Mathematical Programming manuscript No.
(will be inserted by the editor)

Complexity Analysis of Interior Point Algorithms for Non-Lipschitz and Nonconvex Minimization

Wei Bian · Xiaojun Chen · Yinyu Ye

July 25, 2012, Received: 28 March, 2013; 28 October, 2013 / Accepted: date

Abstract We propose a first order interior point algorithm for a class of non-Lipschitz and nonconvex minimization problems with box constraints, which arise from applications in variable selection and regularized optimization. The objective functions of these problems are continuously differentiable typically at interior points of the feasible set. Our first order algorithm is easy to implement and the objective function value is reduced monotonically along the iteration points. We show that the worst-case iteration complexity for finding an ϵ scaled first order stationary point is $O(\epsilon^{-2})$. Furthermore, we develop a second order interior point algorithm using the Hessian matrix, and solve a quadratic program with a ball constraint at each iteration. Although the second order interior point algorithm costs more computational time than that of the first order algorithm in each iteration, its worst-case iteration complexity for finding an ϵ scaled second order stationary point is reduced to $O(\epsilon^{-3/2})$. Note that an ϵ scaled second order stationary point must also be an ϵ scaled first order stationary point.

Keywords constrained non-Lipschitz optimization · complexity analysis · interior point method · first order algorithm · second order algorithm

This work was supported partly by Hong Kong Research Council Grant PolyU5003/10p, The Hong Kong Polytechnic University Postdoctoral Fellowship Scheme, the NSF foundation (11101107,11271099) of China and US AFOSR Grant FA9550-12-1-0396.

Wei Bian
Department of Mathematics, Harbin Institute of Technology, Harbin, China. E-mail: bianweilvse520@163.com.

Xiaojun Chen
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. E-mail: maxjchen@polyu.edu.hk

Yinyu Ye
Department of Management Science and Engineering, Stanford University, Stanford, CA 94305. E-mail: yinyu-ye@stanford.edu

Mathematics Subject Classification (2000) 90C30 · 90C26 · 65K05 · 49M37

1 Introduction

In this paper, we consider the following optimization problem:

$$\begin{aligned} \min \quad & f(x) = H(x) + \lambda \sum_{i=1}^{n} \varphi(x_i^p) \\ \text{s.t.} \quad & x \in \Omega = \{x : 0 \le x \le b\}, \end{aligned} \tag{1}$$

where $H : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $\varphi : [0, +\infty) \to [0, +\infty)$ is continuous and concave, $\lambda > 0$, $0 < p < 1$, $b = (b_1, b_2, \dots, b_n)^T$ with $b_i \in (0, +\infty) \cup \{+\infty\}$, and $0 \le x \le b$ means that if $b_i < +\infty$, $x_i \in [0, b_i]$, otherwise $x_i \ge 0, i = 1, 2, \dots, n$. Moreover, φ is continuously differentiable in $(0, +\infty)$ and $\varphi(0) = 0$. Without loss of generality, we assume that a minimizer of (1) exists and $\min_\Omega f(x) \ge 0$.

Problem (1) is nonsmooth, nonconvex, and non-Lipschitz, which has been extensively used in image restoration, signal processing and variable selection; see, e.g., [2,10,14,15,19,20,24,31]. The function $H(x)$ is often used as a data fitting term, while the function $\sum_{i=1}^n \varphi(x_i^p)$ is used as a regularization term. The feasible set Ω of (1) includes $\mathbb{R}^n_+ = \{x : x \ge 0\}$ as a special case. Moreover, the unconstrained problem

$$\min \quad H(x) + \lambda \sum_{i=1}^{n} \varphi(|x_i|^p), \tag{2}$$

where the non-Lipschitz points are in the interior of the feasible region, can be equivalently reformulated as the constrained problem

$$\begin{aligned} \min \quad & H(x^+ - x^-) + \lambda \sum_{i=1}^{n} \varphi((x_i^+)^p) + \lambda \sum_{i=1}^{n} \varphi((x_i^-)^p) \\ \text{s.t.} \quad & x^+ \ge 0,\ x^- \ge 0 \end{aligned} \tag{3}$$

by using variable splitting $x = x^+ - x^-$. In Section 3, we show that the scaled first and second order stationary points of problems (2) and (3) are in one-one correspondence. The advantage of (3) is that non-smooth points are only at the boundary of the feasible set which allows us to use the gradient and Hessian of the objective functions in the interior point methods.

Numerical algorithms for nonconvex optimization problems have been studied extensively. However, little theoretical complexity or convergence speed analysis of the algorithms is known, in contrast to the complexity study of convex optimization in the past thirty years. We now review few results on complexity analysis of nonconvex optimization problems.

Smooth, nonconvex. Using an interior point algorithm, Ye [29] proved that an ϵ KKT or first order stationary point of a general quadratic program

$$\min \frac{1}{2}x^T Qx + c^T x \quad \text{s.t. } Ax = q,\ x \geq 0 \tag{4}$$

can be computed in $O(\epsilon^{-1} \log \epsilon^{-1})$ iterations, where each iteration would solve a ball-constrained or trust-region quadratic program that is equivalent to a simple convex minimization problem. Here $Q \in \mathbb{R}^{n\times n}, c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}, q \in \mathbb{R}^m$. Ye [29] also proved that, as $\epsilon \to 0$, the iterative sequence converges to a point satisfying the second order necessary optimality condition. More precisely, the least eigenvalue of Q in the null space of all active constraints is greater than $-\epsilon$.

For general unconstrained nonconvex optimization, it was shown in [17, 21] that the standard steepest descent method with line search or trust-region can find an ϵ first order stationary point in $O(\epsilon^{-2})$ iterations. In [22], Nesterov and Polyak showed that a Newton-type method based on cubic regularization requires at most $O(\epsilon^{-3/2})$ iterations to find an ϵ first order stationary point. Instead of using Hessian and its global Lipschitz constant, Cartis, Gould and Toint [4–6] further proposed an adaptive regularization with cubics (ARC) method in which approximations of Hessian and its Lipschitz constant are updated at each iteration. They have also showed that the ARC takes at most $O(\epsilon^{-3/2})$ iterations to find an ϵ first order stationary point. Then, an algorithm with one-dimensional global optimization of the cubic model is given and the sharpness of the complexity bound $O(\epsilon^{-3/2})$ is derived in [7].

Lipschitz continuous, nonconvex. Cartis, Gould and Toint [3] estimated the worst-case complexity of a first order trust-region or quadratic regularization method for solving the following unconstrained nonsmooth, nonconvex minimization problem

$$\min \quad \Phi_h(x) := H(x) + h(c(x)), \tag{5}$$

where $h : \mathbb{R}^m \to \mathbb{R}$ is convex but may be nonsmooth and $c : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable. Their method takes at most $O(\epsilon^{-2})$ iterations to reduce the size of a first order criticality measure below ϵ, which is as the same order as the worst-case complexity of steepest-descent methods applied to unconstrained smooth nonconvex optimization.

Garmanjani and Vicente [15] proposed a class of smoothing direct-search methods for a general unconstrained nonsmooth optimization by applying a direct-search method to the smoothing function $\tilde{f}$ of the objective function f [8]. Such approach can be considered as the zero order methods because only function values are used. When f is locally Lipschitz, the smoothing direct-search method [15] took at most $O(\epsilon^{-3} \log \epsilon^{-1})$ iterations to find an x such that $\|\nabla \tilde{f}(x, \mu)\| \leq \epsilon$ and $\mu \leq \epsilon$, where μ is the smoothing parameter. When $\mu \to 0$, $\tilde{f}(x, \mu) \to f(x)$ and $\nabla \tilde{f}(x, \mu) \to v$ with $v \in \partial f(x)$.

Non-Lipschitz, nonconvex. Ge, Jiang and Ye [16] extended the complexity result of [29] to the following concave non-Lipschitz minimization

$$\min \sum_{i=1}^{n} x_i^p \quad \text{s.t. } Ax = q,\ x \geq 0 \tag{6}$$

and showed that finding an ϵ scaled first order stationary point or global minimizer requires at most $O(\epsilon^{-1} \log \epsilon^{-1})$ iterations. Recently, Bian and Chen [1] proposed a smoothing quadratic regularization (SQR) algorithm for solving unconstrained non-Lipshchitz minimization problem (2). At each iteration, the SQR algorithm solves a strongly convex quadratic minimization problem with a diagonal Hessian matrix, which has a simple closed form solution. The SQR algorithm is easy to implement and its worst-case complexity of reaching an ϵ scaled first order stationary point is $O(\epsilon^{-2})$. To overcome the nonsmoothness of the second term in the objective function of (2), smoothing methods were used in the SQR algorithm.

In this paper, we propose a first order interior point method and a second order interior point method for solving the constrained non-Lipschitz nonconvex optimization problem (1), using the smoothness of f in $\{x : 0 < x \leq b\}$ and keeping all iterates in it. The former uses the gradient of f to derive a quadratic overestimation and has the worst-case complexity $O(\epsilon^{-2})$ for finding an ϵ scaled first order stationary point of (1). The latter uses the gradient and the Hessian to derive a cubic overestimation and has the worst-case iteration complexity $O(\epsilon^{-3/2})$ for finding an ϵ scaled second order stationary point of a special version of (1) as the following

$$\min_{x \geq 0} H(x) + \lambda \sum_{i=1}^{n} x_i^p. \tag{7}$$

To our best knowledge, these two methods are the first methods with the state of the art iteration complexity bounds for constrained, non-Lipschitz and nonconvex optimization. Specially, the second order interior point algorithm is the first method to find an ϵ scaled second order stationary point or an ϵ global minimizer of non-Lipschitz, nonconvex optimization problem (7) in no more than $O(\epsilon^{-3/2})$ iterations.

The above results are summarized in Table 1.

