SMOOTHING SQP ALGORITHM FOR NON-LIPSCHITZ OPTIMIZATION WITH COMPLEXITY ANALYSIS

WEI BIAN[*] AND XIAOJUN CHEN[†]

6 February 2012

Abstract. In this paper, we propose a smoothing sequential quadratic programming (SSQP) algorithm for solving a class of nonsmooth nonconvex, perhaps even non-Lipschitz minimization problems, which has wide applications in statistics and sparse reconstruction. At each step, the SSQP algorithm solves a strongly convex quadratic minimization problem with a diagonal Hessian matrix, which has a simple closed-form solution. The SSQP algorithm is easy to implement and has almost no time cost to solve the convex quadratic minimization subproblems. We show that the worst-case complexity of reaching an ε scaled stationary point is $O(\varepsilon^{-2})$. Moreover, if the objective function is locally Lipschitz, the SSQP algorithm with a slightly modified updating scheme can obtain an ε Clarke stationary point at most $O(\varepsilon^{-3})$ steps.

Key words. Nonsmooth nonconvex optimization, smoothing approximation, sequential quadratic programming (SQP), convergence, complexity.

AMS subject classifications. 90C30, 90C26, 65K05, 49M37

1. Introduction.

The convexity and Lipschitz continuity are two important conditions in optimization. However, some real-world applications are often modeled by nonconvex or even non-Lipschitz optimization problems. In this paper, we concentrate on the following unconstrained nonsmooth nonconvex optimization problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) := H(x) + \sum_{i=1}^{n} \varphi(|x_i|^p), \tag{1.1}$$

where $H : \mathbb{R}^n \to \mathbb{R}_+$ is continuously differentiable and its gradient ∇H is globally Lipschitz with a Lipschitz constant $\beta > 0$, $0 < p \le 1$, and $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a given penalty function satisfying the following assumption.

(A_φ) φ is continuously differentiable, nondecreasing, $\nabla\varphi$ is locally Lipschitz, and there is a positive constant α such that for all $t \in (0,\infty)$,

$$0 \le \nabla\varphi(t) \le \alpha, \quad |\xi| \le \alpha \quad \text{and} \quad |\xi|t \le \alpha, \quad \forall \xi \in \partial(\nabla\varphi(t)),$$

where ∂ means the Clarke generalized gradient [9].

To illustrate that the application of (1.1) is not restricted by assumption (A_φ), six widely used penalty functions φ in statistics and sparse reconstruction are given in Appendix.

Numerical algorithms for solving nonsmooth optimization have been studied for decades, but most algorithms assume that objective functions are Lipschitz continuous in the convergence analysis and complexity estimation. The subgradient methods are the first order numerical algorithms for nonsmooth convex minimization problems and

[*] Department of Mathematics, Harbin Institute of Technology, Harbin, China (bianweilvse520@163.com). Current address: Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. The author's work was supported by the Hong Kong Polytechnic University Postdoctoral Fellowship Scheme and the NSF foundation (11101107,10971043) of China.

[†] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (maxjchen@polyu.edu.hk). The author's work was supported in part by Hong Kong Research Grant Council grant (PolyU5005/09P).

the complexity of these algorithms is proved to be of the order $O(\varepsilon^{-2})$ [15]. Based on a special smoothing technique for the maximal function, Nesterov [20] improves the traditional complexity of the gradient algorithms to $O(\varepsilon^{-1})$ for nonsmooth convex optimization problems. A gradient sampling algorithm is proposed by Burke, Lewis and Overton in [3] for finding a Clarke ε stationary point of a locally Lipschitz function with probability 1, where ε is a fixed sampling radius. Most recently, by means of the first order methods, Catis, Gould and Toint [4] estimate the function evaluation worst-case complexity of minimizing the following function

$$\Phi_h(x) := H(x) + h(c(x)), \tag{1.2}$$

where $h : \mathbb{R}^m \to \mathbb{R}$ is convex but may be nonsmooth, $H : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable but Φ_h may be nonsmooth nonconvex. They prove that it takes at most $O(\varepsilon^{-2})$ steps to reduce a first order criticality measure below ε in a first order trust region method or a quadratic regularization method, where the complexity result is the same in order as the function evaluation complexity of steepest descent methods applied to the case that Φ_h is differentiable. In [12], Garmanjani and Vicente propose a smoothing direct search (DS) algorithm basing on smoothing techniques and derivative free methods to solve a general unconstrained nonsmooth nonconvex optimization problem. The smoothing DS algorithm can obtain a Clarke stationary point of a locally Lipschitz function f and it takes at most $O(-\varepsilon^{-3}\log\varepsilon)$ function evaluations to find an x such that $\|\nabla\tilde{f}(x,\mu)\|_\infty \le \varepsilon$ and $\mu \le \varepsilon$, where $\tilde{f}$ is a smoothing function of f and $\mu > 0$ is a parameter. In [13], Ge, Jiang and Ye develop an interior-point potential reduction algorithm for solving the following non-Lipschitz constrained optimization

$$\begin{array}{ll} \min & \sum_{i=1}^n x^p \\ \text{s.t.} & Ax = x, \quad x \ge 0, \end{array}$$

and show that the interior-point algorithm returns a scaled ε-KKT point in no more than $O(n\varepsilon^{-1}\log\varepsilon^{-1})$ steps.

The main purpose of this paper is to construct an efficient algorithm for solving (1.1) with the worst-case complexity estimation. We propose a smoothing sequential quadratic programming (SSQP) algorithm for solving (1.1). The SSQP algorithm is easy to implement and has almost no time cost to solve a strongly convex quadratic program with a diagonal Hessian matrix at each step. Indeed, the quadratic program has a simple closed-form solution. We show that the worst-case complexity of finding an ε scaled stationary point is $O(\varepsilon^{-2})$. Moreover, if the objective function is locally Lipschitz, the SSQP algorithm with a slightly modified updating scheme can obtain an ε Clarke stationary point at most $O(\varepsilon^{-3})$ steps. To the best of our knowledge, the SSQP algorithm is the first algorithm with the worst-case complexity for non-Lipschitz unconstrained optimization. Moreover, the modified SSQP algorithm for locally Lipschitz minimization has a better complexity bound than the smoothing DS algorithm proposed in [12]. Note that many penalty functions cannot be written as $h(c(x))$ in (1.2), for example, the logistic penalty function, $\varphi(|x_i|) = \log(1 + \alpha|x_i|)$. Hence the first order trust region method proposed in [4] cannot be applied to solve (1.1).

Penalty functions play an important role in statistical modeling, particularly in variable selection. When estimating the vector x of regression coefficients, which is sparse in the sense that many of its elements are zero, one often minimizes the penalized objective function in (1.1). The performance of resulting estimator depends on the choice of the penalty functions. A widely used penalty function is the l_1-norm, with which the minimization model is often called LASSO [19]. Three principles (unbiasedness, sparsity and continuity) for a good penalty function are proposed in [1, 11]. Fan and Li [11] show that the smoothly clipped absolute deviation (SCAD) penalty function proposed in [10] has better properties than the l_1 penalty function in parametric and nonparametric models. More recently, a minimax concave penalty function (MCP) is proposed by Zhang in [24]. The SCAD and MCP penalty functions satisfy assumption (A_φ) with $p = 1$. When $0 < p < 1$, the penalty function in (1.1) is non-Lipschitz, which includes the l_p penalty function as a special case. For the l_p penalty function, Fan and Li [11] point out that the oracle property does not hold for the l_1 penalty, while it continues to hold for the l_p penalty with $0 < p < 1$ by suitable choice of the parameters in it. In [17], Huang, Horowitz and Ma provide some conditions under which the l_p penalized least square problem with $0 < p < 1$ can correctly distinguish nonzero and zero coefficients in sparse high-dimensional settings. Moreover, the l_p penalized least square model with $0 < p < 1$ can also be used for variable selection at the group and individual variable levels simultaneously, while the l_1 penalized least square model can only work for individual variable selection [18].

Besides the applications in statistics, minimization problem (1.1) is often used in sparse reconstruction, which provides an efficient model to extract the essential features of general solutions, e.g. in the context of data compression and order reduction with applications to signal and image analysis, inverse scattering, de-convolution and tomography problems. In this kind of applications, the original goal is to find a fitting solution with fewer nonzero elements for an underdetermined linear or nonlinear system $H(x) = 0$. Such problems can be modeled as l_0-norm minimization, which is difficult to solve by virtue of the noncontinuous structure of the l_0-norm. In order to overcome this difficult, many researchers solve these problems by using regularized minimization (1.1) with nonconvex and nonsmooth penalty functions. In [7, 21], it is shown that nonconvex and nonsmooth minimization yields better edge preservation than convex minimization. Moreover, it appears that the non-Lipschitz penalty l_p with $0 < p < 1$ is more robust in image restoration with blurring and noises.

This paper is organized as follows. In Section 2, smoothing approximations for the nonsmooth function f in (1.1) are studied. In Section 3, the SSQP algorithm for solving (1.1) and theoretical analysis including the convergence and complexity results are given, where the worst-case complexity of reaching an ε scaled stationary point is $O(\varepsilon^{-2})$. In Section 4, we consider the SSQP algorithm with a modified updating scheme to solve the nonsmooth, nonconvex but locally Lipschitz optimization problem (1.1) with $p = 1$, where the worst-case complexity of reaching an ε Clarke stationary point is $O(\varepsilon^{-3})$. In Section 5, two numerical examples are given to show the efficiency of the SSQP algorithm.

Let $I = \{1, 2, \ldots, n\}$ and $\mathbb{N} = \{0, 1, \ldots\}$. For a column vector $x \in \mathbb{R}^n$, both $[x]_i$ and x_i denote the ith component of x and $[x_i]_{i=1}^n := x$. For a constant a, $\lceil a \rceil$ indicates the smallest positive integer such that $\lceil a \rceil \geq a$.

2. Smoothing Approximations. Smoothing approximations for nonsmooth optimization have been studied for decades [2, 5, 20, 22]. In this section, based on the special structure of the nonsmooth function f in problem (1.1), we use a smoothing function for the absolute value function $|\cdot|$ to construct a smoothing function $\tilde{f}$ of f. Using the special structure of the smoothing function, we derive some properties of $\tilde{f}$, which provide theoretical basis for constructing the SSQP algorithm.

