Sparse estimation via ℓ_q optimization method in high-dimensional linear regression

Xin Li[*] Yaohua Hu[†] Chong Li[‡] Xiaoqi Yang[§] Tianzi Jiang[¶]

Abstract

In this paper, we discuss the statistical properties of the ℓ_q optimization methods ($0 < q \leq 1$), including the ℓ_q minimization method and the ℓ_q regularization method, for estimating a sparse parameter from noisy observations in high-dimensional linear regression with either a deterministic or random design. For this purpose, we introduce a general q-restricted eigenvalue condition (REC) and provide its sufficient conditions in terms of several widely-used regularity conditions such as sparse eigenvalue condition, restricted isometry property, and mutual incoherence property. By virtue of the q-REC, we exhibit the stable recovery property of the ℓ_q optimization methods for either deterministic or random designs by showing that the ℓ_2 recovery bound $O(\epsilon^2)$ for the ℓ_q minimization method and the oracle inequality and ℓ_2 recovery bound $O(\lambda^{\frac{2}{2-q}}s)$ for the ℓ_q regularization method hold respectively with high probability. The results in this paper are nonasymptotic. The numerical results verify the established statistical property and demonstrate the advantages of the ℓ_q regularization method over the classical Lasso.

Keywords: sparse estimation, lower-order optimization method, restricted eigenvalue condition, ℓ_2 recovery bound, oracle property

[*] School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, P. R. China (11435017@zju.edu.cn).

[†] College of Mathematics and Statistics, Shenzhen University, Shenzhen 518060, P. R. China (mayhhu@szu.edu.cn).

[‡] School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, P. R. China (cli@zju.edu.cn).

[§] Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (mayangxq@polyu.edu.hk).

[¶] Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P.R. China (jiangtz@nlpr.ia.ac.cn).

1 Introduction

In various areas of applied sciences and engineering, a fundamental problem is to estimate an unknown parameter $\beta^* \in \mathbb{R}^n$ of a linear regression model

$$y = X\beta^* + e, \tag{1}$$

where $X \in \mathbb{R}^{m\times n}$ is a design matrix, $e \in \mathbb{R}^m$ is a vector containing random measurement noise, and thus $y \in \mathbb{R}^m$ is the corresponding vector of the noisy observations. According to the context of practical applications, the design matrix could be either deterministic or random.

The curse of dimensionality always occurs in the high-dimensional regime of many application fields. For example, in magnetic resonance imaging [9], remote sensing [2], systems biology [30], one is typically only able to collect far fewer samples than the number of variables due to physical or economical constraints, i.e., $m \ll n$. Under the high-dimensional scenario, estimating the true underlying parameter of model (1) is a vital challenge in contemporary statistics, whereas the classical ordinary least squares (OLS) does not work well in this scenario because the corresponding linear system is seriously ill-conditioned.

1.1 ℓ_1 Optimization Problems

Fortunately, in practical applications, a wide class of problems usually have certain special structures, employing which could eliminate the nonidentifiability of model (1) and enhance the predictability. One of the most popular structures is the sparsity structure, that is, the underlying parameter β^* in the high-dimensional space is sparse. One common way to measure the degree of sparsity is the ℓ_q norm, which for $0 < q \le 1$ is defined as

$$\|\beta\|_q := \left(\sum_{i=1}^{n} |\beta_i|^q\right)^{1/q},$$

while $\|\beta\|_0$ is defined as the number of nonzero entries of β. We first review the literature of sparse estimation for the case when the design matrix X is deterministic. Considering a bounded noise (i.e., $\|e\|_2 \le \epsilon$), it is natural to solve the following (constrained) ℓ_0 minimization problem

$$(\mathrm{CP}_{0,\epsilon}) \qquad \min \|\beta\|_0 \qquad \text{s.t.} \quad \|y - X\beta\|_2 \le \epsilon.$$

Unfortunately, it is NP-hard to compute its global solution due to the nonconvex and combinational natures [28].

To deal with this obstacle, a common technique is to use the (convex) ℓ_1 norm to approach the ℓ_0 norm:

$$(\mathrm{CP}_{1,\epsilon}) \qquad \min \|\beta\|_1 \qquad \text{s.t.} \quad \|y - X\beta\|_2 \le \epsilon,$$

which can be efficiently solved by several standard methods; see [14, 21] and references therein. The stable statistical properties of $(\mathrm{CP}_{1,\epsilon})$ have been explored under the regularity conditions. One of the most important stable statistical properties is the ℓ_2 recovery bound property, which is to estimate the upper bound of the error between the optimal solution of the optimization problem and the true underlying parameter in terms of the noise level ϵ. More specifically, let $s \ll n$ and β^* be an s-sparse parameter (i.e., $\|\beta^*\|_0 \le s$) satisfying the linear regression model (1). The ℓ_2 recovery bound for $(\mathrm{CP}_{1,\epsilon})$ was provided in [18] and [9] under the mutual incoherence property (MIP) or the restricted isometry property (RIP)[1], respectively:

$$\|\bar{\beta}_{1,\epsilon} - \beta^*\|_2 = O(\epsilon),$$

where $\bar{\beta}_{1,\epsilon}$ stands for the optimal solution of $(\mathrm{CP}_{1,\epsilon})$.

Motivated by the computational issue, another popular technique for s-parse estimation is to solve the (unconstrained) ℓ_1 regularization problem:

$$(\mathrm{RP}_{1,\lambda}) \qquad \min \frac{1}{2m}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

where $\lambda > 0$ is the regularization parameter, providing a tradeoff between data fidelity and sparsity. The ℓ_1 regularization model, also named the Lasso estimator [37], has attracted a great deal of attention in parameter estimation in the high-dimensional scenario, because its convexity structure is beneficial in designing exclusive and efficient algorithms and gaining wide applications; see [3, 15] and references therein. For the noise-free case, the ℓ_2 recovery bound for $(\mathrm{RP}_{1,\lambda})$ was provided in [39] under the RIP or the restricted eigenvalue condition (REC)[2]:

$$\|\hat{\beta}_{1,\lambda} - \beta^*\|_2^2 = O(\lambda^2 s),$$

where $\hat{\beta}_{1,\lambda}$ denotes the optimal solution of $(\mathrm{RP}_{1,\lambda})$. Furthermore, assuming that the noise in model (1) is normally distributed $e \sim \mathscr{N}(0, \sigma^2\mathbb{I}_m)$, it

[1] It was claimed in [7] that the RIP [10] is implied by the MIP [19], while the restricted isometry constant (RIC) is more difficult to be calculated than the mutual incoherence constant (MIC).

[2] It was reported in [4] that the REC is implied by the RIP, and in [32] that a broad class of correlated Gaussian design matrices satisfy the REC but violate the RIP with high probability.

was established in [4, 6, 44] that the following ℓ_2 recovery bound with high probability

$$\|\hat{\beta}_{1,\lambda} - \beta^*\|_2^2 = O\left(\sigma^2 s \frac{\log n}{m}\right),$$

when the regularization parameter is chosen as $\lambda = \sigma\sqrt{\frac{\log n}{m}}$ and under the RIP, REC or other regularity conditions, respectively. However, the ℓ_1 minimization and regularization problems suffer several dissatisfactions in both theoretical and practical applications. In particular, it was reported by extensive theoretical and empirical studies that the ℓ_1 minimization and regularization problems suffer from significant estimation bias when parameters have large absolute values; the induced solutions are much less sparse than the true parameter, they cannot recover a sparse signal with the least samples when applied to compressed sensing, and that they often result in sub-optimal sparsity in practice; see, e.g., [12, 20, 41, 40, 45]. Therefore, there is a great demand for developing the alternative sparse estimation technique that enjoys nice statistical theory and successful applications.