Our original goal was to produce the $O(\epsilon^{-1} \log \epsilon^{-1})$ worst-case complexity bound for problem (1), as it was established for problems (4) in [29] and (6) in [16]. But we failed even when $H(x)$ is quadratic and we leave this as an open problem. In developing the $O(\epsilon^{-1} \log \epsilon^{-1})$ bound, potential reduction techniques were used, where the quadratic objective or the concavity of $\sum_{i=1}^{n} x_i^p$ has played a key role. In either case, a quadratic overestimation of the potential function can be constructed and it is to be minimized, which makes the analysis considerably simpler and the convergence rate better. However, when the two objectives in (4) and (6) are added together, which is a special case

	Smooth	Lipschitz continuous	Non-Lipschitz
$O(\epsilon^{-1}\log\epsilon^{-1})$	[Ye 1998] (4)		[Ge et al 2011](6)
$O(\epsilon^{-3/2})$	[Nesterov et al 2006]; [Cartis et al 2011]		**this paper** for (7)
$O(\epsilon^{-2})$	[Nesterov 2004]; [Gratton et al 2008]	[Cartis et al 2011](5)	[Bian et al 2012](2); **this paper** for (1)
$O(\epsilon^{-3}\log\epsilon^{-1})$		[Garmanjani et al 2012]	

Table 1: Worst-case complexity results for nonconvex optimization

of (1) namely

$$\min_{x\geq 0}\frac{1}{2}x^TQx + c^Tx + \lambda\sum_{i=1}^n x_i^p,$$

a quadratic overestimation of the potential function is no longer achievable.

Our paper is organized as follows. In Section 2, a first order interior point algorithm is proposed for solving (1), which only uses ∇f and a Lipschitz constant of H on Ω and is easy to implement. Any iteration point $x^k > 0$ belongs to Ω and the objective function is monotonically decreasing along the generated sequence $\{x^k\}$. Moreover, the algorithm produces an ϵ scaled first order stationary point of (1) in at most $O(\epsilon^{-2})$ iterations. In Section 3, a second order interior point algorithm is given to solve (7), which can generate an ϵ scaled second order stationary point in at most $O(\epsilon^{-3/2})$ iterations. Since our problem has constraints, the ϵ scaled first order and second order stationary points resemble the complementarity condition for all inequality constrained optimization problems. In Section 4, we present numerical results to illustrate the efficiency of the algorithms and the complexity bound.

Throughout this paper, $\mathbb{K} = \{0, 1, 2, \dots\}$, $\mathbb{I} = \{1, 2, \dots, n\}$, $\mathbb{I}_b = \{i \in \{1, 2, \dots, n\} : b_i < +\infty\}$ and $e_n = (1, 1, \dots, 1)^T \in \mathbb{R}^n$. For $x \in \mathbb{R}^n$, $A = (a_{ij})_{m\times n} \in \mathbb{R}^{m\times n}$ and $p > 0$, $\|x\| = \|x\|_2$, $\mathrm{diag}(x) = \mathrm{diag}(x_1, x_2, \dots, x_n)$, $|A|^p = (|a_{ij}|^p)_{m\times n}$. For two matrices $A, B \in \mathbb{R}^{n\times n}$, we denote $A \succeq B$ when $A - B$ is positive semi-definite.

2 First Order Method

In this section, we propose a first order interior point algorithm for solving (1), which uses the first order reduction technique and keeps all iterates $x^k > 0$ in the feasible set Ω. We show that the objective function value $f(x^k)$ is monotonically decreasing along the sequence $\{x^k\}$ generated by the algorithm, and the worst-case complexity of the algorithm for generating an ϵ global minimizer or ϵ scaled first order stationary point of (1) is $O(\epsilon^{-2})$, which is the same in the worst-case complexity order of the steepest-descent methods for smooth nonconvex optimization problems. Moreover, it is worth noting that the proposed first order interior point algorithm is easy to implement, and the computation cost at each iteration is little.

Throughout this section, we need the following assumptions.

Assumption 2.1: ∇H is globally Lipschitz on Ω with a Lipschitz constant $\beta \geq 1$.

Specially, when $\mathbb{I}_b \neq \emptyset$ we choose β such that $\beta \geq \max_{i \in \mathbb{I}_b} \frac{1}{b_i}$.

Assumption 2.2: For any given $x^0 \in \Omega$, there is $R \geq 1$ such that $\sup\{\|x\|_\infty : f(x) \leq f(x_0), x \in \Omega\} \leq R$.

When $H(x) = \frac{1}{2}\|Ax - q\|^2$, Assumption 2.1 holds with $\beta = \|A^T A\|$. Assumption 2.2 holds, if Ω is bounded or $H(x) \geq 0$ for all $x \in \Omega$ and $\varphi(s) \to \infty$ as $s \to \infty$.

2.1 First Order Necessary Condition

Note that for problem (1), when $0 < p < 1$, the Clarke generalized gradient of $\varphi(s^p)$ does not exist at 0. Inspired by the first and second order necessary conditions for local minimizers of unconstrained non-Lipschitz optimization in [11, 12], we give the scaled first and second order necessary condition for the local minimizers of constrained non-Lipschitz optimization (1) in this section and the next section, respectively. Then, for any $\epsilon \in (0, 1]$, the ϵ scaled first order and second order stationary point of (1) can be deduced directly.

First, for $\epsilon > 0$, an ϵ global minimizer of (1) is defined as a feasible solution $0 \leq x_\epsilon \leq b$ and

$$f(x_\epsilon) - \min_{0 \leq x \leq b} f(x) \leq \epsilon.$$

It has been proved in [9] that finding a global minimizer of the unconstrained l_2-l_p minimization problem is strongly NP hard. For any fixed $x \in \mathbb{R}^n$, denote $X = \mathrm{diag}(x)$. Any local minimizer x of the unconstrained l_2-l_p optimization modeled by (2) with $H(x) = \frac{1}{2}\|Ax - q\|^2$ and $\varphi(s) = s$ satisfies the first order necessary condition [12]

$$2XA^T(Ax - q) + \lambda p|x|^p = 0.$$

Similarly, using X as a scaling matrix, if x is a local minimizer of (1), then $x \in \Omega$ satisfies the first order necessary condition

$$[X\nabla f(x)]_i = x_i[\nabla H(x)]_i + \lambda p \varphi'(s)_{s=x_i^p} x_i^p = 0 \qquad \text{if } x_i < b_i; \tag{8a}$$

$$[\nabla f(x)]_i = [\nabla H(x)]_i + \lambda p \varphi'(s)_{s=x_i^p} x_i^{p-1} \leq 0 \qquad \text{if } x_i = b_i. \tag{8b}$$

For (1), if x at which f is differentiable satisfies the first order necessary condition given above, then $x \in \Omega$ satisfies

(i) $[\nabla f(x)]_i = 0$ if $x_i < b_i$;
(ii) $[\nabla f(x)]_i \leq 0$ if $x_i = b_i$.

Moreover, although $[\nabla f(x)]_i$ does not exist when $x_i = 0$, one can see that, as $x_i \to 0+$, $[\nabla f(x)]_i \to +\infty$.

Now we can define an ϵ scaled first order stationary point of (1).

Definition 1 For a given $\epsilon \in (0,1]$, we call $x \in \Omega$ an ϵ scaled first order stationary point of (1), if

(i) $|[X\nabla f(x)]_i| \leq \epsilon$ if $x_i < (1-\frac{1}{2}\epsilon)b_i$;
(ii) $[\nabla f(x)]_i \leq \epsilon$ if $x_i \geq (1-\frac{1}{2}\epsilon)b_i$.

Definition 1 is consistent with the first order necessary conditions in (8a)-(8b) with $\epsilon = 0$.

2.2 First Order Interior Point Algorithm

Note that for any $x, x^+ \in (0,b]$, Assumption 2.1 implies that

$$H(x^+) \leq H(x) + \langle \nabla H(x), x^+ - x\rangle + \frac{\beta}{2}\|x^+ - x\|^2.$$

Since φ is concave on $[0,+\infty)$, then for any s, $t \in (0,+\infty)$,

$$\varphi(t) \leq \varphi(s) + \langle \nabla\varphi(s), t-s\rangle.$$

Thus, for any x, $x^+ \in (0,b]$, we obtain

$$f(x^+) \leq f(x) + \langle \nabla f(x), x^+ - x\rangle + \frac{\beta}{2}\|x^+ - x\|^2.$$

Let $x^+ = x + Xd_x$. We obtain

$$f(x^+) \leq f(x) + \langle X\nabla f(x), d_x\rangle + \frac{\beta}{2}\|Xd_x\|^2. \tag{9}$$

To achieve a reduction of the objective function at each iteration, we minimize a quadratic function subject to a box constraint at each iteration when $x > 0$, which is

$$\begin{aligned} \min \quad & \langle X\nabla f(x), d_x\rangle + \frac{\beta}{2}d_x^T X^2 d_x \\ \text{s.t.} \quad & -\frac{1}{2}e_n \leq d_x \leq X^{-1}(b-x). \end{aligned} \tag{10}$$

For any fixed $x \in (0,b]$, the objective function in (10) is strongly convex and separable about every element of x, and thus the unique solution of (10) has a closed form as

$$d_x = P_{\mathcal{D}_x}[-\frac{1}{\beta}X^{-1}\nabla f(x)],$$

where $\mathcal{D}_x = [-\frac{1}{2}e_n, X^{-1}(b-x)]$ and $P_{\mathcal{D}_x}$ is the orthogonal projection operator on the box $\mathcal{D}_x$.

Denote $X_k = \mathrm{diag}(x^k)$ and use d_k and $\mathcal{D}_k$ to denote d_{x^k} and $\mathcal{D}_{x^k}$, respectively.

First Order Interior Point Algorithm

Give $\epsilon \in (0,1]$ and choose $x^0 \in (0,b]$.

For $k \geq 0$, set

$$d_k = P_{\mathcal{D}_k}[-\frac{1}{\beta}X_k^{-1}\nabla f(x^k)], \tag{11a}$$

$$x^{k+1} = x^k + X_k d_k. \tag{11b}$$

The first order interior point algorithm presented in the current paper is significantly different from the classical interior point methods [28]. The current method is based on a simple gradient projection into the interior of the nonnegative orthant, while the classical one follows the central path of the logarithmic barrier function via the Newton method.

