SMOOTHING QUADRATIC REGULARIZATION METHODS FOR BOX CONSTRAINED NON-LIPSCHITZ OPTIMIZATION IN IMAGE RESTORATION

WEI BIAN* AND XIAOJUN CHEN†

7 September 2014

Abstract. We propose a smoothing quadratic regularization (SQR) method for solving box constrained optimization problems with a non-Lipschitz regularization term that includes the l_p norm $(0 < p < 1)$ of the gradient of the underlying image in the l_2-l_p problem as a special case. At each iteration of the SQR algorithm, a new iterate is generated by solving a strongly convex quadratic problem with box constraints. We prove that any cluster point of ϵ scaled first order stationary points with $\epsilon > 0$ satisfies a first order necessary condition for a local minimizer as ϵ goes to 0, and the worst-case iteration complexity of the SQR algorithm for finding an ϵ scaled first order stationary point is $O(\epsilon^{-2})$. Numerical examples are given to show good performance of the SQR algorithm for image restoration.

Key words. Smoothing quadratic regularization method, image restoration, total-variation regularization, non-Lipschitz optimization, worst-case complexity,

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

1. Introduction.

In this paper, we consider the following minimization problem

$$\min_{l\le x\le u} \quad f(x) := \Theta(x) + \sum_{i=1}^{m} \varphi(|d_i^T x|^p), \tag{1.1}$$

where $\Theta : \mathbb{R}^n \to \mathbb{R}_+$ is continuously differentiable, $0 < p \le 1$, $D = (d_1, d_2, \dots, d_m)^T \in \mathbb{R}^{m\times n}$, $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous with $\varphi(0) = 0$ and $l \in \mathbb{R}^n \cup \{-\infty\}^n$, $u \in \mathbb{R}^n \cup \{+\infty\}^n$ with $l < u$. Problem (1.1) has many important applications in medical and astronomical image restoration [5, 10, 12, 17, 18, 28, 29, 30] and film restoration [22]. In (1.1), the first term measures how well the restored image is fitting the observed data under the imaging system, the second term induces special properties of the restored image, and the constraint can improve the restored image using a priori information. Using a nonconvex nonsmooth non-Lipschitz regularization function $\sum_{i=1}^m \varphi(|d_i^T x|^p)$ in the second term has remarkable advantages for the restoration of piecewise constant images [10, 17, 29]. Typical choices of D for the potential function include the identity matrix, first order difference operator, second order difference operator or some overcomplete dictionary [24].

The success of model (1.1) with $0 < p < 1$ in sparse optimization is related to the non-Lipschitz property of the objective function. Local minimizers of (1.1) with $0 < p < 1$ have various nice properties over the minimizers of it with $p = 1$. In image restoration, (i) it is shown in [6] that (1.1) with $0 < p < 1$ promotes a better gradient sparsity than it with $p = 1$; (ii) (1.1) with $0 < p < 1$ is also more robust with respect to noise; (iii) theoretical and numerical results show that local minimizers of (1.1) with $0 < p < 1$ have advantages in distinguishing zero and nonzero

*Department of Mathematics, Harbin Institute of Technology, Harbin, China. (bianweilvse520@163.com). The author's work was supported in part by the NSF foundation (11101107) of China.

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. (maxjchen@polyu.edu.hk). The author's work was supported in part by Hong Kong Research Grant Council grant (PolyU5001/12P).

entries of coefficients in sparse high-dimensional approximation [1, 6, 7, 19] and bring the restored image closed contours and neat edges [10, 17]. Moreover, in variable selection, the l_p potential function with $0 < p < 1$ owns the oracle property [15, 23] in statistics, while l_1 does not; the problem (1.1) with $0 < p < 1$ can be used for variable selection at the group and individual variable levels simultaneously, while (1.1) with $p = 1$ can only work for individual variable selection [20]. Thus, we focus on the minimization problem (1.1) with $0 < p < 1$ in this paper.

Theory and algorithms for some special cases of problem (1.1) with $0 < p < 1$ have been developed in the last few years. The lower bound theory [10, 11, 28, 30] ensures that each component of Dx at any local minimizer of problem (1.1) is either zero or not less than a positive constant, which implies the restored image closed contours and neat edges in the imaging system. For the unconstrained version of problem (1.1), the reweighted algorithms [25, 26] is proved to be globally convergent. Moreover, the R-regularized Newton scheme is superlinearly convergent [18], and the trust region Newton method converges to a scaled second order stationary point [9]. For problem (1.1) with D being the identity matrix, the smoothing quadratic regularization (SQR) algorithm for (1.1) without constraints and the interior point algorithm for (1.1) with the nonnegative constraint converge to an ϵ scaled first order stationary point with worst-case iteration complexity $O(\epsilon^{-2})$ [2, 3]. However, to the best of our knowledge, for problem (1.1) with an arbitrary matrix D in the regularization term and arbitrary vectors l and u in the constraint, there is no algorithm which can always find an ϵ scaled first order stationary point in no more than $O(\epsilon^{-2})$ iterations.

The use of D and l, u in (1.1) brings problem (1.1) many advantages in image restoration and reconstruction [9, 10, 21, 28, 29, 30, 31]. However, because the non-Lipschitz potential function in (1.1) is neither separable with respect to components of x nor concave in the feasible set, (1.1) is harder to solve than the problem with D being the identity matrix and $l = 0$. Algorithms in [2, 3, 9, 18, 25, 26, 30] cannot be directly extended to solve (1.1), and the lower bound theory in [10, 11, 28, 30] and the definitions of the ϵ scaled stationary point in [2, 3] are invalid. Thus, (1.1) gives many challenging problems in developing effective algorithms with nice convergence theorems and computational complexity bounds.

In this paper, we generalize the subspace idea for unconstrained optimization in [9] and define the first order necessary optimality condition and an ϵ scaled first order stationary point of problem (1.1). We prove that any cluster point of ϵ scaled first order stationary points satisfies the first order necessary optimality condition as ϵ tending to 0. Moreover, a new SQR algorithm with the worst-case complexity $O(\epsilon^{-2})$ is given for solving (1.1). The updating techniques in the algorithm for approximating the Lipschitz constant of the gradient of Θ is adopted from [4]. Though the framework of the SQR algorithm in this paper is adopted from [2], the construction of the quadratic subproblem and the updating scheme of the smoothing parameter are entirely different, since the SQR algorithm in [2] can only be applied to the unconstrained problem with $\varphi(|d_i^T x|^p) = \varphi(|x_i|^p)$, and its convergence and complexity analysis uses the separability of variables in the term $\sum_{i=1}^n \varphi(|x_i|^p)$ without considering the feasibility of iterates.

The rest of this paper is organized as follows. In section 2, using a smoothing function of the objective function f in (1.1), we present an SQR algorithm for solving (1.1). In section 3, we derive a first order necessary condition for a local minimizer of (1.1), and define an ϵ scaled first order stationary point of (1.1). We prove that any cluster point of ϵ scaled first order stationary points of (1.1) satisfies the first order necessary condition as ϵ tends to 0. In section 4, we show that the worst-case complexity of the SQR algorithm for finding an ϵ scaled first order stationary point of (1.1) is $O(\epsilon^{-2})$. In section 5, we report numerical experiments with one randomly generated test example and two image restoration problems to validate the theoretical results and show good performance of the proposed SQR algorithm.

Notations: Denote $\mathbb{K} = \{0, 1, 2, \ldots\}$ and $e = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$. For $x, l, u \in \mathbb{R}^n$, let $\|x\| := \|x\|_2$, $x^2 = (x_1^2, \ldots, x_n^2)^T$ and $R = \max_{l \le x \le u} \|x\|$. For a matrix $A \in \mathbb{R}^{r \times n}$ and an index set $J \subseteq \{1, \ldots, r\}$, A_J denotes the submatrix of A whose rows are indexed by J. For a subspace $\mathbb{S} \subseteq \mathbb{R}^n$, orthon($\mathbb{S}$) is a subset of $\mathbb{R}^{n \times \dim(S)}$, in which the columns of each matrix form an orthonormal basis of $\mathbb{S}$ if dim($\mathbb{S}$)≥ 1 and orthon($\mathbb{S}$) $= 0_{n\times 1}$ if $\mathbb{S} = \{0\}$.

2. Smoothing quadratic regularization method. In this section, we present an SQR method for solving (1.1). The non-Lipschitz points of (1.1) are $\{x : d_i^T x = 0, \text{ for some } i \in \{1, \ldots, m\}\}$. Throughout this paper, we need the following assumptions on Θ and φ:

- $\Theta : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient $\nabla\Theta$ is globally Lipschitz with a Lipschitz constant $\hat{\beta}$;
- φ is differentiable and concave in $(0, \infty)$, φ' is locally Lipschitz continuous and there is a positive constant α such that for all $t \in (0, \infty)$,

$$0 \le \varphi'(t) \le \alpha \quad \text{and} \quad |\xi| \le \alpha \quad \forall \xi \in \partial(\varphi'(t)), \tag{2.1}$$

  where ∂ means the Clarke generalized gradient [13].

Many data fitting functions and penalty functions in sparse image restoration and reconstruction satisfy these conditions [10, 12, 17, 28, 29, 30]

2.1. Smoothing approximation. We define a smoothing function of the objective function f in (1.1). Using the gradient of the smoothing function we can construct a quadratic approximation function of f [8, 27, 32].

