WORST-CASE COMPLEXITY OF SMOOTHING QUADRATIC REGULARIZATION METHODS FOR NON-LIPSCHITZIAN OPTIMIZATION

WEI BIAN* AND XIAOJUN CHEN†

6 February 2012, Revised 2 January, 26 May 2013

Abstract. In this paper, we propose a smoothing quadratic regularization (SQR) algorithm for solving a class of nonsmooth nonconvex, perhaps even non-Lipschitzian minimization problems, which has wide applications in statistics and sparse reconstruction. The proposed SQR algorithm is a first order method. 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 that is inexpensive to calculate. We show that the worst-case complexity of reaching an ϵ scaled stationary point is $O(\epsilon^{-2})$. Moreover, if the objective function is locally Lipschitz continuous, the SQR algorithm with a slightly modified updating scheme for the smoothing parameter and iterate can obtain an ϵ Clarke stationary point in at most $O(\epsilon^{-3})$ iterations.

Key words. Nonsmooth nonconvex optimization, smoothing approximation, quadratic regularization, convergence, worst-case complexity, stationary point.

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

1. Introduction. Convexity and Lipschitz continuity are two important conditions in optimization. However, some real-world applications are often modeled by nonconvex or even non-Lipschitzian 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 $H : \mathbb{R}^n \to [0, +\infty)$ is continuously differentiable and its gradient ∇H is globally Lipschitz continuous with a Lipschitz constant $L_{\nabla H} > 0$, $0 < p \leq 1$, and $\varphi : [0, +\infty) \to [0, +\infty)$ is a given penalty function satisfying the following assumption.

(A_φ) φ is continuously differentiable, nondecreasing, φ' is locally Lipschitz continuous, and there is a positive constant α such that for all $t \in (0, \infty)$,

$$0 \leq \varphi'(t) \leq \alpha, \quad |\xi| \leq \alpha \quad \text{and} \qquad |\xi| t \leq \alpha, \quad \forall \xi \in \partial(\varphi'(t)),$$

where ∂ means the Clarke generalized gradient [15].

When $p = 1$, the objective function f in (1.1) is locally Lipschitz continuous, and globally Lipschitz continuous if H is globally Lipschitz continuous. Lipschitz continuity is important to ensure the existence of the Clarke generalized gradient at every point [15]. However, f may be non-Lipschitz continuous for $0 < p < 1$.

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 A.

*Department of Mathematics, Harbin Institute of Technology, Harbin, China (bianweilvse520@163.com). The author's work was supported by the Hong Kong Polytechnic University Postdoctoral Fellowship Scheme and the NSF foundation (11101107,11271099) 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 (PolyU5003/10P).

Numerical algorithms for solving nonsmooth optimization have been studied for decades, but most algorithms assume the Lipschitzian continuity of the objective function in convergence and worst-case complexity analysis. When one element in the generalized gradient of the objective function can be found at every point, the worst-case complexity of the subgradient methods is proved to be of the order $O(\epsilon^{-2})$ for the globally Lipschitz continuous and convex optimization [21]. Based on a special smoothing technique for the maximal function, Nesterov [26] improves the traditional worst-case complexity of the gradient algorithms to $O(\epsilon^{-1})$ for nonsmooth convex constrained optimization. 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. Most recently, by means of the first order methods, Cartis, Gould and Toint [6] 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 and globally Lipschitz continuous 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 $O(\epsilon^{-2})$ iterations to reduce a first order criticality measure below ϵ in a first order trust region method or a quadratic regularization method, where the worst-case 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 [19], Garmanjani and Vicente propose a smoothing direct search (DS) algorithm based on smoothing techniques and derivative free methods to solve a general unconstrained nonsmooth nonconvex, Lipschitzian minimization problem. The smoothing DS algorithm can be seen as a zero order method, where the gradient is not calculated in the algorithm. The smoothing DS algorithm can find an x such that $\|\nabla \tilde{f}(x,\mu)\|_\infty \leq \epsilon$ and $\mu \leq \epsilon$ in at most $O(\epsilon^{-3}\log\epsilon^{-1})$ function evaluations, where $\tilde{f}$ is a smoothing function of f and $\mu > 0$ is a parameter. In [20], Ge, Jiang and Ye develop an interior-point potential reduction algorithm for solving the following non-Lipschitzian constrained optimization

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

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

In this paper, we present a smoothing quadratic regularization (SQR) algorithm for solving (1.1) with the worst-case complexity estimation. The SQR algorithm uses the smoothing functions [2, 10, 19, 26, 30] and regularization methods [6, 7, 8, 28]. 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 that is inexpensive to calculate. We show that the worst-case complexity of finding an ϵ scaled stationary point is $O(\epsilon^{-2})$. Moreover, if the objective function is locally Lipschitz continuous, the SQR algorithm with a slightly modified updating scheme for the smoothing parameter and iterate z^k can obtain an ϵ Clarke stationary point in at most $O(\epsilon^{-3})$ steps. To the best of our knowledge, the SQR algorithm is the first algorithm with the worst-case complexity for non-Lipschitzian unconstrained optimization. As expectation, when applying the gradient of $\tilde{f}$, the modified SQR algorithm for locally Lipschitz continuous minimization has a better complexity result than the smoothing DS algorithm proposed in [19], in which the gradient information

is not used. 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 methods proposed in [6] cannot be applied to solve (1.1).

Nonsmooth nonconvex penalty functions play an important role in sparse reconstruction and statistical modeling, particularly in variable selection. Penalty functions satisfying (A_φ) in (1.1) provide efficient models to extract the essential features of solutions which are sparse in the sense that they have many zero entries [2, 9, 10, 12, 14, 16, 18, 23, 24, 29, 33]. Finding a solution with few nonzero entries for an underdetermined linear or nonlinear system can be modeled as a minimization problem with the l_0-norm penalty function $\|x\|_0$ defined as the number of nonzero entries in x. Such problem is difficult to solve due to the discontinuity of $\|x\|_0$. Nonsmooth penalty functions for finding desired sparse solutions have been studied extensively in the last decades. Three principles (unbiasedness, sparsity and continuity) for a good penalty function are introduced in [1, 18]. A widely used penalty function is the l_1-norm, especially the l_2-l_1 minimization problem which is often called LASSO [25] and whose solutions are in the solution set of the corresponding l_2-l_0 problem under the restricted isometry property [4]. Fan and Li [18] show that the smoothly clipped absolute deviation (SCAD) penalty function proposed in [17] has better properties than the l_1-norm penalty function in parametric and nonparametric models. More recently, Zhang propose a minimax concave penalty function (MCP) [33]. In [14, 29], it is shown that logistic and fraction penalty functions yields better edge preservation than convex penalty functions. All penalty functions mentioned in this paragraph are Lipschitz continuous and satisfy assumption (A_φ).

When $0 < p < 1$, the penalty function in (1.1) is non-Lipschitzian, which includes the l_p-norm penalty $\|x\|_p^p$ as a special case. Fan and Li [18] point out that the oracle property does not hold for the l_1-norm penalty, while it continues to hold for the l_p-norm penalty with $0 < p < 1$ by suitable choice of the parameters in it, where the oracle property means that when the true parameters have some zero components, they are estimated as 0 with probability tending to 1, and the nonzero components are estimated as well as when the correct submodel is known. In [9], Chartrand and Staneva show that by replacing the l_1-norm in the l_2-l_1 minimization with the l_p-norm, exact reconstruction is possible with substantially fewer measurements. In [23], 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 [24]. Numerical methods for solving l_2-l_p minimization problems have been proposed and analyzed, including reweighted minimization algorithms [4] and smoothing methods [2, 12, 13].

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 SQR algorithm for solving (1.1) and theoretical analysis including the convergence and worst-case complexity results are given, where the worst-case complexity of reaching an ϵ scaled stationary point is $O(\epsilon^{-2})$. In Section 4, we slightly modify the SQR algorithm for solving (1.1) with $p = 1$ and the worst-case complexity of reaching an ϵ Clarke stationary point is $O(\epsilon^{-3})$. In Section 5, numerical examples are given to show the worst-case efficiency of the SQR algorithm.

Let $I = \{1, 2, \ldots, n\}$ and $\mathbb{N} = \{0, 1, \ldots\}$. For a column vector $x \in \mathbb{R}^n$, x_i denotes

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 Approximation. Smoothing approximations for nonsmooth optimization have been studied for decades [2, 10, 26, 30]. 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.