Lemma 1 *The proposed First Order Interior Point Algorithm is well defined, which means that* $0 < x^k \leq b$, $\forall k \in \mathbb{K}$.

Proof We only need to prove that if $0 < x^k \leq b$, then $0 < x^{k+1} \leq b$.

On the one hand, by $d_k \leq X_k^{-1}(b - x^k)$, we have

$$x^{k+1} = x^k + X_k d_k \leq x^k + (b - x^k).$$

On the other hand, using $d_k \geq -\frac{1}{2}e_n$, we obtain

$$x^{k+1} = x^k + X_k d_k \geq x^k - \frac{1}{2}x^k = \frac{1}{2}x^k > 0.$$

Hence, $0 < x^{k+1} \leq b$.

Lemma 2 *Let* $\{x^k\}$ *be the sequence generated by the First Order Interior Point Algorithm, then the sequence* $\{f(x^k)\}$ *is monotonely decreasing and satisfies*

$$f(x^{k+1}) - f(x^k) \leq -\frac{\beta}{2}\|X_k d_k\|^2 = -\frac{\beta}{2}\|x^{k+1} - x^k\|^2.$$

Moreover, we have $\|x^k\|_\infty \leq R$.

Proof From the KKT condition of (10), the solution d_k of quadratic programming (10) satisfies the necessary and sufficient condition as follows

$$\begin{aligned} &\beta X_k^2 d_k + X_k \nabla f(x^k) - \mu_k + \nu_k = 0, \quad -\frac{1}{2}e_n \leq d_k \leq X_k^{-1}(b - x^k), \\ &M_k(d_k + \frac{1}{2}e_n) = 0,\ N_k(d_k - X_k^{-1}(b - x^k)) = 0, \end{aligned} \tag{12}$$

where $M_k = \mathrm{diag}(\mu_k)$ and $N_k = \mathrm{diag}(\nu_k)$ with $M_k,\ N_k \succeq 0$.

Moreover, from (9), we obtain

$$\begin{aligned} f(x^{k+1}) - f(x^k) \leq & \langle X_k \nabla f(x^k), d_k \rangle + \frac{\beta}{2} \|X_k d_k\|^2 \\ = & \langle -\beta X_k^2 d_k + \mu_k - \nu_k, d_k \rangle + \frac{\beta}{2} \|X_k d_k\|^2 \\ = & -\frac{\beta}{2} \|X_k d_k\|^2 + \mu_k^T d_k - \nu_k^T d_k \\ = & -\frac{\beta}{2} \|X_k d_k\|^2 - \frac{1}{2} \mu_k^T e_n - \nu_k^T X_k^{-1} (b - x^k). \end{aligned} \tag{13}$$

By $\mu_k, \nu_k \geq 0$ and $0 < x^k \leq b$, we obtain the inequality given in this lemma, which implies that $f(x^k) \leq f(x^0)$, $k \in \mathbb{K}$.

From Assumption 2.2, we obtain $\|x^k\|_\infty \leq R$.

Different from some other potential reduction methods, the objective function is monotonelly decreasing along the sequence generated by the First Order Interior Point Algorithm.

Theorem 1 *For any $\epsilon \in (0, 1]$, the First Order Interior Point Algorithm obtains an ϵ scaled first order stationary point or ϵ global minimizer of (1) in no more than $O(\epsilon^{-2})$ iterations.*

Proof Let $\{x^k\}$ be the sequence generated by the proposed First Order Interior Point Algorithm. Then $x^k \in (0, b]$ and $\|x^k\|_\infty \leq R$, $\forall k \in \mathbb{K}$. Without loss of generality, we suppose that $R \geq 1$.

In the following, we will consider four cases.

Case 1: $\|X_k d_k\| \geq \frac{1}{4\beta R} \epsilon$.

From Lemma 2, we obtain that

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2 \times 4^2 R^2 \beta} \epsilon^2 = -\frac{1}{32 R^2 \beta} \epsilon^2.$$

Case 2: $\mu_k^T e_n \geq \frac{1}{4\beta} \epsilon^2$.

From Lemma 2, we get

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{8\beta} \epsilon^2.$$

Case 3: $\nu_k^T X_k^{-1} (b - x^k) \geq \frac{1}{4} \epsilon^2$.

From (13), we get

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{4} \epsilon^2.$$

Case 4: $\|X_k d_k\| < \frac{1}{4\beta R} \epsilon$, $\mu_k^T e_n < \frac{1}{4\beta} \epsilon^2$ and $\nu_k^T X_k^{-1} (b - x^k) \leq \frac{1}{4} \epsilon^2$.

From the first condition in (12), we obtain that

$$\beta X_k d_k + \nabla f(x^k) - X_k^{-1} \mu_k + X_k^{-1} \nu_k = 0, \tag{14}$$

which implies that

$$X_k\nabla f(x^k) = -\beta X_k^2 d_k + \mu_k - \nu_k. \tag{15}$$

Then, we obtain

$$|[X_k\nabla f(x^k)]_i| \le \beta\|X_k\|_\infty\|X_k d_k\| + \|\mu_k\|_\infty + |[\nu_k]_i| \le \frac{1}{2}\epsilon + |[\nu_k]_i|. \tag{16}$$

If $i \notin \mathbb{I}_b$, $[\nu_k]_i = 0$ and we have $|[X_k\nabla f(x^k)]_i| \le \frac{1}{2}\epsilon$. Fix $i \in \mathbb{I}_b$. We consider two subclasses in this case.

Case 4.1: $x_i^k < b_i - \frac{b_i}{2}\epsilon$.

Then, $\frac{b_i - x_i^k}{x_k^i} \ge \frac{b_i\epsilon}{2b_i} \ge \frac{\epsilon}{2}$, which gives $[\nu_k]_i \le \frac{\epsilon}{2}$ from $\nu_k^T X_k^{-1}(b - x^k) \le \frac{1}{4}\epsilon^2$

Then, (16) gives $|[X_k\nabla f(x^k)]_i| \le \epsilon$.

Case 4.2: $x_i^k \ge b_i - \frac{b_i}{2}\epsilon$.

Then, $x_i^k \ge \frac{b_i}{2}$. By $\|\mu_k\|_\infty \le \frac{1}{4\beta}\epsilon^2$, we have $|[X_k^{-1}\mu_k]_i| \le \frac{\epsilon^2}{2\beta b_i} \le \frac{\epsilon}{2}$. By $[\nu_k]_i \ge 0$, we obtain

$$\begin{aligned}
[\nabla f(x^k)]_i =& -[\beta X_k d_k - X_k^{-1}\mu_k + X_k^{-1}\nu_k]_i \\
\le& \beta\|X_k d_k\|_\infty + |[X_k^{-1}\mu_k]_i| - [X_k^{-1}\nu_k]_i \\
\le& \beta\|X_k d_k\|_\infty + \frac{1}{2}\epsilon - [X_k^{-1}\nu_k]_i \le \epsilon - [X_k^{-1}\nu_k]_i \le \epsilon.
\end{aligned}$$

Therefore, from the analysis in Cases 4.1 - 4.2, x^k is an ϵ scaled first order stationary point of (1).

Basing on the above analysis in Cases 1 - 3, at least one of the following two facts holds at the kth iteration:

(i) $f(x^{k+1}) - f(x^k) \le -\frac{1}{32R^2\beta}\epsilon^2$;

(ii) x^k is an ϵ scaled first order stationary point of (1).

Therefore, we would produce an ϵ global minimizer or an ϵ scaled first order stationary point of (1) in at most $32f(x^0)R^2\beta\epsilon^{-2}$ iterations.

Remark 1 When $\lambda = 0$ or $p = 1$ in (1), the KKT condition of (1) can be written as

$$\begin{cases} [\nabla f(x)]_i \ge 0 & \text{if } x_i = 0, \\ [\nabla f(x)]_i = 0 & \text{if } 0 < x_i < b_i, \\ [\nabla f(x)]_i \le 0 & \text{if } x_i = b_i. \end{cases} \tag{17}$$

which are sufficient but not necessary for the conditions in (8).

Denote $\bar\epsilon = \min\{1, \min_{i\in\mathbb{I}_b}\frac{b_i}{2}\}$. Similar to the analysis in Theorem 1, from (14), for $\epsilon \in (0, \bar\epsilon]$, if $\|X_k d_k\| < \frac{1}{4\beta R}\epsilon$, $\mu_k^T e_n < \frac{1}{4\beta}\epsilon^2$ and $\nu_k^T X_k^{-1}(b - x^k) \le \frac{1}{4}\epsilon^2$, we can obtain the following estimation

$$\begin{cases} [\nabla f(x^k)]_i \ge -\dfrac{1}{2}\epsilon & \text{if } 0 \le x_i^k \le \epsilon, \\ |[\nabla f(x^k)]_i| \le \dfrac{1}{2}\epsilon + \dfrac{1}{2b_i}\epsilon & \text{if } \epsilon < x_i^k < (1 - \frac{\epsilon}{2})b_i \\ [\nabla f(x^k)]_i \le \epsilon & \text{if } x_i^k \ge (1 - \frac{\epsilon}{2})b_i. \end{cases} \tag{18}$$

When $\epsilon = 0$, the conditions in (18) are consistent with the KKT condition of (1) in (17). Thus, we can state that the First Order Interior Point Algorithm can be extended to (1) with $\lambda = 0$ or $p = 1$ for finding an ϵ KKT point with the worst-case complexity $O(\epsilon^{-2})$, which recovers the complexity bound of the steepest descent method for constrained nonconvex optimization.

3 Second Order Interior Point Algorithm

In this section, we consider problem (7), a special case of (1) with $\Omega = \{x : x \geq 0\}$, under Assumption 2.2 and the following assumption on H.

Assumption 3.1: H is twice continuously differentiable and $\nabla^2 H$ is globally Lipschitz on Ω with Lipschitz constant γ.