DEFINITION 2.1. *[8] 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 $\tilde{h}(\cdot, \mu)$ is continuously differentiable for any fixed $\mu > 0$ and $\lim_{z \to x, \mu \downarrow 0} \tilde{h}(z, \mu) = h(x)$ holds for any $x \in \mathbb{R}^n$.*

According to the assumptions on Θ and φ, we can define a smoothing function of f by using a smoothing function of the absolute value function. In particular, we define a smoothing function of f as the following

$$\tilde{f}(x, \mu) = \Theta(x) + \sum_{i=1}^m \varphi(\theta^p(d_i^T x, \mu)) \quad \text{with} \quad \theta(t, \mu) = \begin{cases} |t| & \text{if } |t| > \mu \\ \dfrac{t^2}{2\mu} + \dfrac{\mu}{2} & \text{if } |t| \le \mu. \end{cases}$$

The function $\theta(t, \mu)$ is a smoothing function of $|t|$, nondecreasing with respect to μ, and

$$0 = \arg\min_{t \in \mathbb{R}} |t| = \arg\min_{t \in \mathbb{R}} \theta(t, \mu) = \arg\min_{t \in \mathbb{R}} \varphi(\theta^p(t, \mu)), \quad \forall \mu \in (0, \infty).$$

Moreover, we have

$$|\nabla_t \theta^p(t, \mu)| \le p\theta^{p-1}(t, \mu), \ |\nabla_t^2 \theta^p(t, \mu)| \le p\theta^{p-2}(t, \mu) \text{ when } |t| \ne \mu. \tag{2.2}$$

Specially, when $|t| < \mu$, $\nabla_t^2 \theta^p(t, \mu) > 0$, which means that $\theta^p(t, \mu)$ is a convex smoothing function of $|t|^p$ in $(-\mu, \mu)$.

Since $\varphi'(t) \geq 0$ for all $t \in (0, \infty)$, $\tilde{f}(x, \mu)$ is nondecreasing with respect to μ for any fixed $x \in \mathbb{R}^n$. Denote

$$\Phi(s) := \varphi(|s|^p), \quad \tilde{\Phi}(s, \mu) := \varphi(\theta^p(s, \mu)), \tag{2.3a}$$

$$g(x, \mu) := \nabla_x \tilde{f}(x, \mu) = \nabla\Theta(x) + \sum_{i=1}^{m} \tilde{\Phi}'(d_i^T x, \mu) d_i, \tag{2.3b}$$

where $\tilde{\Phi}'(s, \mu) := \nabla_s \varphi(\theta^p(s, \mu))$. From $0 \leq \theta^p(t, \mu) - |t|^p \leq \theta^p(0, \mu) \leq (\frac{\mu}{2})^p$ and (2.1), we have

$$0 \leq \varphi(\theta^p(t, \mu)) - \varphi(|t|^p) \leq \alpha(\frac{\mu}{2})^p,$$

which gives

$$0 \leq \tilde{f}(x, \mu) - f(x) \leq \sum_{|d_i^T x| < \mu} \alpha(\frac{\mu}{2})^p \quad \forall x \in \mathbb{R}^n,\ \mu \in (0, \infty). \tag{2.4}$$

When $|t| \neq \mu$ and φ is twice continuously differentiable at t, we have $\varphi(\theta^p(t, \mu))$ is twice continuously differentiable at t and

$$\nabla_t^2 \varphi(\theta^p(t, \mu)) = \varphi''(s)_{s=\theta^p(t,\mu)} (\nabla_t \theta^p(t, \mu))^2 + \varphi'(s)_{s=\theta^p(t,\mu)} \nabla_t^2 \theta^p(t, \mu).$$

Thus, by (2.2) and the assumptions on φ, the following estimation on the elements in $\partial_t(\nabla_t \varphi(\theta^p(t, \mu)))$ holds

$$\max\{|\xi| : \xi \in \partial_t(\nabla_t \varphi(\theta^p(t, \mu)))\} \leq 8\alpha p \mu^{p-2} \quad \forall t \in \mathbb{R},\ \mu \in (0, 1].$$

Inspired by Talyer's expansion, for any $x^+, x \in \mathbb{R}^n$,

$$\begin{aligned} &\varphi(\theta^p(d_i^T x^+, \mu)) - \varphi(\theta^p(d_i^T x, \mu)) \\ \leq &\nabla_t \varphi(\theta^p(t, \mu))_{t=d_i^T x} d_i^T (x^+ - x) + 4\alpha p \mu^{p-2} (d_i^T (x^+ - x))^2. \end{aligned} \tag{2.5}$$

Notice that $\varphi(|t|^p)$ is concave on $\mathbb{R}_+$ and $\mathbb{R}_-$, respectively. We have

$$\varphi(|\hat{t}|^p) \leq \varphi(|t|^p) + \nabla\varphi(|t|^p)(\hat{t} - t)$$

for any $\hat{t}, t \in \mathbb{R}$ such that $\hat{t}t > 0$. Hence, for any $x, x^+ \in \mathbb{R}^n$ satisfying

$$(d_i^T x^+)(d_i^T x) > 0,\ |d_i^T x^+| \geq \mu \text{ and } |d_i^T x| \geq \mu,$$

by $\theta(s, \mu) = |s|$ when $|s| \geq \mu$, we have

$$\varphi(\theta^p(d_i^T x^+, \mu)) \leq \varphi(\theta^p(d_i^T x, \mu)) + \nabla_t \varphi(\theta^p(t, \mu))_{t=d_i^T x} d_i^T (x^+ - x). \tag{2.6}$$

LEMMA 2.2. *For $x \in [l, u]$, $s \in \mathbb{R}^n$ and $\mu \in (0, 1]$, if $l \leq x + s \leq u$, $s^2 \leq \|D\|_\infty^{-2} \mu^{2p} e$ and the following inequality holds*

$$\Theta(x + s) \leq \Theta(x) + \langle \nabla\Theta(x), s \rangle + \frac{\beta}{2} \|s\|^2 \tag{2.7}$$

with $\beta > 0$, then

$$\tilde{f}(x + s, \mu) - \tilde{f}(x, \mu) \leq \langle g(x, \mu), s \rangle + \frac{\beta}{2} \|s\|^2 + 4\alpha p \mu^{p-2} \sum_{|d_i^T x| \leq 2\mu^p} (d_i^T s)^2. \tag{2.8}$$

Proof. From $s^2 \le \|D\|_\infty^{-2}\mu^{2p}e$, we have

$$|d_i^T s| \le \|Ds\|_\infty \le \|D\|_\infty \|s\|_\infty \le \mu^p \quad i = 1, 2, \ldots, m.$$

Then, $|d_i^T x| > 2\mu^p$ implies $|d_i^T(x+s)| > \mu^p$ and $(d_i^T x)(d_i^T(x+s)) > 0$ for $i = 1, 2, \ldots, m$, which together with (2.5) and (2.6) gives

$$\begin{aligned}
&\sum_{i=1}^m \varphi(\theta^p(d_i^T(x+s), \mu)) - \sum_{i=1}^m \varphi(\theta^p(d_i^T x, \mu)) \\
\le& \langle \sum_{i=1}^m \nabla_t \varphi(\theta^p(t,\mu))_{t=d_i^T x}, d_i^T s\rangle + 4\alpha p \mu^{p-2} \sum_{|d_i^T x| \le 2\mu^p} (d_i^T s)^2.
\end{aligned}$$

Thus, we obtain (2.8) from (2.7). □

For $x \in [l, u]$ and $\mu \in (0, 1]$, to achieve a potential reduction, we solve the following strongly convex quadratic program in $\mathbb{R}^n$:

$$\begin{aligned}
\min \quad & \langle g(x,\mu), s\rangle + \frac{\beta}{2}\|s\|^2 + 4\alpha p \mu^{p-2} \sum_{|d_i^T x| \le 2\mu^p} (d_i^T s)^2 \\
\text{s.t.} \quad & s^2 \le \delta^2 \mu^{2p} e, \; l - x \le s \le u - x,
\end{aligned} \tag{2.9}$$

where $\delta = \|D\|_\infty^{-1}$ and $g(x, \mu)$ is defined in (2.3b).

SQR Algorithm.

Step 0: Initialization: Choose $x^0 \in [l, u]$, $0 < \mu_0 \le 1$, $\beta_0 \ge 1$, $0 < \sigma < 1$ and $\eta > 1$. Set $k = 0$.

Step 1: New point calculation: Solve (2.9) with $x = x^k$, $\mu = \mu_k$ and $\beta = \beta_k$ for s^k, and let $y^k = x^k + s^k$.

Step 2: Updating the regularization weight: If

$$\Theta(y^k) - \Theta(x^k) > \langle \nabla\Theta(x^k), y^k - x^k\rangle + \frac{1}{2}\beta_k \|x^k - y^k\|^2,$$

let

$$\beta_{k+1} = \eta\beta_k, \; x^{k+1} = x^k, \; \mu_{k+1} = \mu_k$$

and return to Step 1; otherwise, let

$$\beta_{k+1} = \beta_k, \; x^{k+1} = y^k$$

and go to Step 3.

Step 3: Updating the smoothing parameter: Let

$$\mu_{k+1} = \begin{cases} \mu_k & \text{if } \tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) < -\mu_k^{2p} \\ \sigma\mu_k & \text{otherwise.} \end{cases} \tag{2.10}$$

Step 4: Constructing convergence sequence: Let

$$z^{k+1} = \begin{cases} x^k & \text{if } \mu_{k+1} = \sigma\mu_k \\ z^k & \text{otherwise.} \end{cases} \tag{2.11}$$

Increment k by one and return to Step 1.

The proposed SQR algorithm is for the constrained optimization problem (1.1) with an arbitrary D in the potential function φ, while the SQR method in [2] is for unconstrained optimization problems with D being the identity matrix in the potential function φ. Hence, the two SQR algorithms are entirely different. Specially, the quadratic programs in the two algorithms have essential distinctions. In [2], the quadratic program is an unconstrained problem and can be split into n one dimensional problems to get a simple closed form solution. In this paper, the quadratic program is a constrained problem and cannot have a closed form solution. The convergence and complexity analysis for the SQR algorithm in this paper is more comprehensive due to the existence of an arbitrary D in the non-Lipschitz potential function and l, u in the constraints.

From Lemma 2.2, for $x^k \in [l,u]$ and $\mu_k \in (0,1]$, by $l - x^k \le s_k \le u - x^k$ and $\sigma \in (0,1)$, we have $x^{k+1} \in [l,u]$ and $\mu_{k+1} \in (0,1]$. Then, the proposed SQR algorithm is well defined, and $x^k \in [l,u]$, $\mu_k \in (0,1]$ for all $k \in \mathbb{K}$.