To address the bias and the sub-optimal issues induced by the ℓ_1 norm, several nonconvex regularizers have been proposed such as the smoothly clipped absolute deviation (SCAD) [20], minimax concave penalty (MCP) [41], ℓ_0 norm [43], ℓ_q norm ($0 < q < 1$) [22], and capped ℓ_1 norm [27]; specifically, the SCAD and MCP fall into the category of folded concave penalized (FCP) methods. It was studied in [43] that the global solution of the FCP sparse linear regression enjoys the oracle property under the sparse eigenvalue condition; see Remark 4(iii) for details.

It is worth noting that the ℓ_q norm regularizer ($0 < q < 1$) has been recognized as an important technique for sparse optimization and gained successful applications in various applied science fields; see, e.g., [12, 30, 40]. In the present paper, we focus on the statistical property of the ℓ_q optimization method, which is beyond the category of the FCP. Throughout the whole paper, we always assume that $0 < q \leq 1$ unless otherwise specified.

1.2 ℓ_q Optimization Problems

Due to the fact that $\lim_{q\to 0^+} \|\beta\|_q^q = \|\beta\|_0$, the ℓ_q norm has also been adopted as another alternative sparsity promoting penalty function of the ℓ_0 and ℓ_1 norms. The following ℓ_q optimization problems have attracted a great amount of attention and gained successful applications in a wide range of fields (see [12, 30, 40] and references therein):

$$(\mathrm{CP}_{q,\epsilon}) \qquad \min \|\beta\|_q \qquad \text{s.t.} \quad \|y - X\beta\|_2 \leq \epsilon,$$

and

$$(\mathrm{RP}_{q,\lambda}) \qquad \min \frac{1}{2m}\|y - X\beta\|_2^2 + \lambda\|\beta\|_q^q.$$

In particular, [12] and [40] showed that the ℓ_q minimization and the $\ell_{\frac{1}{2}}$ regularization admit a significantly stronger sparsity promoting capability than the ℓ_1 minimization and the ℓ_1 regularization, respectively; that is, they allow to obtain a more sparse solution from a smaller amount of samplings. [30] revealed that the $\ell_{\frac{1}{2}}$ regularization achieved a more reliable biological solution than the ℓ_1 regularization in the field of systems biology.

The advantage of the lower-order optimization problem has also been shown in theory that it requires a weaker regularity condition to guarantee the stable statistical property than the classical ℓ_1 optimization problem. In particular, let $\bar{\beta}_{q,\epsilon}$ and $\hat{\beta}_{q,\lambda}$ denote the optimal solution of $(\mathrm{CP}_{q,\epsilon})$ and $(\mathrm{RP}_{q,\lambda})$, respectively. The ℓ_2 recovery bound for $(\mathrm{CP}_{q,\epsilon})$ was established in [16] and [36] under MIP and RIP respectively:

$$\|\bar{\beta}_{q,\epsilon} - \beta^*\|_2 = O(\epsilon), \tag{2}$$

where the MIP or RIP is weaker than the one used in the study of $(\mathrm{CP}_{1,\epsilon})$. [24] established an ℓ_2 recovery bound for $(\mathrm{RP}_{q,\lambda})$ in the noise-free case:

$$\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 = O(\lambda^{\frac{2}{2-q}}s) \tag{3}$$

under the introduced q-REC, which is strictly weaker than the classical REC. However, the theoretical study of the ℓ_q optimization problem is still limited; particularly, there is still no paper devoted to establishing the statistical property of the ℓ_q minimization problem when the noise is randomly distributed, and that of the ℓ_q regularization problem in the noise-aware case.

1.3 Contributions of This Paper

The main contribution of the present paper is the establishment of the s-tatistical properties for the ℓ_q optimization problems, including $(\mathrm{CP}_{q,\epsilon})$ and $(\mathrm{RP}_{q,\lambda})$, in the noise-aware case; specifically, in the case when the linear regression model (1) involves a Gaussian noise $e \sim \mathscr{N}(0, \sigma^2\mathbb{I}_m)$. For this purpose, we extend the q-REC [24] to a more general one, which is one of the weakest regularity conditions for estimating the ℓ_2 recovery bounds of sparse estimation models, and provide some sufficient conditions for guaranteeing the general q-REC in terms of REC, RIP, MIP (with a less restrictive constant); see Propositions 1 and 2. Under the general q-REC, we show that the ℓ_2 recovery bound (2) holds for (CP$_{q,\epsilon}$) with high probability, and that

$$\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 = O\left(\left(\sigma^2 \frac{\log n}{m}\right)^{\frac{1}{2-q}} s\right),$$

as well as the estimation of prediction loss and the oracle property, hold for (RP$_{q,\lambda}$) with high probability; see Theorems 1 and 2, respectively. These results provide a unified framework of the statistical properties of the ℓ_q optimization problems, and improve the ones of the ℓ_q minimization problem [16, 36] and the ℓ_1 regularization problem [4, 6, 44] under the q-REC; see Remark 4. They are not only of independent interest in establishing statistical properties for the lower-order optimization problems with randomly noisy data, but also provide a useful tool for the study of the case when the design matrix X is random.

Another contribution of the present paper is to explore the ℓ_2 recovery bounds for the ℓ_q optimization problems with a random design matrix X and random noise e, which is more realistic in the real-world applications; e.g., compressed sensing [8], signal processing [9], statistical learning [1]. As reported in [32], the key issue for studying the statistical properties of a sparse estimation model with a random design matrix is to provide suitable conditions on the population covariance matrix Σ of X, which can guarantee the regularity conditions with high probability; see, e.g., [9, 32]. Motivated by the real-world applications, we consider the standard case when X is a Gaussian random design with i.i.d. $\mathscr{N}(0,\Sigma)$ rows and the linear regression model (1) involves a Gaussian noise, explore a sufficient condition for ensuring the q-REC of X with high probability in terms of the q-REC of Σ, and apply the preceding results to establish the ℓ_2 recovery bounds (2) for (CP$_{q,\epsilon}$), and (3), as well as the predication loss and the oracle inequality, for (RP$_{q,\lambda}$), respectively; see Theorems 3 and 4. These results provide a unified framework of the statistical properties of the ℓ_q optimization problems with a Gaussian random design under the q-REC, which cover the ones of the ℓ_1 optimization problems (see [46, Theorem 3.1]) as special cases; see Corollaries 3 and 4. To the best of our knowledge, most results presented in this paper are new, either for the deterministic or random design matrix.

We also carry out the numerical experiments on the standard simulated data. The preliminary numerical results verify the established statistical properties and show that the ℓ_q optimization methods possess better recovery performance than the ℓ_1 optimization method, which coincides with existing numerical studies [24, 40] on the ℓ_q regularization problem.

The remainder of this paper is organized as follows. In section 2, we introduce the lower-order REC and discuss its sufficient conditions. In section 3, we establish the ℓ_2 recovery bounds for $(\text{CP}_{q,\epsilon})$ and $(\text{RP}_{q,\lambda})$ with a deterministic design matrix. The extension to the linear regression model with a Gaussian random design and preliminary numerical results are presented in sections 4 and 5, respectively.