By using the cubic overestimation idea [4–6, 18, 22, 23], a second order interior point algorithm is proposed for solving (7), which uses the Hessian of H. We show that the worst-case complexity of the second order interior point algorithm for finding an ϵ scaled second order stationary point of (7) is $O(\epsilon^{-3/2})$. Comparing with the first order interior point algorithm proposed in Section 2, the worst-case complexity of the second order interior point algorithm is better and the generated point satisfies stronger optimality conditions. However, a quadratic program with ball constraint has to be solved at each iteration.

3.1 Second Order Necessary Condition for (7)

Based on the scaled second order necessary condition for unconstrained non-Lipschitz optimization in [11, 12], we know that if x is a local minimizer of (7), then $x \in \Omega$ satisfies

$$X\nabla^2 f(x)X = X\nabla^2 H(x)X + \lambda p(p-1)X^p \succeq 0. \tag{19}$$

If $x > 0$, then f is differentiable at x and (19) implies that $\nabla^2 f(x) \succeq 0$.

Now we give the definition of the ϵ scaled second order stationary point of (7) as follows.

Definition 2 For a given $\epsilon \in (0, 1]$, we call $x \in \Omega$ an ϵ scaled second order stationary point of (7), if

$$\|X\nabla f(x)\|_\infty \leq \epsilon \quad \text{and} \quad X\nabla^2 f(x)X \succeq -\sqrt{\epsilon} I_n.$$

Definition 2 is consistent with the scaled first and second order necessary conditions given above when $\epsilon = 0$.

3.2 Second Order Interior Point Algorithm

From the cubic overestimation idea, for any $x, x^+ \in \Omega$, Assumption 3.1 implies that

$$\begin{aligned} H(x^+) \leq & H(x) + \langle \nabla H(x), x^+ - x\rangle \\ & + \frac{1}{2}\langle \nabla^2 H(x)(x^+ - x), x^+ - x\rangle + \frac{1}{6}\gamma\|x^+ - x\|^3. \end{aligned}$$

Similarly, for any $t, s > 0$,

$$\begin{aligned} t^p \leq & s^p + \langle ps^{p-1}, t - s\rangle \\ & + \frac{p(p-1)}{2}s^{p-2}(t-s)^2 + \frac{p(p-1)(p-2)}{6}s^{p-3}(t-s)^3. \end{aligned}$$

Thus, for any x, $x^+ > 0$, we obtain

$$\begin{aligned} f(x^+) - f(x) \leq & \langle \nabla f(x), x^+ - x\rangle + \frac{1}{2}\langle \nabla^2 f(x)(x^+ - x), x^+ - x\rangle \\ & + \frac{1}{6}\gamma\|x^+ - x\|^3 + \lambda\frac{p(p-1)(p-2)}{6}\sum_{i=1}^n x_i^{p-3}(x_i^+ - x_i)^3, \end{aligned}$$

which can also be expressed by

$$\begin{aligned} f(x^+) - f(x) \leq & \langle X\nabla f(x), d_x\rangle + \frac{1}{2}\langle X\nabla^2 f(x)Xd_x, d_x\rangle \\ & + \frac{1}{6}\gamma\|Xd_x\|^3 + \frac{\lambda p(p-1)(p-2)}{6}\sum_{i=1}^n x_i^p[d_x]_i^3. \end{aligned} \tag{20}$$

Since $p(1-p) \leq \frac{1}{4}$, then $p(1-p)(2-p) \leq \frac{1}{2}$. Combining this estimation with $\sum_{i=1}^n [d_x]_i^3 \leq \|d_x\|^3$, if $\|x\|_\infty \leq R$, then (20) implies

$$\begin{aligned} f(x^+) - f(x) \leq & \langle X\nabla f(x), d_x\rangle + \frac{1}{2}\langle X\nabla^2 f(x)Xd_x, d_x\rangle \\ & + \frac{1}{6}(\gamma R^3 + \frac{1}{2}\lambda R^p)\|d_x\|^3. \end{aligned} \tag{21}$$

At $x > 0$, we minimize a quadratic function subject to a ball constraint to achieve a sufficient reduction. For a given $\epsilon \in (0, 1]$, we solve the following problem

$$\begin{aligned} \min \quad & q(d_x) = \langle X\nabla f(x), d_x\rangle + \frac{1}{2}\langle X\nabla^2 f(x)Xd_x, d_x\rangle \\ \text{s.t.} \quad & \|d_x\|^2 \leq \vartheta^2\epsilon \end{aligned} \tag{22}$$

where $\vartheta = \frac{1}{2}\min\{\frac{1}{\gamma R^3 + \frac{1}{2}\lambda R^p}, 1\}$ and R is the constant in Assumption 2.2.

To solve (22), we consider two cases.

Case 1: $X\nabla^2 f(x)X$ is positive semi-definite.

From the KKT condition of (22), the solution d_x of (22) satisfies the following necessary and sufficient conditions

$$X\nabla^2 f(x)Xd_x + X\nabla f(x) + \rho_x d_x = 0, \tag{23a}$$

$$\rho_x \geq 0, \quad \|d_x\|^2 \leq \vartheta^2\epsilon, \quad \rho_x(\|d_x\|^2 - \vartheta^2\epsilon) = 0. \tag{23b}$$

In this case, (22) is a convex quadratic program with ball constraint, which can be solved effectively in polynomial time, see [27] and references therein.

Case 2: $X\nabla^2 f(x)X$ has at least one negative eigenvalue.

From [26], the solution d_x of (22) satisfies the following necessary and sufficient conditions

$$X\nabla^2 f(x)Xd_x + X\nabla f(x) + \rho_x d_x = 0, \tag{24a}$$

$$\|d_x\|^2 = \vartheta^2\epsilon, \quad X\nabla^2 f(x)X + \rho_x I_n \quad \text{is positive semi-definite.} \tag{24b}$$

In this case, (22) is a nonconvex quadratic programming with ball constraint. However, by the results in [29,30], (24) can be solved effectively in polynomial time with the worst-case complexity $O(\log(\log(\epsilon^{-1})))$. The double log is established based on a globally convergent Newton method. In the first phase the method selects a point from at most $-\log\log(\epsilon)$ many candidate points using the bisection method. Then it is proved that, starting from the selected point, the quadratic convergence of the Newton method is guaranteed.

From (23) and (24), if d_x solves (22), then

$$\begin{aligned} q(d_x) =& \langle X\nabla f(x), d_x\rangle + \frac{1}{2}\langle X\nabla^2 f(x)Xd_x, d_x\rangle \\ =& \langle -X\nabla^2 f(x)Xd_x - \rho_x d_x, d_x\rangle + \frac{1}{2}\langle X\nabla^2 f(x)Xd_x, d_x\rangle \\ =& -\rho_x\|d_x\|^2 - \frac{1}{2}d_x^T X\nabla^2 f(x)Xd_x. \end{aligned} \tag{25}$$

Since $X\nabla^2 f(x)X + \rho_x I_n$ is always positive semi-definite,

$$d_x^T X\nabla^2 f(x)Xd_x \geq -\rho_x\|d_x\|^2,$$

and thus

$$q(d_x) \leq -\frac{1}{2}\rho_x\|d_x\|^2.$$

Therefore, from (21), we have

$$f(x^+) - f(x) \leq -\frac{1}{2}\rho_x\|d_x\|^2 + \frac{1}{6}(\gamma R^3 + \frac{1}{2}\lambda R^p)\|d_x\|^3. \tag{26}$$

Second Order Interior Point Algorithm

```
Choose $x^0 \in \mathrm{int}(\Omega)$ and $\epsilon \in (0,1]$.
For $k \geq 0$,
        solve (22) with $x = x^k$ for $d_k$        (27a)
        $x^{k+1} = x^k + X_k d_k$.                  (27b)
```

The Second Order Interior Point Algorithm is related to the classical interior point methods [28]. The computational work of each iteration is identical to the algorithm in [29]. However, in [29] the objective is a quadratic overestimation of the Karmarkar-type potential function, and in the current paper the objective is just the Taylor quadratic expansion of the original objective function. The main differences are: 1) the current paper deals with more general objective function and [29] deals with only quadratic function, which is why different convergence rates are established. For quadratic functions, there would be no third-order errors so that the analysis was considerably simpler and a better convergence rate was achieved in [29]. In fact, it is a surprise to us that we are able to establish the current convergence rate for such general objectives. 2) [29] needs a prior known lower bound for the objective function in order to construct a valid potential function, but the current method does not need such information.

From the definition of ϑ in (22), $\|d_k\|_\infty \leq \frac{1}{2}$, similar to the analysis in Lemma 1, the Second Order Interior Point Algorithm is also well defined. Let $\{x^k\}$ be the sequence generated by it, then $x^k > 0$. In what follows, we will prove that there is $\kappa > 0$ such that either $f(x^{k+1}) - f(x^k) \leq -\kappa\epsilon^{\frac{3}{2}}$ or x^{k+1} is an ϵ scaled second order stationary point of (7).

Lemma 3 *If $\rho_k > \frac{2}{9\vartheta}\|d_k\|$ holds for all $k \in \mathbb{K}$, then the Second Order Interior Point Algorithm produces an ϵ global minimizer of (1) in at most $O(\epsilon^{-3/2})$ iterations. Moreover, $\|x^k\|_\infty \leq R$, $\forall k \in \mathbb{K}$.*

Proof If $\rho_k > \frac{2}{9\vartheta}\|d_k\|$, then $\rho_k > \frac{4}{9}(\gamma R^3 + \frac{1}{2}\lambda R^p)\|d_k\|$ and we have that

$$\frac{1}{6}(\gamma R^3 + \frac{1}{2}\lambda R^p)\|d_k\| < \frac{3}{8}\rho_k, \tag{28}$$

and from (23b) and (24b), we obtain $\|d_k\|^2 = \vartheta^2\epsilon$.

From (26) and (28), we have

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\leq -\frac{1}{2}\rho_k\|d_k\|^2 + \frac{1}{6}(\gamma R^3 + \frac{1}{2}\lambda R^p)\|d_k\|^3 \\ &< -\frac{1}{2}\rho_k\|d_k\|^2 + \frac{3}{8}\rho_k\|d_k\|^2 = -\frac{1}{8}\rho_k\|d_k\|^2, \end{aligned}$$

which means $f(x^{k+1}) < f(x^k)$.