Let $\{x^k\}$, $\{y^k\}$, $\{z^k\}$, $\{\mu_k\}$ and $\{\beta_k\}$ be the sequences generated by the SQR algorithm. Denote

$$N_s = \{k \in \mathbb{K} : \beta_{k+1} = \beta_k\}, \quad \mathcal{T} = \{k \in \mathbb{K} : \ \mu_{k+1} = \sigma\mu_k\}. \tag{2.12}$$

We call the kth iteration is successful if $k \in N_s$. Note that $\mathcal{T} \subseteq N_s$.

Lemma 2.3. *The sequence $\{\tilde{f}(x^k, \mu_k)\}$ is non-increasing. Moreover, when $k \in N_s$, there are nonnegative vectors ν^k, γ^k and ρ_k such that*

$$\begin{aligned}\tilde{f}(x^{k+1}, \mu^{k+1}) - \tilde{f}(x^k, \mu^k) \le & -\frac{\beta_k}{2}\|s^k\|^2 - 4\alpha p \mu_k^{p-2} \sum_{|d_i^T x^k| \le 2\mu_k^p} (d_i^T s^k)^2 \\ & - \delta\mu_k e^T \nu^k - (x^k - l)^T \gamma^k - (u - x^k)^T \rho^k.\end{aligned}$$

Proof. From the KKT condition of (2.9), the solution s^k of the strongly convex quadratic program (2.9) satisfies

$$g(x^k, \mu_k) + \beta_k s^k + 8\alpha p \mu_k^{p-2} \sum_{|d_i^T x^k| \le 2\mu_k^p} (d_i^T s^k) d_i + M_k s^k - \gamma^k + \rho^k = 0, \tag{2.13a}$$

$$(s^k)^2 \le \delta^2 \mu_k^{2p} e, \ l - x^k \le s^k \le u - x^k, \tag{2.13b}$$

$$M_k((s^k)^2 - \delta^2\mu_k^{2p} e) = 0, \ (s^k + x^k - l)^T \gamma^k = 0, \ (s^k - u + x^k)^T \rho^k = 0, \tag{2.13c}$$

where $M_k = \mathrm{diag}(\nu^k)$ with ν^k, γ^k, $\rho^k \ge 0$.

On the one hand, when $k \in N_s$, $x^{k+1} = y^k$, from Lemma 2.2, (2.13a) and (2.13c), we have

$$\begin{aligned}&\tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) \\ \le& \langle g(x^k, \mu_k), s^k\rangle + \frac{\beta_k}{2}\|s^k\|^2 + 4\alpha p \mu^{p-2} \sum_{|d_i^T x^k| \le 2\mu_k^p} |d_i^T s^k|^2 \\ =& -\frac{\beta_k}{2}\|s^k\|^2 - 4\alpha p \mu_k^{p-2} \sum_{|d_i^T x^k| \le 2\mu_k^p} (d_i^T s^k)^2 - (s^k)^T M_k s^k + (\gamma^k)^T s^k - (\rho^k)^T s^k \\ =& -\frac{\beta_k}{2}\|s^k\|^2 - 4\alpha p \mu_k^{p-2} \sum_{|d_i^T x^k| \le 2\mu_k^p} (d_i^T s^k)^2 - \delta^2 \mu_k^{2p} e^T \nu^k - (x^k - l)^T \gamma^k - (u - x^k)^T \rho^k \\ \le& 0.\end{aligned} \tag{2.14}$$

Since $\mu_{k+1} \le \mu_k$ and $\tilde{f}(x, \mu)$ is non-decreasing about μ for any fixed $x \in \mathbb{R}^n$, the results in this lemma holds for $k \in N_s$.

On the other hand, when $k \notin N_s$, $x^{k+1} = x^k$ and $\mu_{k+1} = \mu_k$, which implies $\tilde{f}(x^{k+1}, \mu_{k+1}) = \tilde{f}(x^k, \mu_k)$. Thus, $\tilde{f}(x^k, \mu_k)$ is nonincreasing for all $k \in \mathbb{K}$. □

LEMMA 2.4. *For all $k \in N_s$, if*

$$\tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) > -\mu_k^{2p}, \tag{2.15}$$

then x^k satisfies

$$\begin{aligned} &\left\| Z_k^T(\nabla\Theta(x^k) + \sum_{|d_i^T x^k| > 2\mu_k^p} \Phi'(d_i^T x^k) d_i - \gamma^k + \rho^k) \right\|_\infty \le (\sqrt{2n\beta_k} + \delta^{-1})\mu_k^p, \\ &(x^k - l)^T \gamma^k \le \mu_k^{2p}, \quad (u - x^k)^T \rho^k \le \mu_k^{2p}, \\ &l \le x^k \le u, \ \gamma^k \ge 0, \ \rho^k \ge 0, \end{aligned} \tag{2.16}$$

*where $Z_k \in$*orthon$((\text{span}\{d_i : |d_i^T x^k| \le 2\mu_k^p\})^\perp)$.

Proof. If one of the following five inequalities fails

$$\frac{\beta_k}{2}\|s^k\|^2 < \mu_k^{2p}, \quad 4\alpha p \mu_k^{p-2} \sum_{|d_i^T x^k| \le 2\mu_k^p} |d_i^T s^k|^2 < \mu_k^{2p}, \tag{2.17a}$$

$$\delta^2 \mu_k^{2p} e^T \nu^k < \mu_k^{2p}, \quad (x^k - l)^T \gamma^k < \mu_k^{2p} \quad \text{and} \quad (u - x^k)^T \rho^k < \mu_k^{2p}, \tag{2.17b}$$

by (2.14), we obtain $\tilde{f}(x^{k+1}, \mu_k) - \tilde{f}(x^k, \mu_k) \le -\mu_k^{2p}$. Hence, (2.15) implies all inequalities in (2.17) hold and then we only need to prove the estimation in (2.16) under the conditions in (2.17).

First, from $l \le x^k \le u$, $\gamma^k \ge 0$, $\rho^k \ge 0$, the second and third inequalities in (2.17b) give

$$0 \le (x^k - l)^T \gamma^k \le \mu_k^{2p}, \quad 0 \le (u - x^k)^T \rho^k \le \mu_k^{2p}. \tag{2.18}$$

The first inequality in (2.17b) gives $\|\nu^k\|_1 \le \delta^{-2}$. Then we obtain

$$\|M_k s^k\|_1 = \sum_{i=1}^n |\nu_i^k s_i^k| \le \|\nu^k\|_1 \|s^k\|_\infty \le \delta^{-1}\mu_k^p. \tag{2.19}$$

Moreover, (2.13a) can be rewritten as

$$g(x^k, \mu_k) - \gamma^k + \rho^k = -\beta_k s^k - 8\alpha p \mu_k^{p-2} \sum_{|d_i^T x^k| \le 2\mu_k^p} (d_i^T s^k) d_i - M_k s^k. \tag{2.20}$$

Multiplying Z_k^T to the both sides of (2.20), from the properties of Z_k, we have

$$Z_k^T(g(x^k, \mu_k) - \gamma^k + \rho^k) = Z_k^T(\nabla\Theta(x^k) + \sum_{|d_i^T x^k| > 2\mu_k^p} \Phi'(d_i^T x^k) d_i - \gamma^k + \rho^k), \tag{2.21a}$$

$$Z_k^T(-\beta_k s^k - 8\alpha p \mu_k^{p-2} \sum_{|d_i^T x^k| \le 2\mu_k^p} (d_i^T s^k) d_i - M_k s^k) = -\beta_k Z_k^T s^k - Z_k^T M_k s^k. \tag{2.21b}$$

Then, from (2.20) and (2.21), we obtain

$$
\begin{aligned}
&\left\| Z_k^T(\nabla\Theta(x^k) + \sum_{|d_i^T x^k|>2\mu_k^p} \Phi'(d_i^T x^k)d_i - \gamma^k + \rho^k) \right\|_\infty \\
=&\left\|\beta_k Z_k^T s^k + Z_k^T M_k s^k\right\|_\infty \\
\le&\beta_k\|Z_k^T s^k\|_\infty + \|Z_k^T M_k s^k\|_\infty \\
\le&\beta_k\|s^k\|_1 + \|M_k s^k\|_1 \le (\sqrt{2n\beta_k} + \delta^{-1})\mu_k^p,
\end{aligned}
\tag{2.22}
$$

where we use that the columns of Z_k are orthonormal. By (2.18) and (2.22), we obtain the results in this lemma. ∎

The following lemma presents some properties of the sequences $\{\beta_k\}$, $\{\mu_k\}$ and $\{f(x^k)\}$.

LEMMA 2.5. *The following statements hold.*

(i) $\beta_k \le \bar{\beta} := \max\{\beta_0, \eta\hat{\beta}\}$ *for all* $k \in \mathbb{K}$*;*

(ii) $\lim_{k\to\infty} \mu_k = 0$*;*

(iii) $\lim_{k\to\infty} f(x^k)$ *exists.*

Proof. By Step 2 in the SQR algorithm, β_k is updated when $\beta_k \le \hat{\beta}$, then statement (i) can be easily proved by the assumption on Θ.

From (2.12), we have

$$
\sum_{k\in\mathcal{T}} \mu_k^{2p} < \sum_{k=1}^{\infty} \mu_0^{2p}\sigma^{2p(k-1)} = \frac{\mu_0^{2p}}{1-\sigma^{2p}}. \tag{2.23}
$$

Note that when $k \in N_s\backslash\mathcal{T}$, from (2.10), we have

$$
\mu_k^{2p} < \tilde{f}(x^k, \mu_k) - \tilde{f}(x^{k+1}, \mu_{k+1}) \quad \text{and} \quad \mu_{k+1} = \mu_k.
$$

This, together with the nonincreasing property of $\tilde{f}(x^k, \mu_k)$ and (2.4), gives

$$
\sum_{k\in N_s\backslash\mathcal{T}} \mu_k^{2p} < \sum_{k\in N_s\backslash\mathcal{T}} (\tilde{f}(x^k, \mu_k) - \tilde{f}(x^{k+1}, \mu_{k+1})) \le \tilde{f}(x^0, \mu_0) - \min_{x\in[l,u]} f(x). \tag{2.24}
$$

Adding (2.23) and (2.24), we have

$$
\sum_{k\in N_s} \mu_k^{2p} < \tilde{f}(x^0, \mu_0) - \min_{x\in[l,u]} f(x) + \frac{\mu_0^{2p}}{1-\sigma^{2p}}. \tag{2.25}
$$

If there are finite elements in N_s, then there is $\bar{k} \in \mathbb{K}$ such that $k \notin N_s$ $\forall k \ge \bar{k}$, which implies that $\beta_k \ge \beta_0\eta^{k-\bar{k}}$ $\forall k \ge \bar{k}$. By $\eta > 1$, $\lim_{k\to\infty} \beta_k = \infty$, which leads a contradiction with the boundedness of $\{\beta_k\}$ given in (i). Thus, there are infinite elements in N_s, which together with (2.25) gives $\lim_{k\to\infty} \mu_k = 0$.