Definition 2.1. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function. We call $\tilde{h} : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ a* smoothing function *of h, if $\tilde{h}$ satisfies the following conditions.*

(i) For any fixed $\mu > 0$, $\tilde{h}(\cdot, \mu)$ is continuously differentiable in $\mathbb{R}^n$.

(ii) For any fixed $x \in \mathbb{R}^n$, $\lim_{z \to x, \mu \downarrow 0} \tilde{h}(z, \mu) = h(x)$.

Smoothing functions have been widely used in nonsmooth optimization with rich theory and applications. In this section, we focus on a class of smoothing functions constructed by the following smoothing function of $|\cdot|$,

$$\theta(s, \mu) = \begin{cases} |s| & \text{if } |s| > \mu \\ \dfrac{s^2}{2\mu} + \dfrac{\mu}{2} & \text{if } |s| \le \mu. \end{cases}$$

It is not difficult to find the following properties of the smoothing function θ.

Lemma 2.2.

(i) $|\nabla_s \theta(s, \mu)| \le 1$, $\forall\, s \in \mathbb{R}$, $\mu \in (0, \infty)$.

(ii) $\frac{\mu}{2} \le \theta(s, \mu) \le \mu$, $\forall |s| \le \mu$.

(iii) $0 \le \theta^p(s, \mu) - |s|^p \le \theta^p(0, \mu) = (\frac{\mu}{2})^p$, $\forall\, s \in \mathbb{R}$, $\mu \in [0, \infty)$, $p \in (0, 1]$.

Due to the continuous differentiability of functions φ and θ, we can easily obtain that $\varphi(\theta^p(s, \mu))$ is a smoothing function of $\varphi(|s|^p)$ and $\varphi(\theta^p(s, \cdot))$ is nondecreasing in $(0, \infty)$ for any fixed $s \in \mathbb{R}$. The following proposition presents estimations for elements in the generalized Hessian [9] of $\varphi(\theta^p(\cdot, \mu))$ for any fixed $\mu > 0$.

Proposition 2.3. *For any fixed $\mu > 0$ and $\xi \in \partial_s(\nabla_s \varphi(\theta^p(s, \mu)))$, it follows that*

$$|\xi| \le 8\alpha p \begin{cases} |\dfrac{s}{2}|^{p-2} & |s| > 2\mu \\ \mu^{p-2} & |s| \le 2\mu. \end{cases}$$

Proof. The first derivative of $\varphi(\theta^p(s, \mu))$ with respect to s is given by

$$\nabla_s \varphi(\theta^p(s, \mu)) = \nabla \varphi(t)_{t=\theta^p(s,\mu)} \nabla_s \theta^p(s, \mu).$$

From the expression of $\theta^p(s, \mu)$, for any fixed $\mu > 0$, it derives that

$$\begin{cases} \text{if } p = 1 \text{ and } |s| \le \mu, & |\eta| \le \mu^{-1} \quad \forall \eta \in \partial_s(\nabla_s \theta(s, \mu)) \\ \text{if } p = 1 \text{ and } |s| > \mu, & \nabla_s^2 \theta(s, \mu) = 0 \\ \text{if } p < 1 \text{ and } |s| \le \mu, & |\eta| \le p(1-p)\mu^{p-2} \quad \forall \eta \in \partial_s(\nabla_s \theta^p(s, \mu)) \\ \text{if } p < 1 \text{ and } |s| > \mu, & \nabla_s^2 \theta^p(s, \mu) = p(1-p)|s|^{p-2}, \end{cases}$$

which follows that $\nabla_s \theta^p(s, \mu)$ is globally Lipschitz respect to s for any fixed $\mu > 0$.

Since $\nabla \varphi$ is locally Lipschitz, $\nabla_s \varphi(\theta^p(s, \mu))$ is locally Lipschitz respect to s for any fixed $\mu > 0$. For a fixed $\mu > 0$, denote $\mathcal{D}_\mu$ the set of points at which $\varphi(\theta^p(\cdot, \mu))$ is differentiable.

Based on the property that a locally Lipschitz function is differentiable almost everywhere (in the sense of Lebesgue measure), the measure of $\mathbb{R} \setminus \mathcal{D}_\mu$ is 0.

First, we consider the case that $p = 1$. When $|s| > \mu$ and $s \in \mathcal{D}_\mu$,

$$|\nabla_s^2 \varphi(\theta^p(s, \mu))| = |\nabla^2 \varphi(t)_{t=\theta(s,\mu)} (\nabla_s \theta(s, \mu))^2| \le \alpha \theta^{-1}(s, \mu) = \alpha |s|^{-1}. \tag{2.1}$$

On the other hand, when $|s| \le \mu$ and $s \in \mathcal{D}_\mu$, it follows that

$$\begin{aligned} |\nabla_s^2 \varphi(\theta^p(s, \mu))| =& |\nabla^2 \varphi(t)_{t=\theta(s,\mu)} (\nabla_s \theta(s, \mu))^2 + \nabla \varphi(t)_{t=\theta(s,\mu)} \nabla_s^2 \theta(s, \mu)| \\ =& |\nabla^2 \varphi(t)_{t=\theta(s,\mu)} (\frac{s}{\mu})^2 + \nabla \varphi(t)_{t=\theta(s,\mu)} \frac{1}{\mu}| \\ \le& \alpha \theta^{-1}(s, \mu) + \alpha \mu^{-1} \le 3\alpha \mu^{-1}, \end{aligned} \tag{2.2}$$

where the last inequality uses $\theta(s,\mu)=\frac{s^2}{2\mu}+\frac{\mu}{2}\geq\frac{\mu}{2}$.

Next, we consider the case that $0<p<1$. For any fixed $\mu>0$, from the chain rule, when $s\in\mathcal{D}_\mu$, the second derivative of $\varphi(\theta^p(s,\mu))$ respect to s is calculated by

$$\begin{aligned}\nabla_s^2\varphi(\theta^p(s,\mu))=&\,p^2\nabla^2\varphi(t)_{t=\theta^p(s,\mu)}\theta^{2p-2}(s,\mu)(\nabla_s\theta(s,\mu))^2\\&+p\nabla\varphi(t)_{t=\theta^p(s,\mu)}\theta^{p-1}(s,\mu)\nabla_s^2\theta(s,\mu)\\&+p(p-1)\nabla\varphi(t)_{t=\theta^p(s,\mu)}\theta^{p-2}(s,\mu)(\nabla_s\theta(s,\mu))^2.\end{aligned}\tag{2.3}$$

When $|s|\leq\mu$, we have $\frac{\mu}{2}\leq\theta(s,\mu)\leq\mu$ and $|\nabla_s\theta(s,\mu)|\leq\frac{|s|}{\mu}\leq1$. From assumption (A_φ) and (2.3), we obtain that for $s\in\mathcal{D}_\mu$ with $|s|\leq\mu$,

$$\begin{aligned}&|\nabla_s^2\varphi(\theta^p(s,\mu))|\\\leq&p^2|\nabla^2\varphi(t)_{t=\theta^p(s,\mu)}\theta^p(s,\mu)|\theta^{p-2}(s,\mu)\frac{s^2}{\mu^2}+p\nabla\varphi(t)_{t=\theta^p(s,\mu)}\theta^{p-1}(s,\mu)\frac{1}{\mu}\\&+p(1-p)|\nabla\varphi(t)_{t=\theta^p(s,\mu)}|\theta^{p-2}(s,\mu)\frac{s^2}{\mu^2}\\\leq&\alpha p^2\theta^{p-2}(s,\mu)+\alpha p\frac{\theta(s,\mu)}{\mu}\theta^{p-2}(s,\mu)+p(1-p)\alpha\theta^{p-2}(s,\mu)\\\leq&2\alpha p\theta^{p-2}(s,\mu)\leq8\alpha p\mu^{p-2}.\end{aligned}\tag{2.4}$$

Similarly, when $|s|>\mu$ and $s\in\mathcal{D}_\mu$, we get

$$\begin{aligned}|\nabla_s^2\varphi(\theta^p(s,\mu))|\leq&p^2|\nabla^2\varphi(t)_{t=|s|^p}|s|^p||s|^{p-2}+p(1-p)|\nabla\varphi(t)_{t=|s|^p}||s|^{p-2}\\\leq&\alpha p^2|s|^{p-2}+\alpha p(1-p)|s|^{p-2}\leq\alpha p|s|^{p-2}.\end{aligned}\tag{2.5}$$

It is easy to see that (2.1), (2.2), (2.4) and (2.5) imply that for $s\in\mathcal{D}_\mu$ and $0<p\leq1$,

$$|\nabla_s^2\varphi(\theta^p(s,\mu))|\leq8\alpha p\begin{cases}|\frac{s}{2}|^{p-2} & |s|>2\mu\\\mu^{p-2} & |s|\leq2\mu.\end{cases}$$

Combining this inequality with the definition of the Clarke generalized gradient, we complete the proof. □

Using the smoothing function $\theta(\cdot,\mu)$ of $|\cdot|$, we define the following smoothing function

$$\tilde{f}(x,\mu)=H(x)+\sum_{i=1}^n\varphi(\theta^p(x_i,\mu))$$

of f, which satisfies the two conditions of Definition 2.1. Let

$$\tilde{g}(x,\mu)=[\tilde{g}_1(x,\mu),\ldots,\tilde{g}_n(x,\mu)]^T:=\nabla_x\tilde{f}(x,\mu).$$

3. Smoothing SQP Algorithm.

Sequential quadratic programming (SQP) methods are popular iterative methods for solving smooth optimization problems [14, 16, 23], which solve a quadratic programming problem at each step. In this section, we propose a smoothing SQP method to solve the nonsmooth optimization problem (1.1). At each step, we construct a convex quadratic approximation of the smoothing function $\tilde{f}(\cdot,\mu_k)$ around x^k and update μ_k by a simple criterion. In our convergence and complexity analysis, we assume that the function f is level bounded, i.e. for any $\Gamma>0$, the level set $\{x\in\mathbb{R}^n: \ f(x)\le\Gamma\}$ is bounded.