If this always holds, then $f(x^k)$ is strictly monotone decreasing. By Assumption 2.2, $\|x^k\|_\infty \leq R$, $k \in \mathbb{K}$. Moreover,

$$f(x^k) - f(x^0) \leq -\frac{k}{8}\rho_k\|d_k\|^2 \leq -\frac{k}{36}\vartheta^2\epsilon^{3/2},$$

which follows that we would produce an ϵ global minimizer of (1) in at most $36f(x^0)\vartheta^{-2}\epsilon^{-3/2}$ iterations.

In what follows, we prove that x^{k+1} is an ϵ scaled second order stationary point of (7) when $\rho_k \leq \frac{2}{9\vartheta}\|d_k\|$ for some k.

Lemma 4 *If there is $k \in \mathbb{K}$ such that $\rho_k \leq \frac{2}{9\vartheta}\|d_k\|$, then x^{k+1} satisfies*

$$\|X_{k+1}\nabla f(x^{k+1})\|_\infty \leq \epsilon.$$

Proof From (23a) and (24a), the following relation always holds

$$\begin{aligned}-\rho_k d_k =& X_k\nabla^2 f(x^k)X_k d_k + X_k\nabla f(x^k)\\ =& X_k\nabla^2 H(x^k)X_k d_k + \lambda p(p-1)X_k^{p-2}X_k^2 d_k + X_k\nabla H(x^k) + \lambda p X_k(x^k)^{p-1}\\ =& X_k(\nabla^2 H(x^k)X_k d_k + \lambda p(p-1)X_k^{p-2}X_k d_k + \nabla H(x^k) + \lambda p(x^k)^{p-1}),\end{aligned}$$

which implies

$$\nabla^2 H(x^k)X_k d_k + \lambda p(p-1)X_k^{p-2}X_k d_k + \nabla H(x^k) + \lambda p(x^k)^{p-1} + \rho_k X_k^{-1}d_k = 0.$$

Thus, there is $\tau \in [0,1]$ such that

$$\begin{aligned}&\nabla H(x^{k+1}) + \lambda p(x^{k+1})^{p-1}\\ =&\nabla H(x^{k+1}) + \lambda p(x^{k+1})^{p-1} - \nabla^2 H(x^k)X_k d_k - \lambda p(p-1)X_k^{p-2}X_k d_k - \nabla H(x^k)\\ &- \lambda p(x^k)^{p-1} - \rho_k X_k^{-1}d_k\\ =&\nabla^2 H(\tau x^k + (1-\tau)x^{k+1})X_k d_k - \nabla^2 H(x^k)X_k d_k\\ &+ \lambda p(x^{k+1})^{p-1} - \lambda p(p-1)X_k^{p-2}X_k d_k - \lambda p(x^k)^{p-1} - \rho_k X_k^{-1}d_k.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\|X_{k+1}(\nabla H(x^{k+1}) + \lambda p(x^{k+1})^{p-1})\|_\infty\\ \leq&\|X_{k+1}(\nabla^2 H(\tau x^k + (1-\tau)x^{k+1})X_k d_k - \nabla^2 H(x^k)X_k d_k)\|_\infty\\ &+ \|X_{k+1}(\lambda p(x^{k+1})^{p-1} - \lambda p(p-1)X_k^{p-2}X_k d_k - \lambda p(x^k)^{p-1})\|_\infty\\ &+ \rho_k\|X_{k+1}X_k^{-1}d_k\|_\infty.\end{aligned} \tag{29}$$

From Assumption 3.1, we estimate the first term after the inequality in (29) as follows

$$\begin{aligned}&\|X_{k+1}\|_\infty\|\nabla^2 H(\tau x^k + (1-\tau)x^{k+1})X_k d_k - \nabla^2 H(x^k)X_k d_k\|_\infty\\ \leq&\gamma(1-\tau)\|X_{k+1}\|_\infty\|X_k d_k\|_\infty^2 \leq \gamma\|X_{k+1}\|_\infty\|X_k\|_\infty^2\|d_k\|_\infty^2 \leq \gamma R^3\|d_k\|_\infty^2.\end{aligned} \tag{30}$$

From $X_k D_k = X^{k+1} - X^k$, we have

$$X_k^{-1}X_{k+1} = D_k + I_n, \tag{31}$$

with $D_k = \mathrm{diag}(d_k)$, which implies

$$X_k^{-1}x^{k+1} = d_k + e_n.$$

Then, we consider the second term in (29),

$$\begin{aligned}&X_{k+1}(\lambda p(x^{k+1})^{p-1}-\lambda p(p-1)X_k^{p-2}X_k d_k-\lambda p(x^k)^{p-1})\\=&\lambda p X_k^p(X_k^{-p}(x^{k+1})^p+(1-p)X_k^{-1}X_{k+1}d_k-X_k^{-1}x^{k+1})\\=&\lambda p X_k^p((d_k+e_n)^p+(1-p)(D_k+I_n)d_k-(d_k+e_n))\\=&\lambda p X_k^p((d_k+e_n)^p+(1-p)d_k^2-pd_k-e_n).\end{aligned} \tag{32}$$

Using the Taylor expansion that $1+pt-\frac{p(1-p)}{2}t^2\le(1+t)^p\le 1+pt$, we obtain

$$e_n+pd_k-\frac{p(1-p)}{2}d_k^2\le(d_k+e_n)^p\le e_n+pd_k. \tag{33}$$

Adding $(1-p)d_k^2-pd_k-e_n$ into (33), we have

$$0\le(d_k+e_n)^p+(1-p)d_k^2-pd_k-e_n\le(1-p)d_k^2.$$

Thus,

$$\|(d_k+e_n)^p+(1-p)d_k^2-pd_k-e_n\|_\infty\le(1-p)\|d_k\|_\infty^2. \tag{34}$$

Then, from (32) and (34), we get

$$\begin{aligned}&\|X_{k+1}(\lambda p(x^{k+1})^{p-1}-\lambda p(p-1)X_k^{p-2}X_k d_k-\lambda p(x^k)^{p-1})\|_\infty\\\le&\lambda p\|X_k^p\|_\infty\|(d_k+e_n)^p+(1-p)d_k^2-pd_k-e_n\|_\infty\\\le&\lambda p(1-p)\|X_k^p\|_\infty\|d_k\|_\infty^2\le\frac{1}{2}\lambda R^p\|d_k\|_\infty^2.\end{aligned} \tag{35}$$

From $\|d_k\|_\infty\le\frac{1}{2}$, we have

$$\rho_k\|X_{k+1}X_k^{-1}d_k\|_\infty\le\quad\rho_k\|d_k\|_\infty(1+\|d_k\|_\infty)\le\frac{3}{3}\rho_k\|d_k\|_\infty. \tag{36}$$

Therefore, from (29), (30), (35) and (36), we obtain

$$\begin{aligned}&\|X_{k+1}\nabla H(x^{k+1})+\lambda p(x^{k+1})^p\|_\infty\\\le&(\gamma R^3+\frac{1}{2}\lambda R^p)\|d_k\|_\infty^2+\frac{3}{2}\rho_k\|d_k\|_\infty\\=&\frac{1}{2}\epsilon+\frac{1}{6}\epsilon\le\frac{2}{3}\epsilon.\end{aligned}$$

Lemma 5 *Under the assumptions in Lemma 4, x^{k+1} satisfies*

$$X_{k+1}\nabla^2 f(x^{k+1})X_{k+1}\succeq-\sqrt{\epsilon}I_n.$$

Proof From (23b) and (24b), we know

$$X_k\nabla^2 f(x^k)X_k+\rho_k I_n \text{ is positive semi-definite.}$$

Then,

$$\nabla^2 H(x^k)+\lambda p(p-1)X_k^{p-2}\succeq-\rho_k X_k^{-2}; \tag{37}$$

From Assumption 3.1, we obtain

$$\|\nabla^2 H(x^{k+1}) - \nabla^2 H(x^k)\| \leq \gamma \|x^k - x^{k+1}\| \leq \gamma \|X_k\|_\infty \|d_k\| \leq \gamma R \|d_k\|. \quad (38)$$

Note that $\nabla^2 H(x^{k+1})$ and $\nabla^2 H(x^k)$ are symmetric, (38) gives

$$\nabla^2 H(x^{k+1}) - \nabla^2 H(x^k) \succeq -\gamma R \|d_k\| I_n. \quad (39)$$

Adding (37) and (39), we get

$$\nabla^2 H(x^{k+1}) \succeq -\lambda p(p-1) X_k^{p-2} - \rho_k X_k^{-2} - \gamma R \|d_k\| I_n. \quad (40)$$

Adding $\lambda p(p-1) X_{k+1}^{p-2}$ into the both sides of (40), we obtain

$$\begin{aligned} & X_{k+1} \nabla^2 H(x^{k+1}) X_{k+1} + \lambda p(p-1) X_{k+1} X_{k+1}^{p-2} X_{k+1} \\ \succeq & -\lambda p(p-1) X_{k+1} X_k^{p-2} X_{k+1} - \rho_k X_{k+1} X_k^{-2} X_{k+1} \\ & - \gamma R \|d_k\| X_{k+1}^2 + \lambda p(p-1) X_{k+1} X_{k+1}^{p-2} X_{k+1}. \end{aligned} \quad (41)$$

Using (31) again, we get

$$-\rho_k X_{k+1} X_k^{-2} X_{k+1} = -\rho_k (D_k + I_n)^2 \succeq -\frac{9}{4} \rho_k I_n. \quad (42)$$

On the other hand, using (31), we have

$$\begin{aligned} & X_{k+1} X_{k+1}^{p-2} X_{k+1} - X_{k+1} X_k^{p-2} X_{k+1} = X_{k+1}^p - X_{k+1}^2 X_k^{p-2} \\ = & X_{k+1}^p (I_n - (X_{k+1}^{2-p} X_k^{p-2})) = X_{k+1}^p (I_n - (I_n + D_k)^{2-p}). \end{aligned} \quad (43)$$