We end this section by presenting the notations adopted in this paper. We use Greek lowercase letters α, β, δ to denote the vectors, capital letters J, T to denote the index sets, and script captical letters $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ to denote the random events. For $\beta \in \mathbb{R}^n$ and $J \subseteq \{1, 2, \ldots, n\}$, we use β_J to denote the vector in $\mathbb{R}^n$ that $(\beta_J)_i = \beta_i$ for $i \in J$ and zero elsewhere, $|J|$ to denote the cardinality of J, $J^c := \{1, 2, \ldots, n\} \setminus J$ to denote the complement of J, and $\text{supp}(\beta)$ to denote the support of β, i.e., the index set of nonzero entries of β. Particularly, $\mathbb{I}_m$ stands for the identity matrix in $\mathbb{R}^m$, and $\mathbb{P}(\mathscr{A})$ and $\mathbb{P}(\mathscr{A}|\mathscr{B})$ denote the probability that event $\mathscr{A}$ happens and the conditional probability that event $\mathscr{A}$ happens given that event $\mathscr{B}$ happens, respectively.

2 Restricted Eigenvalue Conditions

This section aims to discuss some regularity conditions imposed on the design matrix X that are needed to guarantee the stable statistical properties of $(\text{CP}_{q,\epsilon})$ and $(\text{RP}_{q,\lambda})$.

In statistics, the ordinary least squares (OLS) is a classical technique for estimating the unknown parameters in a linear regression model and has favourable properties if some regularity conditions are satisfied; see, e.g., [31]. For example, the OLS always requires the positive definiteness of the Gram matrix $\Gamma(X) := X^\top X$, that is,

$$\min_{\beta \in \mathbb{R}^n : \beta \neq 0} \frac{(\beta^\top \Gamma(X) \beta)^{1/2}}{\|\beta\|_2} = \min_{\beta \in \mathbb{R}^n : \beta \neq 0} \frac{\|X\beta\|_2}{\|\beta\|_2} > 0. \tag{4}$$

However, in the high-dimensional setting, the OLS does not work well; in fact, the matrix $\Gamma(X)$ is seriously degenerate, i.e.,

$$\min_{\beta \in \mathbb{R}^n : \beta \neq 0} \frac{\|X\beta\|_2}{\|\beta\|_2} = 0.$$

To deal with the challenges caused by the high-dimensional data, the Lasso (least absolute shrinkage and selection operator) estimator was introduced by [37]. Since then the Lasso estimator has gained a great success in the sparse representation and machine learning of high-dimensional data; see, e.g., [4, 38, 44] and references therein. It was pointed out that Lasso requires a weak condition, called the restricted eigenvalue condition (REC) [4], to ensure the nice statistical properties; see, e.g., [39, 26, 29]. In the definition of REC, the minimum in (4) is replaced by a minimum over a restricted set of vectors measured by an ℓ_1 norm inequality, and the norm $\|\beta\|_2$ in the denominator is replaced by the ℓ_2 norm of only a part of β. The notion of REC was extended to the group-wised lower-order REC in [24], which was used there to explore the oracle property and ℓ_2 recovery bound of the $\ell_{p,q}$ regularization problem in a noise-free case.

Inspired by the ideas in [4, 24], we here introduce a lower-order REC for the ℓ_q optimization problems, similar to but more general than the one in [24], where the minimum is taken over a restricted set of vectors measured by an ℓ_q norm inequality. To proceed, we shall introduce some useful notations. For the remainder of this paper, let $a > 0$ and (s, t) be a pair of integers such that

$$1 \leq s \leq t \leq n \quad \text{and} \quad s + t \leq n. \tag{5}$$

For $\delta \in \mathbb{R}^n$ and $J \subseteq \{1, 2, \ldots, n\}$, we define by $J(\delta; t)$ the index set corresponding to the first t largest coordinates in absolute value of δ in J^c. For $X \in \mathbb{R}^{m\times n}$, its q-restricted eigenvalue modulus relative to (s, t, a) is defined by

$$\phi_q(s,t,a,X) := \min\left\{\frac{\|X\delta\|_2}{\|\delta_{J\cup J(\delta;t)}\|_2} : |J| \leq s, \|\delta_{J^c}\|_q^q \leq a\|\delta_J\|_q^q\right\}. \tag{6}$$

The lower-order REC is defined as follows.

Definition 1. *Let $0 \leq q \leq 1$ and $X \in \mathbb{R}^{m\times n}$. X is said to satisfy the q-restricted eigenvalue condition relative to (s, t, a) (q-REC(s, t, a) in short) if*

$$\phi_q(s, t, a, X) > 0.$$

Remark 1. (i) *Clearly, the q-REC(s, t, a) provides a unified framework of the REC-type conditions, e.g., it includes the classical REC in [4] (when $q = 1$) and the q-REC(s, t) in [24] (when $a = 1$) as special cases.*

(ii) *The restricted eigenvalue modulus (with $q = 1$) defined in (6) is slightly different from the one for the classical REC in [4], in which the factor $\sqrt{m}$ appears in the denominator there. For example, if the matrix X has i.i.d. Gaussian entries, the restricted eigenvalue modulus in [4] scales as a constant independent of s, m, and n and thus $\phi_q(s, t, a, X)$ given by (6) scales as $\sqrt{m}$, whenever $\frac{s}{m}\log n$ is bounded.*

It is natural to study the relationships between the q-RECs and other types of regularity conditions. To this end, we first recall some basic properties of the ℓ_q norm in the following lemmas; particularly, Lemma 1 is taken from [23, Section 8.12] and [24, Lemmas 1 and 2].

Lemma 1. *Let $\alpha, \beta \in \mathbb{R}^n$. Then the following relations are true:*

$$\|\beta\|_{q_2} \le \|\beta\|_{q_1} \le n^{\frac{1}{q_1}-\frac{1}{q_2}} \|\beta\|_{q_2} \quad \textit{for any } 0 < q_1 \le q_2 < +\infty, \tag{7}$$

$$\|\alpha\|_q^q - \|\beta\|_q^q \le \|\alpha+\beta\|_q^q \le \|\alpha\|_q^q + \|\beta\|_q^q \quad \textit{for any } 0 < q \le 1. \tag{8}$$

Lemma 2. *Let $p \ge 1$, $n_1, n_2 \in \mathbb{N}$, $\alpha \in \mathbb{R}_+^{n_1}$, $\beta \in \mathbb{R}_+^{n_2}$ and $c > 0$ be such that*

$$\max_{1\le i\le n_1} \alpha_i \le \min_{1\le j\le n_2} \beta_j \quad \textit{and} \quad \sum_{i=1}^{n_1} \alpha_i \le c \sum_{j=1}^{n_2} \beta_j. \tag{9}$$

Then

$$\sum_{i=1}^{n_1} \alpha_i^p \le c \sum_{j=1}^{n_2} \beta_j^p. \tag{10}$$

Proof. Let $\alpha_{\max} := \max\limits_{1\le i\le n_1} \alpha_i$ and $\beta_{\min} := \min\limits_{1\le j\le n_2} \beta_j$. Then it holds that

$$\alpha_{\max} \sum_{i=1}^{n_1} \alpha_i^p \le \alpha_{\max}^p \sum_{i=1}^{n_1} \alpha_i \quad \text{and} \quad \beta_{\min}^p \sum_{j=1}^{n_2} \beta_j \le \beta_{\min} \sum_{j=1}^{n_2} \beta_j^p. \tag{11}$$

