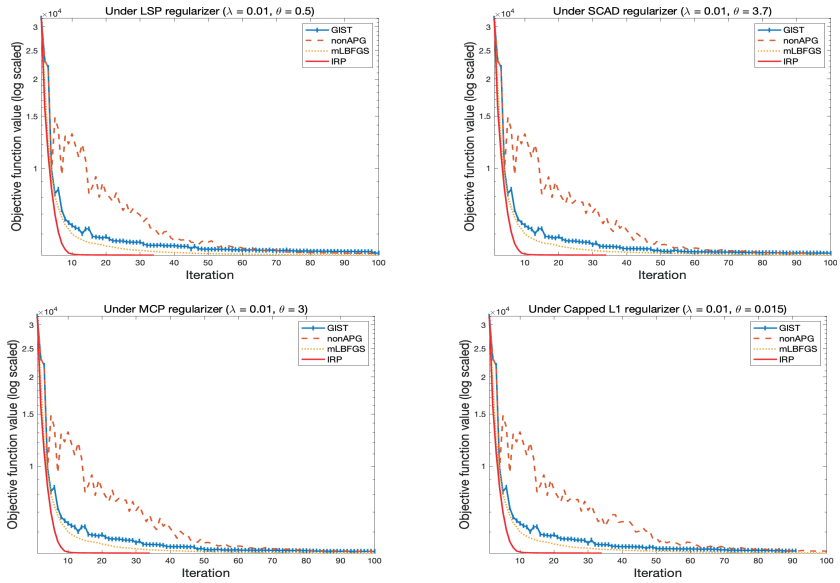

Figure 1: Objective function values versus iteration numbers on wa2.t data set.

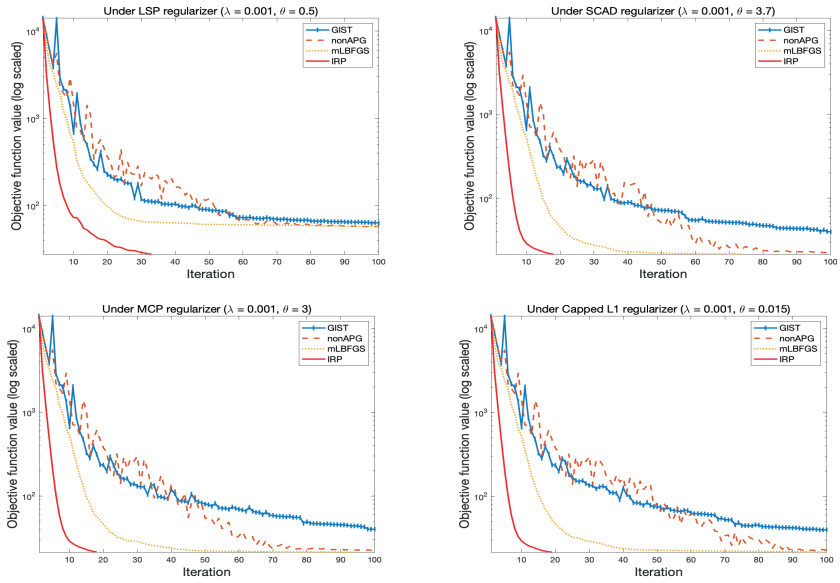

Figure 2: Objective function values versus iteration numbers on rcv1.binary data set.