From the Taylor expansion that $(1+t)^{2-p} \leq 1 + (2-p)t + \frac{(2-p)(1-p)}{2} t^2$, we get

$$(I_n + D_k)^{2-p} \preceq I_n + (2-p) D_k + \frac{1}{2}(2-p)(1-p) D_k^2. \quad (44)$$

Applying (44), $D_k \preceq \frac{1}{2} I_n$ and $0 < p < 1$ to (43), we derive

$$\begin{aligned} & \lambda p(1-p)(X_{k+1} X_{k+1}^{p-2} X_{k+1} - X_{k+1} X_k^{p-2} X_{k+1}) \\ = & \lambda p(1-p) X_{k+1}^p (I_n - (I_n + D_k)^{2-p}) \\ \succeq & -\lambda p(1-p) X_{k+1}^p ((2-p) D_k + \frac{1}{2}(2-p)(1-p) D_k^2) \\ \succeq & -\lambda p(1-p) \|x_{k+1}\|_\infty^p ((2-p)\|d_k\| + \frac{1}{2}(2-p)(1-p)\|d_k\|^2) I_n \\ \succeq & -\lambda p(1-p) R^p (2 - p + \frac{(2-p)(1-p)}{2}) \|d_k\| I_n \\ \succeq & -\lambda R^p \frac{p(1-p)(2-p)(3-p)}{2} \|d_k\| I_n \\ \succeq & -\frac{\lambda}{2} R^p \|d_k\| I_n. \end{aligned} \quad (45)$$

From (41), (42), (43), (45), and $\rho_k \le \frac{2}{9\vartheta}\|d_k\|$, we obtain

$$\begin{aligned}
&X_{k+1}\nabla^2 f(x^{k+1})X_{k+1}\\
=&X_{k+1}\nabla^2 H(x^{k+1})X_{k+1} + \lambda p(p-1)X_{k+1}X_{k+1}^{p-2}X_{k+1}\\
\succeq&-(\frac{9}{4}\rho_k + \gamma R^3\|d_k\| + \frac{1}{2}\lambda R^p\|d_k\|)I_n\\
\succeq&-(\frac{1}{2\vartheta} + \gamma R^3 + \frac{1}{2}\lambda R^p)\|d_k\|I_n\\
\succeq&-\frac{1}{\vartheta}\|d_k\|I_n \succeq -\sqrt{\epsilon}I_n.
\end{aligned}$$

According to Lemmas 3 - 5, we can obtain the worst-case complexity of the Second Order Interior Point Algorithm for finding an ϵ scaled second order stationary point of (7).

Theorem 2 *For any $\epsilon \in (0,1]$, the proposed Second Order Interior Point Algorithm obtains an ϵ scaled second order stationary point or ϵ global minimizer of (7) in no more than $O(\epsilon^{-3/2})$ iterations.*

Remark 2 When $H(x) = \frac{1}{2}\|Ax-q\|^2$ in (7), then $\gamma = 0$ and ϑ in (22) turns to be

$$\vartheta = \frac{1}{\max\{2, \lambda R^p\}}.$$

The proposed Second Order Interior Point Algorithm obtains an ϵ scaled second order stationary point or ϵ global minimizer of (7) in no more than $36f(x^0)\max\{4, \lambda^2 R^{2p}\}\epsilon^{-3/2}$ iterations.

Remark 3 Through the worst-case complexity result in Theorem 2 can be extended to (7) with $\lambda = 0$ or $p = 1$, the complexity bound is for finding the scaled second order stationary point of (7). When $\lambda = 0$ or $p = 1$, the general second order optimality condition of (7) is defined as

$$x \ge 0, \quad \nabla f(x) \ge 0, \quad x^T\nabla f(x) = 0 \quad \text{and} \quad \nabla^2 H(x) \succeq 0,$$

then we call x satisfies the ϵ second order optimality condition of (7) if

$$x \ge 0, \quad \nabla f(x) \ge -\epsilon, \quad \|X\nabla f(x)\|_\infty \le \epsilon \quad \text{and} \quad \nabla^2 H(x) \succeq -\sqrt{\epsilon}I_n. \qquad (46)$$

Thus, due to the second inequality in (46) and the scaling in the second inequality of Definition 2, the conditions given in Definition 2 is necessary but not sufficient for the ϵ second order optimality conditions in (46).

Moreover, the second order interior point algorithm and the complexity bound given in section can be extended to

$$\min_{x\ge 0} \quad H(x) + \lambda\sum_{i=1}^n \varphi(x^p),$$

where φ is twice continuously differentiable in $(0,+\infty)$ and $\varphi'(t) > 0$, for instance, φ_2 and φ_3 in Section 5.

We end this section by showing our interior point algorithms can be applied to solve the unconstrained problem (2). In particular, we show the scaled first and second order stationary points of (2) and (3) are in one-one correspondence. Denote

$$\begin{aligned}
&\mathcal{F}^n = \{x : x \text{ is a scaled first order stationary point of (2)}\},\\
&\mathcal{F}^{2n}_+ = \{(x^+, x^-) : (x^+, x^-) \text{ is a scaled first order stationary point of (3)}\},\\
&\mathcal{S}^n = \{x : x \text{ is a scaled second order stationary point of (2)}\},\\
&\mathcal{S}^{2n}_+ = \{(x^+, x^-) : (x^+, x^-) \text{ is a scaled second order stationary point of (3)}\}.
\end{aligned}$$

Theorem 3 *(i) Suppose $\varphi'(s) > 0$ for $s > 0$, then*

$$\begin{aligned}
(x^+, x^-) \in \mathcal{F}^{2n}_+ \ &\Rightarrow\ x \in \mathcal{F}^n \quad \textit{with} \quad x = x^+ - x^-;\\
x \in \mathcal{F}^n \ &\Rightarrow\ (x^+, x^-) \in \mathcal{F}^{2n}_+ \quad \textit{with}
\end{aligned}$$

$$x^+ = \max(0, x) \quad \textit{and} \quad x^- = \max(0, -x). \tag{47}$$

(ii) Suppose $\varphi(s) = s$ and H is twice continuously differentiable, then

$$\begin{aligned}
(x^+, x^-) \in \mathcal{S}^{2n}_+ \ &\Rightarrow\ x \in \mathcal{S}^n \quad \textit{with} \quad x = x^+ - x^-;\\
x \in \mathcal{S}^n \ &\Rightarrow\ (x^+, x^-) \in \mathcal{S}^{2n}_+ \quad \textit{with (47).}
\end{aligned}$$

Proof (i) Suppose $(x^+, x^-) \in \mathcal{F}^{2n}_+$, then x^+, $x^- \geq 0$ and

$$\begin{cases} X^+ \nabla H(x^+ - x^-) + \lambda p[\varphi'(s)_{s=(x_i^+)^p} (x_i^+)^p]_{i=1}^n = 0 \\ -X^- \nabla H(x^+ - x^-) + \lambda p[\varphi'(s)_{s=(x_i^-)^p} (x_i^-)^p]_{i=1}^n = 0, \end{cases} \tag{48}$$

where $X^+ = \text{diag}(x^+)$ and $X^- = \text{diag}(x^-)$. Thus,

$$\begin{aligned}
&X^- X^+ \nabla H(x^+ - x^-) + \lambda p[\varphi'(s)_{s=(x_i^+)^p} (x_i^+)^p (x_i^-)]_{i=1}^n\\
&-X^+ X^- \nabla H(x^+ - x^-) + \lambda p[\varphi'(s)_{s=(x_i^-)^p} (x_i^-)^p (x_i^+)]_{i=1}^n = 0,
\end{aligned}$$

which gives

$$\varphi'(s)_{s=(x_i^+)^p} x_i^- (x_i^+)^p + \varphi'(s)_{s=(x_i^-)^p} x_i^+ (x_i^-)^p = 0, \quad \forall i \in \mathbb{I}.$$

By $\varphi'(s) > 0$ for $s > 0$ and the above equation, we obtain that $x_i^+ x_i^- = 0$, $\forall i \in \mathbb{I}$. Thus, adding the two equations in (48) gives

$$(X^+ - X^-) \nabla H(x^+ - x^-) + \lambda p[\varphi'(s)_{s=(x_i^+ - x_i^-)^p} (x_i^+ - x_i^-)^p]_{i=1}^n = 0,$$

which means that $x = x^+ - x^- \in \mathcal{F}^n$.

On the other hand, suppose $x \in \mathcal{F}^n$, then $(x^+)^T x^- = 0$, where x^+ and x^- are with the form in (47). Thus, (48) holds, which implies $(x^+, x^-) \in \mathcal{F}^{2n}_+$.

(ii) Suppose $(x^+, x^-) \in \mathcal{S}_+^{2n}$ and denote $x = x^+ - x^-$, then

$$\begin{pmatrix} X^+ & \\ & X^- \end{pmatrix} \begin{pmatrix} \nabla^2 H(x) & -\nabla^2 H(x) \\ -\nabla^2 H(x) & \nabla^2 H(x) \end{pmatrix} \begin{pmatrix} X^+ & \\ & X^- \end{pmatrix} \\ +\lambda p(p-1) \begin{pmatrix} (X^+)^p & \\ & (X^-)^p \end{pmatrix} \succeq 0, \tag{49}$$

which follows

$$\begin{pmatrix} I_n & I_n \end{pmatrix} \begin{pmatrix} X^+ & \\ & X^- \end{pmatrix} \begin{pmatrix} \nabla^2 H(x) & -\nabla^2 H(x) \\ -\nabla^2 H(x) & \nabla^2 H(x) \end{pmatrix} \begin{pmatrix} X^+ & \\ & X^- \end{pmatrix} \begin{pmatrix} I_n \\ I_n \end{pmatrix} \\ +\lambda p(p-1) \begin{pmatrix} I_n & I_n \end{pmatrix} \begin{pmatrix} (X^+)^p & \\ & (X^-)^p \end{pmatrix} \begin{pmatrix} I_n \\ I_n \end{pmatrix} \succeq 0.$$

Thus,

$$(X^+ - X^-)\nabla^2 H(x)(X^+ - X^-) + \lambda p(p-1)(X^+)^p + \lambda p(p-1)(X^-)^p \succeq 0. \tag{50}$$

From $(X^+)^p + (X^-)^p \succeq |X^+ - X^-|^p$ and (50), we obtain

$$(X^+ - X^-)\nabla^2 H(x)(X^+ - X^-) + \lambda p(p-1)|X^+ - X^-|^p \succeq 0, \tag{51}$$

which implies that $x \in \mathcal{S}_n$ with $x = x^+ - x^-$.