For $s \in \mathbb{R}$, $\mu > 0$ and $x \in \mathbb{R}^n$, define

$$\kappa(s,\mu) = 8\alpha p \begin{cases} |\frac{s}{2}|^{p-2} & \text{if } |s| > 2\mu \\ \mu^{p-2} & \text{if } |s| \leq 2\mu; \end{cases} \quad \theta(s,\mu) = \begin{cases} |s| & \text{if } |s| > \mu \\ \dfrac{s^2}{2\mu} + \dfrac{\mu}{2} & \text{if } |s| \leq \mu; \end{cases} \tag{2.1}$$

$$\begin{aligned} &\tilde{f}(x,\mu) = H(x) + \sum_{i=1}^{n} \varphi(\theta^p(x_i,\mu)); \\ &\tilde{g}(x,\mu) = [\tilde{g}_1(x,\mu),\ldots,\tilde{g}_n(x,\mu)]^T := \nabla_x \tilde{f}(x,\mu); \end{aligned} \tag{2.2}$$

Definition 2.1. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a continuous 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 $x \in \mathbb{R}^n$, $\lim_{z\to x,\mu\downarrow 0} \tilde{h}(z,\mu) = h(x)$.*

It is easy to verify that $\theta(s,\mu)$ in (2.1) is a smoothing function of $|\cdot|$, which satisfies the following properties.

Lemma 2.2.

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

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

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

Due to the continuous differentiability of functions φ and θ, $\varphi(\theta^p(s,\mu))$ is a smoothing function of $\varphi(|s|^p)$ and $\tilde{f}(x,\mu)$ is a smoothing function of $f(x)$. Moreover, from the condition (A_φ) and Lemma 2.2 (iii), we have

$$0 \leq \tilde{f}(x,\mu) - f(x) \leq n\alpha(\frac{\mu}{2})^p, \quad \forall x \in \mathbb{R}^n,\ \mu \in [0,+\infty). \tag{2.3}$$

Note that $\varphi(\theta^p(s,\cdot))$ is nondecreasing in $[0,\infty)$ for any fixed $s \in \mathbb{R}$.

The following proposition presents an estimation on elements in the generalized Hessian [15] 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 $|\xi| \leq \kappa(s,\mu)$, which means that $\kappa(s,\mu)$ is an upper bound for all elements in the generalized Hessian $\partial_s^2 \varphi(\theta^p(s,\mu))$ for any fixed $\mu \in (0,\infty)$.*

Proof. See Appendix B. □

3. Smoothing Quadratic Regularization Algorithm. Quadratic regularization methods are popular iterative methods for solving smooth optimization problems [6, 22, 32], which solve a quadratic programming problem at each iteration. Inspired by smoothing approximations and quadratic regularization methods, we propose a smoothing quadratic regularization (SQR) algorithm for (1.1). At each iteration of the SQR algorithm, a convex quadratic approximation with a diagonal and positive definite Hessian matrix is constructed by using the smoothing function $\tilde{f}$. The quadratic subproblem has a simple closed-form solution that is inexpensive to calculate. The smoothing parameter is updated by a simple criterion. We show that

any accumulation point of the iterates is a scaled stationary point of (1.1) with the worst-case complexity $O(\epsilon^{-2})$.

In our convergence and worst-case complexity analysis, we assume that f is level bounded, i.e. for any $\Gamma > 0$, the level set $\{x \in \mathbb{R}^n : \ f(x) \leq \Gamma\}$ is bounded. Without loss of generality, we assume that $f(x) \geq 0$ for all $x \in \mathbb{R}^n$.

For any fixed $y \in \mathbb{R}^n$, $\mu > 0$ and $\beta > 0$, we define the quadratic approximation of $\tilde{f}(\cdot, \mu)$ around y as the following,

$$Q(x, y, \mu, \beta) = \tilde{f}(y, \mu) + \langle \tilde{g}(y, \mu), x - y \rangle + \frac{1}{2} \sum_{i=1}^{n} \gamma_i(y, \mu, \beta)(x_i - y_i)^2, \tag{3.1}$$

where the functions $\tilde{f}$ and $\tilde{g}$ are given in (2.2),

$$\gamma_i(y, \mu, \beta) = \max \left\{ \beta + \kappa(y_i, \mu), \ \frac{|\tilde{g}_i(y, \mu)|}{\max\{\frac{|y_i|}{2}, \mu\}^{1-\frac{p}{2}} \mu^{\frac{p}{2}}} \right\}, \quad i \in I, \tag{3.2}$$

and the functions κ is given in (2.1).

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

$$H(x) \leq H(y) + \langle \nabla H(y), x - y \rangle + \frac{L_{\nabla H}}{2} \|x - y\|_2^2. \tag{3.3}$$

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 $|s - \hat{s}| \leq \max\{\frac{|\hat{s}|}{2}, \mu\}$, the following inequality holds*

$$\varphi(\theta^p(s, \mu)) \leq \varphi(\theta^p(\hat{s}, \mu)) + \langle \nabla_{\hat{s}} \varphi(\theta^p(\hat{s}, \mu)), s - \hat{s} \rangle + \frac{\kappa(\hat{s}, \mu)}{2} (s - \hat{s})^2. \tag{3.4}$$

Proof. See Appendix B. □

By (3.3) and (3.4), the following lemma gives an important relation between the quadratic function Q and the smoothing function $\tilde{f}$ defined in (2.2).

LEMMA 3.2. *For $x, y \in \mathbb{R}^n$ such that $|x_i - y_i| \leq \max\{\frac{|y_i|}{2}, \mu\}$, $\forall i \in I$, if the following inequality holds with $\beta > 0$,*

$$H(x) \leq H(y) + \langle \nabla H(y), x - y \rangle + \frac{1}{2} \beta \|x - y\|_2^2, \tag{3.5}$$

then

$$\tilde{f}(x, \mu) \leq Q(x, y, \mu, \beta), \tag{3.6}$$

where Q is defined in (3.1).

Proof. By (3.4) and (3.5), inequality (3.6) holds for $\gamma_i(y, \mu, \beta) \equiv \beta + \kappa(y_i, \mu)$, $\forall i \in I$. Thus (3.6) holds with γ_i defined by (3.2) using the max operator. □

The quadratic program in the SQR algorithm is constructed based on Lemma 3.2. For any fixed $y \in \mathbb{R}^n$, $\mu > 0$ and $\beta > 0$, the right hand of inequality (3.6) is a strictly quadratic convex function. Hence we can use $Q(x, y, \mu, \beta)$ as the quadratic regularization to $\tilde{f}(\cdot, \mu)$ around y.

3.1. Proposed Algorithm. Following the methods for updating the regularization weight in [6, 7, 8], we update the regularization weight in the SQR algorithm when the Lipschitz constant of ∇H is unknown. The scheme for updating the smoothing parameter μ is crucial for the smoothing methods, which can affect the convergence and worst-case complexity of the SQR algorithm. A simple and intelligent scheme for updating μ is used in the SQR algorithm.

SQR Algorithm

Step 0: **Initialization:** Choose $x^0 = z^0 \in \mathbb{R}^n$, $\mu_0 > 0$, $\beta_0 \geq 1$, $0 < \sigma, \sigma_1, \sigma_2 < 1$, $\eta > 1$. Set $k = 0$.

Step 1: **New point calculation:** Solve

$$y^k = \arg\min_{x \in \mathbb{R}^n} Q(x, x^k, \mu_k, \beta_k), \tag{3.7}$$

where function Q is given in (3.1).

If $y^k \neq x^k$, define

$$r_k = \frac{H(y^k) - H(x^k) - \langle \nabla H(y^k), x^k - y^k \rangle}{\frac{1}{2}\beta \|x^k - y^k\|^2},$$

else define $r_k = 1$. When $r_k \leq 1$, let $x^{k+1} = y^k$; else let $x^{k+1} = x^k$.

Step 2: **Updating the regularization weight:** Set

$$\beta_{k+1} = \begin{cases} \max\{\beta_0, \sigma_1 \beta_k\} & \text{if } r_k \leq \sigma_2 \qquad [k \text{ very successful}] \\ \beta_k & \text{if } r_k \in (\sigma_2, 1] \qquad [k \text{ successful}] \\ \eta \beta_k & \text{if } r_k > 1 \qquad [k \text{ unsuccessful}] \end{cases} \tag{3.8}$$

If $r_k > 1$, let

$$\mu_{k+1} = \mu_k, \ z^{k+1} = z^k$$

and return to Step 1; otherwise, go to Step 3.

Step 3: **Updating the smoothing parameter:** Let

$$\mu_{k+1} = \begin{cases} \mu_k & \text{if } \tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) < -4\alpha p \mu_k^p \\ \sigma \mu_k & \text{othewise.} \end{cases} \tag{3.9}$$

Step 4: **Constructing convergence point:** Let

$$z^{k+1} = \begin{cases} x^k & \text{if } \mu_{k+1} = \sigma \mu_k \\ z^k & \text{othewise.} \end{cases} \tag{3.10}$$

Increment k by one and return to Step 1.