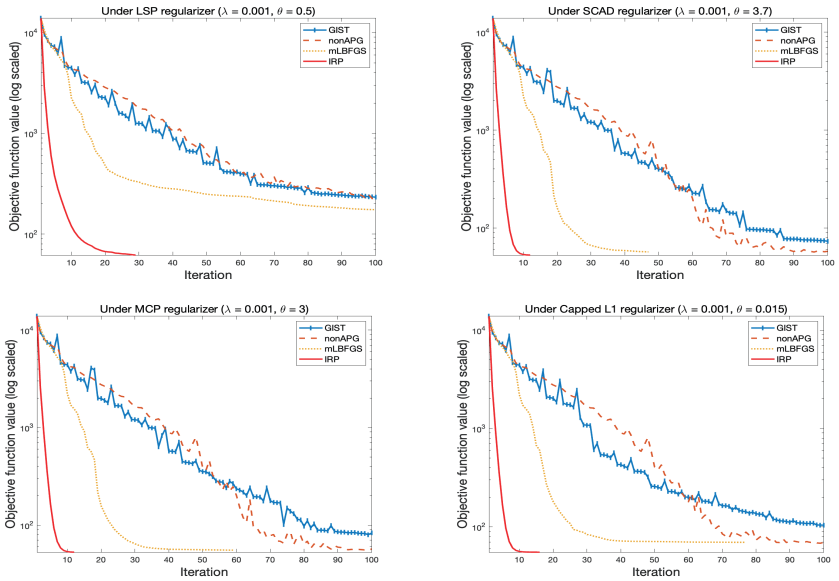

Figure 3: Objective function values versus iteration numbers on news20.binary data set.

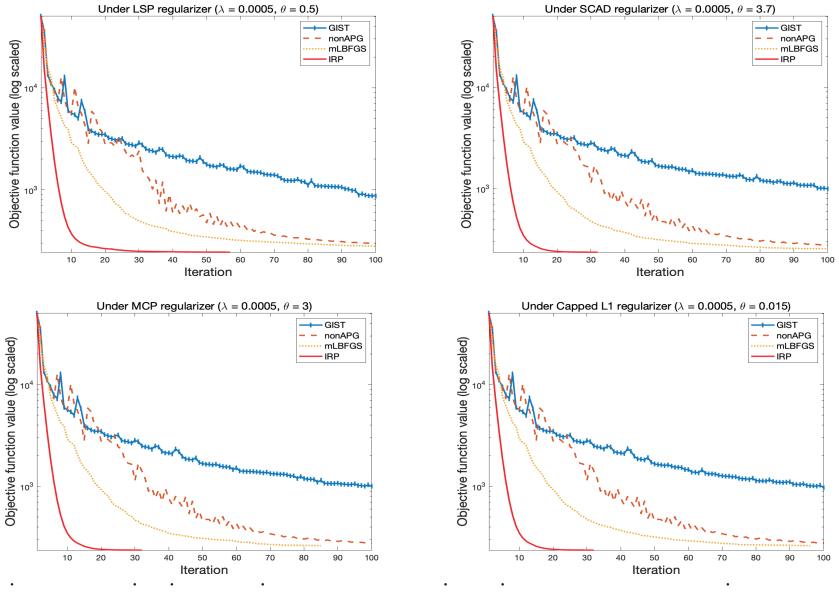

Figure 4: Objective function values versus iteration numbers on real-sim data set.

Figures 1-4 report the changes of the objective function values with the increase of the iteration number under respective four different regularizers. It can be observed that our IRP consistently exhibits the fastest convergence rate in the sense of the objective function value obtained, and has the ability to achieve smaller objective function values than other methods.

Table 3: Numerical comparison on wa2.t dataset with $\lambda = 10^{-2}$

	GIST			nonAPG			mLBFGS			IRP		
Regularizers	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V
LSP	110	0.81	5159.78	162	2.62	5102.05	116	2.29	5097.05	34	0.54	5093.72
SCAD	115	0.83	5146.29	94	1.39	5163.98	126	2.59	5092.25	34	0.58	5089.26
MCP	105	0.80	5155.15	125	1.97	5124.32	126	2.55	5091.90	34	0.55	5089.24
Capped ℓ_1	90	0.69	5173.37	123	2.00	5133.90	129	2.52	5091.81	34	0.56	5089.24

Table 4: Numerical comparison on rcv1.binary dataset with $\lambda = 10^{-3}$

	GIST			nonAPG			mLBFGS			IRP		
Regularizers	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V
LSP	553	12.31	41.38	416	19.53	43.06	239	9.43	46.50	33	2.86	27.57
SCAD	665	14.01	23.39	146	6.97	21.57	74	3.03	21.55	18	1.46	21.20
MCP	120	2.24	36.31	154	7.23	21.56	94	4.00	21.28	18	1.45	21.17
Capped ℓ_1	451	8.81	24.32	204	9.17	21.20	108	4.72	21.15	18	1.45	21.13