On the other hand, suppose $x \in S^n$ and (x^+, x^-) with the form in (47). Then, (51) holds, which follows

$$\begin{pmatrix} D_n^+ \\ D_n^- \end{pmatrix} (X^+ - X^-)\nabla^2 H(x)(X^+ - X^-) \begin{pmatrix} D_n^+ & D_n^- \end{pmatrix} \\ +\lambda p(p-1) \begin{pmatrix} D_n^+ \\ D_n^- \end{pmatrix} |X^+ - X^-|^p \begin{pmatrix} D_n^+ & D_n^- \end{pmatrix} \succeq 0, \tag{52}$$

where $D_n^+ = \text{diag}(\text{sign}(x^+))$ and $D_n^- = \text{diag}(\text{sign}(x^-))$. By the definitions in (47), we obtain

$$\begin{aligned} & D_n^+(X^+ - X^-) = X^+, \ D_n^-(X^+ - X^-) = -X^-, \\ & D_n^+|X^+ - X^-|^p D_n^- = D_n^-|X^+ - X^-|^p D_n^+ = 0_{n\times n}, \\ & D_n^+|X^+ - X^-|^p D_n^+ = |X^+|^p, \ D_n^-|X^+ - X^-|^p D_n^- = |X^-|^p. \end{aligned} \tag{53}$$

From (52) and (53), we obtain (49), which implies that $(x^+, x^-) \in \mathcal{S}_+^{2n}$.

4 Numerical Experiments

In this section, we present three examples to show the good performance and worst-case complexity of the First Order Interior Point Algorithm (FOIPA) proposed in this paper for solving (1). The numerical testing is carried out on a Lenovo PC (3.00GHz, 2.00GB of RAM) with the use of Matlab 7.4. In the following three examples, 'Iteration number' denotes the number of iterations for obtaining an ϵ scaled first order stationary point.

	FOIPA				Lasso	Best subset
λ, p	112.7,0.01	13.94,0.1	7.6,0.3	7.74,0.5		
x_1^*(lcavol)	0.6497	0.6499	0.6487	0.6433	0.533	0.740
x_2^*(lweight)	0.2941	0.2918	0.2856	0.2767	0.169	0.316
x_3^*(age)	0	0	0	0	0	0
x_4^*(lbph)	0	0	0	0	0.002	0
x_5^*(svi)	0.1498	0.1468	0.14	0.1336	0.094	0
x_6^*(lcp)	0	0	0	0	0	0
x_7^*(pleason)	0	0	0	0	0	0
x_8^*(pgg45)	0	0	0	0	0	0
Number of nonzero	3	3	3	3	4	2
Prediction error	0.4194	0.4205	0.4230	0.4261	0.479	0.492
Iteration number	2001	582	411	610		

Table 2: Example 1: Variable selection by FOIPA Lasso and Best subset methods

Example 1 (**Prostate Cancer**) The prostate cancer date is downloaded from the web site http://stat.stanford. edu/tibs/ElemStatLearn/data.html. It consists of the medical records of 97 men who were about to receive a radical prostatectomy, which is divided into a training set with 67 observations and a test set with 30 observations. For more detail of the data set, see [8,11,12, 19]. We use the following constrained $l_2 - l_p$ model to solve this problem

$$\min_{x\geq 0} \|Ax-q\|^2 + \lambda \sum_{i=1}^{n} x_i^p,$$

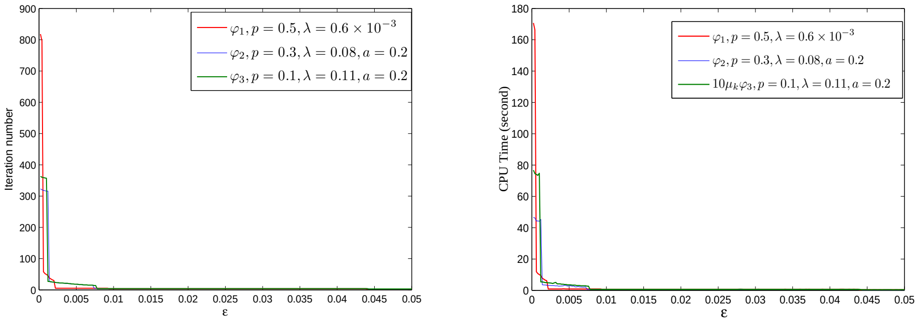
where $A \in \mathbb{R}^{67\times 8}$ and $q \in \mathbb{R}^{67}$ are built by the training set.

Let $x^0 = 0.1e_8$ and $\epsilon = 10^{-3}$. The numerical results of the FOIPA with $\beta = 2\|A^TA\|$ are given in Table 2, in which two well-known methods (Lasso, Best subset) from Table 3.3 in [19] are also listed. Table 2 indicates the FOIPA with $0 < p < 1$ can find fewer main factors with smaller prediction error than the other two methods.

Example 2 (**Nonnegative Compressed Sensing**) In this example, we test the FOIPA with a compressive sensing problem, where the goal is to reconstruct a length-n nonnegative sparse signal from m observations, where $m < n$. The purpose of this example is to show the worst-case complexity result and good performance of the FOIPA.

We use the following code in Matlab to generate the original signal $x_s \in \mathbb{R}^n$, sensing matrix $A \in \mathbb{R}^{m\times n}$ and observation signal $q \in \mathbb{R}^m$ with given positive integers n and T.

```
m=n/4; x_s=zeros(n,1); P=randperm(n); A=randn(m,n);
x_s(P(1:T))=abs(randn(T,1)); A=orth(A')'; q=A*x_s.
```



(a)

(b)

Fig. 1: Example 1: FOIPA for finding an ϵ scaled first order stationary point of (54) (a) Iteration number, (b) CPU Time (second), where ϕ_1, φ_2 and φ_3 are given in Section 5

In this example, we consider the following constrained optimization problem

$$\min_{x \geq 0} \|Ax - q\|^2 + \sum_{i=1}^{n} \varphi(x_i^p). \tag{54}$$

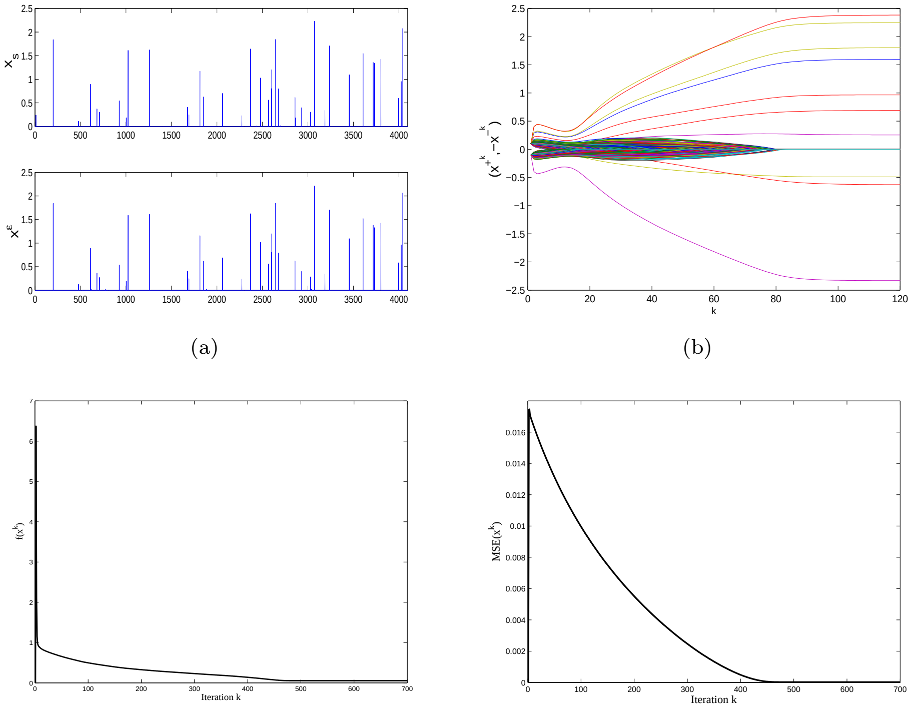
Set $n = 4096$ and $T = 40$ in the Matlab code to generate A, q and x_s. Choose $x^0 = 0.1e_n$. For different choices of φ and p, the 'Iteration number' and 'CPU time' for obtaining an ϵ scaled first order stationary point of (54) are illustrated in Fig. 1. From this figure, we can see that the total iterations for obtaining an ϵ scaled first order stationary point is much smaller than the worst-case estimation $32f(x^0)R^2\beta\epsilon^{-2}$ given in the proof of Theorem 1.

For the case that $\varphi := \varphi_1$ with $\lambda = 0.6 \times 10^{-3}$ and $p = 0.5$, the original signal x_s and ϵ scaled first order stationary point x^ϵ with $\epsilon = 10^{-4}$ are pictured in Fig. 2(a). Meantime, the convergence of x^k, $f(x^k)$ and $MSE(x^k)$ are also illustrated in Fig. 2(b)-2(d), where $MSE(x^k)$ is the mean squared error of x^k to the original signal x_s defined by $MSE(x) = \|x - x_s\|^2/n$.