Noting that the Hessian matrix $\nabla_x^2 Q(x, x^k, \mu_k, \beta_k)$ is a diagonal and positive definite matrix, (3.7) has a simple closed-form solution

$$y_i^k = x_i^k - \frac{\tilde{g}_i(x^k, \mu_k)}{\gamma_i(x^k, \mu_k, \beta_k)}, \quad \forall i \in I. \tag{3.11}$$

The proposed SQR algorithm is a first order method. The sequences $\{x^k\}$, $\{y^k\}$, $\{\beta_k\}$, $\{r_k\}$, $\{\mu_k\}$ and $\{z^k\}$ are well-defined. Specially, when $y^k = x^k$, from (3.11), we

find that $\tilde{g}(x^k, \mu_k) = 0$, which means that x^k is a Clarke stationary point of $\tilde{f}(x, \mu_k)$ for the fixed μ_k. However, x^k may not be a stationary point of $f(x)$. Since we want to find a stationary point of f, we have to decrease the value of μ at this iteration and continue to run the SQR algorithm.

LEMMA 3.3. *For all $k \in \mathbb{N}$, $\beta_k \le \bar{\beta} := \max\{\beta_0, \eta L_{\nabla H}\}$.*

Proof. Since ∇H is globally Lipschitz continuous with Lipschitz constant $L_{\nabla H}$, then when $\beta_k \ge L_{\nabla H}$, k is successful in the sense of (3.8). We set $\beta_{k+1} = \eta\beta_k$ only when $\beta_k < L_{\nabla H}$. Thus, $\beta_k \le \max\{\beta_0, \eta L_{\nabla H}\}$ for all $k \in \mathbb{N}$. □

In particular, if $L_{\nabla H}$ is known, we can let $\beta_0 = L_{\nabla H}$, which follows that $\beta_k = \beta_0$ for all $k \in \mathbb{N}$ and every iteration is successful.

For the sequences $\{\mu_k\}$ and $\{r_k\}$, we denote

$$N^- = \{k \in \mathbb{N} : \ \mu_{k+1} = \sigma\mu_k\} \quad \text{and} \quad N_s = \{k \in \mathbb{N} : \quad r_k \le 1\}.$$

For any $k \in N_s$, the kth iteration is successful. From Step 2 in the SQR algorithm, we find that $N^- \subseteq N_s$ and the sequence $\{z^k\}$ can be written as

$$\begin{cases} z^{k+1} = x^k & \text{if } k \in N^- \\ z^k = x^{N_r^-} & \text{if } N_r^- + 1 \le k \le N_{r+1}^-, \end{cases} \tag{3.12}$$

where N_r^- is the rth smallest elements in N^-. The relationships among x^k, z^k and μ_k with $\mu_0 = 1$ are illustrated in Figure 3.1.

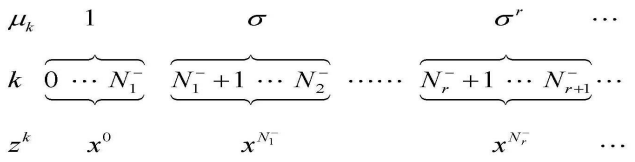


Fig. 3.1: Illustration of the relationship among x^k, z^k and μ_k

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

$$K(x, \mu) = \{i \in I : \ |x_i| \le 2\mu\}, \quad J(x, \mu) = \{i \in I : \ |x_i| > 2\mu\}. \tag{3.13}$$

$K(x, \mu)$ and $J(x, \mu)$ are mutually disjoint and $I = K(x, \mu) \bigcup J(x, \mu)$. For any fixed $\beta > 0$, we divide each of $K(x, \mu)$ and $J(x, \mu)$ into two mutually disjoint sets

$$\begin{aligned} K^+(x, \mu, \beta) &= \{i \in K(x, \mu) : \ |\tilde{g}_i(x, \mu)| \ge \mu(\beta + \kappa(x_i, \mu))\}, \\ J^+(x, \mu, \beta) &= \left\{i \in J(x, \mu) : \ |\tilde{g}_i(x, \mu)| \ge |\frac{x_i}{2}|^{1-\frac{p}{2}} \mu^{\frac{p}{2}} (\beta + \kappa(x_i, \mu))\right\}, \\ K^-(x, \mu, \beta) &= K(x, \mu) \setminus K^+(x, \mu, \beta) \quad \text{and} \quad J^-(x, \mu, \beta) = J(x, \mu) \setminus J^+(x, \mu, \beta). \end{aligned}$$

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

LEMMA 3.4. *The sequence $\{\tilde{f}(x^k, \mu_k)\}$ is monotonically decreasing. In particular, when $k \in N_s$,*

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) \le -\sum_{i=1}^{n} \frac{\tilde{g}_i^2(x^k, \mu_k)}{2\gamma_i(x^k, \mu_k, \beta_k)}. \tag{3.14}$$

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

Proof. Firstly, we prove that

$$|y_i^k - x_i^k| \leq \max\{\frac{|x_i^k|}{2}, \mu_k\}, \quad \forall i \in I. \tag{3.15}$$

From (3.11), for any $i \in I$, we have

$$|y_i^k - x_i^k| = \frac{|\tilde{g}_i(x^k, \mu_k)|}{\gamma_i(x^k, \mu_k, \beta_k)}. \tag{3.16}$$

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

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

which implies that

$$\gamma_i(x^k, \mu_k, \beta_k) = \frac{|\tilde{g}_i(x^k, \mu_k)|}{\mu_k}. \tag{3.18}$$

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

$$|x_i^k| \leq 2\mu_k \quad \text{and} \quad |\tilde{g}_i(x^k, \mu_k)| < \mu_k(\beta_k + \kappa(x_i^k, \mu_k)), \tag{3.19}$$

which implies that

$$\gamma_i(x^k, \mu_k, \beta_k) = \beta_k + \kappa(x_i^k, \mu_k). \tag{3.20}$$

Then, from (3.16)-(3.20), we know

$$|y_i^k - x_i^k| \leq \mu_k, \quad \forall i \in K(x^k, \mu_k).$$

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

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

which implies that

$$\gamma_i(x^k, \mu_k, \beta_k) = \frac{|\tilde{g}_i(x^k, \mu_k)|}{|\frac{x_i^k}{2}|^{1-\frac{p}{2}}\mu_k^{\frac{p}{2}}}. \tag{3.22}$$

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

$$|x_i^k| > 2\mu_k \quad \text{and} \quad |\tilde{g}_i(x^k, \mu_k)| < |\frac{x_i^k}{2}|^{1-\frac{p}{2}}\mu_k^{\frac{p}{2}}(\beta_k + \kappa(x_i^k, \mu_k)), \tag{3.23}$$

which implies that

$$\gamma_i(x^k, \mu_k, \beta_k) = \beta_k + \kappa(x_i^k, \mu_k). \tag{3.24}$$

Then, from (3.16) and (3.22)-(3.24), we have

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

Therefore, we can obtain the estimation in (3.15).

When $k \notin N_s$, $r_k > 1$. From Step 2, $x^{k+1} = x^k$ and $\mu_{k+1} = \mu_k$. Thus, $\tilde{f}(x^{k+1}, \mu_{k+1}) = \tilde{f}(x^k, \mu_k)$.

When $k \in N_s$, $r_k \leq 1$. Then $x^{k+1} = y^k$ and

$$H(x^{k+1}) \leq H(x^k) + \langle \nabla H(x^k), x^{k+1} - x^k \rangle + \frac{1}{2}\beta_k \|x^{k+1} - x^k\|_2^2. \tag{3.25}$$

Applying $|x_i^{k+1} - x_i^k| \leq \max\{\frac{|x_i^k|}{2}, \mu_k\}$, $\forall i \in I$ and (3.25) to Lemma 3.2, it holds that

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

which, together with (3.11), gives that

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

Since $\mu_{k+1} \leq \mu_k$, we can obtain the inequality in (3.14) when $k \in N_s$. Since f is level bounded, from the monotonically decreasing property of $\tilde{f}(x^k, \mu_k)$ and (2.3), we find that there is $R \geq 1$ such that $\|x^k\|_\infty \leq R$, $k \in \mathbb{N}$. □

The objective function f is non-Lipschitzian when $0 < p < 1$. It has been proved in [11] that finding a global minimizer of the unconstrained l_2-l_p minimization problem with $0 < p < 1$ is strongly NP hard. We extend the definition of the scaled first order necessary condition in [13] to define the scaled stationary points of (1.1) with $0 < p \leq 1$. From $0 \leq \varphi'(t) \leq \alpha$ for $t \in (0, \infty)$ in Assumption (A_φ), we have $\lim_{t \downarrow 0} t^p \varphi'(t) = 0$. For simplicity, when $x_i = 0$, we set $|x_i|^p \varphi'(t)_{t=|x_i|^p} = 0$ in the following definition.

DEFINITION 3.5. *Let $G : R^n \to R^n$ be defined by*

$$G(x) := (G_1(x), G_2(x), \ldots, G_n(x)) = X\nabla H(x) + p|X|^p [\varphi'(t)_{t=|x_i|^p}]_{i=1}^n,$$

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

$$\|G(x^*)\|_\infty \leq \epsilon.$$

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

Following the proof of Theorem 2.3 in [13], we can show that any local minimizer of (1.1) is a scaled stationary point of (1.1).

Since the SQR 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 $\mu \to 0$.

