Noname manuscript No.
(will be inserted by the editor)

An Improved Algorithm for the $L_2 - L_p$ Minimization Problem

Dongdong Ge · Rongchuan HE · Simai HE

Received: date / Accepted: date

Abstract In this paper we consider a class of non-Lipschitz and non-convex minimization problems which generalize the $L_2 - L_p$ minimization problem. We propose an iterative algorithm that decides the next iteration based on the local convexity/concavity/sparsity of its current position. We show that our algorithm finds an ϵ-KKT point within $O(\log \epsilon^{-1})$ iterations. The same result is also applied to the problem with general linear constraints under mild conditions.

Keywords Nonsmooth optimization · Nonconvex optimization · Bridge Regression · Complexity analysis

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

1 Introduction

In this paper, we consider the following optimization problem:

Dongdong Ge
Department of Management Science,
Shanghai University of Finance and Economics, Shanghai, China
E-mail: ge.dongdong@mail.shufe.edu.cn

Rongchuan HE
Department of Management Science,
City University of Hong Kong, Hong Kong, China
E-mail: rongchuhe2@gmail.com

Simai HE
Department of Management Science,
Shanghai University of Finance and Economics, Shanghai, China
E-mail: simaihe@mail.shufe.edu.cn

$$\begin{aligned} \text{Minimize} \quad & h(x) = \frac{1}{2}x^T Qx + a^T x + c + \lambda \sum_i x_i^p \\ \text{Subject to} \quad & x \geq 0 \end{aligned} \tag{1}$$

where $Q \in R^{n\times n}, 0 \preceq Q \prec \beta I, a \in R^n, c \in R, \lambda > 0, 0 < p < 1$. This non-Lipschitz and nonconvex problem is a generalization of the $L_2 - L_p$ minimization problem with nonnegative constraints:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\|Ax - b\|^2 + \lambda \sum_i x_i^p \\ \text{Subject to} \quad & x \geq 0 \end{aligned} \tag{2}$$

Chen et al. [5] show the equivalence between Problem (2) and the unconstrained $L_2 - L_p$ minimization problem. A global minimizer of the $L_2 - L_p$ problem is also called a bridge estimator in statistical literature[10]. The bridge regression problem has been studied extensively in variable selection and sparse least squares fitting for high dimensional data. See [3,4,6,8–10,**?**,15–17] and references therein.

Despite the fact that different approaches have been developed to tackle the problem (2), e.g., [3,4,6,14,16], Chen et al.[7] show that the $L_2 - L_p$ minimization problem is strongly NP-Hard for any $p \in [0, 1)$, including its smoothed version. From complexity theory perspective, an NP-hard optimization problem with a polynomially bounded objective function does not admit a polynomial-time algorithm, and a strongly NP-hard optimization problem with a polynomially bounded objective function does not even admit a fully-polynomial-time approximation scheme (FPTAS), unless P=NP [21].

A theoretically (nearly) linear time algorithm for problem (1) still remains unknown yet. The objective function in problem (1) is a combination of a quadratically convex function and a p-norm concave function, with either part of which an optimization problem can be approximated to an ϵ-KKT point in $O(\epsilon^{-1}\log(\epsilon^{-1}))$ steps. In the previous literature of non-convex minimization, Ye [22] proposes a potential reduction algorithm to obtain an ϵ-KKT point in $O(\epsilon^{-1}\log(\epsilon^{-1}))$ iterations for the general quadratic minimization problem with linear constraints, where each iteration solves a trust-region quadratic minimization problem. He also proves that when ϵ goes to 0, the iterative sequence converges to a point which satisfies the second order necessary optimization condition. Ge et al.[13] present a similar potential reduction algorithm on minimizing linearly constrained concave function $\|x\|_p^p$ $(0 < p < 1)$ with the worst case complexity $O(\epsilon^{-1}\log(\epsilon^{-1}))$.

Recently, Bian et al. [1] propose two algorithms for a class of non-Lipschitz and non-convex minimization problems with box constraints. Based on the idea of affine scaling, the solution in each iteration only lays on the interior region in which the objective function is differentiable. Their first order approximation algorithm can obtain an ϵ-KKT point in $O(\epsilon^{-2})$ steps. Moreover, they also propose a second order approximation algorithm with the worst-case

complexity $O(\epsilon^{-\frac{3}{2}})$, whereas a higher computational complexity is required at each iteration. Bian et al. [2] present an smoothing quadratic regularization algorithm for solving a class of unconstrained non-smooth non-convex problems. They show that their method takes at most $O(\epsilon^{-2})$ steps to find an ϵ-KKT solution. To the best of our knowledge, an algorithm with (nearly) linear time complexity is still under the exploration: overcoming the additional difficulty caused by the presence of both the convex and concave functions in an object needs a further analysis and understanding of the problem structure.

In this paper, we propose a breakthrough iterative algorithm to find an ϵ-KKT point for problem (1) with a worst-case complexity $O(\log(\epsilon^{-1}))$. We also show that the same result is applied to the general case with linear constraints under mild conditions. There are two interesting techniques we use in the algorithm. Firstly, we adjust the direction and step length of the next iteration based on the strength of the local convexity and concavity of the current solution. Secondly, we develop a lower bound at each step similar to the work by [6]. At each step, if an entry falls below the lower bound, the algorithm will fix its value to zero and remove all the information of that dimension. The dimension of the decision vector is reduced by one while the objective function value keeps decreasing.

The remaining part of the paper is organized as follows. In section 2, we present the algorithm and briefly explain the main idea. In section 3, we provide a formal complexity analysis of the algorithm. In section 4, we extend the discussion to the general linearly constrained case.

Notations: Throughout the paper, let $I = \{1, 2, \cdots, n\}$. For any column vector $x \in R^n$, x_i denotes the ith component of x. Let $W \subset I$, $x_W := [x_i]_{i \in W}$, which is a subset of vector x, and $x_{-W} := [x_i]_{i \in I \setminus W}$.

2 A 3-Criterion Algorithm

In this section we first present a potential function and analyze its connection with ϵ-KKT points. Then we present the algorithm with a precise explanation of the main idea.

Throughout the paper, we make the following assumptions for the convenience of our analysis.

Assumption 1 *The optimal value of problem (1) is lower bounded by 0.*

Assumption 2 *For any $x^0 \geq 0$, there exists γ such that $sup\{\|x\|_\infty | h(x) \leq h(x^0)\} \leq \gamma$.*

Assumption 1 is natural considering that the optimal value of problem (1) is always lower bounded. It will not change our computational complexity analysis. Assumption 2 always holds for the $L_2 - L_p$ minimization problem given that $\frac{1}{2}\|Ax - b\|^2$ is lower bounded by 0, and $x_i^p \to \infty$ as $x_i \to \infty$.

Denote the support set of x^* by $supp(x^*) = \{i | 1 \leq i \leq n, x_i^* \neq 0\}$. We define an ϵ-KKT point of problem (1) as follows.

Definition 1 For any $\epsilon \in (0,1)$, we call $x^* \geq 0$ an ϵ-KKT point of problem (1) if there exists $y^* \geq 0$, such that

$$x^* \geq 0, \tag{3a}$$

$$\|[\nabla h(x^*) - y^*]_{supp(x^*)}\| \leq \epsilon, \tag{3b}$$

$$y^* \geq 0, \tag{3c}$$

$$(x^*)^T y^* \leq \epsilon. \tag{3d}$$

We only consider the KKT condition on the support set given that $[\nabla h(x)]_i$ doesn't exist when $x_i = 0$, and $[\nabla h(x)]_i \to +\infty$ as $x_i \to 0^+$. We also relax both of the first order conditions and complementary conditions by a scale ϵ. Our definition is consistent with the KKT conditions given in [1, **?**] with $\epsilon = 0$ for the unconstrained $L_2 - L_p$ problem.