Since $\{\tilde{f}(x^k, \mu_k)\}$ is nonincreasing, by (2.4) and $x^k \in [l, u]$, $\lim_{k\to\infty} \tilde{f}(x^k, \mu_k)$ exists. By virtue of $\lim_{k\to\infty} \mu_k = 0$ and (2.4), we have

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

∎

3. First order necessary condition. The scaled first order and second order necessary conditions for unconstrained non-Lipschitz optimization have been studied in [9, 11]. For the constrained non-Lipschitz optimization (1.1) with $D = I_n$, the scaled first and second order stationary points are defined in [3, 16]. Inspired by the subspace idea, we first derive a first order necessary condition for local minimizers of constrained non-Lipschitz optimization (1.1), whereafter the scaled and ϵ scaled first order stationary points of (1.1) are defined. Note that the results established in this paper have no assumption on the matrix D.

First, we give some notations used in this section. for fixed $x \in \mathbb{R}^n$ and $\epsilon > 0$, denote

$$\mathcal{D}_x^\epsilon = \{d_i : i \in \{1, 2, \dots, m\},\ |d_i^T x| \leq \epsilon\}.$$

For simplicity, we denote $\mathcal{D}_x := \mathcal{D}_x^\epsilon$ when $\epsilon = 0$. Obviously, $\dim(\mathrm{span}\mathcal{D}_{\bar{x}}) = n$ implies $\bar{x} = 0$, and f is non-Lipschitz at $\bar{x}$ if $\mathcal{D}_{\bar{x}}$ is nonempty.

Now, we derive a first order necessary condition for local minimizers of (1.1) using two matrices whose columns form orthonormal basis of $\mathrm{span}\mathcal{D}_{\bar{x}}$ and $(\mathrm{span}\mathcal{D}_{\bar{x}})^\perp$, namely,

$$Y_{\bar{x}} \in \mathrm{orthon}(\mathrm{span}\mathcal{D}_{\bar{x}}) \quad \text{and} \quad Z_{\bar{x}} \in \mathrm{orthon}((\mathrm{span}\mathcal{D}_{\bar{x}})^\perp).$$

From the definitions of $\mathcal{D}_{\bar{x}}$ and $Z_{\bar{x}}$, we have $Z_{\bar{x}}^T d_{\bar{x}} = 0\ \forall d_{\bar{x}} \in \mathrm{span}\mathcal{D}_{\bar{x}}$ and there is a unique vector $\bar{z}$ such that

$$\bar{x} = Z_{\bar{x}}\bar{z} \quad \text{and} \quad \bar{z} = Z_{\bar{x}}^T \bar{x}. \tag{3.1}$$

LEMMA 3.1. *If $\bar{x}$ is a local minimizer of (1.1), there are vectors γ, $\rho \in \mathbb{R}^n$ such that $\bar{x}$ satisfies*

$$Z_{\bar{x}}^T(\nabla\Theta(\bar{x}) + \sum_{d_i^T\bar{x}\neq 0} \Phi'(d_i^T\bar{x})d_i - \gamma + \rho) = 0, \tag{3.2a}$$

$$(\bar{x} - l)^T\gamma = 0, \quad (u - \bar{x})^T\rho = 0, \tag{3.2b}$$

$$l \leq \bar{x} \leq u,\ \gamma \geq 0,\ \rho \geq 0. \tag{3.2c}$$

Proof. Suppose $\bar{x}$ is a local minimizer of (1.1). If $\dim(\mathrm{span}\mathcal{D}_{\bar{x}}) = n$, then $\bar{x} = 0$ and $Z_{\bar{x}} = 0_{n\times 1}$, which means the conditions in this lemma naturally holds for $\bar{x}$ with $\gamma = \rho = 0_{n\times 1}$.

Now, we suppose $\dim(\mathrm{span}\mathcal{D}_{\bar{x}}) = n - r < n$, then $Z_{\bar{x}} \in \mathbb{R}^{n\times r}$ is a nonzero matrix, $\bar{x} \in [l, u]$ and there is an $\eta_{\bar{x}} > 0$ such that

$$
\begin{aligned}
f(\bar{x}) =& \min_x \{f(x) : x \in [l,u], \|x-\bar{x}\| \le \eta_{\bar{x}}\} \\
=& \min_{y,z} \{\Theta(Y_{\bar{x}}y + Z_{\bar{x}}z) + \sum_{i=1}^m \varphi(|d_i^T(Y_{\bar{x}}y + Z_{\bar{x}}z)|^p) : \\
& Y_{\bar{x}}y + Z_{\bar{x}}z \in [l,u], \|Y_{\bar{x}}y + Z_{\bar{x}}z - Z_{\bar{x}}\bar{z}\| \le \eta_{\bar{x}}\} \\
\le& \min_z \{\Theta(Y_{\bar{x}}0 + Z_{\bar{x}}z) + \sum_{i=1}^m \varphi(|d_i^T(Y_{\bar{x}}0 + Z_{\bar{x}}z)|^p) : \\
& Y_{\bar{x}}0 + Z_{\bar{x}}z \in [l,u], \|Y_{\bar{x}}0 + Z_{\bar{x}}z - Z_{\bar{x}}\bar{z}\| \le \eta_{\bar{x}}\} \\
=& \min_z \{\Theta(Z_{\bar{x}}z) + \sum_{i=1}^m \varphi(|d_i^T Z_{\bar{x}}z|^p) : Z_{\bar{x}}z \in [l,u], \|Z_{\bar{x}}z - Z_{\bar{x}}\bar{z}\| \le \eta_{\bar{x}}\} \\
=& \min_z \{\Theta(Z_{\bar{x}}z) + \sum_{d_i^T\bar{x}\neq 0} \varphi(|d_i^T Z_{\bar{x}}z|^p) : Z_{\bar{x}}z \in [l,u], \|Z_{\bar{x}}z - Z_{\bar{x}}\bar{z}\| \le \eta_{\bar{x}}\}.
\end{aligned}
$$

In what follows, we will find the first order necessary condition for local minimizers of (1.1) from the reduced optimization problem in $\mathbb{R}^r$:

$$
\min_{Z_{\bar{x}}z\in[l,u]} \quad v(z) = \Theta(Z_{\bar{x}}z) + \sum_{d_i^T\bar{x}\neq 0} \varphi(|d_i^T Z_{\bar{x}}z|^p), \tag{3.3}
$$

where $v(z)$ is continuously differentiable and its gradient is locally Lipschitz continuous around $\bar{z}$.

By (3.1) and (3.3),

$$
v(\bar{z}) = \Theta(Z_{\bar{x}}\bar{z}) + \sum_{i=1}^m \varphi(|d_i^T Z_{\bar{x}}\bar{z}|^p) = f(\bar{x}).
$$

Therefore, $v(\bar{z}) \le \min_z\{v(z) : Z_{\bar{x}}z \in [l,u], \|Z_{\bar{x}}(z-\bar{z})\| \le \eta_{\bar{x}}\}$.

Since $Z_{\bar{x}} \in \mathbb{R}^{n\times r}$ is of full column rank, $\bar{z}$ is a local minimizer of

$$
\min_{Z_{\bar{x}}z\in[l,u]} v(z).
$$

By the KKT condition for a local minimizer of (3.3), $\bar{z}$ satisfies

$$
\nabla v(\bar{z}) - Z_{\bar{x}}^T\gamma + Z_{\bar{x}}^T\rho = 0, \tag{3.4a}
$$

$$
(Z_{\bar{x}}\bar{z} - l)^T\gamma = 0,\ (u - Z_{\bar{x}}\bar{z})^T\rho = 0 \tag{3.4b}
$$

$$
l \le Z_{\bar{x}}\bar{z} \le u,\ \gamma \ge 0,\ \rho \ge 0. \tag{3.4c}
$$

By (3.1), (3.2b) and (3.2c) can be obtained from (3.4b) and (3.4c).

From (3.1) and (3.3), we have

$$
\begin{aligned}
\nabla v(\bar{z}) =& Z_{\bar{x}}^T(\nabla\Theta(y)_{y=Z_{\bar{x}}\bar{z}} + \sum_{d_i^T\bar{x}\neq 0} \Phi'(d_i^T Z_{\bar{x}}\bar{z})d_i) \\
=& Z_{\bar{x}}^T(\nabla\Theta(\bar{x}) + \sum_{d_i^T\bar{x}\neq 0} \Phi'(d_i^T\bar{x})d_i),
\end{aligned} \tag{3.5}
$$

which together with (3.4a) gives (3.2a). □

In view of the first order necessary condition for local minimizers of (1.1) given in Lemma 3.1, we define the scaled and ϵ scaled first order stationary points of (1.1).