PROPOSITION 3.6. *For all $x \in \mathbb{R}^n$, $\mu \in (0, \infty)$ and $0 < p \leq 1$, we have*

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

Proof. See Appendix B. □

The following lemma gives the magnitude of the decreasing of $\tilde{f}(x^k, \mu_k)$ and $\|G(x^k)\|_\infty$ for $k \in N_s$.

LEMMA 3.7. *For all $k \in N_s$, if $K^+(x^k, \mu_k, \beta_k) \bigcup J^+(x^k, \mu_k, \beta_k) \neq \emptyset$, then*

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) < -4\alpha p \mu_k^p; \tag{3.26}$$

otherwise,

$$\|G(x^k)\|_\infty < C\mu_k^{\frac{p}{2}}, \tag{3.27}$$

where $C = \max\left\{2\bar{\beta}\mu_0^{2-\frac{p}{2}} + 19\alpha\mu_0^{\frac{p}{2}}, 2R^2\bar{\beta} + 16\alpha R + 3\alpha\mu_0^{\frac{p}{2}}\right\}$ with $\bar{\beta}$ defined in Lemma 3.3 and $R \geq 1$ such that $R \geq \|x^k\|_\infty$, $\forall k \in \mathbb{N}$.

Proof. Fix $k \in N_s$.We first consider the case $K^+(x^k, \mu_k, \beta_k) \bigcup J^+(x^k, \mu_k, \beta_k) \neq \emptyset$.

If there is an $i \in I$ such that $i \in K^+(x^k, \mu_k, \beta_k)$, from (2.1), (3.14), (3.17) and (3.18), we obtain that

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) \leq -\frac{\mu_k}{2}|\tilde{g}_i(x^k, \mu_k)| < -\frac{\mu_k^2}{2}\kappa(x_i^k, \mu_k) = -4\alpha p \mu_k^p. \tag{3.28}$$

If there is an $i \in I$ such that $i \in J^+(x^k, \mu_k, \beta_k)$, from (2.1), (3.14), (3.21) and (3.22), we have

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) \leq -\frac{1}{2}\mu_k^{\frac{p}{2}}|\frac{x_i^k}{2}|^{1-\frac{p}{2}}|\tilde{g}_i(x^k, \mu_k)| < -4\alpha p \mu_k^p. \tag{3.29}$$

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

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

For $i \in K^-(x^k, \mu_k, \beta_k)$, from (2.1), (3.19) and Lemma 3.3, we obtain

$$\left|x_i^k \tilde{g}_i(x^k, \mu_k)\right| < 2\mu_k^2(\beta_k + 8\alpha p \mu_k^{p-2}) \leq 2(\bar{\beta}\mu_0^{2-\frac{p}{2}} + 8\alpha\mu_0^{\frac{p}{2}})\mu_k^{\frac{p}{2}}, \tag{3.30}$$

where $\bar{\beta}$ is defined as in Lemma 3.3.

If $i \in J^-(x^k, \mu_k, \beta_k)$, from (2.1), (3.23) and Lemma 3.3, we obtain

$$\left|x_i^k \tilde{g}_i(x^k, \mu_k)\right| < 2\mu_k^{\frac{p}{2}}|\frac{x_i^k}{2}|^{2-\frac{p}{2}}(\beta_k + 8\alpha p|\frac{x_i^k}{2}|^{p-2}) \leq 2(R^2\bar{\beta} + 8\alpha R)\mu_k^{\frac{p}{2}}, \tag{3.31}$$

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

Combining (3.30) and (3.31), we have

$$|x_i^k \tilde{g}_i(x^k, \mu_k)| < \max\{2\bar{\beta}\mu_0^{2-\frac{p}{2}} + 16\alpha\mu_0^{\frac{p}{2}}, 2R^2\bar{\beta} + 16\alpha R\}\mu_k^{\frac{p}{2}}, \quad \forall i \in I.$$

From Proposition 3.6, we can conclude the results in this lemma. □

Now, we are ready to present the first convergence result of the SQR algorithm.

LEMMA 3.8. $\lim_{k\to\infty} \mu_k = 0$ *and* $\lim_{k\to\infty} f(x^k) = \lim_{k\to\infty} f(z^k)$ *exists.*

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

$$\sum_{k\in N^-} \mu_k^p < \sum_{r=1}^{\infty} \mu_0\sigma^{p(r-1)} = \frac{\mu_0}{1-\sigma^p}. \tag{3.32}$$

On the other hand, when $k \in N_s \backslash N^-$, from (3.9), we have

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

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

$$\sum_{k \in N_s \backslash N^-} \mu_k^p < \frac{1}{4\alpha p} \sum_{k \in N_s \backslash N^-} [\tilde{f}(x^k, \mu_k) - \tilde{f}(x^{k+1}, \mu_{k+1})] \leq \frac{1}{4\alpha p} \tilde{f}(x^0, \mu_0). \tag{3.33}$$

Adding (3.32) and (3.33), we have

$$\sum_{k \in N_s} \mu_k^p < \frac{1}{4\alpha p} \tilde{f}(x^0, \mu_0) + \frac{\mu_0}{1 - \sigma^p}. \tag{3.34}$$

Next, we will prove that the number of iterations in N_s is infinite. If not, there is $\bar{k} \in \mathbb{N}$ such that $k \notin N_s$, $\forall k \geq \bar{k}$, which implies that $\beta_k \geq \beta_0 \eta^{k-\bar{k}}$, $\forall k \geq \bar{k}$. Since $\eta > 1$, $\lim_{k \to +\infty} \beta_k = +\infty$, which leads a contraction with the result in Lemma 3.3.

Therefore, there are infinite elements in N_s. By the nonincreasing property of μ_k and (3.34), we have $\lim_{k \to \infty} \mu_k = 0$.

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 $\lim_{k \to \infty} \mu_k = 0$, (2.3) and (3.10), we obtain

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

□

Assume the initial point is not an ϵ scaled stationary point of (1.1) for an given $\epsilon \in (0, 1]$. The next theorem proves that there is an iteration such that the generated point z_k of the SQR algorithm is an ϵ scaled stationary point of (1.1) and the worst case complexity is $O(\epsilon^{-2})$. Since $\lim_{k \to \infty} \mu_k = 0$, we suppose $\mu_0 = 1$ in the following complexity estimation, which will not change the complexity order of the SQR algorithm for any $\mu_0 > 0$.

THEOREM 3.9. *Any accumulation point of $\{z^k\}$ generated by the SQR algorithm is a scaled stationary point of (1.1) defined in Definition 3.5. Given any $\epsilon \in (0, 1]$, the total number of successful iterations for obtaining an ϵ scaled stationary point of (1.1) when applying the proposed SQR algorithm is at most*

$$\lceil J_s \epsilon^{-2} \rceil$$

and the total number of iterations is at most

$$\lceil J \epsilon^{-2} \rceil$$

where

$$J_s = \frac{C^2 \tilde{f}(x^0, \mu_0)}{4\alpha p \sigma^p} - \frac{2}{p} \log_\sigma eC + 1 \quad \textit{and} \quad J = J_s + \log_\eta \bar{\beta} - \log_\eta \beta_0 + 1$$

with $\bar{\beta}$ defined in Lemma 3.3 and C defined in Lemma 3.7.

Proof. Let $\epsilon \in (0, 1]$ be a given number. Since $C \geq 1$ and $0 < \sigma < 1$, there is a positive integer $j \geq 2$ such that

$$C\sigma^{\frac{(j-1)p}{2}} < \epsilon \quad \text{and} \quad C\sigma^{\frac{(j-2)p}{2}} \geq \epsilon, \tag{3.35}$$

where C is the constant given in Lemma 3.7. Then from Lemma 3.7 and Lemma 3.8, we have

$$\|G(x^{N_r^-})\|_\infty \le C(\mu_{N_r^-})^{\frac{p}{2}} \le C(\mu_{N_j^-})^{\frac{p}{2}} = C(\sigma^{j-1})^{\frac{p}{2}} < \epsilon, \quad \forall\, r \in \mathbb{N},\, r \ge j. \tag{3.36}$$

From (3.12) and (3.36), we obtain

$$\|G(z^k)\|_\infty < \epsilon, \quad \forall\, k \ge N_j^- + 1. \tag{3.37}$$

In order to let (3.37) hold, it needs to carry out at most $N_j^- + 1$ iterations of the SQR algorithm. From Lemma 3.7, when $k \in N_s \backslash N^-$ and $k \le N_j^-$,

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) < -4\alpha p \mu_k^p \le -4\alpha p \sigma^{(j-1)p} \le -\frac{4\alpha p \sigma^p \epsilon^2}{C^2}. \tag{3.38}$$

Suppose there are k_j successful iterations up to N_j^-, then there are at least $k_j - j + 1$ successful iterations such that inequality (3.38) holds.