When $0<p<1$, (1.1) is a non-Lipschitz optimization problem. It is hard to find an efficient algorithm for solving it. Inspired by smoothing approximations and SQP methods, in this section, we construct a smoothing SQP algorithm for solving (1.1) with $0<p\le 1$. The SSQP algorithm uses a smoothing function $\tilde{f}(x,\mu)$ to approximate the nonsmooth objective function f and solves a strongly convex quadratic subproblems generated from $\tilde{f}(x,\mu_k)$ at each step. The Hessian of the quadratic subproblem is a diagonal and positive definite matrix. Thus, the SSQP algorithm is easy to implement and has almost no time cost in calculating a new iterate and updating the smoothing parameter at each step. We show that any accumulation point of the iterates is a scaled stationary point of optimization problem (1.1) with complexity $O(\varepsilon^{-2})$.

Note that for any $x,y\in\mathbb{R}^n$, the assumptions on H imply

$$H(x)\le H(y)+\langle\nabla H(y),x-y\rangle+\frac{\beta}{2}\sum_{i=1}^{n}(x_i-y_i)^2. \tag{3.1}$$

Following by Proposition 2.3, we define

$$\kappa(s,\mu)=8\alpha p\begin{cases}|\frac{s}{2}|^{p-2} & \text{if } |s|>2\mu\\ \mu^{p-2} & \text{if } |s|\le 2\mu,\end{cases} \tag{3.2}$$

which is an upper bound for all elements in the generalized Hessian $\partial_s^2\varphi(\theta^p(s,\mu))$ for any fixed $\mu\in(0,\infty)$. What follows is an inequality derived by Taylor's formula.

PROPOSITION 3.1. *For any $\mu>0$ and $s,\hat{s}\in\mathbb{R}$ such that $|s-\hat{s}|\le\max\{\frac{|\hat{s}|}{2},\mu\}$, the following inequality holds*

$$\varphi(\theta^p(s,\mu))\le\varphi(\theta^p(\hat{s},\mu))+\langle\nabla_{\hat{s}}\varphi(\theta^p(\hat{s},\mu)),s-\hat{s}\rangle+\frac{\kappa(\hat{s},\mu)}{2}(s-\hat{s})^2. \tag{3.3}$$

Proof. For $|\hat{s}|\le 2\mu$, by Proposition 2.3, $\partial_s^2\varphi(\theta^p(s,\mu))$ can be uniformly bounded by $\kappa(\hat{s},\mu)=8\alpha p\mu^{p-2}$ for any $\mu>0$. Thus, by Taylor's formula, (3.3) holds naturally in this case.

For $|\hat{s}|>2\mu$, since $|s-\hat{s}|\le\max\{\frac{|\hat{s}|}{2},\mu\}=\frac{|\hat{s}|}{2}$, for any $\tau\in[0,1]$,

$$|\tau s+(1-\tau)\hat{s}|\ge|\hat{s}|-\tau|s-\hat{s}|\ge|\hat{s}|-\tau\frac{|\hat{s}|}{2}\ge\frac{|\hat{s}|}{2},$$

which implies $|\tau s+(1-\tau)\hat{s}|^{p-2}\le|\frac{\hat{s}}{2}|^{p-2}$.

From the above inequality, Proposition 2.3 and Taylor's formula, there is $\bar{\tau}\in[0,1]$ such that

$$\begin{aligned}\varphi(\theta^p(s,\mu))&\le\varphi(\theta^p(\hat{s},\mu))+\langle\nabla_{\hat{s}}\varphi(\theta^p(\hat{s},\mu)),s-\hat{s}\rangle+\frac{8\alpha p|\bar{\tau}s+(1-\bar{\tau})\hat{s}|^{p-2}}{2}(s-\hat{s})^2\\&\le\varphi(\theta^p(\hat{s},\mu))+\langle\nabla_{\hat{s}}\varphi(\theta^p(\hat{s},\mu)),s-\hat{s}\rangle+\frac{\kappa(\hat{s},\mu)}{2}(s-\hat{s})^2.\end{aligned}$$

□

Now we give an important inequality for the smoothing function $\tilde{f}$.

LEMMA 3.2. *Suppose* $|x_i - y_i| \leq \max\{\frac{|y_i|}{2}, \mu\}$ *holds for all* $i \in I$, *then*

$$\tilde{f}(x,\mu) \leq \tilde{f}(y,\mu) + \langle \tilde{g}(y,\mu), x-y\rangle + \frac{1}{2}\sum_{i=1}^{n} \gamma_i(y,\mu)(\beta + \kappa(y_i,\mu))(x_i - y_i)^2, \tag{3.4}$$

where

$$\gamma_i(y,\mu) = \max\{1, \frac{|\tilde{g}_i(y,\mu)|}{\max\{\frac{|y_i|}{2},\mu\}^{1-\frac{p}{2}}\mu^{\frac{p}{2}}(\beta + \kappa(y_i,\mu))}\}, \quad i \in I. \tag{3.5}$$

Proof. By (3.1) and (3.3), the inequality (3.4) holds for $\gamma_i(y,\mu) \equiv 1$. Thus (3.4) holds with γ_i defined by (3.5) using the max operator. □

3.1. Proposed Algorithm. The quadratic program in the SSQP algorithm is constructed based on Lemma 3.2. For any fixed $y \in \mathbb{R}^n$ and $\mu > 0$, the right hand of the inequality (3.4) is a strictly quadratic convex function. Let the quadratic approximation to $\tilde{f}(\cdot,\mu)$ at x be

$$Q(x,y,\mu) = \tilde{f}(y,\mu) + \langle \tilde{g}(y,\mu), x-y\rangle + \frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{d_i(y,\mu)}, \tag{3.6}$$

where

$$d_i(y,\mu) = \frac{1}{\gamma_i(y,\mu)(\beta + \kappa(y_i,\mu))}, \quad i \in I.$$

For any fixed $y \in \mathbb{R}^n$ and $\mu \in (0,\infty)$, we consider the following quadratic programming subproblem

$$\min_{x \in \mathbb{R}^n} Q(x,y,\mu). \tag{3.7}$$

Since $Q(\cdot,y,\mu)$ is a strongly quadratic convex function for any fixed y and μ, problem (3.7) has an unique minimizer.

The scheme for updating the smoothing parameter μ is crucial for the efficiency of smoothing methods. It can effect the convergence and complexity of the SSQP algorithm. A simple and intelligent scheme for updating μ_k is used in the following SSQP algorithm.

SSQP Algorithm
Choose $x^0 \in \mathbb{R}^n$, $\mu_0 > 0$ and $\sigma \in (0,1)$. Set $k = 0$ and $z^0 = x^0$.
For $k \geq 0$, set

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} Q(x, x^k, \mu_k), \tag{3.8a}$$

$$\mu_{k+1} = \begin{cases} \mu_k & \text{if } \tilde{f}(x^{k+1},\mu_k) - \tilde{f}(x^k,\mu_k) \leq -4\alpha p \mu_k^p \\ \sigma\mu_k & \text{othewise,} \end{cases} \tag{3.8b}$$

$$z^{k+1} = \begin{cases} x^k & \text{if } \mu_{k+1} = \sigma\mu_k \\ z^k & \text{othewise.} \end{cases} \tag{3.8c}$$

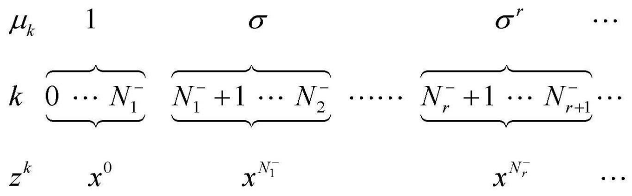


Fig. 3.1: Illustration of the relationship among μ_k, x^k and z^k

Noting that the Hessian matrix $\nabla_x^2 Q(x, x^k, \mu_k)$ is a diagonal and positive definite matrix, (3.8a) has a simple closed-form solution

$$x_i^{k+1} = x_i^k - d_i(x^k, \mu_k)\tilde{g}_i(x^k, \mu_k), \quad \forall i \in I. \tag{3.9}$$

Moreover, there is no line search procedure in the SSQP algorithm. It is easy to implement and there is almost no time losing in performing (3.8a)-(3.8c) at each step.

In what follows, we always denote

$$N^- = \{k \in \mathbb{N} : \ \mu_{k+1} = \sigma\mu_k\} \quad \text{and} \quad N^+ = \{k \in \mathbb{N} : k \notin N^-\},$$

where N_r^- and N_r^+ are the rth smallest elements in N^- and N^+, respectively. The sequence $\{z^k\}$ is well-defined and can be written as

$$\begin{cases} z^{k+1} = x^k & \text{if } k \in N^- \\ z^k = x^{N_r^-} & \text{if } N_r^- + 1 \leq k \leq N_{r+1}^-. \end{cases} \tag{3.10}$$

The relationships among x^k, z^k and μ_k with $\mu_0 = 1$ are illustrated in Figure 3.1.

3.2. Theoretical Results. In this subsection, we will give the theoretical analysis on the SSQP algorithm, including the convergence and complexity results. First, we define some index sets for any fixed $x \in \mathbb{R}^n$ and $\mu > 0$. Let

$$K(x,\mu) = \{i \in I : \ |x_i| \leq 2\mu\}, \quad J(x,\mu) = \{i \in I : \ |x_i| > 2\mu\}. \tag{3.11}$$

$K(x,\mu)$ and $J(x,\mu)$ are mutually disjoint and $I = K(x,\mu) \bigcup J(x,\mu)$. We divide each of $K(x,\mu)$ and $J(x,\mu)$ into two mutually disjoint sets

$$K^+(x,\mu) = \{i \in K(x,\mu) : \ |\tilde{g}_i(x,\mu)| \geq \mu(\beta + \kappa(x_i,\mu))\},$$

$$J^+(x,\mu) = \{i \in J(x,\mu) : \ |\tilde{g}_i(x,\mu)| \geq |\frac{x_i}{2}|^{1-\frac{p}{2}}\mu^{\frac{p}{2}}(\beta + \kappa(x_i,\mu))\},$$

$$K^-(x,\mu) = K(x,\mu) \setminus K^+(x,\mu) \quad \text{and} \quad J^-(x,\mu) = J(x,\mu) \setminus J^+(x,\mu).$$

The following lemma shows that the sequence $\{\tilde{f}(x^k, \mu_k)\}$ is monotonely decreasing and strictly decreasing at (x^k, μ_k) when $\|\tilde{g}(x^k, \mu_k)\|_\infty \neq 0$.