DEFINITION 3.2. *We call $\bar{x}$ a scaled first order stationary point of (1.1), if there are vectors γ, $\rho \in \mathbb{R}^n$ such that $\bar{x}$ satisfies (3.2) in Lemma 3.1.*

DEFINITION 3.3. *For $\epsilon > 0$, we call x^ϵ an ϵ scaled first order stationary point of (1.1), if there are vectors γ^ϵ, $\rho^\epsilon \in \mathbb{R}^n$ such that x^ϵ satisfies*

$$\left\| (Z^\epsilon_{x^\epsilon})^T (\nabla\Theta(x^\epsilon) + \sum_{|d_i^T x^\epsilon| > \epsilon} \Phi'(d_i^T x^\epsilon) d_i - \gamma^\epsilon + \rho^\epsilon) \right\|_\infty \le \epsilon, \tag{3.6a}$$

$$(x^\epsilon - l)^T \gamma^\epsilon \le \epsilon,\ (u - x^\epsilon)^T \rho^\epsilon \le \epsilon, \tag{3.6b}$$

$$l \le x^\epsilon \le u,\ \gamma^\epsilon \ge 0,\ \rho^\epsilon \ge 0. \tag{3.6c}$$

where $Z^\epsilon_{x^\epsilon} \in \mathrm{orthon}((\mathrm{span}\mathcal{D}^\epsilon_{x^\epsilon})^\perp)$.

Definitions 3.2 and 3.3 are consistent at $\epsilon = 0$. The next proposition validates this consistence for ϵ tending to 0.

PROPOSITION 3.4. *Let x^ϵ be an ϵ ($\epsilon > 0$) scaled first order stationary point of (1.1). Then any cluster point of x^ϵ is a scaled first order stationary point of (1.1) as $\epsilon \to 0$.*

Proof. Suppose $\bar{x}$ is a limit point of $\{x^k\}$ as k tending to ∞, where x^k is an ϵ_k scaled first order stationary point of (1.1) and $\lim_{k\to\infty} \epsilon_k = 0$.

If dim(span$\mathcal{D}_{\bar{x}}$)$= n$, then $\bar{x} = 0$ and $Z_{\bar{x}} = 0_{n\times 1}$, which implies that $\bar{x}$ is a scaled first order stationary point. In what follows, we suppose that dim(span$\mathcal{D}_{\bar{x}}$)$< n$.

First, we prove that there is $k_{\bar{x}} \in \mathbb{K}$ such that $\mathcal{D}_k \subseteq \mathcal{D}_{\bar{x}}\ \forall k \ge k_{\bar{x}}$, where $\mathcal{D}_k := \mathcal{D}^{\epsilon_k}_{x^k}$. If not, there is a subsequence $\{x^{k_j}\} \subseteq \{x^k\}$ such that $\lim_{j\to\infty} \epsilon_{k_j} = 0$ and $\mathcal{D}_{k_j} \not\subseteq \mathcal{D}_{\bar{x}}$ for all j, by $\mathcal{D}_{k_j} \subseteq \{d_1, d_2, \ldots, d_m\}$, there is an element $d \in \mathbb{R}^n$ such that $d \in \mathcal{D}_{k_j}$ but $d \notin \mathcal{D}_{\bar{x}}$. Then, $|d^T x^{k_j}| \le \epsilon_{k_j}$, letting j tend to ∞, we have $|d^T \bar{x}| = 0$, which leads a contradiction with $d \notin \mathcal{D}_{\bar{x}}$.

Denote $Z_k \in \mathrm{orthon}((\mathrm{span}\mathcal{D}_k)^\perp)$. Then, we can find matrices Z_k and $Z_{\bar{x}}$ such that Z_k contains all columns of $Z_{\bar{x}}$ for all $k \ge k_{\bar{x}}$.

From $\theta(t, \epsilon_k) = |t|$ when $|t| \ge \epsilon_k$ and by (3.6a), there are vectors $\gamma^k, \rho^k \in \mathbb{R}^n$ such that

$$\left\| Z_k^T (\nabla\Theta(x^k) + \sum_{i=1}^m \tilde{\Phi}'(d_i^T x^k, \epsilon_k) d_i - \gamma^k + \rho^k) \right\|_\infty \le \epsilon_k,$$

from the inclusion property between $Z_{\bar{x}}$ and Z_k, which gives

$$\left\| Z_{\bar{x}}^T (\nabla\Theta(x^k) + \sum_{i=1}^m \tilde{\Phi}'(d_i^T x^k, \epsilon_k) d_i - \gamma^k + \rho^k) \right\|_\infty \le \epsilon_k.$$

By the definition on $Z_{\bar{x}}$, we have

$$\left\| Z_{\bar{x}}^T (\nabla\Theta(x^k) + \sum_{d_i^T \bar{x} \ne 0} \tilde{\Phi}'(d_i^T x^k, \epsilon_k) d_i - \gamma^k + \rho^k) \right\|_\infty \le \epsilon_k. \tag{3.7}$$

If $\{\gamma^k\}$ and $\{\rho^k\}$ are bounded, there exist subsequence $\{k_j\} \subseteq \{k\}$, and γ, $\rho \in \mathbb{R}^n$ such that $\lim_{j\to\infty} \gamma^{k_j} = \gamma$ and $\lim_{j\to\infty} \rho^{k_j} = \rho$.

Letting j tend to ∞, (3.6b), (3.6c) and (3.7) give

$$\begin{aligned}
&\left\| Z_{\bar{x}}^T(\nabla\Theta(\bar{x}) + \sum_{d_i^T\bar{x}\neq 0} \Phi'(d_i^T\bar{x})d_i - \gamma + \rho)\right\|_\infty = 0,\\
&(\bar{x}-l)^T\gamma = 0, \quad (u-\bar{x})^T\rho = 0,\\
&l \le \bar{x} \le u,\ \gamma \ge 0,\ \rho \ge 0,
\end{aligned} \tag{3.8}$$

which means that $\bar{x}$ is a scaled first order stationary point of (1.1).

Next, we consider the case that $\{\gamma^k\}$ or $\{\rho^k\}$ are unbounded.

From (3.6b) and (3.6c), we have

$$\lim_{k\to\infty}(\bar{x}-l)^T\gamma^k = 0, \quad \lim_{k\to\infty}(u-\bar{x})^T\rho^k = 0. \tag{3.9}$$

Denote $J_{\bar{x}} = \{i \in \{1,2,\ldots,n\} : \bar{x}_i = l_i \text{ or } \bar{x}_i = u_i\}$. From (3.9), $l \le \bar{x} \le u$, γ^k, $\rho^k \ge 0$, we have $\lim_{k\to\infty}\gamma_i^k = \lim_{k\to\infty}\rho_i^k = 0\ \forall i \notin J_{\bar{x}}$. If $J_{\bar{x}} = \emptyset$, letting k tend to ∞, we can also have (3.8). Then, we suppose $J_{\bar{x}} \neq \emptyset$ and let $J_{\bar{x}} = \{t+1,\ldots,n\}$ without loss of generality.

Letting k tend to ∞ in (3.7), we have

$$\lim_{k\to\infty} Z_{\bar{x}}^T(\gamma^k - \rho^k) = Z_{\bar{x}}^T(\nabla\Theta(\bar{x}) + \sum_{d_i^T\bar{x}\neq 0}\Phi'(d_i^T\bar{x})d_i),$$

which follows

$$\lim_{k\to\infty}[Z_{\bar{x}}]_{J_{\bar{x}}}^T(\gamma_{J_{\bar{x}}}^k - \rho_{J_{\bar{x}}}^k) = Z_{\bar{x}}^T(\nabla\Theta(\bar{x}) + \sum_{d_i^T\bar{x}\neq 0}\Phi'(d_i^T\bar{x})d_i). \tag{3.10}$$

From (3.10), there exist $y \in \mathbb{R}^{n-t}$ such that

$$[Z_{\bar{x}}]_{J_{\bar{x}}}^T y = Z_{\bar{x}}^T(\nabla\Theta(\bar{x}) + \sum_{d_i^T\bar{x}\neq 0}\Phi'(d_i^T\bar{x})d_i).$$

Then $\bar{x}$ satisfies (3.8) with γ, $\rho \in \mathbb{R}^n$, where $\gamma_{J_{\bar{x}}} = (y)_+$, $\rho_{J_{\bar{x}}} = y - (y)_+$ and the other elements of γ and ρ are 0.

Therefore, $\bar{x}$ is a scaled first order stationary point of (1.1). □

REMARK 3.1. *Proposition 3.4 says that any cluster point of ϵ scaled first order stationary points of (1.1) is a scaled first order stationary point of (1.1) as ϵ tends to 0. Conversely, if $\bar{x}$ is a scaled first order stationary point of (1.1) and $\{x^k\}$ is a sequence converging to $\bar{x}$, then there is a sequence $\{\epsilon_k\}$ such that x^k is an ϵ_k scaled first order stationary point of (1.1) and $\lim_{k\to\infty}\epsilon_k = 0$. Thus, Proposition 3.4 gives some hints on how to find a scaled first order stationary point of (1.1).*

4. Worst-case complexity analysis.

We are now ready to present the worst-case complexity of the SQR algorithm for finding an ϵ scaled first order stationary point of (1.1).

THEOREM 4.1. *Any accumulation point of $\{z^k\}$ is a scaled first order stationary point of (1.1). Moreover, given any $\epsilon \in (0,1]$, the proposed SQR algorithm obtains an ϵ scaled first order stationary point of (1.1) defined in Definition 3.3 in no more than $O(\epsilon^{-2})$ iterations.*

Proof. Without loss of generality, we suppose $\mu_0 = 1$. Let j be the smallest positive integer such that

$$C\sigma^{p(j-1)} \leq \epsilon \quad \text{and} \quad C\sigma^{p(j-2)} > \epsilon, \tag{4.1}$$

where $C = \max\{\sqrt{2n\bar{\beta}} + \delta^{-1}, 2\}$ with $\bar{\beta}$ given in Lemma 2.5.

Denote t_j be the jth element of $\mathcal{T}$ defined in (2.12). Then, we will prove that x^{t_j} is an ϵ scaled first order stationary point of (1.1).

Note that

$$\mu_k = \sigma^{j-1} \quad \forall\ t_{j-1} + 1 \leq k \leq t_j. \tag{4.2}$$

Using $2\mu_{t_j}^p = 2\sigma^{p(j-1)} < \epsilon$, $\mathcal{D}_{t_j} \subseteq \mathcal{D}_{t_j}^\epsilon$, where

$$\mathcal{D}_{t_j} = \{d_i : |d_i^T x^{t_j}| \leq 2\mu_{t_j}^p\}, \quad \mathcal{D}_{t_j}^\epsilon = \{d_i : |d_i^T x^{t_j}| \leq \epsilon\}.$$

Then, we can find Z_{t_j} and $Z_{t_j}^\epsilon$ such that $Z_{t_j} \in \text{orthon}((\text{span}\mathcal{D}_{t_j})^\perp)$ and $Z_{t_j}^\epsilon \in \text{orthon}((\text{span}\mathcal{D}_{t_j}^\epsilon)^\perp)$.