Owing to the monotonically increasing property of $\tilde{f}(x^k, \mu_k)$ and (3.38), we have that

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

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

$$k_j < \frac{C^2 \tilde{f}(x^0, \mu_0)}{4\alpha p \sigma^p \epsilon^2} + j - 1. \tag{3.39}$$

From the second inequality in (3.35), we obtain that

$$j \le \frac{2}{p}(\log_\sigma \epsilon - \log_\sigma C) + 2. \tag{3.40}$$

Combining (3.39) with (3.40), we have

$$k_j < \frac{C^2 \tilde{f}(x^0, \mu_0)}{4\alpha p \sigma^p \epsilon^2} + \frac{2}{p} \log_\sigma \epsilon - \frac{2}{p} \log_\sigma C + 1.$$

Since $\epsilon^2 \log_\sigma \epsilon \le -\frac{1}{\ln \sigma}$, we have

$$k_j < J_s \epsilon^{-2},$$

where $J_s = \frac{C^2 \tilde{f}(x^0, \mu_0)}{4\alpha p \sigma^p} - \frac{2}{p} \log_\sigma eC + 1$.

We deduce that $\beta_{N_j^-} \ge \beta_0 \eta^{N_j^- - k_j}$. By Lemma 3.3, we have

$$\frac{\bar{\beta}}{\beta_0} \ge \eta^{N_j^- - k_j}.$$

Taking the logarithm on the both sides of the above inequality and recalling that $\eta > 1$, we obtain

$$N_j^- \le k_j + \log_\eta \bar{\beta} - \log_\eta \beta_0. \tag{3.41}$$

By (3.41), we obtain

$$N_j^- + 1 < \left\lceil J\epsilon^{-2} \right\rceil,$$

where $J = J_s + \log_\eta \bar{\beta} - \log_\eta \beta_0 + 1$.

Coming back to (3.37), this shows the worst-case complexity of the SQR algorithm for obtaining an ϵ scaled stationary point of (1.1). Let $\epsilon \to 0$, 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). □

In general, we can define

$$\gamma_i(y,\mu,\beta) = \max\left\{\beta + \kappa(y_i,\mu),\ \frac{|\tilde{g}_i(y,\mu)|}{\max\{\frac{|y_i|}{2},\mu\}^{1-\tau p}\mu^{\tau p}}\right\},\quad i \in I, \tag{3.42}$$

with $\tau \in (0,1]$, where the $\gamma_i(y,\mu,\beta)$ in (3.2) is a special case of it with $\tau = \frac{1}{2}$. Similar to the analysis and proof ideas in this section, the worst-case complexity of the SQR algorithm with $\gamma_i(y,\mu,\beta)$ defined in (3.42) is $O(\epsilon^{-\frac{1}{\tau}})$ when $0 < \tau < \frac{1}{2}$ and $O(\epsilon^{-2})$ when $\frac{1}{2} \le \tau \le 1$.

4. Locally Lipschitz optimization. In this section, we consider (1.1) with $p = 1$, which is locally Lipschitz continuous optimization problem. We present a slightly modified SQR algorithm to find a Clarke stationary point of (1.1). We call this algorithm SQR$_1$ algorithm. For fixed $x \in \mathbb{R}^n$, $\mu > 0$ and $\beta > 0$, let

$$K^+(x,\mu,\beta) = \left\{i \in K(x,\mu):\ |\tilde{g}_i(x,\mu)| \ge \frac{\mu^2}{\mu_0}(\beta + \kappa(x_i,\mu))\right\},$$
$$J^+(x,\mu,\beta) = \left\{i \in J(x,\mu):\ |\tilde{g}_i(x,\mu)| \ge |\frac{x_i}{2}|^{\frac{1}{2}}\mu^{\frac{3}{2}}\mu_0^{-1}(\beta + \kappa(x_i,\mu))\right\},$$

where $K(x,\mu)$ and $J(x,\mu)$ are defined as in Section 3. Moreover, the index sets $K^-(x,\mu)$, $J^-(x,\mu)$, N_s and N^- are also defined as in Section 3.

In this section, $Q(x,y,\mu,\beta)$ is with the format in (3.1), where $\gamma_i(y,\mu,\beta)$ is given as

$$\gamma_i(y,\mu,\beta) = \max\left\{\beta + \kappa(y_i,\mu),\ \frac{\mu_0|\tilde{g}_i(y,\mu)|}{\max\{\frac{|y_i|}{2},\mu\}^{\frac{1}{2}}\mu^{\frac{3}{2}}}\right\}.$$

SQR$_1$ Algorithm

Step 0-2 are same as them in the SQR algorithm.

Step 3: **Updating the smoothing parameter:**

$$\mu_{k+1} = \begin{cases}\mu_k & \text{if } \tilde{f}(x^{k+1},\mu_k) - \tilde{f}(x^k,\mu_k) < -4\alpha\mu_0^{-2}\mu_k^3\\ \sigma\mu_k & \text{othewise,}\end{cases} \tag{4.1}$$

Step 4: **Constructing convergence point:** Let

$$z^{k+1} = \begin{cases}x_\mu^k & \text{if } \mu_{k+1} = \sigma\mu_k\\ z^k & \text{othewise,}\end{cases} \tag{4.2}$$

where $[x_\mu^k]_i = \begin{cases}x_i^k & \text{if } |x_i^k| \ge \mu_k\\ 0 & \text{otherwise}\end{cases}$ for $i \in I$.

Increment k by one and return to Step 1.

The SQR algorithm and the SQR_1 algorithm have the same structure, except the updating schemes for μ_k and z^k. Since Step 2 in the SQR_1 algorithm is same as it in the SQR algorithm, the estimation on β_k in Lemma 3.3 also holds for the SQR_1 algorithm. Similar to the proof of Lemma 3.4, we have $|x_i^{k+1} - x_i^k| \leq \max\{\mu_k, |\frac{x_i^k}{2}|\}$. Therefore, the statements in Lemma 3.4 hold for the SQR_1 algorithm.

We define a Clarke stationary point of (1.1) with $p = 1$ as follows.

DEFINITION 4.1. *[15] We call x^* an ϵ Clarke stationary point of (1.1) if there exists $\xi \in \partial f(x^*)$ such that*

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

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

When $p = 1$, the scaled stationary point condition $G(x) = 0$ is a necessary but not sufficient condition for the Clarke stationary point of (1.1). Consider the following example,

$$\min \quad f := (x_1 + 2x_2 - 1)^2 + |x_1| + |x_2|.$$

For any $x \in \mathbb{R}^2$, $\partial f(x) = \{(2(x_1-1)+1, 4(x_1-1)+\tau)^T : \tau \in [-1,1]\}$, where $[-1,1] = \partial|s|$ at $s = 0$ is used. The vector $\bar{x} = (1/2, 0)^T$ is a scaled stationary point, but not a Clarke stationary point of f, since $0 = \text{diag}(\bar{x})(2(\bar{x}_1 - 1) + 1, 4(\bar{x}_1 - 1)) + \tau))^T$ for any $\tau \in [-1,1]$ but $0 \notin \partial f(\bar{x})$. The vector $x^* = (0, 3/8)^T$ is both a scaled stationary point and a Clarke stationary point of f with $0 = \text{diag}(x^*)(2(2x_2^* - 1) + \tau, 4(2x_2^* - 1) + 1))^T$ for any $\tau \in [-1,1]$ and $0 \in \partial f(x^*)$.

Hence the complexity order of the SQR algorithm for obtaining an ϵ scaled stationary point is better than the SQR_1 algorithm for obtaining an ϵ Clarke stationary point.

From the definition of $\tilde{f}$ and the analysis in [10], for any fixed $x \in \mathbb{R}^n$, it follows that

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

PROPOSITION 4.2. *For any $x \in \mathbb{R}^n$ and $\mu > 0$,*

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

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

Proof. See Appendix B. □

Similar to the proof of Lemma 3.7, we have the following estimate.

LEMMA 4.3. *For all $k \in N_s$, if $K^+(x^k, \mu_k, \beta_k) \bigcup J^+(x^k, \mu_k, \beta_k) \neq \emptyset$, then*

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) < -4\alpha\mu_0^{-2}\mu_k^3;$$

otherwise, there exists $\xi^k \in \partial f(z^k)$ such that

$$\|\xi^k\|_\infty < C_1\mu_k,$$

where

$$C_1 = 2\bar{\beta} \max\left\{\sqrt{\frac{R}{\mu_0}}, 1\right\} + 8\alpha \max\left\{\mu_0^{-1}, \mu_0^{-\frac{1}{2}}\right\} + \alpha,$$

with $\bar{\beta}$ defined in Lemma 3.3 and $R \geq 1$ such that $R \geq \|x^k\|_\infty$ for all $k \in \mathbb{N}$.

Moreover, from Lemma 4.3, all the results in Lemma 3.8 hold for the SQR_1 algorithm. Similarly, we also assume the initial point is not an ϵ Clarke stationary point of (1.1) for a given $\epsilon \in (0, 1]$ without loss of generality and suppose $\mu_0 = 1$ for the sake of simplicity in the following complexity analysis of the SQR_1 algorithm.