Lemma 3.3. *The sequence $\{\tilde{f}(x^k, \mu_k)\}$ is monotonely decreasing and satisfies*

$$\begin{aligned} \tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) \leq & \tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) \\ & \leq -\sum_{i=1}^n \frac{d_i(x^k, \mu_k)\tilde{g}_i^2(x^k, \mu_k)}{2} \leq 0. \end{aligned} \tag{3.12}$$

Moreover, there is $R \geq 1$ *such that* $\|x^k\|_\infty \leq R$, $\forall k \in \mathbb{N}$.

Proof. Firstly, we prove that

$$|x_i^{k+1} - x_i^k| \leq \max\{\frac{|x_i^k|}{2}, \mu_k\}, \quad \forall i \in I. \tag{3.13}$$

From (3.9), for any $i \in I$, we have

$$|x_i^{k+1} - x_i^k| = d_i(x^k, \mu_k)|\tilde{g}_i(x^k, \mu_k)|. \tag{3.14}$$

For $i \in K^+(x^k, \mu_k)$, we get

$$|x_i^k| \leq 2\mu_k \quad \text{and} \quad |\tilde{g}_i(x^k, \mu_k)| \geq \mu_k(\beta + \kappa(x_i^k, \mu_k)), \tag{3.15}$$

which implies that

$$\gamma_i(x^k, \mu_k) = \frac{|\tilde{g}_i(x^k, \mu_k)|}{\mu_k(\beta + \kappa(x_i^k, \mu_k))} \quad \text{and} \quad d_i(x^k, \mu_k) = \frac{\mu_k}{|\tilde{g}_i(x^k, \mu_k)|}. \tag{3.16}$$

For $i \in K^-(x^k, \mu_k)$, we get

$$|x_i^k| \leq 2\mu_k \quad \text{and} \quad |\tilde{g}_i(x^k, \mu_k)| < \mu_k(\beta + \kappa(x_i^k, \mu_k)), \tag{3.17}$$

which implies that

$$\gamma_i(x^k, \mu_k) = 1 \quad \text{and} \quad d_i(x^k, \mu_k) = \frac{1}{\beta + \kappa(x_i^k, \mu_k)}. \tag{3.18}$$

Then, from (3.14), (3.16) and (3.18), we know

$$|x_i^{k+1} - x_i^k| \leq \mu_k, \quad \forall i \in K(x^k, \mu_k).$$

Similarly, for $i \in J^+(x^k, \mu_k)$, we get

$$|x_i^k| > 2\mu_k \quad \text{and} \quad |\tilde{g}_i(x^k, \mu_k)| \geq |\frac{x_i^k}{2}|^{1-\frac{p}{2}}\mu_k^{\frac{p}{2}}(\beta + \kappa(x_i^k, \mu_k)), \tag{3.19}$$

which implies that

$$d_i(x^k, \mu_k) = \frac{|\frac{x_i^k}{2}|^{1-\frac{p}{2}}\mu_k^{\frac{p}{2}}}{|\tilde{g}_i(x^k, \mu_k)|}. \tag{3.20}$$

For $i \in J^-(x^k, \mu_k)$, we have

$$|x_i^k| > 2\mu_k \quad \text{and} \quad |\tilde{g}_i(x^k, \mu_k)| < |\frac{x_i^k}{2}|^{1-\frac{p}{2}}\mu_k^{\frac{p}{2}}(\beta + \kappa(x_i^k, \mu_k)), \tag{3.21}$$

which implies that

$$d_i(x^k, \mu_k) = \frac{1}{\beta + \kappa(x_i^k, \mu_k)}. \tag{3.22}$$

Then, from (3.14), (3.20) and (3.22), we have

$$|x_i^{k+1} - x_i^k| \leq |\frac{x_i^k}{2}|^{1-\frac{p}{2}}\mu_k^{\frac{p}{2}} \leq \frac{|x_i^k|}{2}, \quad \forall i \in J(x^k, \mu_k).$$

Therefore, we can obtain the estimation in (3.13).

Applying $|x_i^{k+1} - x_i^k| \le \max\{\frac{|x_i^k|}{2}, \mu_k\}$, $\forall i \in I$ to Lemma 3.2, it holds that

$$\tilde{f}(x^{k+1}, \mu_k) \le Q(x^{k+1}, x^k, \mu_k),$$

which, together with (3.9), gives that

$$\begin{aligned}
&\tilde{f}(x^{k+1}, \mu_k) \\
\le&\tilde{f}(x^k, \mu_k) + \langle \tilde{g}(x^k, \mu_k), x^{k+1} - x^k \rangle + \frac{1}{2}\sum_{i=1}^n \frac{1}{d_i(x^k, \mu_k)}(x_i^{k+1} - x_i^k)^2 \\
=&\tilde{f}(x^k, \mu_k) - \sum_{i=1}^n d_i(x^k, \mu_k)\tilde{g}_i^2(x^k, \mu_k) + \frac{1}{2}\sum_{i=1}^n d_i(x^k, \mu_k)\tilde{g}_i^2(x^k, \mu_k) \\
=&\tilde{f}(x^k, \mu_k) - \frac{1}{2}\sum_{i=1}^n d_i(x^k, \mu_k)\tilde{g}_i^2(x^k, \mu_k).
\end{aligned}$$

Since $\mu_{k+1} \le \mu_k$, we can obtain the inequality in (3.12). Moreover, from $\tilde{f}(x^k, \mu_k) \ge f(x^k)$, we have

$$f(x^k) \le \tilde{f}(x^0, \mu_0).$$

Using the assumption that f is level bounded, we find that there is $R \ge 1$ such that

$$\|x^k\|_\infty \le R, \quad k \in \mathbb{N}.$$

□

The objective function f is non-Lipschitz for $0 < p < 1$. It has been proved that finding a global minimizer of the unconstrained l_2-l_p minimization problem with $0 < p < 1$ is strongly NP hard in [6]. We extend the definition of the scaled first order necessary condition in [8] to define the scaled stationary points of (1.1) with $0 < p \le 1$.

Definition 3.4. *Let $G : R^n \to R^n$ be defined by*

$$G(x) = X\nabla H(x) + p|X|^p[\nabla\varphi(t)_{t=|x_i|^p}]_{i=1}^n,$$

where X =diag$(x_1, \dots, x_n)$ and $|X|^p$ =diag$(|x_1|^p, \dots, |x_n|^p)$. For a given $\varepsilon \ge 0$, we call $x^* \in \mathbb{R}^n$ an ε scaled stationary point of (1.1) if

$$\|G(x^*)\|_\infty \le \varepsilon.$$

And x^* is called a scaled stationary point of (1.1) if $\varepsilon = 0$.

Following the proof of Theorem 2.3 in [8], we can show that any local minimizer of (1.1) is a scaled stationary point of (1.1).

The SSQP algorithm acts on the smoothing approximation function $\tilde{f}$. The following lemma presents that $X\tilde{g}(\cdot, \mu)$ tends to $G(\cdot)$ uniformly with $O(\mu^p)$ as $\mu \to 0$.

Lemma 3.5. *For all $x \in \mathbb{R}^n$, $\mu \in (0, \mu_0]$ and $0 < p \le 1$, we have*

$$\|X\tilde{g}(x, \mu) - G(x)\|_\infty \le 3\alpha p \mu^p.$$

Proof. Let us consider a fixed $i \in I$. For $|x_i| \geq \mu$, it is obtained that

$$x_i \tilde{g}_i(x,\mu) = x_i[\nabla H(x)]_i + p\nabla\varphi(t)_{t=|x_i|^p}|x_i|^p = [G(x)]_i.$$

For $|x_i| < \mu$ and $0 < p < 1$, we have $\frac{\mu}{2} \leq \theta(x_i,\mu)$ and

$$\begin{aligned}
&|x_i\tilde{g}_i(x,\mu) - x_i[\nabla H(x)]_i - p\nabla\varphi(t)_{t=|x_i|^p}|x_i|^p| \\
=&p|\nabla\varphi(t)_{t=\theta^p(x_i,\mu)}\theta^{p-1}(x_i,\mu)\frac{x_i^2}{\mu} - \nabla\varphi(t)_{t=|x_i|^p}|x_i|^p| \\
\leq&2\alpha p\mu^p + \alpha p\mu^p \leq 3\alpha p\mu^p.
\end{aligned}$$

Similarly, for $|x_i| < \mu$ and $p = 1$, it gives

$$\begin{aligned}
&|x_i\tilde{g}_i(x,\mu) - x_i[\nabla H(x)]_i - \nabla\varphi(t)_{t=|x_i|}|x_i|| \\
=&|\nabla\varphi(t)_{t=\theta(x_i,\mu)}\frac{x_i^2}{\mu} - \nabla\varphi(t)_{t=|x_i|}|x_i|| \leq 2\alpha\mu.
\end{aligned}$$

Hence, for all $0 < p \leq 1$, from

$$\|X\tilde{g}(x,\mu) - G(x)\|_\infty = \max_{1\leq i\leq n}|x_i\tilde{g}_i(x,\mu) - x_i[\nabla H(x)]_i - p\nabla\varphi(t)_{t=|x_i|^p}|x_i|^p|,$$

we complete the proof of this lemma. □

The following lemma gives the magnitude of the decreasing of $\tilde{f}$ and $|x_i^k\tilde{g}_i(x^k,\mu_k)|$.

Lemma 3.6. *If $K^+(x^k,\mu_k)\bigcup J^+(x^k,\mu_k) \neq \emptyset$, then*

$$\tilde{f}(x^{k+1},\mu_k) - \tilde{f}(x^k,\mu_k) \leq -4\alpha p\mu_k^p.$$

Otherwise,

$$|x_i^k\tilde{g}_i(x^k,\mu_k)| \leq C_1\mu_k^{\frac{p}{2}}, \quad \forall i \in I,$$

where $C_1 = \max\{(\beta + 8\alpha)\mu_0^{\frac{p}{2}}, 2R^2\beta + 16\alpha pR\}$ with $R \geq 1$ and $R \geq \|x^k\|_\infty$, $\forall k \in \mathbb{N}$.