From Lemma 2.4, (2.10) and (4.2), there are γ^{t_j}, $\rho^{t_j} \in \mathbb{R}^n$ such that x^{t_j} satisfies

$$\left\| Z_{t_j}^T(\nabla\Theta(x^{t_j}) + \sum_{|d_i^T x^{t_j}| > 2\mu_{t_j}^p} \Phi'(d_i^T x^{t_j}) d_i - \gamma^{t_j} + \rho^{t_j}) \right\|_\infty \leq \epsilon, \tag{4.3a}$$

$$(x^{t_j} - l)^T \gamma^{t_j} < \mu_{t_j}^{2p} \leq \epsilon^2, \quad (u - x^{t_j})^T \rho^{t_j} < \mu_{t_j}^{2p} \leq \epsilon^2, \tag{4.3b}$$

$$l \leq x^{t_j} \leq u, \quad \gamma^{t_j} \geq 0, \quad \rho^{t_j} \geq 0. \tag{4.3c}$$

(4.3a) implies that

$$\left\| (Z_{t_j}^\epsilon)^T(\nabla\Theta(x^{t_j}) + \sum_{|d_i^T x^{t_j}| > 2\mu_{t_j}^p} \Phi'(d_i^T x^{t_j}) d_i - \gamma^{t_j} + \rho^{t_j}) \right\|_\infty \leq \epsilon. \tag{4.4}$$

For all d_i such that $|d_i^T x^{t_j}| \leq \epsilon$, by the definition of $Z_{t_j}^\epsilon$, we have $Z_{t_j}^\epsilon d_i = 0$. Then, by $2\mu_{t_j}^p \leq \epsilon$, (4.4) implies

$$\left\| (Z_{t_j}^\epsilon)^T(\nabla\Theta(x^{t_j}) + \sum_{|d_i^T x^{t_j}| > \epsilon} \Phi'(d_i^T x^{t_j}) d_i - \gamma^{t_j} + \rho^{t_j}) \right\|_\infty \leq \epsilon. \tag{4.5}$$

Combining $x^{t_j} \in [l, u]$, $\gamma^{t_j}, \rho^{t_j} \geq 0$, (4.3b), (4.3c) and (4.5), we conclude that x^{t_j} is an ϵ scaled first order stationary point of (1.1) and we need at most t_j iterations to find it.

Suppose there are s_j successful iterations up to the t_jth iteration. From Step 2 in the SQR algorithm and Lemma 2.5 (i), $\bar{\beta} \geq \beta_{t_j} \geq \beta_0 \eta^{t_j - s_j}$, which implies that $\eta^{t_j - s_j} \leq \bar{\beta}/\beta_0$. Then,

$$t_j \leq s_j + \log_\eta \bar{\beta} - \log_\eta \beta_0. \tag{4.6}$$

Thus, in order to evaluate t_j, we only need to evaluate s_j.

From (2.10), when $k \in N_s \backslash \mathcal{T}$,

$$\tilde{f}(x^{k+1}, \mu_{k+1}) - \tilde{f}(x^k, \mu_k) \le -\mu_k^{2p}. \tag{4.7}$$

Since there are at least $s_j - j + 1$ successful iterations before the t_jth iteration such that (4.7) holds, from the nonincreasing of $\tilde{f}(x^k, \mu_k)$, we have

$$\tilde{f}(x^{t_j}, \mu_{t_j}) - \tilde{f}(x^0, \mu_0) \le -(s_j - j + 1)\sigma^{2p(j-1)}. \tag{4.8}$$

By the second inequality in (4.1), we have

$$j \le \frac{1}{p}\log_\sigma \frac{\epsilon}{C} + 2, \quad \sigma^{2p(j-1)} \ge \sigma^{2p}C^{-2}\epsilon^2. \tag{4.9}$$

(4.8) and (4.9) give

$$s_j \le \frac{C^2(\tilde{f}(x^0, \mu_0) - \min_{x\in[l,u]} f(x))}{\sigma^{2p}\epsilon^2} + \frac{1}{p}\log_\sigma \frac{\epsilon}{C} + 1. \tag{4.10}$$

From (4.6) and (4.10), we have

$$t_j \le \frac{C^2(\tilde{f}(x^0, \mu_0) - \min_{x\in[l,u]} f(x))}{\sigma^{2p}\epsilon^2} + \frac{1}{p}\log_\sigma \frac{\epsilon}{C} + \log_\eta \bar{\beta} - \log_\eta \beta_0 + 1.$$

Thus, the worst-case complexity of the proposed SQR algorithm for obtaining an ϵ scaled first order stationary point of (1.1) is $O(\epsilon^{-2})$.

By (2.11), for any fixed $j = 1, 2, \ldots$, $z^k = z^{t_j+1} = x^{t_j}$, $\forall t_j + 1 \le k \le t_{j+1}$. From Lemma 2.5 (ii), (4.3b), (4.3c) and (4.5), any accumulation point of $\{z^k\}$ is a scaled first order stationary point of (1.1). $\square$

REMARK 4.1. *Let $\mu_0 = 1$. Suppose the constant C in the proof of Theorem 4.1 is known. From the proof of Theorem 4.1, if j satisfies (4.1), then $z^{t_j+1} = x^{t_j}$ is an ϵ scaled first order stationary point of (1.1). By (4.1) and (4.2), if $\bar{k} \in \mathbb{K}$ satisfies*

$$C\mu_{\bar{k}}^p \le \epsilon \quad \textit{and} \quad \mu_{\bar{k}+1} = \sigma\mu_{\bar{k}}, \tag{4.11}$$

then z^k is an ϵ scaled first order stationary point of (1.1) for all $k \ge \bar{k} + 1$.

Though it is difficult to judge which iterate is an ϵ scaled first order stationary point of (1.1) from Definition 3.3, we can use (4.11) to find an ϵ scaled first order stationary point of (1.1) satisfying Definition 3.3.

To end this section, we apply the SQR algorithm and our analysis to the unconstrained non–Lipschitz optimization

$$\min_x \quad \Theta(x) + \sum_{i=1}^m \varphi(|d_i^T x|^p). \tag{4.12}$$

From the ideas in Lemma 3.1, if $\bar{x}$ is a local minimizer of (4.12), then $\bar{x}$ satisfies (3.2a) with $\gamma = \rho = 0$, and the ϵ scaled first order stationary point of (4.12) is defined as follows.

DEFINITION 4.2. *For $\epsilon \ge 0$, we call x^ϵ an ϵ scaled first order stationary point of (4.12), if*

$$\left\| (Z_{x^\epsilon}^\epsilon)^T(\nabla\Theta(x^\epsilon) + \sum_{|d_i^T x^\epsilon| > \epsilon} \Phi'(d_i^T x^\epsilon) d_i) \right\|_\infty \le \epsilon, \tag{4.13}$$

and x^ϵ a scaled first order stationary point of (4.12) if (4.13) holds with $\epsilon = 0$, where $Z^\epsilon_{x^\epsilon}$ is defined in Definition 3.3.

The quadratic program (2.9) reduces to

$$\begin{aligned} \min \quad & \langle g(x,\mu), s\rangle + \frac{\beta}{2}\|s\|^2 + 4\alpha p \mu^{p-2} \sum_{|d_i^T x| \le 2\mu^p} |d_i^T s|^2 \\ \text{s.t.} \quad & s^2 \le \delta^2 \mu^{2p} e. \end{aligned} \tag{4.14}$$

Similar to the analysis above, we have the following corollary.

COROLLARY 4.3. *Given any $\epsilon \in (0,1]$, the proposed SQR algorithm obtains an ϵ scaled first order stationary point of (4.12) defined in Definition 4.2 in no more than $O(\epsilon^{-2})$ iterations.*

5. Numerical Experiments. In this section, we report numerical results of three examples to validate the theoretical results and show the good performance of the proposed SQR algorithm. The numerical testing is performed using MATLAB 2009a on a Lenov PC (3.00GHz, 2.00GB of RAM). The strongly convex quadratic subproblem (2.9) is solved by the projected alternating Barzilai-Borwein method in [14] with the zero vector as the initial iterate. Throughout the numerical experiments, we let $\mu_0 = 1$, $\beta_0 = 2$, $\eta = 1.1$ and $\sigma = 0.99$ in the SQR algorithm.

Example 5.1 is a randomly generated test problem to support the iteration complexity bound of the SQR algorithm for finding a scaled first order stationary point of (1.1) given in Theorem 4.1.