THEOREM 4.4. *Any accumulation point of $\{z^k\}$ generated by the SQR_1 algorithm is a Clarke stationary point of (1.1) defined in Definition 4.1. Given any $\epsilon \in (0, 1]$, the total number of successful iterations for obtaining an ϵ Clarke stationary point of (1.1) when applying the proposed SQR_1 algorithm is at most*

$$\lceil J_s \epsilon^{-3} \rceil$$

and the total number of iterations is at most

$$\lceil J \epsilon^{-3} \rceil$$

where

$$J_s = \frac{(2\bar{\beta}R + 9\alpha)^3 \tilde{f}(x^0, \mu_0)}{4\alpha\sigma^3} - \log_\sigma(2e\bar{\beta}R + 9e\alpha) + 1 \quad \textit{and} \quad J = J_s + \log_\eta \bar{\beta} - \log_\eta \beta_0 + 1$$

with $\bar{\beta}$ defined in Lemma 3.3 and $R \geq 1$ such that $R \geq \|x^k\|_\infty$ for all $k \in \mathbb{N}$.

Proof. For a given $\epsilon \in (0, 1]$, there exists a positive integer $j \geq 2$ such that

$$(2\bar{\beta}R + 9\alpha)\sigma^{(j-1)} \leq \epsilon \quad \text{and} \quad (2\bar{\beta}R + 9\alpha)\sigma^{(j-2)} > \epsilon. \tag{4.3}$$

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

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) \geq -4\alpha\mu_k^3. \tag{4.4}$$

If (4.4) holds, from Lemma 4.3, then $K^+(x^{N_j^-}, \mu_{N_j^-}, \beta_{N_j^-}) \bigcup J^+(x^{N_j^-}, \mu_{N_j^-}, \beta_{N_j^-}) = \emptyset$, which follows that there is $\xi^{N_j^-} \in \partial f(z^{N_j^-})$ such that

$$\|\xi^{N_j^-}\|_\infty < (2\bar{\beta}R + 9\alpha)\mu_{N_j^-}.$$

Combining the above inequality and the non-increasing property of μ_k, we have

$$\min\left\{\|\xi\|_\infty : \xi \in \partial f(z^k)\right\} < (2\bar{\beta}R + 9\alpha)\mu_{N_j^-} \leq \epsilon, \quad \forall k \geq N_j^- + 1. \tag{4.5}$$

Similar to the proof of Theorem 3.9, we only need to evaluate N_j^-. From Lemma 4.3, when $k \in N_s \backslash N^-$ and $k \leq N_j^-$,

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

There are at least $k_j - j + 1$ iterations such that the above inequality holds, where k_j is the number of successful iterations up to N_j^-. Hence we obtain

$$0 \leq \tilde{f}(x^{N_j^-}, \mu_{N_j^-}) < \tilde{f}(x^0, \mu_0) - \frac{4\alpha\sigma^3\epsilon^3}{(2\bar{\beta}R + 9\alpha)^3}(k_j - j + 1).$$

Thus,

$$k_j < \frac{(2\bar{\beta}R + 9\alpha)^3}{4\alpha\sigma^3}\tilde{f}(x^0, \mu_0)\epsilon^{-3} + j - 1.$$

Moreover, (4.3) gives $j \leq \log_\sigma \epsilon - \log_\sigma(2\bar{\beta}R + 9\alpha) + 2$. Hence, we have

$$k_j < J_s\epsilon^{-3},$$

where

$$J_s = \frac{(2\bar{\beta}R + 9\alpha)^3}{4\alpha\sigma^3}\tilde{f}(x^0, \mu_0) - \log_\sigma(2e\bar{\beta}R + 9e\alpha) + 1.$$

Coming back to (4.5) and (3.41), we can obtain the estimation of interations in this theorem and the worst-case complexity of the SQR_1 algorithm for finding a Clarke stationary point of (1.1). Let $\epsilon \to 0$, then any accumulation point of z^k is a Clarke stationary point of (1.1). □

Similar to the proof idea of Theorem 4.4, the SQR_1 algorithm takes at most $O(\epsilon^{-3})$ iterations to reduce $\|\nabla\tilde{f}(x, \mu)\|_\infty \leq \epsilon$ and $\mu \leq \epsilon$, while the DS algorithm in [19] needs to take at most $O(\epsilon^{-3}\log\epsilon^{-1})$ iterations.

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

$$\min\{\|\xi\|_\infty, \xi \in \partial f(z^k)\} \leq \epsilon \tag{4.6}$$

directly. However, from the proof analysis in Proposition 4.2, if

$$\|\nabla H(z^{k^*}) + \sum_{i=1}^{n} \nabla_x\varphi(\theta^p(x_i^{k^*}, \mu_{k^*}))\|_\infty + \alpha\mu_{k^*} \leq \epsilon, \tag{4.7}$$

then there exists $\xi^{k^*} \in \partial f(z^{k^*})$ such that $\|\xi^{k^*}\|_\infty \leq \epsilon$. Hence, we can use (4.7) to verify (4.6).

Nesterov and Polyak [28] propose a Newton method based on a cubic regularization for solving a general unconstrained smooth nonconvex optimization, where the Hessian of the objective function is needed. Cartis, Gould, Toint [6] and Nesterov [27] propose quadratic regularization methods for solving nonsmooth nonconvex optimization problems (1.2) with the worst-case complexity $O(\epsilon^{-2})$. However, the objective function in (1.1) with φ_2, φ_3, φ_4, φ_5 and φ_6 in Appendix A cannot be reformed by (1.2). To the best of our knowledge, if the penalty φ is not convex, it is an open problem whether the globally Lipschitz continuity of the objective function in (1.1) can improve the worst-case complexity order of the SQR_1 algorithm for finding an ϵ Clarke stationary point of (1.1) defined in Definition 4.1.

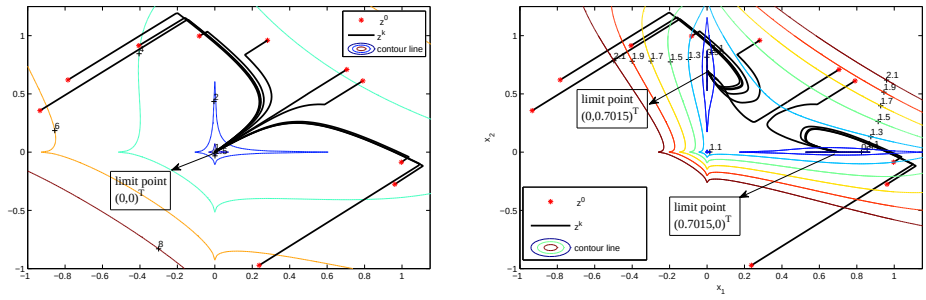
5. Numerical Experiments. In this section, we give two examples to show the performance of the SQR algorithm for solving (1.1) with $p = \frac{1}{2}$ and $p = 1$, respectively. The numerical testing is carried out on a Lenovo PC (3.00GHz, 2.00GB of RAM) with the use of Matlab 7.4. Throughout this section, we always use $\mu_0 = 10$, $\eta = 2$ and $\sigma = \sigma_1 = \sigma_2 = 0.9$. Example 5.1 is used to show that the SQR 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 SQR 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

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

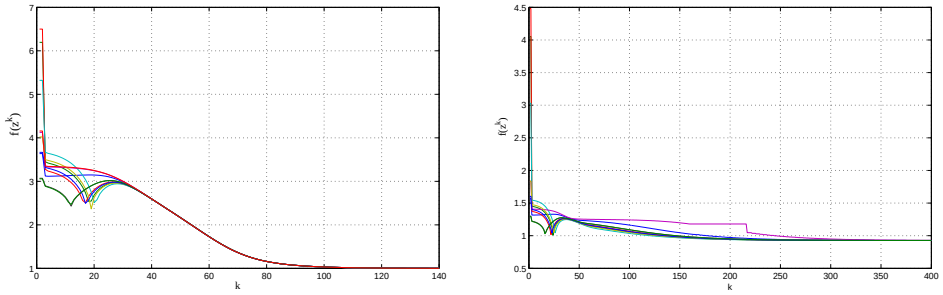
where $\lambda > 0$. This example is used to explain the optimality conditions in [11].

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 [13], 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 $(0.7015, 0)$. We choose $\beta_0 = 4$, then $\beta_k = 4$ for all $k \in \mathbb{N}$ and all iterations are successful. Figure 5.1 shows the tracks of z^k generated by the SQR 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 SQR algorithm for finding a global minimizer of (1.1) with $0 < p < 1$, even such problem is NP hard in general.