Proof. Fix $k \in \mathbb{N}$. We first consider the case that $K^+(x^k,\mu_k)\bigcup J^+(x^k,\mu_k) \neq \emptyset$.

If there is an $i \in I$ such that $i \in K^+(x^k,\mu_k)$, from (3.2), (3.12), (3.15) and (3.16), we obtain that

$$\tilde{f}(x^{k+1},\mu_{k+1}) - \tilde{f}(x^k,\mu_k) \leq -\frac{\mu_k}{2}|\tilde{g}_i(x^k,\mu_k)| \leq -\frac{\mu_k^2}{2}\kappa(x_i^k,\mu_k) \leq -4\alpha p\mu_k^p. \quad (3.23)$$

If there is an $i \in I$ such that $i \in J^+(x^k,\mu_k)$, from (3.2), (3.12), (3.19) and (3.20), we have

$$\tilde{f}(x^{k+1},\mu_{k+1}) - \tilde{f}(x^k,\mu_k) \leq -\frac{1}{2}\mu_k^{\frac{p}{2}}|\frac{x_i^k}{2}|^{1-\frac{p}{2}}|\tilde{g}_i(x^k,\mu_k)| \leq -4\alpha p\mu_k^p. \quad (3.24)$$

Next, we consider the case that $K^+(x^k,\mu_k)\bigcup J^+(x^k,\mu_k) = \emptyset$. Then,

$$I = K^-(x^k,\mu_k)\bigcup J^-(x^k,\mu_k).$$

For $i \in K^-(x^k,\mu_k)$, from (3.2), (3.17) and (3.18), we get

$$|x_i^k\tilde{g}_i(x^k,\mu_k)| \leq \mu_k^2(\beta + 8\alpha p\mu_k^{p-2}) \leq (\beta + 8\alpha)\mu_k^p.$$

If $i \in J^-(x^k, \mu_k)$, from (3.2), (3.21) and (3.22), we obtain

$$\begin{aligned} |x_i^k \tilde{g}_i(x^k, \mu_k)| \leq & 2\mu_k^{\frac{p}{2}} |\frac{x_i^k}{2}|^{2-\frac{p}{2}} (\beta + 8\alpha p |\frac{x_i^k}{2}|^{p-2}) \\ \leq & 2(R^2\beta + 8\alpha p R)\mu_k^{\frac{p}{2}}, \end{aligned}$$

where $R \geq 1$ and $\|x^k\|_\infty \leq R$, $\forall k \in \mathbb{N}$.

Combining the above analysis, we conclude the second inequality in this lemma. □

Now, we are ready to present the convergence theorem of the SSQP algorithm.

THEOREM 3.7. *For all $k \in \mathbb{N}$, if $K^+(x^k, \mu_k) \bigcup J^+(x^k, \mu_k) \neq \emptyset$, then*

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) \leq -4\alpha p \mu_k^p. \tag{3.25}$$

Otherwise,

$$\|G(x^k)\|_\infty \leq C\mu_k^{\frac{p}{2}}, \tag{3.26}$$

where $C = \max\{(\beta + 11\alpha)\mu_0^{\frac{p}{2}}, 2R^2\beta + 16\alpha p R + 3\alpha\mu_0^{\frac{p}{2}}\}$ with $R \geq 1$ and $R \geq \|x^k\|_\infty$, $\forall k \in \mathbb{N}$. Moreover, $\sum_{k=0}^\infty \mu_k^p < \infty$ and $\lim_{k\to\infty} f(x^k) = \lim_{k\to\infty} f(z^k)$ exists.

Proof. From Lemma 3.5 and Lemma 3.6, we can obtain (3.25) and (3.26).

When $N_r^- < k \leq N_{r+1}^-$ with $r \in \mathbb{N}$, $\mu_k = \mu_0\sigma^r$, which is illustrated in Figure 3.1 with $\mu_0 = 1$. If $k \in N^+$, from (3.25), we have

$$4\alpha p \mu_k^p \leq \tilde{f}(x^k, \mu_k) - \tilde{f}(x^{k+1}, \mu_{k+1}) \quad \text{and} \quad \mu_{k+1} = \mu_k.$$

Combining it with the non-increasing property of $\tilde{f}(x^k, \mu_k)$ gives

$$\sum_{k\in N^+} \mu_k^p \leq \frac{1}{4\alpha p} \sum_{k\in N^+} [\tilde{f}(x^k, \mu_k) - \tilde{f}(x^{k+1}, \mu_{k+1})] \leq \frac{1}{4\alpha p} \tilde{f}(x^0, \mu_0). \tag{3.27}$$

From the relationship illustrated in Figure 3.1, when $k = N_r^-$, $\mu_k = \mu_0\sigma^{r-1}$. Then,

$$\sum_{k\in N^-} \mu_k^p \leq \sum_{r=1}^\infty \mu_0\sigma^{p(r-1)} = \frac{\mu_0}{1-\sigma^p}. \tag{3.28}$$

Adding (3.27) and (3.28), we have

$$\sum_{k\in\mathbb{N}} \mu_k^p \leq \frac{1}{4\alpha p}\tilde{f}(x^0, \mu_0) + \frac{\mu_0}{1-\sigma^p}.$$

Therefore, $\lim_{k\to\infty} \mu_k = 0$.

Using Lemma 2.2 (iii) and the middle value theorem, we obtain that

$$|\tilde{f}(x, \mu) - f(x)| \leq n\alpha p \mu^p, \quad \forall x \in \mathbb{R}^n, \mu > 0.$$

Since $\{\tilde{f}(x^k, \mu_k)\}$ is non-increasing and bounded from below, $\lim_{k\to\infty} \tilde{f}(x^k, \mu_k)$ exists. From $\lim_{k\to\infty} \mu_k = 0$ and (3.8c), we find

$$\lim_{k\to\infty} \tilde{f}(x^k, \mu_k) = \lim_{k\to\infty} f(x^k) = \lim_{k\to\infty} f(z^k).$$

□

The next theorem shows the worst order magnitude of the SSQP algorithm for obtaining an ε scaled stationary point of (1.1).

THEOREM 3.8. *Any accumulation point of $\{z^k\}$ generated by the SSQP algorithm is a scaled stationary point of (1.1) with the complexity $O(\varepsilon^{-2})$. Specially, z^k is an ε scaled stationary point of (1.1) when*

$$k \geq \lceil \frac{C^2 \tilde{f}(x^0, \mu_0)}{4\alpha p \sigma^p \varepsilon^2} + \frac{2}{p} \log_\sigma \varepsilon - \frac{2}{p} \log_\sigma C \rceil + 1,$$

where $\mu_0 = 1$ and C is the constant in Theorem 3.7.

Proof. Let $\varepsilon \in (0, 1]$ be a given number and let $j \geq 2$ be the positive integer such that

$$C\sigma^{\frac{(j-1)p}{2}} \leq \varepsilon \quad \text{and} \quad C\sigma^{\frac{(j-2)p}{2}} > \varepsilon, \tag{3.29}$$

where C is the constant in Theorem 3.7. Then from Theorem 3.7, we have

$$\|G(x^{N_r^-})\|_\infty \leq C(\mu_{N_r^-})^{\frac{p}{2}} \leq C(\mu_{N_j^-})^{\frac{p}{2}} \leq C(\sigma^{j-1})^{\frac{p}{2}} \leq \varepsilon, \quad \forall\, r \in \mathbb{N},\, r \geq j. \tag{3.30}$$

From (3.10) and (3.30), we obtain

$$\|G(z^k)\|_\infty \leq \varepsilon, \quad \forall\, k \geq N_j^- + 1. \tag{3.31}$$

In order to let (3.31) hold, it needs to carry out at most $N_j^- + 1$ steps of the SSQP algorithm. From Theorem 3.7, when $k \in N^+$ and $k \leq N_j^-$,

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) \leq -4\alpha p \mu_k^p \leq -4\alpha p \sigma^{(j-1)p} \leq -\frac{4\alpha p \sigma^p \varepsilon^2}{C^2}. \tag{3.32}$$

Moreover, there are at least $N_j^- - j + 1$ steps such that the above inequality holds.

Owning to the non-increasing property of $\tilde{f}(x^k, \mu_k)$ and (3.32), we have that

$$\tilde{f}(x^{N_j^-}, \mu_{N_j^-}) \leq \tilde{f}(x^0, \mu_0) - \frac{4\alpha p \varepsilon^2 \sigma^p (N_j^- - j + 1)}{C^2}.$$

Due to the fact that $\tilde{f}(x, \mu) \geq 0$, $\forall x \in \mathbb{R}^n$, $\mu \geq 0$, we confirm that

$$N_j^- \leq \frac{C^2 \tilde{f}(x^0, \mu_0)}{4\alpha p \sigma^p \varepsilon^2} + j - 1. \tag{3.33}$$

From the first inequality in (3.29), we obtain that

$$j \leq \frac{2}{p}(\log_\sigma \varepsilon - \log_\sigma C) + 1. \tag{3.34}$$

Combining (3.33) with (3.34), we derive

$$N_j^- \leq \lceil \frac{C^2 \tilde{f}(x^0, \mu_0)}{4\alpha p \sigma^p \varepsilon^2} + \frac{2}{p} \log_\sigma \varepsilon - \frac{2}{p} \log_\sigma C \rceil.$$

Coming back to (3.31), this shows the complexity of the SSQP algorithm for obtaining an ε scaled stationary point of (1.1). Let $\varepsilon \to 0$, we obtain that

$$\lim_{k \to \infty} G(z^k) = 0,$$

which shows that any accumulation point of z^k is a scaled stationary point of (1.1). □

4. Locally Lipschitz optimization. In this section, we consider (1.1) with $p = 1$, which is a nonsmooth nonconvex, but locally Lipschitz optimization problem. We present a slightly modified SSQP algorithm to find a Clarke stationary point of (1.1). We call this algorithm SSQP$_1$ algorithm. We show the worst-case complexity of obtaining an ε Clarke stationary point is $O(\varepsilon^{-3})$.