Example 5.2 and Example 5.3 are two often used gray level image restoration problems with intensity values ranging from 0 to 1. They are the Circles image of size 64×64 and the Phantom image of size 256×256. Numerical results show that the proposed SQR algorithm is robust and efficient for image restoration. We use the peak signal-to-noise ratio (PSNR) to evaluate the quality of the restored image, i.e.

$$\mathrm{PSNR}(x^k) = -10 \log_{10} \frac{\|x^k - x_o\|}{n_l \times n_w},$$

where x_o is the original image with the dimension $n_l \times n_w$. Let x_b be the corresponding observed image. The CPU time reported is in seconds. For the regularization term, we use two different potential functions

$$\varphi_1(s) = \lambda s, \quad \varphi_2(s) = \lambda \frac{0.5s}{1+0.5s}, \quad \text{where} \quad \lambda > 0. \tag{5.1}$$

EXAMPLE 5.1. ***A test example for complexity bound.*** In this example, we solve (1.1) with $\Theta(x) = \|x - b\|^2$, $l = -2e$, $u = 2e$, $\varphi(s) := \varphi_1(s)$ with $\lambda = 0.2$, and

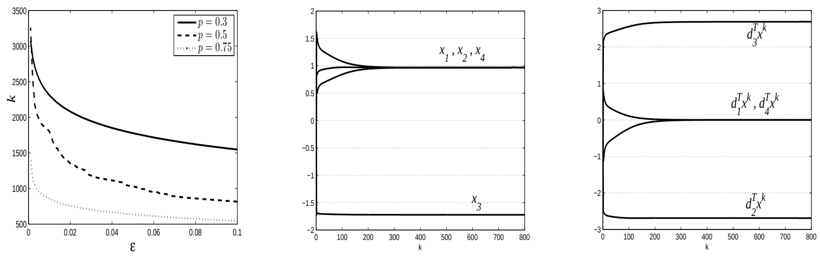
$$D = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$

We generate 20 samples of original signal s and vector b independently as the following

```
s=randn(4,1);s'=median([-2*ones(1,4);s';2*ones(1,4)]);
b=s+normrnd(0,0.05,[4,1]);
```

With these 20 randomly generated vectors b, the average number of iterations for obtaining an ϵ scaled first order stationary point of (1.1) with three different values of

p and the stop criterion in Remark 4.1 is illustrated in Figure 5.1(a), where we define $C = 6.2$. Moreover, the convergence of x^k and Dx^k with $p = 0.5$ are illustrated in Figure 5.1(b)-5.1(c). The numerical presentation in Figures 5.1 is consistent with the theoretical result in Theorem 4.1.



(a) (b) (c)

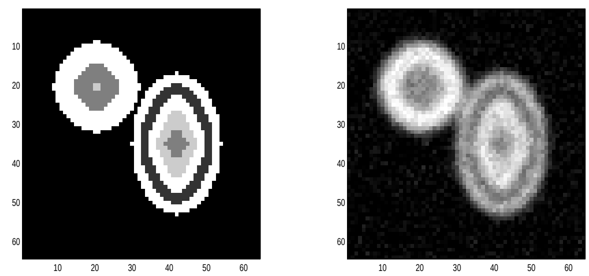
Fig. 5.1: (a) Iteration complexity of the SQR algorithm for (1.1) with different values of ϵ; (b) convergence of x^k; (c) convergence of Dx^k

Example 5.2. ***Circles image with size*** 64×64. In this example, we test the proposed SQR algorithm using the 64×64 Circles image [10, 17, 29]. We discuss the restoration of the Circles image in two parts according to the class of observed images.

A. Observed image with blurring and noise. In this part, the observed image x_b is that all the pixels are blurred by a two dimensional Gaussian function, and then added a Gaussian noise. The blurring function is chosen to be

$$h(i,j) = e^{-2(i/3)^2 - 2(j/3)^2},$$

truncated such that the function has a support of 7×7, and the Gaussian noise is with the mean of 0 and the standard deviation of 0.05. The original image and the observed image are shown in Figure 5.2.



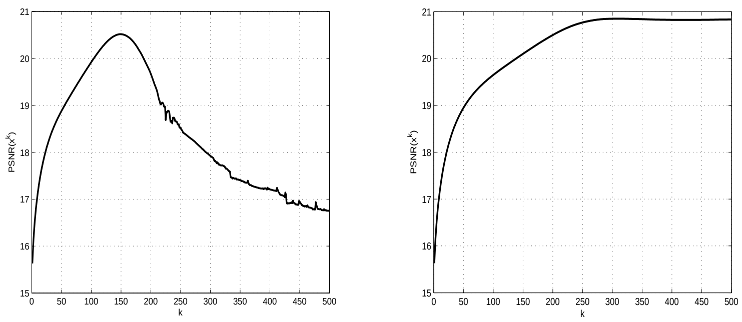
(a) (b)

Fig. 5.2: (a) original image; (b) observed image(PSNR=15.50)

Dependence on D. In order to show the importance of the difference operators in image restoration, we choose $\varphi := \varphi_2$, $l = 0_{n\times 1}$, $u = e$ and $p = 0.5$ in (1.1) and we test the SQR algorithm with $D := D_0$ and $D := D_1$ to restore the Circles image with blurring and noise, where

$$D_0 = I_n, \quad \text{and} \quad D_1 = \begin{pmatrix} L_1 \otimes I \\ I \otimes L_1 \end{pmatrix} \quad \text{with} \quad L_1 = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix}.$$

Figure 5.3(a) shows the convergence of PSNR(x^k) with $D = D_0$ and $\lambda = 0.081$, which is the best choice of λ among $0.0001 : 0.0002 : 0.1$ to let the SQR algorithm find an x^k with the highest PSNR before 500 iterations. Also with $D = D_0$ and



(a) (b)

Fig. 5.3: (a) PSNR with $D := D_0$; (b) PSNR with $D := D_1$

$\lambda = 0.081$, $f(x^{500}) = 33.35$ and $f(x_o) = 36.37$. When $D = D_0$, from our numerical experiments, for almost all λ among $0.0001 : 0.0002 : 0.1$, $\tilde{f}(x^k, \mu_k)$ is monotone decreasing, whereas, the PSNR is not monotonely increasing and $f(x^k)$ can decrease below $f(x_o)$. Thus, the original image is not the optimal solution of (1.1) with $D := D_0$ and λ among $0.0001 : 0.0002 : 0.1$. However, when $D = D_1$ and $\lambda = 0.006$, PSNR(x^k) is monotonely increasing as shown in Figure 5.3(b).

From this numerical experiment, we find that problem (1.1) with the zero order difference operator seems not suitable for the restoration of this image, but using the first order difference operator performs well. This shows the importance of the research in this paper for image sciences.

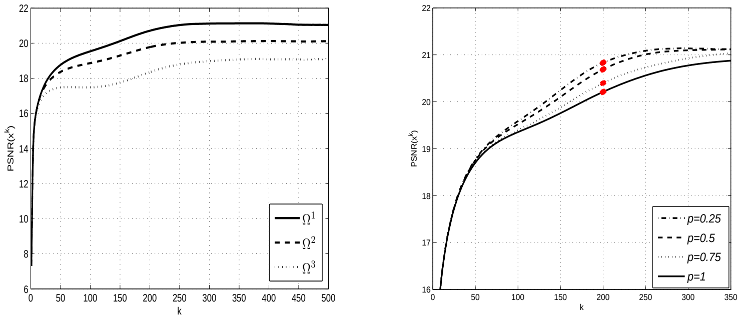
In the sequel parts of this example, we shall choose D in (1.1) to be the first-order difference operator D_1.

Dependence on the constraints. In this part, we let $p = 0.5$, $\varphi := \varphi_2$ with $\lambda = 0.015$ and choose $x^0 = 0_{n\times 1}$. To show the influences of the constraints of (1.1) in image restoration, we test the following three constraints:

$$\Omega^1 = \{x : 0 \le x \le e\}, \quad \Omega^2 = \{x : x \ge 0\}, \quad \Omega^3 = \mathbb{R}^n. \tag{5.2}$$

For these three constraints, the convergence of PSNR(x^k) by the SQR algorithm is shown in Figure 5.4(a), from which we find that the model (1.1) with box constraints Ω^1 is the best.

Dependence on p. In this part, we consider the influence of the value of p in (1.1) to restore the Circles image with blurring and noise. Let $\varphi := \varphi_2$, $l = 0_{n\times 1}$,



(a) (b)

Fig. 5.4: Convergence of PSNR(x^k) for the Circles image: (a) with different constraints; (b) with different values of p

p	1	0.75	0.5	0.25
λ	0.017	0.018	0.016	0.022
PSNR(x^{200})	20.21	20.40	20.69	20.83

Table 5.1: The SQR algorithm for the Circles image with different values of p

$u = e$ and $x^0 = 0$. For each p, the parameter λ is also manually chosen in order to obtain the best PSNR value. The PSNR values of (1.1) with different values of p by the SQR algorithm are plotted in Figure 5.4(b). Moreover, the PSNR values at the 200th iteration are listed in Table 5.1. From Figure 5.4(b) and Table 5.1, we find that solving (1.1) with a smaller value of p by the SQR algorithm can find higher PSNR value.

Independence on initial iterate. In this part, we let $p = 0.5$, $l = 0_{n\times1}$, $u = e$ and $\varphi := \varphi_1$ with $\lambda = 0.006$. We test the SQR algorithm with three different initial iterates: the zero vector denoted by 0, the observed data projected on Ω^1 denoted by $x_b^{\Omega^1}$, and a randomly generated vector in Ω^1 denoted by x_r. For the three initial iterates, the corresponding results at the 300th iteration are given in Table 5.2. The objective value with the original image is $f(x_o) = 13.44$. We observe that the SQR algorithm is stable with respect to the choice of initial iterates, in terms of the PSNR values, objective values and CPU time.

At the end of this part, we should state that the PSNR of the restored image by the SQR algorithm for (1.1) with $D := D_1$ is better than the resorted images in [10, 29] (PSNR=19.03 in [29] and PSNR=19.97 in [10]) for the Circles image with the same blurring and noise.

B. Observed image with Gaussian noise. In this part, we generate the observed image x_b without blurring that all the pixels are contaminated by Gaussian noise with mean of 0 and standard deviation of 0.1. Define $p = 0.5$, $\varphi := \varphi_2$ with $\lambda = 0.15$ and $D := D_1$ in (1.1). Then $f(x_o) = 64.27$, $f(x_b) = 172.43$ and PSNR(x_b) = 20.07.