(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) $\lambda = 1$

EXAMPLE 5.2. We use randomly generated standard testing problems to show the validity of the SQR 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 $b \in \mathbb{R}^m$.

```
n = 100 * n0;  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 : n0),1) = 2 * randn(n0,1);  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 $\|v\|_0 = 10$. 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 A with $a = 3.7$ for φ_4, φ_5, φ_6, and $a = 1$ forφ_1, φ_2, φ_3. From $\varphi(0) = 0$, we have $f(z^0) = H(z^0)$ and $\|G(z^0)\|_\infty = 0$. Moreover, at these given data, $f(z^0) = 3.4939$ and $\|\tilde{g}(z^0,\mu_0)\|_\infty = 0.2071$. We set $\beta_0 = 1$, which is smaller than the Lipschitz constant of the gradient of $\ln(\|Ax-b\|_2^2+1)$.

When $\epsilon = 10^{-3}$, the numerical results using the SQR algorithm for solving (5.2) with $p = \frac{1}{2}$ and $p = 1$ are listed in Table 5.1, where k^* is the smallest integer such that $\mu_k \le \epsilon$ and $\|G(z^k)\|_\infty \le \epsilon$ for all $k \ge k^*$, $\|z^{k^*}\|_0$ is the number of nonzero elements of z^{k^*}.

φ	λ	α	p	k^*	$f(z^{k^*})$	μ_{k^*}	$\|z^{k^*}\|_0$	$\|z^{k^*}-v\|_2$
φ_1	0.3	0.3	0.5	1297	2.3177	3.43E-8	11	0.4609
			1	707	2.3030	1.94E-4	270	1.2412
φ_2	0.3	0.3	0.5	1448	1.9656	2.50E-8	11	0.4612
			1	725	1.8345	2.66E-4	268	1.2390
φ_3	0.3	0.6	0.5	2029	2.1356	2.02E-8	13	0.9542
			1	914	1.9970	3.59E-5	11	0.3552
φ_4	0.5	2	0.5	2015	2.8476	2.42E-10	12	1.1986
			1	826	2.4915	3.18E-6	8	1.2456
φ_5	0.3	0.4111	0.5	1981	1.9007	6.35E-9	11	0.3918
			1	797	1.6808	1.57E-4	304	1.0292
φ_6	0.3	0.3	0.5	1490	3.5033	2.83E-7	17	0.7625
			1	610	4.1913	9.40E-4	285	1.0463

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

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

6. Conclusions. Motivated by quadratic regularization methods [6, 22, 32] and smoothing methods [10, 19], we propose a globally convergent smoothing quadratic regularization (SQR) algorithm with the worst-case complexity $O(\epsilon^{-2})$ for finding ϵ scaled stationary points of a class of nonsmooth nonconvex, perhaps even non-Lipschitzian optimization problems in the form of (1.1). Such class of optimization problems include a number of interesting problems in sparse approximation, signal

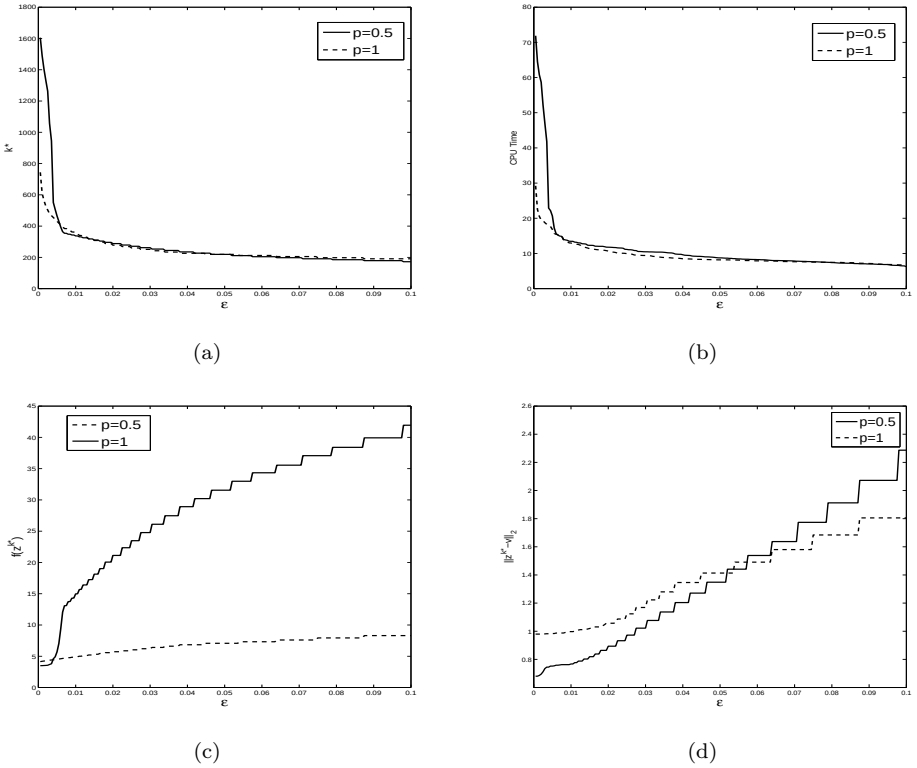


Fig. 5.3: SQR algorithm for (5.2) with φ_6 to find an ϵ scaled stationary point with different values of ϵ: (a) k^*, (b) CPU time, (c) $f(z^{k^*})$, (d) $\|z^{k^*} - v\|_2$

reconstruction, variable selection in recent literature. Moreover, if the objective function f in (1.1) is locally Lipschitz continuous, we can use the SQR_1 algorithm which is a slightly modified version of the SQR algorithm to find an ϵ Clarke stationary point with the worst-case complexity $O(\epsilon^{-3})$. The SQR algorithm and the SQR_1 algorithm are easy to implement, whose each iteration solves a strongly convex quadratic minimization problem with a simple closed-form solution.

7. Acknowledgement. We would like to thank Prof. Nicholas Gould and two anonymous 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] 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. Lean. Syst., 23 (2012), pp. 399-411.

[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] E. Candes and T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203-4215.

[5] E. Candes, M. Wakin and S. Boyd, Enhancing sparsity by reweighted l_1 minimization, J. Fourier Anal. Appl., 14 (2008), pp. 877-905.
[6] 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 (2011), pp. 1721-1739.
[7] C. Cartis, N.I.M. Gould and Ph.L. Toint, *Adaptive cubic regularization methods for unconstrained optimization, Part I: motivation, convergence and numerical results*, Math. Program., 127 (2011), pp. 245-295.
[8] C. Cartis, N.I.M. Gould and Ph.L. Toint,*Adaptive cubic regularization methods for unconstrained optimization, Part II: worst-case function- and derivative-evaluation complexity*, Math. Program., 130 (2011), pp. 295-319.
[9] R. Chartrand and V. Staneva, *Restricted isometry properties and nonconvex compressive sensing*, Inverse Probl., 24 (2008), pp. 1–14.
[10] X. Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Program., 134 (2012), pp. 71-99.
[11] X. Chen, D. Ge, Z. Wang and Y. Ye, *Complexity of unconstrained L_2-L_p minimization*, Math. Program., to appear.
[12] X. Chen, M. Ng and C. Zhang, *Nonconvex l_p regularization and box constrained model for image restoration*, IEEE Trans. Image Processing, 21 (2012), pp. 4709-4721.
[13] 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.
[14] 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.
[15] F.H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[16] D.L. Donoho, *Neighborly polytopes and sparse solutions of underdetermined linear equations*, Technical Report, Standford University, 2005.
[17] J. Fan, *Comments on 'Wavelets in stastics: a review' by A. Antoniadis*, Stat. Method. Appl., 6 (1997), pp. 131–138.
[18] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc., 96 (2001), pp. 1348–1360.
[19] R. Garmanjani and L.N. Vicente, *Smoothing and worst case complexity for direct-search methods in non-smooth optimization*, IMA J. Numer. Anal., to appear, (2012).
[20] D. Ge, X. Jiang and Y. Ye, *A note on the complexity of L_p minimization*, Math. Program., 21 (2011), pp. 1721-1739.
[21] J.-L. Goffin, *On the convergence rate of subgradient optimization methods*, Math. Program., 13 (1977), pp. 329–347.
[22] N.I.M. Gould and D.P. Robinson, *A second derivative SQP method: global convergence*, SIAM J. Optim., 20 (2010), pp. 2023–2048.
[23] 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.
[24] J. Huang, S. Ma, H. Xue and C. Zhang, *A group bridge approach for variable selection*, Biometrika, 96 (2009), pp. 339–355.
[25] R. Tibshirani, *Shrinkage and selection via the Lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[26] Yu. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.
[27] Yu. Nesterov, *Modified Gauss-Newton scheme with worst-case guarantees for global performance*, Optim. Method Softw., 22 (2007), pp. 413–431.
[28] Yu. Nesterov and B.T. Polyak, *Cubic regularization of Newton's method and its global performance*, Math. Program., 108 (2006), pp. 177–205.
[29] 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.
[30] R.T. Rockafellar and R.J-B Wets, *Variational Analysis*, Berlin: Springer, 1998.
[31] R.J. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[32] J.V. Vazirani, *Approximation Algorithms*, Springer, Berlin, 2003.
[33] 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 [23, 25]: $\varphi_1(s) = \lambda s$
- logistic penalty function [29]: $\varphi_2(s) = \lambda \log(1 + as)$
- fraction penalty function [14, 29]: $\varphi_3(s) = \lambda \frac{as}{1+as}$
- hard thresholding penalty function[17]: $\varphi_4(s) = \lambda^2 - (\lambda - s)_+^2$
- smoothly clipped absolute deviation (SCAD) penalty function[17]:

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

- minimax concave penalty (MCP) function [33]:

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

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

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

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

(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

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

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

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

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

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

(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 any $s \geq a\lambda$. $\varphi_5(s)$ is continuously differentiable on $(0, \infty)$ and

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

Hence, the SCAD penalty function is globally Lipschitz continuous 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

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

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

(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

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

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

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

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

Appendix B. Proofs of Propositions.

Proof of Proposition 2.3

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

$$\nabla_s \varphi(\theta^p(s,\mu)) = \varphi'(t)_{t=\theta^p(s,\mu)} \nabla_s \theta^p(s,\mu).$$

From the expression of $\theta^p(s,\mu)$, for any fixed $\mu > 0$, we have

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

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

Since φ' is locally Lipschitz continuous, $\nabla_s\varphi(\theta^p(s,\mu))$ is locally Lipschitz continuous with 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.

According to Rademacher's theorem, a locally Lipschitz continuous function is differentiable almost everywhere (in the sense of Lebesgue measure), i.e. the measure of $\mathbb{R}\setminus\mathcal{D}_\mu$ is 0 [15, Theorem 2.5.1].

First, we consider the case that $p = 1$. When $|s| > \mu$ and $s \in \mathcal{D}_\mu$, by Lemma 2.2 (i),

$$|\nabla_s^2\varphi(\theta^p(s,\mu))| = |\varphi''(t)_{t=\theta(s,\mu)}(\nabla_s\theta(s,\mu))^2| \le \alpha\theta^{-1}(s,\mu) = \alpha|s|^{-1}. \tag{B.1}$$

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

$$\begin{aligned}|\nabla_s^2\varphi(\theta^p(s,\mu))| =&|\varphi''(t)_{t=\theta(s,\mu)}(\nabla_s\theta(s,\mu))^2 + \varphi'(t)_{t=\theta(s,\mu)}\nabla_s^2\theta(s,\mu)|\\ =&|\varphi''(t)_{t=\theta(s,\mu)}(\frac{s}{\mu})^2 + \varphi'(t)_{t=\theta(s,\mu)}\frac{1}{\mu}|\\ \le&\alpha\theta^{-1}(s,\mu) + \alpha\mu^{-1} \le 3\alpha\mu^{-1},\end{aligned} \tag{B.2}$$

where the last inequality uses Lemma 2.2 (ii).

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 $\varphi(\theta^p(s,\mu))$ with respect to s is calculated by

$$\begin{aligned}\nabla_s^2\varphi(\theta^p(s,\mu)) =& p^2\varphi''(t)_{t=\theta^p(s,\mu)}\theta^{2p-2}(s,\mu)(\nabla_s\theta(s,\mu))^2\\ &+p\varphi'(t)_{t=\theta^p(s,\mu)}\theta^{p-1}(s,\mu)\nabla_s^2\theta(s,\mu)\\ &+p(p-1)\varphi'(t)_{t=\theta^p(s,\mu)}\theta^{p-2}(s,\mu)(\nabla_s\theta(s,\mu))^2.\end{aligned} \tag{B.3}$$

When $|s| \le \mu$, we have $|\nabla_s\theta(s,\mu)| \le \frac{|s|}{\mu} \le 1$. From assumption (A_φ), (B.3) and Lemma 2.2 (ii), we obtain that for $s \in \mathcal{D}_\mu$ with $|s| \le \mu$,

$$\begin{aligned}&|\nabla_s^2\varphi(\theta^p(s,\mu))|\\ \le&p^2|\varphi''(t)_{t=\theta^p(s,\mu)}\theta^p(s,\mu)|\theta^{p-2}(s,\mu)\frac{s^2}{\mu^2} + p|\varphi'(t)_{t=\theta^p(s,\mu)}|\theta^{p-1}(s,\mu)\frac{1}{\mu}\\ &+p(1-p)|\varphi'(t)_{t=\theta^p(s,\mu)}|\theta^{p-2}(s,\mu)\frac{s^2}{\mu^2}\\ \le&\alpha p^2\theta^{p-2}(s,\mu) + \alpha p\frac{\theta(s,\mu)}{\mu}\theta^{p-2}(s,\mu) + p(1-p)\alpha\theta^{p-2}(s,\mu)\\ \le&2\alpha p\theta^{p-2}(s,\mu) \le 8\alpha p\mu^{p-2}.\end{aligned} \tag{B.4}$$

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

$$\begin{aligned}|\nabla_s^2\varphi(\theta^p(s,\mu))| \le&p^2|\varphi''(t)_{t=|s|^p}|s|^p||s|^{p-2} + p(1-p)|\varphi'(t)_{t=|s|^p}||s|^{p-2}\\ \le&\alpha p^2|s|^{p-2} + \alpha p(1-p)|s|^{p-2} = \alpha p|s|^{p-2}.\end{aligned} \tag{B.5}$$

It is easy to see that (B.1), (B.2), (B.4) and (B.5) imply that for $s \in \mathcal{D}_\mu$ and $0 < p \le 1$, $|\nabla_s^2\varphi(\theta^p(s,\mu))| \le \kappa(s,\mu)$.

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

Proof of Proposition 3.1

For any fixed $\mu > 0$, when $|\hat{s}| \leq 2\mu$, by Proposition 2.3, $\partial_s^2 \varphi(\theta^p(s,\mu))$ can be uniformly bounded by $\kappa(\hat{s},\mu) = 8\alpha p \mu^{p-2}$. Thus, by Taylor's formula, (3.4) holds naturally in this case.

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

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

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

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

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

Proof of Proposition 3.6

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

$$x_i \tilde{g}_i(x,\mu) = x_i \nabla_{x_i} H(x) + p\varphi'(t)_{t=|x_i|^p}|x_i|^p = G_i(x).$$

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

$$\begin{aligned}&\left|x_i \tilde{g}_i(x,\mu) - x_i \nabla_{x_i} H(x)_i - p\varphi'(t)_{t=|x_i|^p}|x_i|^p\right| \\ =&p|\varphi'(t)_{t=\theta^p(x_i,\mu)}\theta^{p-1}(x_i,\mu)\frac{x_i^2}{\mu} - \varphi'(t)_{t=|x_i|^p}|x_i|^p| \\ \leq&2\alpha p\mu^p + \alpha p\mu^p = 3\alpha p\mu^p.\end{aligned}$$

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

$$\left|x_i \tilde{g}_i(x,\mu) - x_i \nabla_{x_i} H(x) - \varphi'(t)_{t=|x_i|}|x_i|\right| = |\varphi'(t)_{t=\theta(x_i,\mu)}\frac{x_i^2}{\mu} - \varphi'(t)_{t=|x_i|}|x_i|| \leq 2\alpha\mu.$$

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

$$\|X\tilde{g}(x,\mu) - G(x)\|_\infty = \max_{1\leq i\leq n}\left|x_i \tilde{g}_i(x,\mu) - x_i \nabla_{x_i} H(x) - p\varphi'(t)_{t=|x_i|^p}|x_i|^p\right|,$$

we complete the proof.

Proof of Proposition 4.2

Since ∇H is globally Lipschitz continuous with Lipschitz constant $L_{\nabla H}$, we have

$$\|\nabla H(x) - \nabla H(x_\mu)\|_\infty < L_{\nabla H}\mu. \tag{B.6}$$

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

$$\nabla_{x_i}\varphi(\theta(x_i,\mu)) = \begin{cases} \varphi'(t)_{t=\theta(x_i,\mu)}\dfrac{x_i}{\mu} & \text{if } i \in I^-(x,\mu) \\ \varphi'(t)_{t=|x_i|}\text{sign}(x_i) & \text{if } i \in I^+(x,\mu). \end{cases}$$

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

$$\left|\varphi'(t)_{t=\theta(x_i,\mu)}\frac{x_i}{\mu} - \varphi'(0)\frac{x_i}{\mu}\right| \le \left|\varphi'(t)_{t=\theta(x_i,\mu)} - \varphi'(0)\right| \le \alpha\,|\theta(x_i,\mu)| < \alpha\mu.$$

Since $\varphi'(0)\frac{x_i}{\mu} \in \varphi'(0)\cdot[-1,1] = \partial\varphi(|[x_\mu]_i|)$ 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 have

$$\min\left\{\|\sum_{i=1}^n \nabla_x\varphi(\theta(x_i,\mu)) - \eta\|_\infty : \eta \in \sum_{i=1}^n \partial_{x_\mu}\varphi(|[x_\mu]_i|)\right\} < \alpha\mu. \tag{B.7}$$

Combining (B.6) and (B.7), we complete the proof.