Example 3 (**Unconstrained Compressed Sensing**) In order to support the theoretical analysis in Theorem 3, we test the FOIPA into a typical compressive sensing problem, where the goal is to reconstruct a length-n sparse signal (may containing positive and negative signals) from m observations, where $m < n$. The original signal x_s is generated randomly by the code x_s`(P(1:T))=randn(T,1)`, A and q are generated as in Example 2 with $n = 1024$ and $T = 10$. To solve this problem, we construct the following unconstrained $l_2 - l_p$ optimization

$$\min_{x^+, x^- \geq 0} \quad \|A(x^+ - x^-) - q\|^2 + \lambda\|x^+\|_p^p + \lambda\|x^-\|_p^p, \tag{55}$$

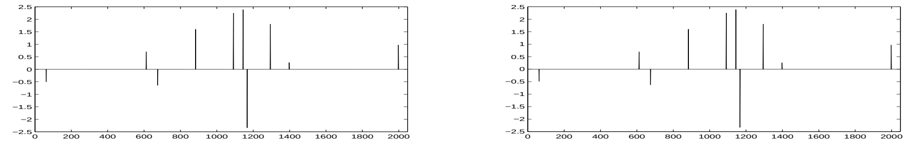
which can be solved by the algorithms proposed in this paper. With initial point ${x^+}^0 = {x^-}^0 = 0.1e_n$ and $\beta = 4\|A^TA\| = 4$, the FOIPA can find an ϵ



(a) (b)

(c) (d)

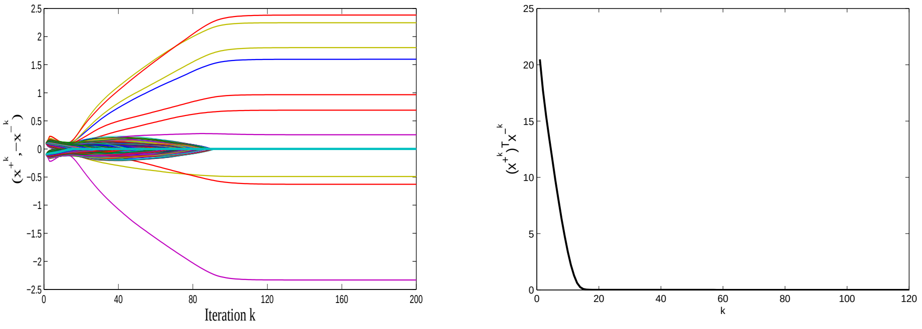
Fig. 2: Example 2: (a) original signal x_s and ϵ scaled stationary point x^ϵ with $\epsilon = 10^{-4}$, (b) convergence of iterate x^k, (c) convergence of function value $f(x^k)$, (d) convergence of mean squared error MSE(x^k)



(a) (b)

Fig. 3: Example 3: Original and reconstructed signal by the FOIPA with $\epsilon = 10^{-3}$ and $x = x^+ - x^-$ (a) original signal, (b) reconstructed signal.

(=10^{-3}) scaled first order stationary point in 131 iterations with CPU time 10.075 seconds. The original signal and the ϵ scaled first order stationary point are described in Fig. 3. The convergence of $(x^{+^k}, -x^{-^k})$ and $(x^{+^k})^T x^{-^k}$ are shown in Fig. 4. From Fig. 3 and Fig. 4, we can see that $x^k = x^{+^k} - x^{-^k}$ converges to the original signal.



(a) (b)

Fig. 4: Example 3: (a) convergence of $(x^{+^k}, -x^{-^k})$, (b) convergence of $(x^{+^k})^T x^{-^k}$

5 Final Remarks

This paper proposes two interior point methods for solving constrained non-Lipschitz, nonconvex optimization problems arising in many important applications. The first order interior point method is easy to implement and its worst-case complexity is $O(\epsilon^{-2})$ which is the same in order as the worst-case complexity of steepest-descent methods applied to unconstrained, nonconvex smooth optimization, and the trust region methods and SQR methods applied to unconstrained, nonconvex nonsmooth optimization [1,3]. The second order interior method has a better complexity order $O(\epsilon^{-3/2})$ for finding an ϵ scaled second order stationary point. It is not answered in this paper that whether the complexity bounds are sharp for the first and second order methods, which gives us an interesting topic for further research.

Assumptions in this paper are standard and applicable to many regularization models in practice. For example, $H(x) = \|Ax - q\|^2$ and φ is one of the following six penalty functions

i) soft thresholding penalty function [20,25]: $\varphi_1(s) = s$
ii) logistic penalty function [24]: $\varphi_2(s) = \log(1 + \alpha s)$
iii) fraction penalty function [13,24]: $\varphi_3(s) = \frac{\alpha s}{1+\alpha s}$
iv) hard thresholding penalty function[14]: $\varphi_4(s) = \lambda - (\lambda - s)_+^2/\lambda$
v) smoothly clipped absolute deviation penalty function[14]:

$$\varphi_5(s) = \int_0^s \min\{1, \frac{(\alpha - t/\lambda)_+}{\alpha - 1}\}\mathrm{d}t$$

vi) minimax concave penalty function [31]:

$$\varphi_6(s) = \int_0^s (1 - \frac{t}{\alpha\lambda})_+ \mathrm{d}t.$$

Here α and λ are two positive parameters, especially, $\alpha > 2$ in $\varphi_5(s)$ and $\alpha > 1$ in $\varphi_6(s)$. These six penalty functions are concave in $[0, \infty)$ and continuously differentiable in $(0, \infty)$, which are often used in statistics and sparse reconstruction.

Acknowledgements We would like to thank Prof. Sven Leyffer, the anonymous Associate Editor and referees for their insightful and constructive comments, which help us to enrich the content and improve the presentation of the results in this paper.

References

1. W. Bian and X. Chen, Worst-case complexity of smoothing quadratic regularization methods for non-Lipschitzian optimization, SIAM J. Optim., 28, 1718-1741 (2013).
2. A.M. Bruckstein D.L. Donoho and M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Review, 51, 34-81 (2009).
3. C. Cartis, N.I.M. Gould and Ph.L. Toint, On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming, SIAM J. Optim., 21, 1721-1739 (2011).
4. C. Cartis, N.I.M. Gould and Ph.L. Toint, Adaptive cubic regularisation methods for unconstrained optimization, Part I: motivation, convergence and numerical results, Math. Program., 127, 245-295 (2011).
5. C. Cartis, N.I.M. Gould and Ph.L. Toint, Adaptive cubic regularisation methods for unconstrained optimization, Part II: worst-case function- and derivative-evaluation complexity, Math. Program., 130, 295-319 (2011).
6. C. Cartis, N.I.M. Gould and Ph.L. Toint, An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity, IMA J. Numer. Anal., doi: 10.1093/imanum/drr035 (2012).
7. C. Cartis, N.I.M. Gould and Ph.L. Toint, Complexity bounds for second-order optimality in unconstrained optimization, J. Complexity, 28, 93-108 (2012).
8. X. Chen, Smoothing methods for nonsmooth, novonvex minimization, Math. Program., 134, 71-99 (2012).
9. X. Chen, D. Ge, Z. Wang and Y. Ye, Complexity of unconstrained L_2-L_p minimization, Math. Program, doi: 10.1007/s10107-012-0613-0 (2012).
10. X. Chen, M. Ng and C. Zhang, Nonconvex l_p regularization and box constrained model for image restoration, IEEE Trans. Image Processing, 21, 4709-4721 (2012).
11. X. Chen, L. Niu and Y. Yuan, Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization, Preprint, 2012.
12. X. Chen, F. Xu and Y. Ye, Lower bound theory of nonzero entries in solutions of l_2-l_p minimization, SIAM J. Sci. Comput., 32, 2832-2852 (2010).
13. X. Chen and W. Zhou, Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization, SIAM J. Imaging Sci., 3, 765-790 (2010).
14. J. Fan, Comments on 'Wavelets in stastics: a review' by A. Antoniadis, Stat. Method. Appl., 6, 131-138 (1997).
15. R. Garmanjani and L.N. Vicente, Smoothing and worst case complexity for direct-search methods in nonsmooth optimization, IMA J. Numer. Anal., doi: 10.1093/imanum/drs027 (2012)
16. D. Ge, X. Jiang and Y. Ye, A note on the complexity of L_p minimization, Math. Program., 21, 1721-1739 (2011).
17. S. Gratton, A. Sartenaer and Ph.L. Toint, Recursive trust-region methods for multiscale nonlinear optimization, SIAM J. Optim., 19, 414-444, (2008).
18. A. Griewank, The modification of Newton's method for unconstrained optimization by bounding cubic terms, Technical Report NA/12 (1981), Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom, 1981.

19. T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer, New York (2009).
20. J. Huang, J.L. Horowitz and S. Ma, Asymptotic properties of bridge estimators in sparse high-dimensional regression models, Ann. Statist., 36, 587-613 (2008).
21. Yu. Nesterov, Introductory Lectures on Convex Optimization, Applied Optimization, Kluwer Academic Publishers, Dordrecht, The netherlands, 2004.
22. Yu. Nesterov and B.T. Polyak, Cubic regularization of Newton's method and its global performance, Math. Program., 108, 177-205 (2006).
23. Yu. Nesterov, Accelerating the cubic regularization of Newton's method on convex problems, Math. Program., 112, 159-181 (2008).
24. M. Nikolova, M.K. Ng, S. Zhang and W.-K. Ching, Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization, SIAM J. Imaging Sci., 1, 2-25 (2008).
25. R. Tibshirani, Shrinkage and selection via the Lasso, J. Roy. Statist. Soc. Ser. B, 58, 267-288 (1996).
26. S.A. Vavasis and R. Zippel, Proving polynomial time for sphere-constrained quadratic programming, Technical Report 90-1182, Department of Computer Science, Cornell University, Ithaca, NY, 1990.
27. S.A. Vavasis, Nonlinear Optimization: Complexity Issues, Oxford Sciences, New York, 1991.
28. Y. Ye, Interior Point Algorithms: Theory and Analysis, John Wiley & Sons, Inc., New York, 1997.
29. Y. Ye, On the complexity of approximating a KKT point of quadratic programming, Math. Program., 80, 195-211 (1998).
30. Y. Ye, A new complexity result on minimization of a quadratic function with a sphere constraint, in Recent Advances in Global Optimization, C. Floudas and P.M. Pardalos, eds., Princeton University Press, Princeton, NJ (1992).
31. C.-H Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist., 38, 894-942 (2010).