For fixed y, μ, $\kappa(s, \mu)$ in $Q(x, y, \mu)$ defined by (3.6) with $p = 1$ has the following form.

$$\kappa(s,\mu) = 8\alpha \begin{cases} |\frac{s}{2}|^{-1} & \text{if } |s| > 2\mu \\ \mu^{-1} & \text{if } |s| \leq 2\mu. \end{cases}$$

SSQP$_1$ Algorithm
Choose $x^0 \in \mathbb{R}^n$, $\mu_0 \in (0, 1]$ and $\sigma \in (0, 1)$. Set $k = 0$ and $z^0 = x^0$.
For $k \geq 0$, let

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} Q(x, x^k, \mu_k), \tag{4.1a}$$

$$\mu_{k+1} = \begin{cases} \mu_k & \text{if } \tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) \leq -4\alpha\mu_k^3 \\ \sigma\mu_k & \text{othewise,} \end{cases} \tag{4.1b}$$

$$z^{k+1} = \begin{cases} x^k_{\mu_k} & \text{if } \mu_{k+1} = \sigma\mu_k \\ z^k & \text{othewise,} \end{cases} \tag{4.1c}$$

where $[x^k_{\mu_k}]_i = \begin{cases} x_i^k & \text{if } |x_i^k| \geq \mu_k \\ 0 & \text{otherwise,} \end{cases} \quad i \in I.$

The SSQP algorithm and the SSQP$_1$ algorithm have the same structure, except the updating scheme for μ_k and z^k. (4.1a) can also be expressed by the format of (3.9) and the results in Lemmas 3.2 - 3.3 hold for the SSQP$_1$ algorithm.

Let us use the index sets $K(x, \mu)$ and $J(x, \mu)$ in (3.11) again. However, we divide each of these two index sets into two different sets. Denote

$$K^+(x, \mu) = \{i \in K(x, \mu) : \ |\tilde{g}_i(x, \mu)| \geq \mu^2(\beta + \kappa(x_i, \mu))\},$$

$$J^+(x, \mu) = \{i \in J(x, \mu) : \ |\tilde{g}_i(x, \mu)| \geq |\frac{x_i}{2}|^{\frac{1}{2}}\mu^{\frac{3}{2}}(\beta + \kappa(x_i, \mu))\},$$

$$K^-(x, \mu) = K(x, \mu) \setminus K^+(x, \mu) \quad \text{and} \quad J^-(x, \mu) = J(x, \mu) \setminus J^+(x, \mu).$$

Then, the following lemma holds.

LEMMA 4.1. *If $K^+(x^k, \mu_k) \bigcup J^+(x^k, \mu_k) \neq \emptyset$, then*

$$\tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) \leq -4\alpha\mu_k^3.$$

Otherwise,

$$\|\tilde{g}(x^k, \mu_k)\|_\infty \leq (\beta R + 8\alpha)\mu_k,$$

where $R \geq 1$ such that $R \geq \|x^k\|_\infty$, $\forall k \in \mathbb{N}$.

Proof. Similar to the proof of Lemma 3.6, the following statements hold.

If there exists $i \in K^+(x^k, \mu_k) \cup J^+(x^k, \mu_k)$, then

$$\tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) \leq -4\alpha\mu_k^3.$$

Otherwise, for any $i \in K^-(x^k, \mu_k)$,

$$|\tilde{g}_i(x^k, \mu_k)| \leq (\beta + 8\alpha)\mu_k,$$

and for any $i \in J^-(x^k, \mu_k)$,

$$|\tilde{g}_i(x^k, \mu_k)| \leq (\beta R + 8\alpha)\mu_k.$$

Hence, we complete the proof of this lemma. □

The objective function f with $p = 1$ is locally Lipschitz, but nonsmooth and nonconvex. We define a Clarke stationary point of (1.1) [9].

Definition 4.2. *We call x^* an ε Clarke stationary of (1.1) if there exists $\xi \in \partial f(x^*)$ such that*

$$\|\xi\|_\infty \leq \varepsilon.$$

And x^* is reduced to a Clarke stationary point of f when $\varepsilon = 0$.

From the definition of $\tilde{f}$ and the analysis in [5], for any fixed $x \in \mathbb{R}^n$, it follows that

$$\{\lim_{z \to x, \mu \downarrow 0} \nabla_z \tilde{f}(z, \mu)\} \subseteq \partial f(x).$$

Lemma 4.3. *For any $x \in \mathbb{R}^n$ and $\mu > 0$,*

$$\min\{\|\nabla \tilde{f}(x, \mu) - \xi\|_\infty, \xi \in \partial f(x_\mu)\} \leq (\alpha + \beta)\mu,$$

where $[x_\mu]_i = \begin{cases} x_i & \text{if } |x_i| \geq \mu \\ 0 & \text{if } |x_i| < \mu, \end{cases} \quad \forall i \in I.$

Proof. Since ∇H is globally Lipschitz with Lipschitz constant β, we have

$$\|\nabla H(x) - \nabla H(x_\mu)\|_\infty \leq \beta\mu. \tag{4.2}$$

Denote $I^+(x, \mu) = \{i \in I : |x_i| \geq \mu\}$ and $I^-(x, \mu) = \{i \in I : |x_i| < \mu\}$. Then,

$$\nabla_{x_i}\varphi(\theta(x_i, \mu)) = \begin{cases} \nabla\varphi(t)_{t=\theta(x_i,\mu)} \dfrac{x_i}{\mu} & \text{if } i \in I^-(x, \mu) \\ \nabla\varphi(t)_{t=|x_i|}\text{sign}(x_i) & \text{if } i \in I^+(x, \mu). \end{cases}$$

From assumption (A_φ) and Lemma 2.2 (ii), for any $i \in I^-(x, \mu)$,

$$|\nabla\varphi(t)_{t=\theta(x_i,\mu)} \frac{x_i}{\mu} - \nabla\varphi(0)\frac{x_i}{\mu}| \leq |\nabla\varphi(t)_{t=\theta(x_i,\mu)} - \nabla\varphi(0)| \leq \alpha|\theta(x_i, \mu)| \leq \alpha\mu.$$

Since $\nabla\varphi(0)\frac{x_i}{\mu} \in \nabla\varphi(0)\cdot[-1, 1] = \partial\varphi(|[x_\mu]_i|)$ for $i \in I^-(x, \mu)$ and $\partial_{x_i}\varphi(\theta(x_i, \mu)) = \partial_{[x_\mu]_i}\varphi(|[x_\mu]_i|)$ for $i \in I^+(x, \mu)$, we derive

$$\min\{\|\sum_{i=1}^n \nabla_x\varphi(\theta(x_i, \mu)) - \eta\|_\infty : \eta \in \sum_{i=1}^n \partial_{x_\mu}\varphi(|[x_\mu]_i|)\} \leq \alpha\mu. \tag{4.3}$$

Combining (4.2) and (4.3), we complete the proof. □

From Lemma 4.1, following the proof of Theorem 3.7 and Theorem 3.8, we can show that the SSQP$_1$ algorithm takes at most $O(\varepsilon^{-3})$ steps to reduce $\|\nabla\tilde{f}(x,\mu)\|_\infty \leq \varepsilon$ and $\mu \leq \varepsilon$, while the DS algorithm in [12] need to take at most $O(-\varepsilon^{-3}\log\varepsilon)$ steps. Moreover, the SSQP$_1$ algorithm takes at most $O(\varepsilon^{-3})$ steps to find an ε Clarke stationary point.

THEOREM 4.4. *Any accumulation point of $\{z^k\}$ generated by the SSQP$_1$ algorithm is a Clarke stationary point of (1.1) and the complexity is $O(\varepsilon^{-3})$. Specially, z^k is an ε Clarke stationary point when*

$$k \geq \lceil C\tilde{f}(x^0,\mu_0)\varepsilon^{-3} + \log_\sigma \varepsilon - \log_\sigma(2\beta R + 9\alpha)\rceil + 1, \tag{4.4}$$

where $C = \frac{(2\beta R+9\alpha)^3}{4\alpha\sigma^3\mu_0^3}$ with $R \geq 1$ and $R \geq \|x^k\|_\infty$, $\forall k \in \mathbb{N}$.

Proof. Similar to the proof of Theorem 3.7, we can obtain that

$$\lim_{t\to\infty} \mu_k = 0.$$

For a given $\varepsilon \in (0,1]$, let $j \geq 2$ be the positive integer such that

$$(2\beta R + 9\alpha)\sigma^{(j-1)} \leq \varepsilon \quad \text{and} \quad (2\beta R + 9\alpha)\sigma^{(j-2)} > \varepsilon. \tag{4.5}$$

From (4.1b), for $k \in N^-$, the following inequality holds

$$\tilde{f}(x^{k+1},\mu_k) - \tilde{f}(x^k,\mu_k) > -4\alpha\mu_k^3. \tag{4.6}$$

If (4.6) holds, from Lemma 4.1, then $K^+(x^{N_j^-},\mu_{N_j^-}) \bigcup J^+(x^{N_j^-},\mu_{N_j^-}) = \emptyset$, which follows that

$$\|\tilde{g}(x^{N_j^-},\mu_{N_j^-})\|_\infty \leq (\beta R + 8\alpha)\mu_{N_j^-}.$$

From Lemma 4.3 and $k \in N^-$, we have

$$\min\{\|\tilde{g}(x^{N_j^-},\mu_{N_j^-}) - \xi\|_\infty, \xi \in \partial f(z^{N_j^-+1})\} \leq (\beta + \alpha)\mu_{N_j^-}.$$

Combining the above two inequalities and the non-increasing property of μ_k, we derive

$$\min\{\|\xi\|_\infty : \xi \in \partial f(z^k)\} \leq (2\beta R + 9\alpha)\mu_{N_j^-} \leq \varepsilon, \quad \forall k \geq N_j^- + 1. \tag{4.7}$$

Similar to the proof of Theorem 3.8, we only need to evaluate N_j^-. From Lemma 4.1, when $k \in N^+$ and $k \leq N_j^-$,

$$\tilde{f}(x^{k+1},\mu_{k+1}) - \tilde{f}(x^k,\mu_k) \leq -4\alpha\mu_k^3 \leq -4\alpha\mu_0^3\sigma^{3(j-1)} \leq -\frac{4\alpha\mu_0^3\sigma^3\varepsilon^3}{(2\beta R + 9\alpha)^3}.$$