Without loss of generality, we assume that $\alpha_{\max} > 0$; otherwise, (10) holds automatically. Thus, by the first inequality of (9) and noting $p \ge 1$, we have that

$$0 < \alpha_{\max}^p \beta_{\min} \le \alpha_{\max} \beta_{\min}^p. \tag{12}$$

Multiplying the inequalities in (11) by $\beta_{\min} \sum\limits_{j=1}^{n_2} \beta_j$ and $\alpha_{\max} \sum\limits_{i=1}^{n_1} \alpha_i$ respectively, we obtain that

$$\begin{aligned}
\alpha_{\max}\beta_{\min} \sum_{i=1}^{n_1} \alpha_i^p \sum_{j=1}^{n_2} \beta_j &\le \alpha_{\max}^p \beta_{\min} \sum_{i=1}^{n_1} \alpha_i \sum_{j=1}^{n_2} \beta_j \\
&\le \alpha_{\max} \beta_{\min}^p \sum_{i=1}^{n_1} \alpha_i \sum_{j=1}^{n_2} \beta_j \\
&\le \alpha_{\max} \beta_{\min} \sum_{i=1}^{n_1} \alpha_i \sum_{j=1}^{n_2} \beta_j^p,
\end{aligned}$$

where the second inequality follows from (12). This, together with the second inequality of (9), yields (10). The proof is complete. □

Extending [24, Proposition 5] to the general q-REC, the following proposition validates the relationship between the q-RECs: the lower the q, the weaker the q-REC. However, the inverse of this implication is not true; see [24, Example 1] for a counter example. We provide the proof so as to make this paper self-contained, although the idea is similar to that of [24, Proposition 5].

Proposition 1. *Let $X \in \mathbb{R}^{m\times n}$, $a > 0$, and (s,t) be a pair of integers satisfying* (5). *Suppose that $0 < q_1 \le q_2 \le 1$ and that X satisfies the q_2-REC(s,t,a). Then X satisfies the q_1-REC(s,t,a).*

Proof. Associated with the q-REC(s,t,a), we define the feasible set

$$C_q(s,a) := \{\delta \in \mathbb{R}^n : \|\delta_{J^c}\|_q^q \le a\|\delta_J\|_q^q \text{ for some } |J| \le s\}. \tag{13}$$

By Definition 1, it remains to show that $C_{q_1}(s,a) \subseteq C_{q_2}(s,a)$. To this end, let $\delta \in C_{q_1}(s,a)$, and let J_0 denote the index set of the first s largest coordinates in absolute value of δ. By the assumption that $\delta \in C_{q_1}(s,a)$ and by the construction of J_0, one has $\|\delta_{J_0^c}\|_{q_1}^{q_1} \le a\|\delta_{J_0}\|_{q_1}^{q_1}$. Then we obtain by Lemma 2 (with q_2/q_1 in place of p) that $\|\delta_{J_0^c}\|_{q_2}^{q_2} \le a\|\delta_{J_0}\|_{q_2}^{q_2}$; consequently, $\delta \in C_{q_2}(s,a)$. Hence, it follows that $C_{q_1}(s,a) \subseteq C_{q_2}(s,a)$, and the proof is complete. □

It is revealed from Proposition 1 that the classical REC is a sufficient condition of the lower-order REC. In the sequel, we will further discuss some other types of regularity conditions: the sparse eigenvalues condition (SEC), the restricted isometry property (RIP), and the mutual incoherence property (MIP), which have been widely used in the literature of statistics and engineering, for ensuring the lower-order REC.

The SEC is a popular regularity condition required to guarantee the nice properties of sparse representation; see [4, 17, 43] and references therein. For $\Delta \in \mathbb{R}^{n\times n}$ and $s \in \mathbb{N}$, the s-sparse minimal eigenvalue and s-sparse maximal eigenvalue of Δ are respectively defined by

$$\sigma_{\min}(s,\Delta) := \min_{\beta\in\mathbb{R}^n : 1\le\|\beta\|_0\le s} \frac{\beta^\top\Delta\beta}{\beta^\top\beta}, \quad \sigma_{\max}(s,\Delta) := \max_{\beta\in\mathbb{R}^n : 1\le\|\beta\|_0\le s} \frac{\beta^\top\Delta\beta}{\beta^\top\beta}. \tag{14}$$

The SEC was first introduced in [17] to show that the optimal solution of $(\mathrm{CP}_{1,\epsilon})$ well approximates that of $(\mathrm{CP}_{0,\epsilon})$ whenever $\sigma_{\min}(2s,\Gamma(X)) > 0$.

The RIP is another well-known regularity condition in the scenario of sparse learning, which was introduced by [10] and has been widely used in the study of the oracle property and ℓ_2 recovery bound for the high-dimensional regression model; see [4, 9, 34] and references therein. Below, we recall the RIP-type notions from [10].

Definition 2. *Let $X \in \mathbb{R}^{m\times n}$ and let $s, t \in \mathbb{N}$ be such that $s + t \leq n$.*

(i) *The s-restricted isometry constant of X, denoted by $\eta_s(X)$, is defined to be the smallest quantity such that, for any $\beta \in \mathbb{R}^n$ and $J \subseteq \{1, \ldots, n\}$ with $|J| \leq s$,*

$$(1 - \eta_s(X))\|\beta_J\|_2^2 \leq \|X\beta_J\|_2^2 \leq (1 + \eta_s(X))\|\beta_J\|_2^2. \tag{15}$$

(ii) *The (s,t)-restricted orthogonality constant of X, denoted by $\theta_{s,t}(X)$, is defined to be the smallest quantity such that, for any $\beta \in \mathbb{R}^n$ and $J, T \subseteq \{1, \ldots, n\}$ with $|J| \leq s$, $|T| \leq t$ and $J \cap T = \emptyset$,*

$$|\langle X\beta_J, X\beta_T\rangle| \leq \theta_{s,t}(X)\|\beta_J\|_2\|\beta_T\|_2. \tag{16}$$

The MIP is also a well-known regularity condition in the scenario of sparse learning, which was introduced by [19] and has been used in [4, 7, 17, 18] and references therein. In the case when each diagonal element of the Gram matrix $\Gamma(X)$ is 1, $\theta_{1,1}(X)$ coincides with the mutual incoherence constant; see [19].

The following lemmas are useful for establishing the relationship between the q-REC and other types of regularity conditions; in particular, Lemmas 3 and 4 are taken from [10, Lemma 1.1] and [39, Lemma 3.1], respectively.