To simplify our analysis, we let $f(x) = \frac{1}{2}\beta x^T x + a^T x + c$ and let $g(x) = \lambda \sum_i x_i^p + \frac{1}{2} x^T (Q - \beta I) x$. Thus the objective function $h(x) = f(x) + g(x)$.

Note that if we fixed a working set $\Lambda \subset \{1, \cdots, n\}$, and let $x^*_{-\Lambda} = 0$, then the ϵ-KKT condition of problem (1) is equivalent to the ϵ-KKT condition of the reduced problem

$$\begin{array}{ll} \text{Minimize} & h(x_\Lambda, x_{-\Lambda} = 0) \\ \text{Subject to} & x_\Lambda \geq 0, \end{array} \tag{4}$$

providing $y^*_{-\Lambda} = 0$. For any feasible solution x, $\nabla h(x)_i$ is undefined at $x_i = 0$. Thus, it is more convenient for us to work on the support set $supp(x)$. To simplify our notation, in the following discussion, we always explicitly assume that we are working on the set supp(x), for the current solution x. The parameters Q and c in the objective function can be adjusted accordingly.

Our algorithm and analysis are heavily based on the idea of potential function reduction. We first introduce a potential function.

The fact that function $g(x)$ is concave implies that, for any $z > 0$ and $x \geq 0$,

$$h(x) \leq f(x) + g(z) + \nabla g(z)^T (x - z).$$

Now we consider the problem:

$$\begin{array}{ll} \text{Minimize} & L_z(x) = f(x) + g(z) + \nabla g(z)^T (x - z) \\ \text{Subject to} & x \geq 0 \end{array} \tag{5}$$

Let $\bar{z}$ be a minimizer of problem (5). The potential function $\Delta L(z) : R^n \to R$ is defined as

$$\Delta L(z) = L_z(z) - L_z(\bar{z}).$$

The objective function $L_z(x)$ in problem (5) is strictly convex and separable on every entry of x. Thus, the unique minimizer of problem (5) has a closed form as

$$\bar{z}_i = \max\{0, -\frac{1}{\beta}[a_i + \lambda p z_i^{p-1} + (Q - \beta I)_i z]\}, \forall i \tag{6}$$

where $(Q - \beta I)_i$ is the ith row of matrix $Q - \beta I$.

We also define a descending direction

$$d_z = \bar{z} - z. \tag{7}$$

Next we show that a positive vector z is an ϵ-KKT point of problem (1) if the potential function value at z is small enough.

Lemma 1 *Given $\epsilon \in (0,1)$, for any $z > 0$, if $\Delta L(z) \leq \frac{\epsilon^2}{2\beta}$, then z is an ϵ-KKT point of problem (1).*

Proof Problem (5) is a convex optimization problem. So its KKT conditions are both necessary and sufficient for the optimality.

$$\nabla f(\bar{z}) + \nabla g(z) - \bar{y} = 0, \tag{8a}$$
$$(\bar{y})^T \bar{z} = 0, \tag{8b}$$
$$\bar{y} \geq 0, \tag{8c}$$
$$\bar{z} \geq 0, \tag{8d}$$

where $\bar{y}$ is an optimal dual variable.

Next, we will show that z is indeed an ϵ-KKT point of problem (1), providing $\bar{y}$ is its dual variable.

Consider the ϵ-KKT conditions for problem (1), and let $x^* = z, y^* = \bar{y}$. For the first order condition (3b), we get

$$\begin{aligned} & \|\nabla f(x^*) + \nabla g(x^*) - y^*\| \\ = \; & \|\nabla f(z) + \nabla g(z) - \bar{y}\| \\ = \; & \|\nabla f(\bar{z}) + \nabla g(z) - \bar{y} + \nabla f(z) - \nabla f(\bar{z})\| \\ = \; & \|\beta d_z\|, \end{aligned}$$

where the third equality follows from (8a).

For the complementary condition (3d),by (8b) and (8a), we have

$$(y^*)^T x^* = (\bar{y})^T z = (\bar{y})^T (\bar{z} + z - \bar{z}) = -\nabla L_z(\bar{z})^T d_z. \tag{9}$$

Since $z \geq 0$ and $\bar{y} \geq 0$, it only remains to show that $\|\beta d_z\| \leq \epsilon$ and $-\nabla L_z(\bar{z})^T d_z \leq \epsilon$.

From the definition of $\Delta L(z)$, we know that

$$\Delta L(z) = L_z(z) - L_z(\bar{z}) = -\nabla L_z(\bar{z})^T d_z + \frac{\beta}{2}(d_z)^T d_z \geq -\nabla L_z(\bar{z})^T d_z.$$

Note that $\bar{z} \in \arg\min_{x \geq 0} L_z(x)$. Due to the first order optimality condition, we have

$$-\nabla L_z(\bar{z})^T d_z \geq 0,$$

which together with (9) imply

$$\Delta L(z) \geq \frac{\beta}{2}(d_z)^T d_z.$$

By the assumption that $\Delta L(z) \leq \frac{\epsilon^2}{2\beta}$, we get

$$\|\beta d_z\| \leq \sqrt{2\beta \Delta L(z)} \leq \epsilon, \quad -\nabla L_z(\bar{z})^T d_z \leq \Delta L(z) \leq \epsilon.$$

Therefore, z is an ϵ-KKT point of problem (1). □

According to Lemma 1, the potential function value $\Delta L(z)$ is lowered bounded by $\frac{\epsilon^2}{2\beta}$ to reach an ϵ-KKT point z. Notice that there are two other possible criteria to measure the convergence and the optimality of an algorithm: the number of zero entries(at most n), the objective function value(lower bounded by 0). We design an iterative algorithm by taking all these 3 criteria into account: at each step, the algorithm will either reduce the problem dimension by one, or decrease the objective value by a constant, or shrink the potential function value at a constant exponential rate.

Before presenting the details of the algorithm, let us give a high level overview of our approach. We start from an initial point x^0, and then we compute the next point based on the situation of the current point x^k. There are three possible exclusive cases that could happen at the current point x^k and we use different strategies to deal with them.

At step k we always check the value of every entry.

- Case 1: an entry of x^k falls below our calculated lower bound. We let the entry be zero, reduce the problem dimension by one and remove all the information related to this dimension in (Q, a) and update x^k accordingly. When no such a sufficiently small entry exists, we turn to Case 2 or Case 3 by checking the local convexity of the objective function at point x^k.
- Case 2: the objective function at x is not locally "strongly" convex. We update x^k by a line search method, the newly updated objective function value can be shown to be decreased by at least a constant value M by taking advantage of the concavity of L_p function.
- Case 3: the objective function is locally "strongly" convex. A modified Newton method is applied. It leads to a constant exponential rate reduction on the potential function value while obtaining a lesser objective function value.

The algorithm terminates only if $x^k = 0$ or an ϵ-KKT point has been found. Vector 0 is a KKT point as well.

A summary of the algorithm is presented in Table 1. One can observe that case 1 and Case 2 would only happen finite times in our algorithm, since the number of decision variables is limited and the objective function is lower bounded. In conjunction with Case 3, we will show that the algorithm obtains an ϵ-KKT point in no more than $O((n + \lceil \frac{h(x^0)}{M} \rceil) \log \frac{1}{\epsilon})$ steps.