Table 5: Numerical comparison on news20.binary dataset with $\lambda = 10^{-3}$

	GIST			nonAPG			mLBFGS			IRP		
Regularizers	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V
LSP	546	98.47	106.25	348	124.91	93.43	804	264.00	85.39	29	11.37	61.56
SCAD	314	47.10	57.20	141	42.40	55.47	47	16.98	57.02	12	2.46	52.86
MCP	340	46.52	57.16	169	49.59	54.46	59	23.79	55.31	12	2.46	52.68
Capped ℓ_1	305	41.16	66.10	123	35.43	67.36	77	33.37	69.64	16	3.85	55.25

Table 6: Numerical comparison on real-sim dataset with $\lambda = 5 \times 10^{-4}$

	GIST			nonAPG			mLBFGS			IRP		
Regularizers	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V
LSP	270	15.22	453.82	1703	184.58	256.14	123	11.11	277.16	57	18.60	245.09
SCAD	555	30.81	338.11	142	15.37	259.11	120	11.67	250.63	32	10.03	236.77
MCP	483	26.53	335.41	142	15.33	259.10	85	7.66	261.44	32	10.06	236.77
Capped ℓ_1	707	39.40	306.34	142	15.39	259.25	96	8.65	260.81	32	10.07	236.87

Tables 3 to 6 list the number of iterations, CPU time and the objective function values of four methods for solving problem (1.1) under four different regularizers. We found that in all instances, the IRP is superior to other methods from the perspective of objective function value and number of iterations.

3.2 Solving problem (1.1) with nonconvex loss function f

In this subsection, we test the performance of the proposed IRP method on nonconvex problems. Specifically, we consider the loss function $f(x) = (1/2)\sum_{i=1}^{n} \log(1+\beta(x_i - u_i^0)^2)$ [21] with $n = 10000$, $u^0 = (1, 1, \cdots, 1)^T$, and $\beta = 100$. Clearly, f is nonconvex and has a Lipschitz continuous gradient with the components $(\nabla f(x))_i = \beta(x_i - u_i^0)/(1+\beta(x_i - u_i^0)^2)$ for $i = 1, \ldots, n$. The DC terms are given in Table 1.

We compared the proposed IRP method with the methods GIST, nonAPG, and mLBFGS. While conducting numerical experiments, we always set the initial point x_0 as the original point in $\mathbb{R}^n$, and take $|F(x_{k+1} - F(x_k))| < 10^{-5}$ as the stopping criterion. In the IRP, we choose $c = 0.619$ and $B_k = \nabla^2 f(x_k) + \max(0, -\lambda_{\min}(\nabla^2 f(x_k)))I_n$, where $\lambda_{\min}(\nabla^2 f(x_k))$ means the smallest eigenvalue of the Hessian $\nabla^2 f(x_k)$.

The results are depicted in Figure 5 and summarized in Table 7. We can observe that IRP, nonAPG, and mLBFGS achieved nearly identical objective values within 20 iterations, whereas GIST failed to meet the termination criterion even after 1000 iterations. Furthermore, in terms of the number of iterations and CPU time consumption, the performance of IRP was superior to nonAPG and mLBFGS.