Lemma 3. *Let $X \in \mathbb{R}^{m\times n}$ and $s, t \in \mathbb{N}$ be such that $s + t \leq n$. Then*

$$\theta_{s,t}(X) \leq \eta_{s+t}(X) \leq \theta_{s,t}(X) + \max\{\eta_s(X), \eta_t(X)\}.$$

Lemma 4. *Let $\alpha, \beta \in \mathbb{R}^n$ and $0 < \tau < 1$ be such that $-\langle \alpha, \beta\rangle \leq \tau\|\alpha\|_2^2$. Then $(1 - \tau)\|\alpha\|_2 \leq \|\alpha + \beta\|_2$.*

For the sake of simplicity, a partition structure and some notations are presented. For a vector $\delta \in \mathbb{R}^n$ and an index set $J \subseteq \{1, 2, \ldots, n\}$, we use $\text{rank}(\delta_i; J^c)$ to denote the rank of the absolute value of δ_i in J^c (in a decreasing order) and $J_k(\delta; t)$ to denote the index set of the k-th batch of the first t largest coordinates in absolute value of δ in J^c. That is,

$$J_k(\delta; t) := \{i \in J^c : \text{rank}(\delta_i; J^c) \in \{kt + 1, \ldots, (k + 1)t\}\} \quad \text{for each } k \in \mathbb{N}. \tag{17}$$

Lemma 5. *Let $X \in \mathbb{R}^{m\times n}$, $0 < q \le 1$, $a > 0$, and (s,t) be a pair of integers satisfying* (5)*. Then the following relations are true:*

$$\phi_q(s,t,a,X) \ge \sqrt{\sigma_{\min}(s+t,\Gamma(X))} - a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\sqrt{\sigma_{\max}(t,\Gamma(X))}, \quad (18)$$

$$\phi_q(s,t,a,X) \le \sqrt{\sigma_{\max}(s+t,\Gamma(X))} + a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\sqrt{\sigma_{\max}(t,\Gamma(X))}. \quad (19)$$

Proof. Fix $\delta \in C_q(s,a)$, as defined by (13). Then there exists $J \subseteq \{1,2,\ldots,n\}$ such that

$$|J| \le s \quad \text{and} \quad \|\delta_{J^c}\|_q^q \le a\|\delta_J\|_q^q. \quad (20)$$

Write $r := \lceil \frac{n-s}{t} \rceil$ (where $\lceil u \rceil$ denotes the largest integer not greater than u), $J_k := J_k(\delta;t)$ (defined by (17)) for each $k \in \mathbb{N}$ and $J_* := J \cup J_0$. Then it follows from [24, Lemma 7] and (20) that

$$\sum_{k=1}^{r} \|\delta_{J_k}\|_2 \le t^{\frac{1}{2}-\frac{1}{q}}\|\delta_{J^c}\|_q \le a^{\frac{1}{q}}t^{\frac{1}{2}-\frac{1}{q}}\|\delta_J\|_q \le a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\|\delta_J\|_2 \quad (21)$$

(due to (7)). Noting by (17) and (20) that $|J_*| \le s+t$ and $|J_k| \le t$ for each $k \in \mathbb{N}$, one has by (14) that

$$\sqrt{\sigma_{\min}(s+t,\Gamma(X))}\|\delta_{J_*}\|_2 \le \|X\delta_{J_*}\|_2 \le \sqrt{\sigma_{\max}(s+t,\Gamma(X))}\|\delta_{J_*}\|_2,$$

$$\|X\delta_{J_k}\|_2 \le \sqrt{\sigma_{\max}(t,\Gamma(X))}\|\delta_{J_k}\|_2 \quad \text{for each} \quad k \in \mathbb{N}.$$

These, together with (21), imply that

$$\begin{aligned}\|X\delta\|_2 &\ge \|X\delta_{J_*}\|_2 - \sum_{k=1}^{r}\|X\delta_{J_k}\|_2 \\ &\ge \left(\sqrt{\sigma_{\min}(s+t,\Gamma(X))} - a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\sqrt{\sigma_{\max}(t,\Gamma(X))}\right)\|\delta_{J_*}\|_2.\end{aligned}$$

Since δ and J satisfying (20) are arbitrary, (18) is shown to hold by (6) and the fact that $J_* = J \cup J(\delta;t)$. One can prove (19) in a similar way, and thus, the details are omitted. □

The following proposition provides the sufficient conditions for the q-REC in terms of the SEC, RIP and MIP; see (a), (b) and (c) below respectively.

Proposition 2. *Let $X \in \mathbb{R}^{m\times n}$, $0 < q \le 1$, $a > 0$, and (s,t) be a pair of integers satisfying* (5)*. Then X satisfies the q-*REC(s,t,a) *provided that one of the following conditions:*

(a) $\sigma_{\min}(s+t,\Gamma(X)) > a\left(\frac{as}{t}\right)^{\frac{2}{q}-1}\sigma_{\max}(t,\Gamma(X))$.

(b) $\eta_t(X)+\theta_{s,t}(X)+a^{\frac{1}{2}}\left(\frac{as}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X) < 1$.

(c) *each diagonal element of* $\Gamma(X)$ *is* 1 *and* $\theta_{1,1}(X) < \left(\left(1+2a\left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right)(s+t)\right)^{-1}$.

Proof. It directly follows from Lemma 5 (cf. (18)) that X satisfies the q-REC(s,t,a) provided that condition (a) holds. Fix $\delta\in C_q(s,a)$, and let J, r, J_k (for each $k\in\mathbb{N}$) and J_* be defined, respectively, as in the beginning of the proof of Lemma 5. Then (21) follows directly and it follows from [24, Lemma 7] and (17) that

$$\|\delta_{J_*^c}\|_1=\sum_{k=1}^r\|\delta_{J_k}\|_1\le t^{1-\frac{1}{q}}\|\delta_{J^c}\|_q\le a^{\frac{1}{q}}t^{1-\frac{1}{q}}\|\delta_J\|_q\le a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-1}\|\delta_J\|_1. \tag{22}$$

Suppose that condition (b) is satisfied. By Definition 2 (cf. (16)), one has that

$$|\langle X\delta_{J_*},X\delta_{J_*^c}\rangle|\le\sum_{k=1}^r|\langle X\delta_{J_*},X\delta_{J_k}\rangle|\le\theta_{t,s+t}(X)\|\delta_{J_*}\|_2\sum_{k=1}^r\|\delta_{J_k}\|_2.$$

Then it follows from (21) that

$$\begin{aligned}|\langle X\delta_{J_*},X\delta_{J_*^c}\rangle| &\le a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X)\|\delta_{J_*}\|_2\|\delta_J\|_2\\ &\le \frac{a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}\|X\delta_{J_*}\|_2^2\end{aligned} \tag{23}$$

(by (15)). Since $s\le t$ (by (5)), one has by Definition 2(i) that $\eta_s(X)\le\eta_t(X)$, and then by Lemma 3 that $\eta_{s+t}(X)\le\theta_{s,t}(X)+\eta_t(X)$. Then it follows from (b) that

$$0<\frac{a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}\le\frac{a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X)}{1-(\eta_t(X)+\theta_{s,t}(X))}<1. \tag{24}$$

This, together with (23), shows that Lemma 4 is applicable (with $X\delta_{J_*}$,

$X\delta_{J_*^c}$, $\frac{a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}$ in place of α, β, τ) to concluding that

$$\begin{aligned}\|X\delta\|_2^2 &\geq \left(1-\frac{a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}\right)^2\|X\delta_{J_*}\|_2^2\\ &\geq (1-\eta_{s+t}(X))\left(1-\frac{a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}\right)^2\|\delta_{J_*}\|_2^2\end{aligned}$$

(due to (15)). Since δ and J satisfying (20) are arbitrary, we derive by (6) and (24) that

$$\phi_q(s,t,a,X)\geq\sqrt{1-\eta_{s+t}(X)}\left(1-\frac{a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}\right)>0;$$

consequently, X satisfies the q-REC(s,t,a).