The algorithm is presented in Algorithm 1. For simplicity, we let d^k denote d_{x^k}, $L^k(x)$ denote $L_{x^k}(x)$ and ΔL^k denote $\Delta L(x^k)$. That is, $L^k(x)$ is the objective function of problem (5) where $z = x^k$, and ΔL^k is the potential function value at point x^k. We let $x_{-i} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

We also observe that when case 1 happens, we reduce the problem dimension by 1 and update the problem information accordingly. The objective

Algorithm 1 3-Criterion Algorithm

Require: : choose $\epsilon \in (0,1)$, $x^0 \geq 0$; Define $s > 0, \tau > 0$ and $L > 0$ (Specify later)
$k = 0$; $Q^0 = Q; a^0 = a$.
while Not Stop **do**
  Case 1:
  if $x_i^k \leq L$ for some index i, **then**
    $x^{k+1} = x_{-i}^k$, i.e., define x^{k+1} by removing the ith entry from x^k.
    Update Q^{k+1} by removing the ith row and column of Q^k, and $a^{k+1} = a_{-i}^k$.
  end if
  Case 2:
  if $x_i^k > L$ for all index i and $(d^k)^T \nabla^2 h(x^k) d^k \leq \tau \|d^k\|^2$ **then**
    $\zeta^k = \max\{t | x^k + td^k \geq 0, x^k - td^k \geq 0\}$
    $x^{k+1} = \arg\min_{x \in \{x^k + \zeta^k d^k, x^k - \zeta^k d^k\}} h(x)$
    $Q^{k+1} = Q^k$, and $a^{k+1} = a^k$.
  end if
  Case 3:
  if $x^k > L$ for all index i and $(d^k)^T \nabla^2 h(x^k) d^k > \tau \|d^k\|^2$ **then**
    $x^{k+1} = x^k + sd^k$
    $Q^{k+1} = Q^k$, and $a^{k+1} = a^k$.
  end if
  if $x^k = 0$ or $\Delta L^k \leq \frac{\epsilon^2}{2\beta}$ **then**
    $x^* = x^k$
    Stop
  else
    $k = k + 1$
  end if
end while

Table 1 Summary in Three Criteria

	Function value $h(x^k)$	Potential function value ΔL_k	Cardinality of x^k
Case 1: A nearly-0 component	nonincreasing	upper bounded by $h(x^0)$	decreased by 1
Case 2: Not-strongly convex	decreased by M	upper bounded by $h(x^0)$	nonincreasing
Case 3: Strongly convex	nonincreasing	shrink at a rate $(1 - s\delta)$	nonincreasing

function, the potential function and all related information should be also updated accordingly. We will prove that this change does not affect our analysis of the algorithm.

3 Computational Complexity Analysis

In this section, we discuss three cases respectively and then conclude the main theorem.

We first consider Case 1. When there exists entry i less than a given lower bound in the current decision vector x^k, we can simplify the problem by removing all the information related to the ith dimension without increasing the objective function value.

Lemma 2 *Let* $0 < L \leq \min\left\{(\frac{1}{\lambda}(n\|Q_i\|\gamma + |a_i|))^{\frac{1}{p-1}}, \forall i\right\}$. *At iteration* k *in Algorithm 1, if there exists an index* i *such that* $0 \leq x_i^k \leq L$, *then by following Algorithm 1, we let* $x^{k+1} = x_{-i}^k$. *And we also update* (Q^k, a^k) *accordingly. Then we have* $h(x^k) - h(x^{k+1}) \geq 0$.

Proof If $x_i^k = 0$, $h(x^k) - h(x^{k+1}) = 0$. For $x_i^k > 0$, recall that $h(x^k) = \frac{1}{2}x^T Q^k x + (a^k)^T x + c + \lambda \sum_i x_i^p$. By $\|x^k\|_\infty \leq \gamma$, we can derive

$$\begin{aligned} & h(x^k) - h(x^{k+1}) \\ = \; & x_i^k Q_i^k x^{k+1} + x_i^k a_i^k + \frac{1}{2}(x_i^k)^2 Q_{ii}^k + \lambda (x_i^k)^p \\ \geq \; & x_i^k \|Q_i^k\|(-n\gamma) + x_i^k a_i^k + \lambda (x_i^k)^p \\ = \; & x_i^k \left[\lambda (x_i^k)^{p-1} - n\|Q_i^k\|\gamma + a_i^k\right]. \end{aligned}$$

Since $0 < x_i^k \leq L \leq \min\left\{(\frac{1}{\lambda}(n\|Q_i\|\gamma + |a_i|))^{\frac{1}{p-1}}, \forall i\right\}$, and $-1 < p - 1 < 0$, we have

$$\lambda (x_i^k)^{p-1} - n\|Q_i\|\gamma + a_i \geq 0.$$

Therefore,

$$h(x^k) - h(x^{k+1}) \geq 0.$$

□

Now consider the case that the objective function is not locally strongly convex. We show that the objective function would decease at least a constant value M by a line search along direction d^k.

Lemma 3 *For any* $k \geq 0$ *and* $L > 0$, *if* $0 < \tau < \frac{2p(1-p)(2-p)(3-p)L^p}{4!n\gamma^2}$, $x_i^k > L$ *for all index* i, $(d^k)^T \nabla^2 h(x^k) d^k \leq \tau \|d^k\|^2$, $\|x^k\|_\infty \leq \gamma$, *and let* $x^{k+1} = \arg\min_{x \in \{x^k + \zeta^k d^k, x^k - \zeta^k d^k\}} h(x)$, *where* $\zeta^k = \max\{t | x^k + td^k \geq 0, x^k - td^k \geq 0\}$, *then*

$$h(x^k) - h(x^{k+1}) \geq M > 0,$$

where $M = \frac{1}{4!}p(1-p)(2-p)(3-p)L^p - \frac{1}{2}\tau n\gamma^2$.

Proof We show it by two steps.

Step 1, when $Q^k = Q, a^k = a$, i.e., case 1 never happens before iteration k.

We start from the Taylor expansion. Since $h(x^{k+1}) = min\{h(x^k + \zeta^k d^k), h(x^k - \zeta^k d^k)\}$, it is sufficient to show that

$$\frac{h(x^k + \zeta^k d^k) + h(x^k - \zeta^k d^k)}{2} \leq h(x^k) - M.$$

From Taylor expansion, we get

$$\begin{aligned} & \frac{h(x^k+\zeta^k d^k)+h(x^k-\zeta^k d^k)}{2} \\ = \; & f(x^k)+g(x^k) \\ & + \frac{1}{2}(\zeta^k)^2(d^k)^T\nabla^2 f(x^k)d^k + \frac{1}{2}(\zeta^k)^2(d^k)^T\nabla^2 g(x^k)d^k \\ & + \sum_{q=2}^{+\infty}\sum_{i=1}^{n}\frac{1}{2q!}[\prod_{j=0}^{2q-1}(p-j)](x_i^k)^{p-2q}(\zeta^k d_i^k)^{2q}(-1)^{2q}. \end{aligned} \tag{10}$$

For the second term of (10), since $(d^k)^T\nabla^2 h(x^k)d^k \le \tau\|d^k\|^2$, we have

$$\begin{aligned} & \frac{1}{2}(\zeta^k)^2\left((d^k)^T\nabla^2 f(x^k)d^k + \frac{1}{2}(d^k)^T\nabla^2 g(x^k)d^k\right) \\ \le \; & \frac{1}{2}(\zeta^k)^2\tau\|d^k\|^2 \\ = \; & \frac{1}{2}\tau\sum_{i=1}^{n}(x_i^k)^2(\frac{\zeta^k d_i^k}{x_i^k})^2. \end{aligned}$$

By the definition of ζ^k, $\frac{\zeta^k|d_i^k|}{x_i^k} \le 1$. Then it follows

$$\frac{1}{2}(\zeta^k)^2\left((d^k)^T\nabla^2 f(x^k)d^k + \frac{1}{2}(d^k)^T\nabla^2 g(x^k)d^k\right) = \frac{1}{2}\tau\sum_{i=1}^{n}(x_i^k)^2 \le \frac{1}{2}\tau n\gamma^2, \tag{11}$$

where the last inequality uses the assumption that $\|x^k\|_\infty \le \gamma$.