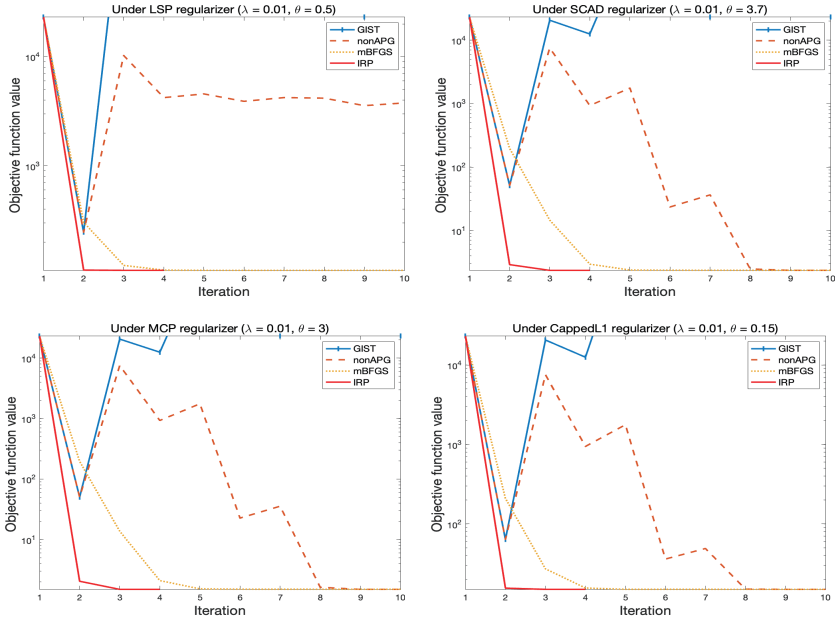

Figure 5: Objective function values versus iteration numbers under four different regularizers.

Table 7: Numerical comparison with $\lambda = 10^{-2}$

	GIST			nonAPG			mLBFGS			IRP		
Regularizers	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V	Iter	Time	Obj_V
LSP	1000*	7.2340	245.35	20	0.2369	109.86	20	0.0321	109.86	3	0.0071	109.86
SCAD	1000*	0.5554	51.13	10	0.0081	2.35	11	0.0305	2.36	3	0.0056	2.35
MCP	1000*	0.5117	50.28	10	0.0070	1.50	11	0.0295	1.51	3	0.0059	1.50
Capped ℓ_1	1000*	2.2919	63.78	10	0.0065	15.00	11	0.0281	15.01	3	0.0053	15.00

The symbol "*" indicates that the stopping condition is not satisfied until the maximum iteration is reached.

4 Conclusion

We developed an inexact regularized proximal Newton-type method (Algorithm 1) for solving general nonconvex nonsmooth composite optimization problems and demonstrated its global convergence. For minimizing the sum of the convex or nonconvex loss function and

four different DC regularizers, our numerical results have shown that the proposed algorithm outperforms other three existing methods in the most of the testing instances. The convergence rate of the proposed method remains unknown. It is likely that the method is linearly convergent only due to the use of linear approximation to the nonconvex term. It is important to improve the convergence rate of the method. It is also very interesting to extend the method to solve problems where f is not twice continuously differentiable.

Acknowledgments

The authors are grateful to the associate editor and two anonymous referees for their comments which helped us to improve our manuscript essentially.

References

[1] A. Beck, *First-order methods in optimization*, Society for Industrial and Applied Mathematics, 2017.

[2] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences* 2 (2009) 183–202.

[3] S. Becker and M.-J. Fadili, A quasi-Newton proximal splitting method, in: *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 2627–2635.

[4] R.H. Byrd, J. Nocedal and F. Oztoprak, An inexact successive quadratic approximation method for L-1 regularized optimization, *Mathematical Programming* 157 (2016) 375–396.

[5] E.J. Candès, M.B. Wakin and S.P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *Journal of Fourier Analysis and Applications* 14 (2008) 877–905.

[6] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* 2 (2011) 1–27.

[7] F.H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.

[8] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* 96 (2001) 1348–1360.

[9] J. Fan, H. Liu, Q. Sun and T. Zhang, I-LAMM for sparse learning: simultaneous control of algorithmic complexity and statistical error, *The Annals of Statistics* 46 (2018) 814–841.

[10] J. Friedman, T. Hastie, H. Höfling and R. Tibshirani, Pathwise coordinate optimization, *The Annals of Applied Statistics* 1 (2007) 302–332.

[11] P. Gong, C. Zhang, Z. Lu, J. Huang and J. Ye, A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems, in: *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, ICML'13, 2013, pp. 37–45.

[12] C.-J. Hsieh, I.S. Dhillon, P. Ravikumar and M.A.Sustik, Sparse inverse covariance matrix estimation using quadratic approximation, in: *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 2330–2338.