Suppose that (c) is satisfied. Then we have by (22) and Definition 2 (cf. (16)) that

$$\begin{aligned}\|X\delta\|_2^2 &= \|X\delta_{J_*}\|_2^2+2\langle X\delta_{J_*},X\delta_{J_*^c}\rangle+\|X\delta_{J_*^c}\|_2^2\\ &\geq \|X\delta_{J_*}\|_2^2-2|\langle X\delta_{J_*},X\delta_{J_*^c}\rangle|\\ &\geq \|X\delta_{J_*}\|_2^2-2\theta_{1,1}(X)\|\delta_{J_*}\|_1\|\delta_{J_*^c}\|_1\\ &\geq \|X\delta_{J_*}\|_2^2-2a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-1}\theta_{1,1}(X)\|\delta_{J_*}\|_1^2.\end{aligned}\tag{25}$$

Separating the diagonal and off-diagonal terms of the quadratic form $\delta_{J_*}^TX^TX\delta_{J_*}$, one has by (7) and (c) that

$$\begin{aligned}\|X\delta_{J_*}\|_2^2 &= \sum_{i=1}^n(X^TX)_{i,i}(\delta_{J_*})_i(\delta_{J_*})_i+\sum_{j\neq k}^n(X^TX)_{j,k}(\delta_{J_*})_j(\delta_{J_*})_k\\ &= \|\delta_{J_*}\|_2^2+\sum_{j\neq k}^n\langle X_{\cdot j}(\delta_{J_*})_j,X_{\cdot k}(\delta_{J_*})_k\rangle\\ &\geq \|\delta_{J_*}\|_2^2-\theta_{1,1}(X)\|\delta_{J_*}\|_1^2\\ &\geq (1-(s+t)\theta_{1,1}(X))\|\delta_{J_*}\|_2^2.\end{aligned}$$

Combining this inequality with (25), we get that

$$\|X\delta\|_2^2\geq\left(1-\left(1+2a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-1}\right)(s+t)\theta_{1,1}(X)\right)\|\delta_{J_*}\|_2^2.$$

Since δ and J satisfying (20) are arbitrary, we derive by (6) and (c) that

$$\phi_q(s,t,a,X) \geq 1 - \left(1 + 2a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-1}\right)(s+t)\theta_{1,1}(X) > 0;$$

consequently, X satisfies the q-REC(s,t,a). The proof is complete. □

Remark 2. *It was established in [4, Lemma 4.1(ii)], [39, Corollary 7.1 and 3.1] and [4, Assumption 5] that X satisfies the classical* REC *under one of the following conditions:*

(a') $\sigma_{\min}(s+t,\Gamma(X)) > \frac{s}{t}a^2\sigma_{\max}(t,\Gamma(X))$.

(b') $\eta_t(X) + \theta_{s,t}(X) + \left(\frac{s}{t}\right)^{\frac{1}{2}} a\theta_{t,s+t}(X) < 1$.

(c') *each diagonal element of* $\Gamma(X)$ *is* 1 *and* $\theta_{1,1}(X) < ((1+2a)(s+t))^{-1}$.

Proposition 2 extends the existing results to the general case when $0 < q \leq 1$ *and partially improves them; in particular, each of conditions* (a)-(c) *in Proposition 2 required for the* q-REC *is less restrictive than the corresponding one of conditions* (a')-(c') *required for the classical* REC *in the situation when* $t > as$*, which usually occurs in the high-dimensional scenario (see, e.g., [4, 9, 46]). Moreover, by Propositions 1 and 2, we achieve that the* q-REC(s,t,a) *is satisfied provided that one of the following conditions:*

(a°) $\sigma_{\min}(s+t,\Gamma(X)) > \min\left\{1, \left(\frac{as}{t}\right)^{\frac{2}{q}-2}\right\}\frac{s}{t}a^2\sigma_{\max}(t,\Gamma(X))$.

(b°) $\eta_t(X) + \theta_{s,t}(X) + \min\left\{1, \left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right\}\left(\frac{s}{t}\right)^{\frac{1}{2}} a\theta_{t,s+t}(X) < 1$.

(c°) *each diagonal element of* $\Gamma(X)$ *is* 1 *and* $\theta_{1,1}(X) < \left(\left(1 + 2a\min\left\{1, \left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right\}\right)(s+t)\right)^{-1}$.

3 Recovery Bounds for Deterministic Design

This section is devoted to establishing the ℓ_2 recovery bounds for $(\text{CP}_{q,\epsilon})$ and $(\text{RP}_{q,\lambda})$ in the case that X is deterministic. Throughout this paper, we assume that the linear regression model (1) involves a Gaussian noise, i.e., $e \sim \mathscr{N}(0,\sigma^2\mathbb{I}_m)$, and adopt the following notations:

let β^* be a solution of (1), $J := \text{supp}(\beta^*)$, $s := |J|$, and let $t \in \mathbb{N}$ satisfy (5).

The ℓ_2 recovery bound of the ℓ_1 regularization problem (i.e., Lasso estimator) was established in [4] under the assumption of the classical REC.

The deduction of the ℓ_2 recovery bound is based on an important property of the optimal solution. More precisely, let $\bar{\beta}_{1,\epsilon}$ and $\hat{\beta}_{1,\lambda}$ be the solutions of the ℓ_1 minimization and the ℓ_1 regularization problems, respectively. It was reported in [9, Eq. (2.2)] and [4, Corollary B.2] that the corresponding residuals satisfy the following dominant properties, with high probability,

$$\|(\bar{\beta}_{1,\epsilon} - \beta^*)_{J^c}\|_1 \leq \|(\bar{\beta}_{1,\epsilon} - \beta^*)_J\|_1$$

and

$$\|(\hat{\beta}_{1,\lambda} - \beta^*)_{J^c}\|_1 \leq 3\|(\hat{\beta}_{1,\lambda} - \beta^*)_J\|_1$$

for the ℓ_1 minimization and the ℓ_1 regularization problems, respectively.

In the study of the ℓ_q minimization and the ℓ_q regularization problems, a natural question arises whether the residuals of solutions of (CP$_{q,\epsilon}$) or (RP$_{q,\lambda}$) satisfy such a dominant property on the support of the true underlying parameter of linear regression (1) with high probability. Below, we provide a positive answer for this question in Propositions 3 and 4. To this end, we present some preliminary lemmas to measure the probabilities of random events related to the linear regression model (1), in which Lemma 6 is taken from [46, Lemma C.1].

Lemma 6. *Let $0 \leq \theta < 1$ and $b \geq 0$. Suppose that*

$$\max_{1 \leq j \leq n} \|X_{\cdot j}\|_2 \leq (1+\theta)\sqrt{m}. \tag{26}$$

Then

$$\mathbb{P}\left(\frac{\|X^\top e\|_\infty}{m} \geq \sigma(1+\theta)\sqrt{\frac{2(1+b)\log n}{m}}\right) \leq \left(n^b\sqrt{\pi \log n}\right)^{-1}.$$

Lemma 7. *Let $d \geq 5$. Then*

$$\mathbb{P}\left(\|e\|_2^2 \geq dm\sigma^2\right) \leq \exp\left(-\frac{d-1}{4}m\right).$$

Proof. Recall that $e = (e_1, \ldots, e_m)^\top \sim \mathscr{N}(0, \sigma^2\mathbb{I}_m)$. Let $u_i := \frac{1}{\sigma}e_i$ for $i = 1, \ldots, m$. Then one has that $u_1, \ldots, u_m$ are i.i.d. Gaussian variables with $u_i \sim \mathscr{N}(0,1)$ for $i = 1, \ldots, m$. Let $u := (u_1, \ldots, u_m)^\top$. Clearly, $\|u\|_2^2 = \frac{1}{\sigma^2}\|e\|_2^2$ is a chi-square random variable with m degrees of freedom (see, e.g., [35, Section 5.6]). Then it follows from standard tail bounds of chi-square random variable (see, e.g., [33, Appendix I]) that

$$\mathbb{P}\left(\frac{\|u\|_2^2 - m}{m} \geq d-1\right) \leq \exp\left(-\frac{d-1}{4}m\right)$$

(as $d \geq 5$). Consequently, we obtain that

$$\mathbb{P}\left(\|e\|_2^2 \geq dm\sigma^2\right) = \mathbb{P}\left(\|u\|_2^2 \geq dm\right) \leq \exp\left(-\frac{d-1}{4}m\right).$$

The proof is complete. □

Recall that β^* satisfies the linear regression model (1).