For the third term of (10), we notice that

$$\frac{1}{2q!}[\prod_{j=0}^{2q-1}(p-j)](x_i^k)^{p-2q}(\zeta^k d_i^k)^{2q}(-1)^{2q} \le 0, \forall q \ge 1,$$

which implies

$$\begin{aligned} & \sum_{q=2}^{+\infty}\sum_{i=1}^{n}\frac{1}{2q!}[\prod_{j=0}^{2q-1}(p-j)]x_i^{p-2q}(\zeta^k d_i^k)^{2q}(-1)^{2q} \\ \le \; & -\frac{1}{4!}p(1-p)(2-p)(3-p)\sum_{i=1}^{n}(x_i^k)^{p-4}(\zeta^k d_i^k)^4. \end{aligned} \tag{12}$$

By the definition of ζ^k, we have $\max\{\frac{\zeta^k|d_i^k|}{x_i^k}\} = 1$. Let $j = argmax_i\{\frac{\zeta^k|d_i^k|}{x_i^k}\}$, we derive

$$\begin{aligned} & -\frac{1}{4!}p(1-p)(2-p)(3-p)\sum_{i=1}^{n}(x_i^k)^{p-4}(\zeta^k d_i^k)^4 \\ \le \; & -\frac{1}{4!}p(1-p)(2-p)(3-p)(x_j^k)^p \\ < \; & -\frac{1}{4!}p(1-p)(2-p)(3-p)L^p \end{aligned} \tag{13}$$

where the last inequality follows from $-(x_j^k)^p < -L^p$.

Upon substituting (12), (13) and (11) into (10), we obtain

$$\begin{aligned} & \frac{h(x^k+\zeta^k d^k)+h(x^k-\zeta^k d^k)}{2} \\ < \quad & h(x^k)+\frac{1}{2}\tau n\gamma^2-\frac{1}{4!}p(1-p)(2-p)(3-p)L^p \\ = \quad & h(x^k)-M. \end{aligned}$$

Due to $0<\tau<\frac{2p(1-p)(2-p)(3-p)L^p}{4!n\gamma^2}$, we have $M>0$. Therefore

$$h(x^k)-h(x^{k+1})\geq M>0.$$

Step 2: case 1 happens previously. Suppose the problem dimension is $m(<n)$ now. We can still follow the algorithm and prove the similar result, i.e., we choose $\tau\leq\tau'<\frac{2p(1-p)(2-p)(3-p)L^p}{4!m\gamma^2}$ and define constant M' and prove accordingly. Then

$$h(x^k)-h(x^{k+1})\geq M'>0.$$

One can observe that we can always use the same τ as the one in Step 1 since $\tau\leq\tau'$. And from the definition process we can easily observe that $M'\geq M>0$.

So in every case, we always have

$$h(x^k)-h(x^{k+1})\geq M>0.$$

□

Next, Lemma 4 shows that when the objective function is locally strongly convex, taking a small step along d^k will lead to a lesser objective function value and a constant exponential rate shrink on the potential function value.

Lemma 4 *For any $k\geq 0,\tau>0$ and $L>0$, if $x_i^k>L$ for all index i, $(d^k)^T\nabla^2 h(x^k)d^k>\tau\|d^k\|^2$,and let $x^{k+1}=x^k+sd^k$, where $0<s<\min\{\frac{1}{\mu}(\tau-\frac{\delta\beta}{2}),\nu,1\}$, $0<\delta<\min\{\frac{2\tau}{\beta},1\}$, $0<\nu<1$, $\mu=\frac{\beta}{2}+\frac{1}{\beta}\{[\lambda p(1-p)]^2[L(1-\nu)]^{2p-4}+\|\beta I-Q\|^2\}$, then we have*

1. *$h(x^k)-h(x^{k+1})\geq 0$,*
2. *$\Delta L^{k+1}\leq(1-s\delta)\Delta L^k$.*

Proof We prove it by two steps similar to Lemma 3.

Step 1. Case 1 never happened before.

1. Firstly, we prove that $h(x^k)-h(x^{k+1})\geq 0$. Recall that $L^k(x)=f(x)+\nabla g(x^k)^T(x-x^k)+g(x^k)$. It follows $L^k(x^k)=h(x^k)$. Combining with the fact that $L^k(x^{k+1})\geq h(x^{k+1})$, we have

$$L^k(x^k)-L^k(x^{k+1})\leq h(x^k)-h(x^{k+1}),$$

Then we show that $L_k(x^k) - L_k(x^{k+1}) \geq 0$ by the optimality of $\bar{x}^k$. Using Taylor expansion, we can obtain

$$L^k(x^k) - L^k(x^{k+1}) = -s\nabla L^k(\bar{x}^k)^T d^k + s(1 - \frac{s}{2})\beta\|d^k\|^2.$$

Since $\bar{x}^k \in \arg\min_{x\geq 0} L^k(x)$, the first order optimality condition implies $\nabla L^k(\bar{x}^k)^T(-d^k) \geq 0$. Therefore,

$$0 \leq L^k(x^k) - L^k(x^{k+1}) \leq h(x^k) - h(x^{k+1}).$$

2. Secondly, we prove that $\Delta L_{k+1} \leq (1 - s\delta)\Delta L_k$. In order to bound ΔL_{k+1} by $(1 - s\delta)\Delta L_k$, we first write ΔL_{k+1} into a combination of $L_k(x)$ and $\nabla g(x)$. By the definition of ΔL_k and $L_k(x)$, we get

$$\begin{aligned}
\Delta L^{k+1} = \;& L^{k+1}(x^{k+1}) - L^{k+1}(\bar{x}^{k+1}) \\
= \;& f(x^{k+1}) - f(\bar{x}^{k+1}) - \nabla g(x^{k+1})^T(\bar{x}^{k+1} - x^{k+1}) \\
= \;& L^k(x^{k+1}) - L^k(\bar{x}^{k+1}) \\
& + \left[\nabla g(x^k) - \nabla g(x^{k+1})\right]^T (\bar{x}^{k+1} - x^{k+1}) \\
= \;& L^k(x^{k+1}) - L^k(\bar{x}^k) \\
& + L^k(\bar{x}^k) - L^k(\bar{x}^{k+1}) + [\nabla g(x^k) - \nabla g(x^{k+1})]^T(\bar{x}^{k+1} - x^{k+1}).
\end{aligned} \tag{14}$$

Then we estimate $L^k(x^{k+1}) - L^k(\bar{x}^k)$ and $L^k(\bar{x}^k) - L^k(\bar{x}^{k+1})$ separately. For the first term of (14),

$$\begin{aligned}
& L^k(x^{k+1}) - L^k(\bar{x}^k) \\
= \;& -\nabla L^k(\bar{x}^k)^T(\bar{x}^k - x^{k+1}) + \frac{\beta}{2}(\bar{x}^k - x^{k+1})^T(\bar{x}^k - x^{k+1}) \\
= \;& -(1-s)\nabla L^k(\bar{x}^k)^T d^k + \frac{\beta}{2}(1-s)^2\|d^k\|^2 \\
= \;& (1-s)[-\nabla L^k(\bar{x}^k)^T d^k + \frac{\beta}{2}\|d^k\|^2 - \frac{\beta}{2}\|d^k\|^2] + \frac{\beta}{2}(1-s)^2\|d^k\|^2 \\
= \;& (1-s)\left[L^k(x^k) - L^k(\bar{x}^k)\right] - \frac{\beta}{2}s(1-s)\|d^k\|^2 \\
= \;& (1-s)\Delta L^k - \frac{\beta}{2}s(1-s)\|d^k\|^2,
\end{aligned} \tag{15}$$

where both the first and fourth equalities follow from the Taylor expansion.