There are at least $N_j^- - j + 1$ steps such that the above inequality holds. Hence we obtain

$$0 \leq \tilde{f}(x^{N_j^-},\mu_{N_j^-}) \leq \tilde{f}(x^0,\mu_0) - C^{-1}\varepsilon^3(N_j^- - j + 1),$$

with $C = \frac{(2\beta R+9\alpha)^3}{4\alpha\sigma^3\mu_0^3}$. Thus,

$$N_j^- \le C\tilde{f}(x^0,\mu_0)\varepsilon^{-3} + j - 1.$$

Moreover, (4.5) gives $j \le \log_\sigma \varepsilon - \log_\sigma(2\beta R + 9\alpha) + 1$. Hence, we have

$$N_j^- \le C\tilde{f}(x^0,\mu_0)\varepsilon^{-3} + \log_\sigma \varepsilon - \log_\sigma(2\beta R + 9\alpha).$$

Coming back to (4.7), we can obtain the estimation of steps in (4.4) and the complexity of the SSQP_1 algorithm for finding a Clarke stationary point of (1.1). Let $\varepsilon \to 0$, then any accumulation point of z^k is a Clarke stationary point of (1.1). □

For given $\varepsilon \in (0,1]$, it is difficult to verify the following inequality

$$\min\{\|\xi\|_\infty, \xi \in \partial f(z^k)\} \le \varepsilon \tag{4.8}$$

directly. However, from (4.7), if

$$\|\tilde{g}(x^{k^*}, \mu_{k^*})\|_\infty + (\alpha+\beta)\mu^{k^*} \le \varepsilon, \tag{4.9}$$

then for all $k \ge k^*$, there exists $\xi^k \in \partial f(z^k)$ such that $\|\xi^k\|_\infty \le \varepsilon$. Hence, we can use (4.9) to verify (4.8).

5. Numerical Experiments. In this section, we give two examples to show the performance of the SSQP algorithm for solving (1.1) with $p = \frac{1}{2}$ and $p = 1$, respectively. The numerical testing is carried out on a Lenovo PC (3.00GHz, 2.00GB of RAM) with the use of Matlab 7.4. Throughout this section, we always use $\mu_0 = 10$ and $\sigma = 0.9$. Example 5.1 is used to show that the SSQP algorithm can find a global minimizer of (1.1). Since $x = 0$ is a trivial scaled stationary point and a local minimizer of (1.1) with $p \in (0,1)$, some first order methods may stop at $x = 0$. We use Example 5.2 to show that the SSQP algorithm with starting point $x^0 = 0$ can find a nonzero scaled stationary point of (1.1) with $p \in (0,1]$. Moreover, at these nonzero stationary points, the function values are less than that at $x = 0$.

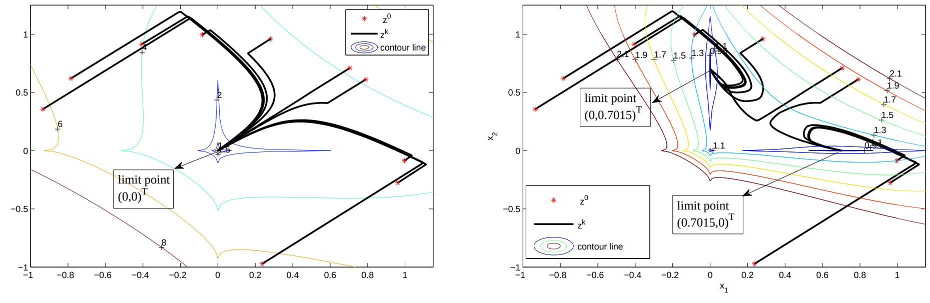
Example 5.1. Consider the following l_2-$l_{\frac{1}{2}}$ optimization problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) := (x_1 + x_2 - 1)^2 + \lambda(\sqrt{|x_1|} + \sqrt{|x_2|}), \tag{5.1}$$

where $\lambda > 0$. This example is used to explain the optimality conditions in [6].

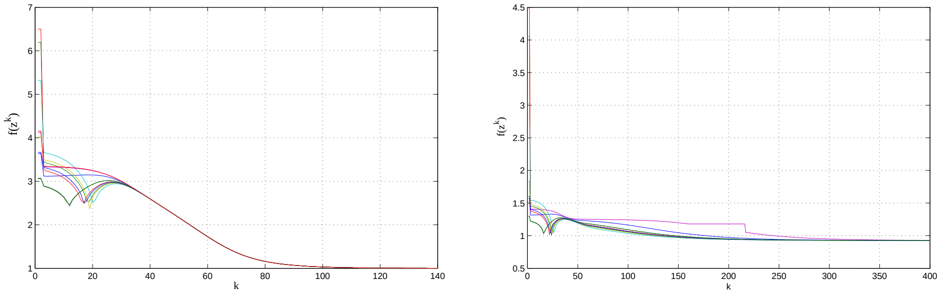
When $\lambda = \frac{8}{3\sqrt{3}}$, $(1/3, 0)$ and $(0, 1/3)$ are two nonzero vectors satisfying the first and second order necessary conditions given in [8], while $(0,0)$ is the unique global minimizer of (5.1). When $\lambda = 1$, the global minimum of f is 0.927 with two minimizers $(0, 0.7015)$ and $(0.7015, 0)$. Figure 5.1 shows the tracks of z^k generated by the SSQP algorithm with 10 different random initial points x^0 and $\lambda = \frac{8}{3\sqrt{3}}$ and $\lambda = 1$. Any sequence $\{z^k\}$ started from one of the 10 initial points converges to one of these minimizers. Figure 5.2 shows the convergence of the corresponding function values $f(z^k)$. This example shows the possibility of the SSQP algorithm for finding a global minimizer of (1.1) with $0 < p < 1$, even such problem is NP hard in general.

Example 5.2. We use randomly generated standard testing problems to show the validity of the SSQP algorithm for finding a scaled stationary point of (1.1). For a given positive integer n_0, we use the following Matlab code to generate $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.



(a) (b)

Fig. 5.1: Tracks of z^k: (a) $\lambda = \frac{8}{3\sqrt{3}}$; (b) $\lambda = 1$



(a) (b)

Fig. 5.2: Convergence of $f(z^k)$: (a) $\lambda = \frac{8}{3\sqrt{3}}$; (b) $\lambda = 1$

```
n = 100 * n_0;  m = n/4;  v = zeros(n,1);  A = randn(m,n);  P = randperm(n);
for j = 1 : n
A(:,j) = A(:,j)/norm(A(:,j));
end
v(P(1 : n_0), 1) = 2 * randn(n_0, 1);  b = A * v − 0.1 * randn(m, 1)
```

We set $n_0 = 10$ in the Matlab code and choose $x^0 = 0 \in \mathbb{R}^n$. Then $\|v\|_0 = 10$, and from $\varphi(0) = 0$, we have $f(z^0) = H(z^0)$ and $\|G(z^0)\|_\infty = 0$. Moreover, at these given data, $f(z^0) = 3.4939$, $\|\tilde{g}(z^0, \mu_0)\|_\infty = 0.2071$ and $\beta = 2\|A^T A\|_2 = 18.2177$.

We consider the optimization problem

$$\min_{x \in \mathbb{R}^n} \quad \ln(\|Ax - b\|_2^2 + 1) + \sum_{i=1}^{n} \varphi(|x_i|^p), \tag{5.2}$$

where φ is defined by φ_i, $i = 1, 2, \ldots, 6$ in Appendix with $a = 3.7$ for φ_1, φ_5, φ_6, and $a = 1$ for φ_2, φ_3, φ_4. To show the robustness of the SSQP algorithm, we choose $\lambda = 0.3$ for all of the 6 penalty functions.

The numerical results using the SSQP algorithm for solving (5.2) with $p = \frac{1}{2}$ and $p = 1$ are listed in Table 5.1, where k^* is the number such that $\mu_k \leq \varepsilon$ and

φ	α	p	k^*	$f(z^{k^*})$	μ^{k^*}	$\|z^{k^*}\|_0$	$\|z^{k^*}-v\|_2$
φ_1	0.3	0.5	1163	2.3652	7.16E-8	9	0.9967
		1	1715	2.2894	1.72E-10	191	1.2253
φ_2	0.3	0.5	1378	2.0435	5.22E-6	9	0.9972
		1	2101	1.8081	4.85E-6	182	1.2222
φ_3	0.6	0.5	6551	1.7901	1.43E-10	19	0.8247
		1	2539	1.5076	6.25E-10	175	1.1938
φ_4	2	0.5	9152	2.6110	1.74E-11	73	2.2405
		1	2198	1.5362	6.65E-6	32	0.4915
φ_5	0.4111	0.5	1453	1.9782	7.96E-8	9	0.9855
		1	1893	1.6617	1.88E-6	180	1.0415
φ_6	0.3	0.5	2189	3.3313	1.66E-7	15	0.6613
		1	1271	3.4939	1.45E-9	194	1.0962

Table 5.1: The SSQP algorithm for finding an $\varepsilon(=10^{-3})$ scaled stationary point of (5.2) with $p=0.5$ and $p=1$

$\|G(z^k)\|_\infty \le \varepsilon$ for all $k \ge k^*$, $\|z^{k^*}\|_0$ is the number of nonzero elements of z^{k^*}, and $\varepsilon = 10^{-3}$.

For $\varphi := \varphi_6$ and different values of ε, the steps k^* and CPU time that are needed to obtain an ε scaled stationary point z^{k^*} are reported in Figures 5.3 with $f(z^{k^*})$ and $\|z^{k^*}-v\|_2$.

REFERENCES

[1] A. Antoniadias and J. Fan, *Regularization of wavelets approximations*, J. Amer. Statist. Assoc., 96 (2001), pp. 929–967.

[2] W. Bian and X. Chen, *Smoothing neural network for constrained non-Lipschitz optimization with applications*, IEEE Trans. Neural Netw., to appear.

[3] J.V. Burke, A.S. Lewis and M.L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751-779.

[4] C. Catis, N. I. M. Gould and P. Toint , *On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming*, SIAM J. Optim., 21 (2011), pp. 1721-1739.

[5] X. Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, preprint (2011).