[13] J.D. Lee, Y. Sun and M.A. Saunders, Proximal Newton-type methods for minimizing composite functions, *SIAM Journal on Optimization* 24 (2014) 1420–1443.

[14] D.-H. Li, M. Fukushima, L. Qi and N. Yamashita, Regularized Newton methods for convex minimization problems with singular solutions, *Computational Optimization and Applications* 28 (2004) 131–147.

[15] H. Li and Z. Lin, Accelerated proximal gradient methods for nonconvex programming, in: *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 379–387.

[16] X. Li, D. Sun and K.-C. Toh, A highly efficient semismooth Newton augmented lagrangian method for solving lasso problems, *SIAM Journal on Optimization* 28 (2018). 433–458.

[17] T. Liu and A. Takeda, An inexact successive quadratic approximation method for a class of difference-of-convex optimization problems, *Computational Optimization and Applications* 82 (2022) 141–173.

[18] Z. Lu, Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming, *Mathematical Programming* 147 (2014) 277–307.

[19] B. S. Mordukhovich, X. Yuan, S. Zeng and J. Zhang, A globally convergent proximal Newton-type method in nonsmooth convex optimization, *Mathematical Programming* 198 (2023) 899–936.

[20] S. Nakayama, Y. Narushima and H. Yabe, Inexact proximal DC Newton-type method for nonconvex composite functions, *Computational Optimization and Applications* 87 (2024) 611–640.

[21] P. Ochs, Y. Chen, T. Brox and T. Pock, iPiano: inertial proximal algorithm for nonconvex optimization, *SIAM Journal on Imaging Sciences* 7 (2014) 1388–1419.

[22] P.A. Olsen, F. Oztoprak, J. Nocedal and S.J. Rennie, Newton-like methods for sparse inverse covariance estimation, in: *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 764–772.

[23] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[24] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer, 1998.

[25] M. Schmidt, E. Berg, M. Friedlander and K. Murphy, Optimizing costly functions with simple constraints: a limited-memory projected quasi-Newton algorithm, in: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, vol. 5, PMLR, 2009, pp. 456–463.

[26] P.D. Tao and L.H. An, Convex analysis approach to D.C. programming: theory, algorithms and applications, *Acta Mathematica Vietnamica* 22 (1997) 289–355.

[27] R. Tibshirani, *Regression shrinkage and selection via the lasso*, *Journal of the Royal Statistical Society*: Series B, 58 (1996) 267–288.

[28] J.F. Toland, A duality principle for non-convex optimisation and the calculus of variations, *Archive for Rational Mechanics and Analysis* 71 (1979) 41–61.

[29] G.-X. Yuan, C.-H. Ho, C.-J. Lin and S.S. Keerthi, An improved GLMNET for L1-regularized logistic regression, *Journal of Machine Learning Research* 13 (2012) 1999–2030.

[30] M.-C. Yue, Z. Zhou and A.M.-C. So, A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property, *Mathematical Programming* 174 (2019) 327–358.

[31] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics* 38 (2010) 894–942.

[32] T. Zhang, Analysis of multi-stage convex relaxation for sparse regularization, *Journal of Machine Learning Research* 11 (2010) 1081–1107.

[33] K. Zhong, I.E.-H. Yen, I.S. Dhillon and P.K. Ravikumar, Proximal quasi-Newton for computationally intensive ℓ_1-regularized M-estimators, in: *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 2375–2383.

[34] H. Zou, The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* 101 (2006) 1418–1429.

Manuscript received 25 June 2023
revised 31 October 2023
accepted for publication 7 December 2023

Can Wu
School of Mathematics and Statistics, Hainan University
Haikou, P.R. China
Department of Applied Mathematics, The Hong Kong Polytechnic University
Hung Hom, Hong Kong
E-mail address: 2019010105@m.scnu.edu.cn

Danqi Zhu
School of Mathematical Sciences, South China Normal University
Guangzhou, P.R. China
E-mail address: 2021021979@m.scnu.edu.cn

Dong-Hui Li
School of Mathematical Sciences, South China Normal University
Guangzhou, P.R. China
E-mail address: lidonghui@m.scnu.edu.cn