Then we estimate the second term of (14). Let $w = \bar{x}^{k+1} - \bar{x}^k$.

$$\begin{aligned}
& L^k(\bar{x}^k) - L^k(\bar{x}^{k+1}) + \left[\nabla g(x^k) - \nabla g(x^{k+1})\right]^T (\bar{x}^{k+1} - x^{k+1}) \\
= & -\nabla L^k(\bar{x}^k)^T w - \frac{\beta}{2} w^T w \\
& + \left[\nabla g(x^k) - \nabla g(x^{k+1})\right]^T (w + \bar{x}^k - x^{k+1}) \\
\leq & -\frac{\beta}{2} w^T w + \left[\nabla g(x^k) - \nabla g(x^{k+1})\right]^T w \\
& + (1-s)(d^k)^T \left[\nabla g(x^k) - \nabla g(x^{k+1})\right] \\
\leq & \frac{1}{2\beta}\|\nabla g(x^k) - \nabla g(x^{k+1})\|^2 + (1-s)(d^k)^T \left[\nabla g(x^k) - \nabla g(x^{k+1})\right],
\end{aligned} \tag{16}$$

where the first inequality uses the fact that the first order optimality condition implies $-\nabla L^k(\bar{x}^k)^T w \leq 0$ because $\bar{x}^k \in \arg\min_{x \geq 0} L^k(x)$; and where the second inequality follows from the fact that

$$\begin{aligned}
\frac{1}{2\beta}\|\nabla g(x^k) - \nabla g(x^{k+1})\|^2 \quad + \quad & \frac{\beta}{2}\|w\|^2 - \left[\nabla g(x^k) - \nabla g(x^{k+1})\right]^T w \\
= \quad & \frac{1}{2\beta}\| \left[\nabla g(x^k) - \nabla g(x^{k+1})\right] - \beta w\|^2 \geq 0.
\end{aligned}$$

Upon substituting (15) and (16) into (14), we have

$$\begin{aligned}
\Delta L^{k+1} \quad \leq \quad & (1-s)\Delta L^k - \frac{\beta}{2} s(1-s)\|d^k\|^2 + \frac{1}{2\beta}\|\nabla g(x^k) \\
& - \nabla g(x^{k+1})\|^2 + (1-s)(d^k)^T \left[\nabla g(x^k) - \nabla g(x^{k+1})\right].
\end{aligned}$$

Note that $\nabla g(x)$ is Lipschitz continuous when x is away from the boundary. Then by Lemma 8 (see Appendix),

$$\|\nabla g(x^k) - \nabla g(x^{k+1})\|^2 \leq 2s^2\{[\lambda p(1-p)]^2 [L(1-\nu)]^{2p-4} + \|Q - \beta I\|^2\}\|d^k\|^2. \tag{17}$$

Since $g(x)$ is the sum of L_p function and a concave quadratic function, Lemma 9 shows that

$$(d^k)^T \left[\nabla g(x^k) - \nabla g(x^{k+1})\right] \leq \frac{-s}{1-s}(d^k)^T \nabla^2 g(x^k) d^k. \tag{18}$$

Let $\eta = [\lambda p(1-p)]^2[L(1-\nu)]^{2p-4} + \|Q - \beta I\|^2$. Combining (17) and (18) with $(d^k)^T\nabla^2 h(x^k)d^k > \tau\|d^k\|^2$, we end with

$$\begin{aligned}
\Delta L^{k+1} \leq\ & (1-s)\Delta L^k - \frac{\beta}{2}s(1-s)\|d^k\|^2 \\
& + \frac{s^2}{\beta}\eta\|d^k\|^2 - s(d^k)^T\nabla^2 g(x^k)d^k \\
=\ & \Delta L^k - s\left(\Delta L^k + (d^k)^T\nabla^2 g(x^k)d^k + \frac{\beta}{2}\|d^k\|^2\right) \\
& + s^2(\frac{\beta}{2} + \frac{\eta}{\beta})\|d^k\|^2 \\
<\ & \Delta L^k - s\left(\Delta L^k - \frac{\beta}{2}\|d^k\|^2 + \tau\|d^k\|^2\right) \\
& + s^2(\frac{\beta}{2} + \frac{\eta}{\beta})\|d^k\|^2.
\end{aligned}$$

Note that $\mu = \frac{\beta}{2} + \frac{\eta}{\beta}$. By Lemma 10 (see Appendix) that if $0 < s < \min\{\frac{1}{\mu}(\tau - \frac{\delta\beta}{2}), 1\}$, then

$$\Delta L^k - s(\Delta L^k - \frac{\beta}{2}\|d^k\|^2 + \tau\|d^k\|^2) + s^2(\mu\|d^k\|^2) \leq (1 - s\delta)\Delta L^k.$$

We conclude that

$$\Delta L^{k+1} \leq (1 - s\delta)\Delta L^k.$$

Step 2. Case 1 happened before.
We can define μ' accordingly. Notice that in this case

$$\mu' = \frac{\beta}{2} + \frac{1}{\beta}\left\{[\lambda p(1-p)]^2[L(1-\nu)]^{2p-4} + \|\beta I - Q^k\|^2\right\}.$$

And $\beta I - Q^k \succeq 0$. So we know $\|\beta I - Q\|^2 \geq \|\beta I - Q^k\|^2$, i.e., $\mu' \leq \mu$. Then we can always find a larger $s' > s$ such that

$$\Delta L^{k+1} \leq (1 - s'\delta)\Delta L^k \leq (1 - s\delta)\Delta L^k.$$

So in this case,the potential function value shrinks with a larger exponential rate. We complete the proof. □

In summary, if $\|x^k\|_\infty \leq \gamma$, and s, τ, and L are defined accordingly in lemma 2, lemma 3, and lemma 4, then the claims above all hold and can be summarized as follows.

- Case 1 would at most happen n times.
- Case 2 would happen at most $\lfloor\frac{h(x^0)}{M}\rfloor$ times because of two facts: the objective value of problem (1) is lower bounded by 0; and in any case, the objective function keeps non-increasing.
- Each time Case 3 happens, the potential function value shrinks by a constant exponential rate.

– In case 1 and case 2, the potential function value is always bounded by $h(x_0)$ according to its definition.

From the observations above, we first have the following lemma:

Lemma 5 *Starting from a given initial solution x^0, Case 1 and Case 2 of the algorithm would happen no more than $n + \lfloor \frac{h(x^0)}{M} \rfloor$ times.*

We are now ready to present the main result of this paper:

Theorem 1 *For any $\epsilon \in (0,1)$, Algorithm 1 obtains an ϵ-KKT point of problem (1) in no more than $O((n + \lfloor \frac{h(x^0)}{M} \rfloor + 1) \left\lceil \ln \frac{h(x^0)2\beta}{\epsilon^2} \ln \frac{1}{1-s\delta} \right\rceil)$ steps.*

Proof For any fixed number $\epsilon \in (0,1)$, let $\theta = \left\lceil \frac{\ln \frac{h(x^0)2\beta}{\epsilon^2}}{-\ln(1-s\delta)} \right\rceil$. We will show that there exists a $k \leq (n + \lfloor \frac{h(x^0)}{M} \rfloor + 1)(\theta + 1)$, such that x^k is an ϵ-KKT point of problem (1).