Lemma 8. *Let $\hat{\beta}_{q,\lambda}$ be an optimal solution of* $(\mathrm{RP}_{q,\lambda})$. *Then*

$$\frac{1}{2m}\|X\beta^* - X\hat{\beta}_{q,\lambda}\|_2^2 \leq \lambda\|\beta^*\|_q^q - \lambda\|\hat{\beta}_{q,\lambda}\|_q^q + \frac{1}{m}\|\hat{\beta}_{q,\lambda} - \beta^*\|_1\|X^\top e\|_\infty.$$

Proof. Since $\hat{\beta}_{q,\lambda}$ is an optimal solution of $(\mathrm{RP}_{q,\lambda})$, it follows that

$$\frac{1}{2m}\|y - X\hat{\beta}_{q,\lambda}\|_2^2 + \lambda\|\hat{\beta}_{q,\lambda}\|_q^q \leq \frac{1}{2m}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_q^q.$$

This, together with (1), yields that

$$\begin{aligned}
\lambda\|\hat{\beta}_{q,\lambda}\|_q^q - \lambda\|\beta^*\|_q^q &\leq \frac{1}{2m}\|y - X\beta^*\|_2^2 - \frac{1}{2m}\|y - X\hat{\beta}_{q,\lambda}\|_2^2 \\
&= \frac{1}{m}\left\langle X(\hat{\beta}_{q,\lambda} - \beta^*), e\right\rangle - \frac{1}{2m}\|X\beta^* - X\hat{\beta}_{q,\lambda}\|_2^2 \\
&\leq \frac{1}{m}\|\hat{\beta}_{q,\lambda} - \beta^*\|_1\|X^\top e\|_\infty - \frac{1}{2m}\|X\beta^* - X\hat{\beta}_{q,\lambda}\|_2^2.
\end{aligned}$$

The proof is complete. □

Below, we present some notations that are useful for the following discussion of the ℓ_2 recovery bounds. Recall that β^* is a solution of (1). Throughout the remainder of this paper, let

$$a > 1, \quad 0 \leq \theta < 1, \quad b \geq 0, \tag{27}$$

unless otherwise specified, and let $r > 0$ be such that

$$r \geq \|\beta^*\|_q. \tag{28}$$

Let

$$\epsilon := \sigma\sqrt{5m} \quad \text{and} \quad \rho := \left(\frac{5\sigma^2}{2\lambda} + r^q\right)^{1/q}, \tag{29}$$

and select the regularization parameter in $(\mathrm{RP}_{q,\lambda})$ as

$$\lambda := \max\left\{\frac{a+1}{a-1}\sigma(1+\theta)2^{1-q}(1+r^q)^{\frac{1-q}{q}}\sqrt{\frac{2(1+b)\log n}{m}},\ \frac{5}{2}\sigma^2\right\}. \tag{30}$$

Define the following two random events relative to linear regression model (1) by

$$\mathscr{A} := \{e : \|e\|_2 \le \epsilon\} \tag{31}$$

and

$$\mathscr{B} := \left\{ e : \frac{a+1}{(a-1)m}(2\rho)^{1-q}\|X^\top e\|_\infty \le \lambda \right\}. \tag{32}$$

The following lemma estimates the probabilities of events $\mathscr{A}$ and $\mathscr{B}$.

Lemma 9. *The probability of event $\mathscr{A}$ satisfies*

$$\mathbb{P}(\mathscr{A}) \ge 1 - \exp(-m). \tag{33}$$

Moreover, suppose that (26) *is satisfied. Then*

$$\mathbb{P}(\mathscr{B}) \ge 1 - \left(n^b\sqrt{\pi \log n}\right)^{-1}, \tag{34}$$

$$\mathbb{P}(\mathscr{A} \cap \mathscr{B}) \ge 1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1}. \tag{35}$$

Proof. By (29) and (31), Lemma 7 is applicable (with $d = 5$) to showing that $\mathbb{P}(\mathscr{A}^c) \le \exp(-m)$, that is, (33) is proved. Then it remains to show (34) and (35). For this purpose, we have by (30) that $\lambda \ge \frac{5}{2}\sigma^2$, and noting that $0 < q \le 1$,

$$\begin{aligned}\lambda &\ge \frac{a+1}{a-1}\sigma(1+\theta)2^{1-q}\left(\frac{5\sigma^2}{2\lambda} + r^q\right)^{\frac{1-q}{q}}\sqrt{\frac{2(1+b)\log n}{m}} \\ &= \frac{a+1}{a-1}\sigma(1+\theta)(2\rho)^{1-q}\sqrt{\frac{2(1+b)\log n}{m}}\end{aligned}$$

(due to (29)). Then one has by (32) that

$$\begin{aligned}\mathbb{P}(\mathscr{B}^c) &\le \mathbb{P}\left(\frac{a+1}{(a-1)m}(2\rho)^{1-q}\|X^\top e\|_\infty \ge \frac{a+1}{a-1}\sigma(1+\theta)(2\rho)^{1-q}\sqrt{\frac{2(1+b)\log n}{m}}\right) \\ &= \mathbb{P}\left(\frac{\|X^\top e\|_\infty}{m} \ge \sigma(1+\theta)\sqrt{\frac{2(1+b)\log n}{m}}\right).\end{aligned}$$

Hence, by assumption (26), Lemma 6 is applicable to ensuring (34). Moreover, it follows from the elementary probability theory that

$$\mathbb{P}(\mathscr{A} \cap \mathscr{B}) \ge \mathbb{P}(\mathscr{A}) - \mathbb{P}(\mathscr{B}^c) \ge 1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1}.$$

The proof is complete. □

We show in the following two propositions that the optimal solution $\hat{\beta}$ of the ℓ_q minimization problem (CP$_{q,\epsilon}$) or the ℓ_q regularization problem (RP$_{q,\lambda}$) satisfies the following dominant property on the support of the true underlying parameter of (1) with high probability:

$$\|(\hat{\beta}-\beta^*)_{J^c}\|_q^q \leq c\|(\hat{\beta}-\beta^*)_J\|_q^q \tag{36}$$

with $c=1$ or $c=a$, respectively.