[6] X. Chen, D. Ge, Z. Wang and Y. Ye, *Complexity of unconstrained L_2-L_p minimization*, submitted to Math. Program., under revision.

[7] X. Chen and W. Zhou, Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization, SIAM J. Imaging Sci., 3 (2010), pp. 765-790.

[8] 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 (2010), pp. 2832–2852.

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

[10] J. Fan, *Comments on 'Wavelets in stastics: a review' by A. Antoniadis*, Stat. Method. Appl., 6 (1997), pp. 131–138.

[11] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc., 96 (2001), pp. 1348–1360.

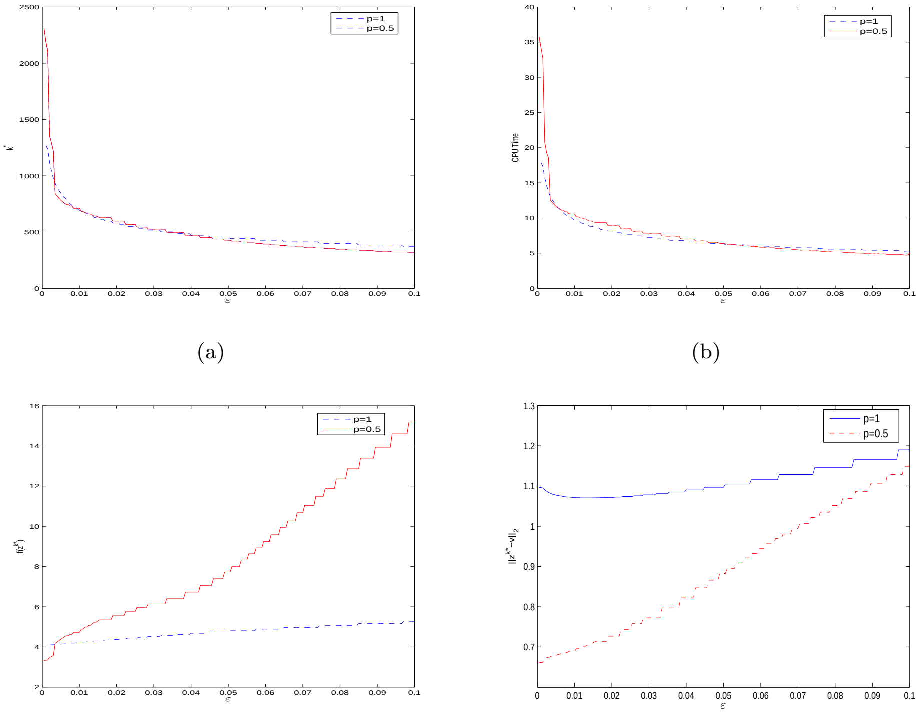
[12] R. Garmanjani and L.N. Vicente, *Smoothing and worst case complexity for direct-search methods in non-smooth optimization*, preprint, 2011.

[13] D. Ge, X. Jiang and Y. Ye, *A note on the complexity of L_p minimization*, Math. Program., 21 (2011), pp. 1721-1739.

[14] P.E. Gill, W. Murray and M.A. Saunders, *SNOPT: an SQP algorithm for large-scale constrained optimization*, SIAM J. Optim., 12 (2002), pp. 979–1006.

[15] J.-L. Goffin, *On the convergence rate of subgradient optimization methods*, Math. Program., 13 (1977), pp. 329–347.

[16] N.I.M. Gould and D.P. Robinson, *A second derivative SQP method: global convergence*,



(a)

(b)

(c)

(d)

Fig. 5.3: SSQP algorithm for (5.2) with φ_6 to find an ε scaled stationary point with different values of ε: (a) k^*, (b) CPU time, (c) $f(z^{k^*})$, (d) $\|z^{k^*} - v\|_2$

SIAM J. Optim., 20 (2010), pp. 2023–2048.

[17] J. Huang, J. L. Horowitz and S. Ma, *Asymptotic properties of bridge estimators in sparse high-dimensional regression models*, Ann. Statist., 36 (2008), pp. 587–613.

[18] J. Huang, S. Ma, H. Xue and C. Zhang, *A group bridge approach for variable selection*, Biometrika, 96 (2009), pp. 339–355.

[19] R. Tibshirani, *Shrinkage and selection via the Lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.

[20] Yu. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

[21] 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 (2008), pp. 2–25.

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

[23] J.V. Vazirani, *Approximation Algorithms*, Springer, Berlin, 2003.

[24] C.-H Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Ann. Statist., 38 (2010), pp. 894–942.

Appendix A. Penalty functions and assumption (A_φ).

We consider the following six penalty functions which are often used in statistics and sparse reconstruction.

- soft thresholding penalty function [17, 19]: $\varphi_1(s) = \lambda s$
- logistic penalty function [21]: $\varphi_2(s) = \lambda \log(1 + as)$
- fraction penalty function [7, 21]: $\varphi_3(s) = \lambda \frac{as}{1+as}$

- hard thresholding penalty function[10]: $\varphi_4(s) = \lambda^2 - (\lambda - s)_+^2$
- smoothly clipped absolute deviation (SCAD) penalty function[10]:

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

- minimax concave penalty (MCP) function [24]:

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

where a and λ are two positive parameters, especially, $a > 2$ in the SCAD penalty function and $a > 1$ in the MCP function.

(1) For the penalty function φ_1, assumption (A_φ) holds with $\alpha \geq \lambda$.

(2) The minimum of $\varphi_2(s)$ is 0 obtained at 0 and $\lim_{s\to\infty} \varphi_2(s) = \infty$. $\varphi_2(s)$ is twice continuously differentiable on $(0, \infty)$, and

$$\nabla\varphi_2(s) = \frac{\lambda a}{(1+as)}, \quad \nabla^2\varphi_2(s) = -\frac{\lambda a^2}{(1+as)^2},$$

which follows that $|\nabla\varphi_2(s)| \leq \lambda a$, $|\nabla^2\varphi_2(s)| \leq \lambda a^2$ and $|\nabla^2\varphi_2(s)|s \leq \lambda a$. Hence, assumption (A_φ) holds with $\alpha \geq \max\{\lambda a, \lambda a^2\}$ for φ_2.

(3) The minimum of $\varphi_3(s)$ is 0 obtained at 0 and $\lim_{s\to\infty} \varphi_3(s) = \lambda$. $\varphi_3(s)$ is twice continuously differentiable on $(0, \infty)$, and

$$\nabla\varphi_3(s) = \frac{\lambda a}{(1+as)^2}, \quad \nabla^2\varphi_3(s) = -\frac{2\lambda a^2}{(1+as)^3},$$

which follows that $|\nabla\varphi_3(s)| \leq \lambda a$, $|\nabla^2\varphi_3(s)| \leq 2\lambda a^2$ and $|\nabla^2\varphi_3(s)|s \leq 2\lambda a$. Hence, assumption (A_φ) holds with $\alpha \geq \max\{2\lambda a^2, 2\lambda a\}$ for φ_3.

(4) For the penalty function φ_4, we can easily obtain that

$$\nabla\varphi_4(s) = 2(\lambda - s)_+ \text{ and } \partial(\nabla\varphi_4(s)) = \begin{cases} -2 & \text{if } s < \lambda \\ [-2, 0] & \text{if } s = \lambda \\ 0 & \text{if } s > \lambda. \end{cases}$$

Hence, assumption (A_φ) holds with $\alpha \geq \max\{2\lambda, 2\}$ for φ_4.

(5) The SCAD penalty function can be expressed by the form

$$\varphi_5(s) = \begin{cases} \lambda s & \text{if } s \leq \lambda \\ \dfrac{2a\lambda s - s^2 - \lambda^2}{2(a-1)} & \text{if } \lambda < s \leq a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & \text{if } a\lambda < s. \end{cases}$$

The minimum of the SCAD penalty function on $[0, \infty)$ is 0 obtained at 0 and its maximum is $\frac{(a+1)\lambda^2}{2}$ obtained at all $s \geq a\lambda$. $\varphi_5(s)$ is continuously differentiable on $(0, \infty)$ and

$$\nabla\varphi_5(s) = \min\{\lambda, \frac{(a\lambda - s)_+}{a-1}\}.$$

Hence, the SCAD penalty function is globally Lipschitz with Lipschitz constant λ. Moreover, $\varphi_5(s)$ is twice continuously differentiable for $s \in (0, \lambda) \bigcup (\lambda, a\lambda) \bigcup (a\lambda, \infty)$, and

$$\partial(\nabla \varphi_5(s)) = \begin{cases} 0 & \text{if } 0 < s < \lambda \quad \text{or} \quad s > a\lambda \\ [-\dfrac{1}{a-1}, 0] & \text{if } s = \lambda \text{ and } s = a\lambda \\ -\dfrac{1}{a-1} & \text{if } \lambda < s < a\lambda. \end{cases}$$

Hence, assumption (A_φ) holds with $\alpha \geq \max\{\lambda, \frac{1}{a-1}, \frac{a\lambda}{a-1}\}$ for φ_5.

(6) The MCP function can be expressed by the form

$$\varphi_6(s) = \begin{cases} \lambda s - \dfrac{s^2}{2a} & \text{if } s < a\lambda \\ \dfrac{a\lambda^2}{2} & \text{if } s \geq a\lambda. \end{cases}$$

The minimum of MCP is 0 obtained at 0 and its maximum is $\frac{a\lambda^2}{2}$ obtained at all $s \geq a\lambda$. $\varphi_6(s)$ is continuously differentiable in $(0, \infty)$ and

$$\nabla \varphi_6(s) = (\lambda - \frac{s}{a})_+, \quad \forall s \in (0, \infty).$$

Hence, the MCP function is globally Lipschitz with Lipschitz constant λ. Moreover, $\varphi_6(s)$ is twice continuously differentiable for $s \in (0, a\lambda) \bigcup (a\lambda, \infty)$, and

$$\partial(\nabla \varphi_6(s)) = \begin{cases} -\dfrac{1}{a} & \text{if } 0 < s < a\lambda \\ [-\dfrac{1}{a}, 0] & \text{if } s = a\lambda \\ 0 & \text{if } s > a\lambda. \end{cases}$$

Hence, assumption (A_φ) holds with $\alpha \geq \max\{\lambda, \frac{1}{a}\}$ for φ_6.