1. If there exists an l such that $l \leq (n + \lfloor \frac{h(x^0)}{M} \rfloor + 1) * (\theta + 1),\ x^l = 0$, then x^l is an ϵ-KKT point of problem (1).
2. If $\forall l \leq (n + \lfloor \frac{h(x^0)}{M} \rfloor + 1) * (\theta + 1), x^l \neq 0$:
   Let $N = (n + \lfloor \frac{h(x^0)}{M} \rfloor + 1) * (\theta + 1)$
   We divide the iteration sequence $\{x^1, x^2, \cdots, x^N\}$ into $(n + \lfloor \frac{h(x^0)}{M} \rfloor + 1)$ contiguous segments. Segment j consists of elements $\{x^{j(\theta+1)+1}, x^{j(\theta+1)+2}, \cdots, x^{(j+1)(\theta+1)}\}$.
   We can observe that the sequence is divided into $(n + \lfloor \frac{h(x^0)}{M} \rfloor + 1)$ segments. By Lemma 5, x^i's at which Case 1 or Case 2 happens would appear at most $(n + \lfloor \frac{h(x^0)}{M} \rfloor)$ times. Therefore there must exists one segment in which neither Case 1 nor Case 2 would happen by the Pigeonhole Principle.
   Suppose this segment includes sequence $x^{k-\theta}, x^{k-\theta+1}, \ldots, x^k$, and $k \leq (n + \lfloor \frac{h(x^0)}{M} \rfloor + 1) * (\theta + 1)$. From lemma 4 , we have $\Delta L_{l+1} \leq (1 - s\delta)\Delta L_l$ for any l in this segment. Therefore,

$$\begin{aligned} \Delta L_k \leq & \ \Delta L_{k-\theta} * (1 - s\delta)^\theta \\ \leq & \ h(x^0) * \frac{\epsilon^2}{h(x^0)2\beta} \\ \leq & \ \frac{\epsilon^2}{2\beta}. \end{aligned}$$

   By lemma 1, we have x^k is an ϵ-KKT point of problem (1). Therefore the algorithm obtains an ϵ-KKT point of problem (1) in no more than $O(n \log \epsilon^{-1})$ steps.

□

4 The Case with Linear Constraints

In this section, we consider a more general case of problem (1) by adding affine linear equality constraints.

$$\begin{array}{ll} \text{Minimize} & h(x) \\ \text{Subject to} & Ax = b \\ & x \geq 0 \end{array} \tag{19}$$

where $h(x)$ is defined the same as problem (1), $A \in R^{m*n}$, and $b \in R^m$. Also let F_p denote the feasible region of problem (19). In this section, we will show that under mild assumptions, the analysis of our algorithm can also be applied to problem (19).

We first make the following assumptions.

Assumption 3 *The optimal value of problem (19) is lower bounded by 0.*

Assumption 4 *F_p is bounded and there exists an r such that $sup\{\|x\|_\infty | x \in F_p\} \leq \gamma$.*

Similar to problem (1), Assumption 3 is only for the simplicity of our analysis. It will not change our complexity result if the optimal value of problem (19) is lower bounded by any other value. Assumption 4 is usual in linearly constrained problems and it can be satisfied in many situations, for example the simplex constraint, i.e. $e^T x = 1$.

We define the ϵ-KKT conditions of problem (19) as follows:

Definition 2 For any $\epsilon \in (0, 1)$, we call $x^* \in F_p$ an ϵ-KKT point of problem (19), if there is y^* , such that

$$(\nabla h(x^*) - A^T y^*)^T x^* \leq \epsilon, \tag{20a}$$

$$\nabla h(x^*) - A^T y^* \geq 0. \tag{20b}$$

Note that this ϵ-KKT condition is little stronger than that in definition 1, since only the complementary condition has been relaxed by ϵ.

Following the idea in problem (1), for any $z \in F_p$, we consider the problem:

$$\begin{array}{ll} \text{Minimize} & L_z(x) = f(x) + g(z) + \nabla g(z)^T (x - z) \\ \text{Subject to} & x \in F_p \end{array} \tag{21}$$

Let $\bar{z}$ be minimizer of problem (21), then the potential function $\Delta L(z) : R^n \to R$ is defined as

$$\Delta L(z) = L_z(z) - L_z(\bar{z}).$$

$\Delta L(z)$ also has a similar relationship with ϵ-KKT point, which is given as follows:

Lemma 6 *For any $z \in F_p$, if $\Delta L(z) \leq \frac{\epsilon^2}{2n\beta\gamma^2}$, then z is an ϵ-KKT point of problem (19).*

Proof First, we consider the following linear minimization problem:

$$\begin{array}{ll} \text{Minimize} & \nabla h(z)(x-z) \\ \text{Subject to} & x \in F_p \end{array} \tag{22}$$

Since F_p is compact, there must be a minimizer $\hat{z}$ of problem (22). Let $\hat{y}$ be the optimal dual variable corresponding to $\hat{z}$, then the KKT conditions are given as follows:

$$(\nabla h(z) - A^T\hat{y})^T\hat{z} = 0, \tag{23a}$$

$$\nabla h(z) - A^T\hat{y} \geq 0. \tag{23b}$$

Consider the ϵ-KKT conditions of problem (19) and Let $x^* = z, y^* = \hat{y}$. Then for the first order condition (20a), we get

$$\begin{aligned} & (\nabla h(x^*) - A^T y^*)^T x^* \\ = \; & (\nabla h(z) - A^T\hat{y})^T z \\ = \; & \left(\nabla h(z) - A^T\hat{y}\right)^T (\hat{z} + z - \hat{z}) \\ = \; & \nabla h(z)^T(z-\hat{z}) - y^T A(z-\hat{z}) \\ = \; & -\nabla h(z)^T(\hat{z}-z), \end{aligned}$$

where the last equality follows from (23a). We only need to show that $-\nabla h(z)^T$ $(\hat{z}-z) \leq \epsilon$, then combining with $\nabla h(z) - A^T\hat{y} \geq 0$ and $z \in F_p$ we could conclude z is an ϵ-KKT point of problem (19).

Since $\Delta L(z) = L_z(z) - L_z(\bar{z})$, and $\bar{z} \in \arg\min_{x\in F_p} L_z(x)$, the following condition holds for any $0 \leq s \leq 1$:

$$\Delta L(z) \geq L_z(z) - L_z(z + s(\hat{z}-z)) = -s\nabla h(z)^T(\hat{z}-z) - \frac{\beta}{2}s^2(\hat{z}-z)^T(\hat{z}-z).$$

Suppose -$\nabla h(z)^T(\hat{z}-z) > \epsilon$ (for contradiction), then we have

$$\begin{aligned} \Delta L(z) > \; & s\epsilon - \frac{\beta}{2}s^2(\hat{z}-z)^T(\hat{z}-z) \\ \geq \; & s\epsilon - \frac{1}{2}s^2\beta n\gamma^2. \end{aligned}$$

where the last inequality uses the assumption 4 that $sup\{\|x\|_\infty | x \in F_p\} \leq \gamma$. Let $s = \frac{\epsilon}{n\beta\gamma^2}$. Thus

$$\Delta L(z) > \frac{\epsilon^2}{2n\beta\gamma},$$

which contradicts the assumption. Therefore, z is an ϵ-KKT point of problem (19). □

The following theorem presents a lower bound condition under which the problem (19) can be solved by Algorithm 1 with the same worst-case complexity as the problem (1).

Theorem 2 *If we have a constant number $L > 0$ such that for any $x \in F_p$, if $\exists j, x_j \leq L$, we could find an $x^+ \in F_p$, such that $\exists i, x_i^+ = 0$ and $h(x) - h(x^+) \geq 0$, in polynomial times, then for any $\epsilon \in (0,1)$, we could use Algorithm 1 to obtain an ϵ-KKT point of problem (19) in no more than $O(\log(\frac{1}{\epsilon}))$ steps.*

The proof of theorem 2 use almost the same logic as that of theorem 1, and is omitted for brevity. Note that the procedure in Case 1 of Algorithm 1 should be adjusted for the new lower bound described in the theorem.