With the three initial iterates used in Table 5.2, the numerical results of the SQR algorithm for solving (1.1) to restore the Circles image with Gaussian noise are

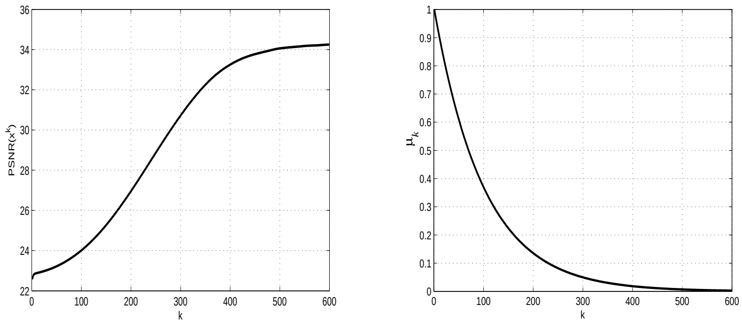
x^0	PSNR(x^0)	PSNR(x^{300})	$f(x^0)$	$f(x^{300})$	CPU
0	7.30	21.02	585.41	15.17	10.21
$x_b^{\Omega^1}$	15.63	21.01	27.34	15.15	10.09
x_r	5.04	20.72	799.50	15.23	10.16

Table 5.2: The SQR algorithm for the Circles image with different initial iterates

x^0	PSNR(x^0)	PSNR(x^{600})	$f(x^0)$	$f(x^{600})$	CPU
0	7.30	34.24	794.42	71.89	80.49
$x_b^{\Omega^1}$	22.57	34.26	139.82	71.59	79.85
x_r	5.04	34.25	1.56e+3	71.89	73.94

Table 5.3: Stability of the SQR algorithm with different initial iterates for the Circles image without blurring

given in Table 5.3, which shows that the SQR algorithm is stable with respect to the initial iterates. Moreover, with $x^0 = x_b^{\Omega^1}$, the convergence of μ_k and PSNR(x^k) are illustrated in Figure 5.5. From the results in Table 5.3 and Figure 5.5, the PSNR of the restored image by the SQR algorithm is also higher than the resorted images in [17, 29] (PSNR=31.03 in [17] and PSNR=31.28 in [29]) for the Circles image with the same Gaussian noise.



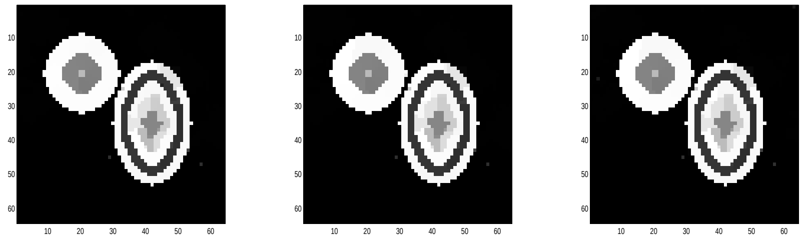
(a) (b)

Fig. 5.5: Convergence of PSNR(x^k) and μ_k for the Circles image with Gaussian noise: (a) PSNR(x^k); (b) μ_k

For the three constraints in (5.2), the restored images by the SQR algorithm with $x^0 = 0$ at the 600th iteration are shown in Figure 5.6. We see that the quality of the restored image with box constraint Ω^1 is better than the other two restored images.

EXAMPLE 5.3. ***Phantom image with size*** 256×256. In this example, we test the proposed SQR algorithm using the 256×256 Phantom image with blurring and Gaussian noise as Example 5.2. Figure 5.7 gives the original and observed images.

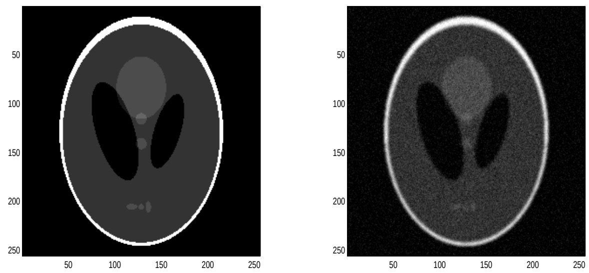
Define $\varphi := \varphi_1$, $p = 0.5$, $D := D_1$ and $\lambda = 0.009$ in (1.1). With the three different constraints in (5.2), we show the restored images by the SQR algorithm with $x^0 = 0$



(a) (b) (c)

Fig. 5.6: Restored images with different constraints: (a) Ω_1(PSNR(x^{600})=34.26); (b) Ω_2 (PSNR(x^{600})=33.89); (c) Ω_3(PSNR(x^{600})=33.01)

in Figure 5.8, and the convergence of PSNR(x^k) with $x^0 = x_b^{\Omega^1}$ in Figure 5.9(a). Similarly as the performance in Example 5.2, the box constraint Ω^1 can provide a better image restoration with higher PSNR value. Figure 5.9(b) shows convergence of $f(x^k)$ and $\tilde{f}(x^k, \mu_k)$ generated by the SQR algorithm with $x^0 = x_b^{\Omega^1}$ for this 256×256 image.



(a) (b)

Fig. 5.7: Phantom image: (a) original image; (b) observed image(PSNR=15.50)

REFERENCES

[1] W. Bian and X. Chen, *Smoothing neural network for constrained non-Lipschitz optimization with applications*, IEEE Trans. Neural Netw. Learn. Syst., 23(2012), pp. 399-411.

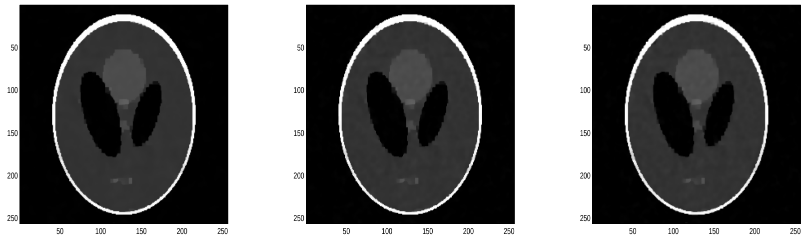
[2] W. Bian and X. Chen, *Worst-case complexity of smoothing quadratic regularization methods for non-Lipschitzian optimization*, SIAM J. Optim., 23(2013), pp. 1718-1741.

[3] W. Bian, X. Chen and Y. Ye, *Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization*, to appear in Math. Program.

[4] 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.

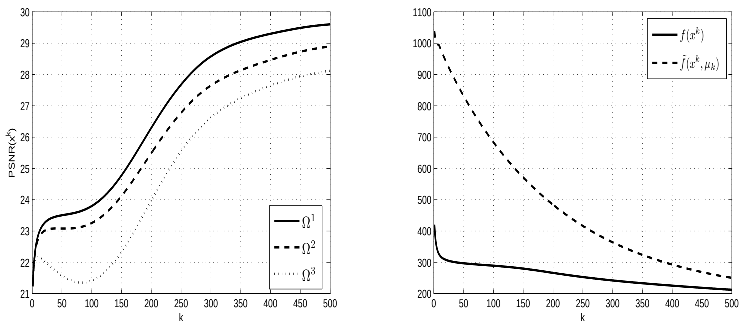
[5] R.H. Chan and K. Chen, *A multilevel algorithm for simultaneously denoising and deblurring images*, SIAM J. Sci. Comput., 32(2010), pp. 1043-1063.

[6] R. Chartrand, *Exact reconstruction of sparse signals via nonconvex minimization*, IEEE Signal Process. Lett., 14(2007), pp. 707-710.



(a) (b) (c)

Fig. 5.8: Restored images for the Phantom image with different constraints: (a) Ω_1 (PSNR(x^{500})=29.60); (b) Ω_2 (PSNR(x^{500})=28.90); (c) Ω_3 (PSNR(x^{500})=28.12)



(a) (b)

Fig. 5.9: Convergence for the Phantom image: (a) PSNR(x^k) with different constraints; (b) $f(x^k)$ and $\tilde{f}(x^k, \mu_k)$ with $\Omega = \Omega_1$.

[7] R. Chartrand and V. Staneva, *Restricted isometry properties and nonconvex compressive sensing*, Inverse Prob., 24(2008), pp. 1-14.

[8] X. Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Program., 134(2012), pp. 71-99.

[9] X. Chen, L. Niu and Y. Yuan, *Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization*, SIAM J. Optim., 23(2013), pp. 1528-1552.

[10] 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.

[11] 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.

[12] 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.

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

[14] Y.H. Dai and R. Fletcher, *Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programing*, Numer. Math., 100(2005), pp. 21-47.

[15] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Stat. Soc., 96(2001), pp. 1348-1360.

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

[17] M. Hintermuller and T. Wu, *Nonconvex TV^q-models in image restoration: analysis and a trust-region regularization based superlinearly convergent solver*, SIAM J. Imaging Sci.,

6(2013), pp. 1385-1415.
[18] M. Hintermuller, T. Wu, *A superlinearly convergent R-regularized Newton scheme for variational models with concave sparsity-promoting priors*, Computat. Optim. Appl., 57(2014), pp. 1-25.
[19] J. Huang, J. L. Horowitz and S. Ma, *Asymptotic properties of bridge estimators in sparse high-dimensional regression models*, Ann. Stat., 36(2008), pp. 587-613.
[20] J. Huang, S. Ma, H. Xue and C. Zhang, *A group bridge approach for variable selection*, Biometrika, 96(2009), pp. 339-355.
[21] P. Huber, Robust Estimation, New York: Wiley, 1981.
[22] A. Jain, *Fundamentals of Digital Image Processing*, Upper Saddle River, NJ: Prentice-Hall, 1989.
[23] K. Knight and W.J. Fu, *Asymptotics for lasso-type estimators*, Ann. Stat., 28(2000), pp. 1356-1378.
[24] K. Kunisch, T. Pock, *A bilevel optimization approach for parameter learning in variational models*, SIAM J. Imaging Sci., 6(2013), pp. 938-983.
[25] M. Lai and Y. Wang,*An unconstrained lq minimization with* $0 < q < 1$ *for sparse solution of under-determined linear systems*, SIAM J. Optim., 21(2011), pp. 82-101.
[26] Z. Lu, *Iterative reweighted minimization methods for* l_p *regularized unconstrained nonlinear programming*, to appear in Math. Program.
[27] Yu. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103(2005), pp. 127-152.
[28] M. Nikolova, *Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares*, Multiscale Model. Simul., 44(2005), pp. 960-991.
[29] M. Nikolova, M.K. Ng, S.Q. 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] M. Nikolova, M.K. Ng and C.P. Tam, *Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction*, IEEE Trans. Image Process., 19(2010), pp. 3073-3088.
[31] M. Ng, P. Weiss and X. Yuan, *Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods*, SIAM J. Sci. Comput., 32(2010), pp. 2710-2736.
[32] R.T. Rockafellar and R.J-B Wets, *Variational Analysis*, Berlin: Springer, 1998.