The conditions in the theorem 2 hold for many problems. For example,

$$\begin{array}{ll} \text{Minimize} & h(x) \\ \text{s.t} & e^T x = 1 \\ & x \geq 0. \end{array} \tag{24}$$

The following lemma provides a lower bound for problem (24).

Lemma 7 *For the problem (24), if*

$$0 < L \leq min\{[(1-p)\frac{|Q_{ii} - Q_{jj}|}{2} + \|Q_i - Q_j\|\sqrt{n}r - (a_i - a_j)]^{\frac{1}{p-1}} | i \neq j\},$$

and $L \leq 1$, then for any $x \in \{x | e^T x = 1, x \geq 0\}$ such that $\exists j, x_j \leq L$, we could find a x^+ such that $\exists i, x_i^+ = 0$ and $h(x) - h(x^+) \geq 0$.

Proof We select i such that $x_i = min\{x_j | x_j \leq L\}$, and randomly select a index j which is different from i. Let $\rho = x_i$, and

$$x^+ = x - \rho e_i + \rho e_j.$$

Then, from the definition of $h(x)$, we have

$$\begin{aligned} & h(x) - h(x^+) \\ = \; & \frac{1}{2}\rho^2(Q_{ii} - Q_{jj}) + \rho(Q_i - Q_j)(x - \rho e_i) + \rho(a_i - a_j) \\ & + \rho^p + \left(x_j^p - (x_j + \rho)^p\right). \end{aligned}$$

By the mean-value theorem, there exists a $x_j' \in [x_j, x_j + \rho]$ such that $x_j^p - (x_j + \rho)^p = -p\rho(x_j')^{p-1}$. Due to $x_j' \geq x_j \geq \rho$, we have

$$x_j^p - (x_j + \rho)^p \geq -p\rho^p.$$

Also, we know that $0 \leq \rho \leq L \leq 1$ and $\|x - \rho e_i\| \leq \sqrt{n}r$, which imply

$$\begin{aligned} & h(x) - h(x^+) \\ \geq \; & -\frac{1}{2}\rho|Q_{ii} - Q_{jj}| - \rho\|Q_i - Q_j\|\sqrt{n}r + \rho(a_i - a_j) \\ & + \rho^p(1-p). \end{aligned}$$

In conjunction with $\rho \leq L \leq [(1-p)\frac{|Q_{ii}-Q_{jj}|}{2} + \|Q_i - Q_j\|\sqrt{n}r - (a_i - a_j)]^{\frac{1}{p-1}}$ and $-1 < p - 1 < 0$, we have $h(x) - h(x^+) \geq 0$. □

One application of problem (24) is the sparse portfolio selection problem, which aims to select a limit number of securities with minimal estimated variance and maximal expected return. Given the estimated covariance matrix $Q \in R^{n*n}$, and the estimated return vector $r \in R^n$ of a set of securities, this problem can be modelled as problem (19).

$$\begin{array}{ll} \text{Minimize} & \frac{1}{2}x^T Qx - ur^T x + \lambda \sum_i x_i^p \\ \text{Subject to} & e^T x = 1 \\ & x \geq 0 \end{array}$$

where $u \in (0, +\infty)$ and $\lambda \in (0, +\infty)$. Interested reader may refer to [5].

References

1. W. Bian, X. Chen and Y. Ye, Complexity Analysis of Interior Point Algorithms for Non-Lipschitz and Nonconvex Minimization, *Preprint*, 2012.
2. W. Bian, X. Chen, Worst-Case Complexity of Smoothing Quadratic Regularization Methods for Non-Lipschitzian Optimization, *SIAM Journal on Optimization*, 2013.
3. R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, *IEEE Signal Processing Letters*, 14 (2007), 707-710.
4. R. Chartrand and V. Staneva, Restricted isometry properties and nonconvex compressive sensing, *Inverse Problem*, 24 (2008), 1-14.
5. C. Chen, X. Li, C. Tolman, S. Wang and Y. Ye, Sparse Portfolio Selection via Quasi-Norm Regularization, Working paper, 2013. http://arxiv.org/abs/1312.6350
6. X. Chen, F. Xu and Y. Ye, Lower Bound Theory of Nonzero Entries in Solutions of ℓ_2-ℓ_p Minimization, *SIAM Journal on Scientific Computing*, 32(5), 2832-2852, 2010.
7. X. Chen, D. Ge, Z. Wang and Y. Ye, Complexity of unconstrained $$L_2-L_p$$ minimization, *Mathematical Programming*, 143(1-2), 371-383, 2014.
8. J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of American Statistical Society*, 96 (2001), 1348-1360.
9. S. Foucart and M. J. Lai, Sparsest solutions of under-determined Linear Systems via l_q minimization for $0 < q \leq 1$, *Applied and Computational Harmonic Analysis*, 26 (2009), 395-407.
10. I. E. Frank and J. H. Freidman, A statistical view of some chemometrics regression tools (with discussion), *Technometrics*, 35(1993), 109-148.
11. M. R. Garey and D. S. Johnson, "Strong" NP-Completeness results: motivation, examples, and implications, *Journal of the Association of Computing Machinery*, 25 (1978), 499-508.
12. M. R. Garey and D. S. Johnson, Computers and Intractability; A Guide to the Theory of NP-Completeness, W. H. Freeman, New York, 1979.
13. D. Ge, X. Jiang and Y. Ye, A note on the complexity of L_p minimization, *Mathematical Programming*, 129(2011), 285-299.
14. J. Huang, J. L. Horowitz and S. Ma, Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *The Annals of Statistics*, 36 (2008), 587-613.
15. K. Knight and W.J. Fu, Asymptotics for lasso-type estimators, *The Annals of Statistics*, 28 (2000), 1356-1378.
16. M. Lai and Y. Wang, An unconstrained l_q minimization with $0 < q < 1$ for sparse solution of under-determined linear systems, *SIAM J. Optimization*, 21 (2011), 82-101.
17. B. K. Natarajan, Sparse approximate solutions to linear systems, *SIAM J. Computing*, 24 (1995), 227-234.
18. J. Nocedal and S.J. Wright, Numerical Optimization, 2nd Edition, Springer, New York, 2006.

19. R. Tibshirani, Regression shrinkage and selection via the Lasso, *J Royal Statistical Society B*, 58 (1996), 267-288.
20. Y. Nesterov and B. T. Polyak, Cubic regularization of Newton method and its global performance, *Mathematical Programming*, 108(1), 177-205, 2006.
21. V. Vazirani, Approximation Algorithms, Springer, Berlin, (2003).
22. Y. Ye, On the complexity of approximating a KKT point of quadratic programming, *Mathematical Programming* Vol. 80, No. 2, Pg. 1998, 2009.

Appendix

Let $g_1(x) = \lambda \sum_i x_i^p$ and $g_2(x) = \frac{1}{2}x^T(Q - \beta I)x$. Note that $g(x) = g_1(x) + g_2(x)$.

Lemma 8 *If $x + d \geq 0$, $x_i \geq L$ for all index i, and $0 < s \leq \nu < 1$, then*

$$\|\nabla g(x + sd) - \nabla g(x)\|^2 \leq 2s^2\{[\lambda p(1-p)]^2[L(1-\nu)]^{2p-4} + \|Q - \beta I\|^2\}\|d\|^2$$

Proof We consider $g_1(x)$ and $g_2(x)$ separately.

By the definition of $g_1(x)$, we have

$$\begin{aligned}
&\|\nabla g_1(x + sd) - \nabla g_1(x)\| \\
\leq &\int_0^1 \|\nabla^2 g_1(x + tsd)sd\| \, \mathrm{d}t \\
= &s\lambda p(1-p) \int_0^1 \sqrt{\sum_{i=1}^n (x_i + std_i)^{2p-4} d_i^2} \, \mathrm{d}t.
\end{aligned}$$

Since $x_i + tsd_i \geq (1-s)L$, for all index i, we have

$$\int_0^1 \sqrt{\sum_{i=1}^n (x_i + std_i)^{2p-4} d_i^2} \mathrm{d}t \leq ((1-s)L)^{p-2} \|d\|.$$

In conjunction with $s \leq \nu$ and $0 < p < 1$, we get

$$\|\nabla g_1(x + sd) - \nabla g_1(x)\|^2 \leq s^2[\lambda p(1-p)]^2[L(1-\nu)]^{2p-4}\|d\|^2.$$

For $g_2(x)$, we have

$$\|\nabla g_2(x + sd) - \nabla g_2(x)\|^2 \leq s^2\|Q - \beta I\|^2\|d\|^2.$$

Thus,

$$\begin{aligned}
&\|\nabla g(x + sd) - \nabla g(x)\|^2 \\
\leq &2\left(\|\nabla g_1(x + sd) - \nabla g_1(x)\|^2 + \|\nabla g_2(x + sd) - \nabla g_2(x)\|^2\right) \\
\leq &2s^2\left\{[\lambda p(1-p)]^2[L(1-\nu)]^{2p-4} + \|Q - \beta I\|^2\right\}\|d\|^2.
\end{aligned}$$

□

Remark: Lemma 8 is the consequence of that $g(x)$'s gradient is Lipschitz continuous in the region that $x_i \geq L(1-\nu)$ for all index i.

Lemma 9 *If $x+d \geq 0$, $x \geq 0$ and $0 < s \leq 1$, then $d^T\left(\nabla g(x) - \nabla g(x+sd)\right) \leq \frac{-s}{1-s} d^T \nabla^2 g(x) d$.*

Proof We consider $g_1(x)$ and $g_2(x)$ separately. From the definition of $g_1(x)$, we have

$$d^T(\nabla g_1(x) - \nabla g_1(x+sd)) = \lambda p \sum_i (d_i(x_i^{p-1} - (x_i + sd_i)^{p-1})),$$

and

$$-\frac{s}{1-s} d^T \nabla^2 g_1(x) d = \frac{s}{1-s} \lambda p(1-p) \sum_i x_i^{p-2} d_i^2.$$

To prove that $d^T\left(\nabla g_1(x) - \nabla g_1(x+sd)\right) \leq \frac{-s}{1-s} d^T \nabla^2 g_1(x) d$, it suffices to show that $d_i(x_i^{p-1}) - (x_i + sd_i)^{p-1} \leq \frac{s}{1-s}(1-p) x_i^{p-2} d_i^2, \forall i$.

If $d_i \geq 0$, by the mean-value theorem, there exists a $x_i' \in [x_i, x_i + sd_i]$, such that $x_i^{p-1} - (x_i + sd_i)^{p-1} = x_i'^{p-2}(1-p) sd_i$. Since $0 < p < 1$, we have

$$d_i x_i'^{p-2}(1-p) sd_i \leq s(1-p) x_i^{p-2} d_i^2 \leq \frac{s}{1-s}(1-p) x_i^{p-2} d_i^2.$$

If $d_i < 0$, We let $v = -\frac{d_i}{x_i}, 0 < v \leq 1$. By multiplying $\frac{x_i^{1-p}}{-d_i}$ on both sides,

$$d_i \left(x_i^{p-1} - (x_i + sd_i)^{p-1}\right) \leq \frac{s}{1-s}(1-p) x_i^{p-2} d_i^2,$$

can be simplified as

$$(1-sv)^{p-1} - 1 \leq \frac{sv}{1-v}(1-p).$$

Since

$$\frac{sv}{1-v}(1-p) \leq \frac{sv}{1-sv}(1-p),$$

letting $u = sv$, it is sufficient to show that $(1-u)^{p-1} - 1 \leq \frac{u}{1-u}(1-p)$. Consider function $\omega(u) = (1-u)^p - (1-u) - u(1-p)$, $0 < u \leq 1$. A simple calculation shows that $\omega(u)$ is a decreasing function. Hence $\omega(u) \leq \lim_{u \to 0} \omega(u) = 0$. It follows that $(1-u)^{p-1} - 1 \leq \frac{u}{1-u}(1-p), \forall 0 < u \leq 1$. Thus, we obtain

$$p d_i \left(x_i^{p-1} - (x_i + sd_i)^{p-1}\right) \leq \frac{s}{1-s} p(1-p) d_i x_i^{p-2} d_i.$$

For $g_2(x)$, since $0 < s \leq 1$, we have

$$\begin{aligned} & d^T\left(\nabla g_2(x) - \nabla g_2(x+sd)\right) \\ = \; & d^T(\beta I - Q) d \\ \leq \; & \frac{1}{1-s} d^T(\beta I - Q) d \\ = \; & \frac{-s}{1-s} d^T \nabla^2 g_2(x) d. \end{aligned}$$

Therefore,

$$\begin{aligned} & d^T(\nabla g(x) - \nabla g(x+sd)) \\ = \; & d^T\left(\nabla g_1(x) - \nabla g_1(x+sd) + \nabla g_2(x) - \nabla g_2(x+sd)\right) \\ \leq \; & \frac{-s}{1-s} d^T \nabla^2 g_1(x) d + \frac{-s}{1-s} d^T \nabla^2 g_2(x) d \\ = \; & \frac{-s}{1-s} d^T \nabla^2 g(x) d. \end{aligned}$$

□

Remark: Lemma 9 is a result of the special structure of function $g(x)$, which is the sum of L_p function and a concave quadratic function.

Lemma 10 *If $\Delta L \geq \frac{\beta}{2}\|d\|^2$, then*

$$\Delta L - s(\Delta L - \frac{\beta}{2}\|d\|^2 + \tau\|d\|^2) + s^2(\mu\|d\|^2) \leq (1 - s\delta)\Delta L,$$

where $\tau > 0$, $\mu > 0$, $0 < \delta < min\{\frac{2\tau}{\beta}, 1\}$, and $0 < s \leq min\{\frac{1}{u}(\tau - \frac{\delta\beta}{2}), 1\}$

Proof

$$\begin{aligned} & \Delta L - s(\Delta L - \frac{\beta}{2}\|d\|^2 + \tau\|d\|^2) + s^2\mu\|d\|^2 - (1 - s\delta)\Delta L \\ = & - s(\Delta L - \frac{\beta}{2}\|d\|^2 + \tau\|d\|^2) + s^2\mu\|d\|^2 + s\delta\Delta L \\ = & s[(s\mu + \frac{\beta}{2} - \tau)\|d\|^2 - (1-\delta)\Delta L] \\ \leq & s[(s\mu + \frac{\beta}{2} - \tau)\|d\|^2 - \frac{(1-\delta)\beta}{2}\|d\|^2] \\ = & s\|d\|^2(s\mu - \tau + \frac{\delta\beta}{2}), \end{aligned}$$

where the inequality follows from $\frac{\beta}{2}\|d\|^2 \leq \Delta L$. Due to $0 < \delta < \frac{2\tau}{\beta}$ and $s \leq \frac{1}{\mu}(\tau - \frac{\delta\beta}{2})$, we have

$$s\mu - \tau + \frac{\delta\beta}{2} \leq 0.$$

Therefore, $\Delta L - s(\Delta L - \frac{\beta}{2}\|d\|^2 + \tau\|d\|^2) + s^2(\mu\|d\|^2) \leq (1 - s\delta)\Delta L$. □