Proposition 3. *Let $\bar{\beta}_{q,\epsilon}$ be an optimal solution of* (CP$_{q,\epsilon}$) *with ϵ given by* (29)*. Then it holds under the event $\mathscr{A}$ that*

$$\|(\bar{\beta}_{q,\epsilon}-\beta^*)_{J^c}\|_q \leq \|(\bar{\beta}_{q,\epsilon}-\beta^*)_J\|_q. \tag{37}$$

Proof. Let $e\in\mathscr{A}$. Recall that β^* satisfies the linear regression model (1), one has that $\|y-X\beta^*\|_2=\|e\|_2\leq\epsilon$ (under the event $\mathscr{A}$), and so, β^* is a feasible vector of (CP$_{q,\epsilon}$). Consequently, by the optimality of $\bar{\beta}_{q,\epsilon}$ for (CP$_{q,\epsilon}$), it follows that $\|\bar{\beta}_{q,\epsilon}\|_q\leq\|\beta^*\|_q$. Write $\delta:=\bar{\beta}_{q,\epsilon}-\beta^*$. Then we obtain that

$$\|\beta^*\|_q^q \geq \|\beta^*+\delta\|_q^q = \|\beta^*+\delta_J+\delta_{J^c}\|_q^q = \|\beta^*+\delta_J\|_q^q+\|\delta_{J^c}\|_q^q, \tag{38}$$

where the last equality holds because $\beta^*_{J^c}=0$. On the other hand, one has by (8) that $\|\beta^*+\delta_J\|_q^q\geq\|\beta^*\|_q^q-\|\delta_J\|_q^q$. This, together with (38), implies (37). The proof is complete. □

Proposition 4. *Let $\hat{\beta}_{q,\lambda}$ be an optimal solution of* (RP$_{q,\lambda}$) *with λ given by* (30)*. Suppose that* (26) *is satisfied. Then*

$$\|\hat{\beta}_{q,\lambda}-\beta^*\|_1 \leq (2\rho)^{1-q}\|\hat{\beta}_{q,\lambda}-\beta^*\|_q^q \tag{39}$$

under the event $\mathscr{A}$, and

$$\|(\hat{\beta}_{q,\lambda}-\beta^*)_{J^c}\|_q^q \leq a\|(\hat{\beta}_{q,\lambda}-\beta^*)_J\|_q^q \tag{40}$$

under the event $\mathscr{A}\cap\mathscr{B}$.

Proof. Let $e\in\mathscr{A}$. Since $\hat{\beta}_{q,\lambda}$ is an optimal solution of (RP$_{q,\lambda}$), one has that

$$\frac{1}{2m}\|y-X\hat{\beta}_{q,\lambda}\|_2^2+\lambda\|\hat{\beta}_{q,\lambda}\|_q^q \leq \frac{1}{2m}\|y-X\beta^*\|_2^2+\lambda\|\beta^*\|_q^q.$$

Then, by (1) and (28), it follows that

$$\|\hat{\beta}_{q,\lambda}\|_q^q \leq \frac{1}{2m\lambda}\|y-X\beta^*\|_2^2+\|\beta^*\|_q^q \leq \frac{1}{2m\lambda}\|e\|_2^2+r^q \leq \rho^q$$

(due to (29) and (31)). Write $\delta := \hat{\beta}_{q,\lambda} - \beta^*$. Then, we obtain by (7) and (28) that

$$\|\delta\|_1 \le \|\hat{\beta}_{q,\lambda}\|_1 + \|\beta^*\|_1 \le \|\hat{\beta}_{q,\lambda}\|_q + \|\beta^*\|_q \le \rho + r < 2\rho.$$

Consequently, noting that $0 < q \le 1$, one sees that $\frac{\|\delta\|_1}{2\rho} \le \left(\frac{\|\delta\|_1}{2\rho}\right)^q$, and then, by (7) that

$$\|\delta\|_1 \le (2\rho)^{1-q}\|\delta\|_1^q \le (2\rho)^{1-q}\|\delta\|_q^q. \tag{41}$$

This shows that (39) is proved. Then it remains to claim (40). To this end, noting that $\beta^*_{J^c} = 0$, we derive by Lemma 8 that

$$\begin{aligned}-\frac{1}{m}\|\delta\|_1\|X^\top e\|_\infty &\le \lambda\|\beta^*\|_q^q - \lambda\|\beta^* + \delta\|_q^q \\ &= \lambda\|\beta^*_J\|_q^q - \lambda\|\beta^*_J + \delta_J\|_q^q - \lambda\|\delta_{J^c}\|_q^q \\ &\le \lambda\left(\|\delta_J\|_q^q - \|\delta_{J^c}\|_q^q\right)\end{aligned}$$

(by (8)). This, together with (41), yields that

$$\lambda\left(\|\delta_J\|_q^q - \|\delta_{J^c}\|_q^q\right) \ge -\frac{1}{m}(2\rho)^{1-q}\|\delta\|_q^q\|X^\top e\|_\infty.$$

Then, under the event $\mathscr{A} \cap \mathscr{B}$, we obtain by (32) that

$$(a+1)\left(\|\delta_J\|_q^q - \|\delta_{J^c}\|_q^q\right) \ge -(a-1)\|\delta\|_q^q = -(a-1)(\|\delta_J\|_q^q + \|\delta_{J^c}\|_q^q),$$

which yields (40). The proof is complete. □

Remark 3. *By Lemma 9, Propositions 3 and 4 show that* (37) *holds with probability at least* $1 - \exp(-m)$*, and* (40) *holds with probability at least* $1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1}$ *if* (26) *is satisfied, respectively.*

By virtue of Lemma 9 and Proposition 3, one of the main theorems of this section is as follows, in which we establish the ℓ_2 recovery bound for the ℓ_q minimization problem $(\mathrm{CP}_{q,\epsilon})$ under the q-REC. This theorem provides a unified framework to show that one can stably recover the underlying parameter with high probability via solving the ℓ_q minimization problem when the design matrix satisfies the weak q-REC.

Theorem 1. *Let* $\bar{\beta}_{q,\epsilon}$ *be an optimal solution of* $(\mathrm{CP}_{q,\epsilon})$ *with* ϵ *given by* (29)*. Suppose that* X *satisfies the* q-$\mathrm{REC}(s,t,1)$*. Then, with probability at least* $1 - \exp(-m)$*, we have that*

$$\|\bar{\beta}_{q,\epsilon} - \beta^*\|_2^2 \le \frac{1 + \left(\frac{s}{t}\right)^{\frac{2}{q}-1}}{\phi_q^2(s,t,1,X)} 4\epsilon^2. \tag{42}$$

Proof. Write $\delta := \bar{\beta}_{q,\epsilon} - \beta^*$, and let $J_* := J \cup J_0(\delta;t)$ (defined by (17)). Fix $e \in \mathscr{A}$. Then it follows from [24, Lemma 7] and Proposition 3 that

$$\|\delta_{J_*^c}\|_2^2 \le t^{1-\frac{2}{q}} \|\delta_{J^c}\|_q^2 \le t^{1-\frac{2}{q}} \|\delta_J\|_q^2 \le \left(\frac{s}{t}\right)^{\frac{2}{q}-1} \|\delta_J\|_2^2 \le \left(\frac{s}{t}\right)^{\frac{2}{q}-1} \|\delta_{J_*}\|_2^2$$

(by (7)), and so

$$\|\delta\|_2^2 = \|\delta_{J_*}\|_2^2 + \|\delta_{J_*^c}\|_2^2 \le \left(1 + \left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right) \|\delta_{J_*}\|_2^2. \tag{43}$$

Recalling that β^* satisfies the linear regression model (1), we have that $\|y - X\beta^*\|_2 = \|e\|_2 \le \epsilon$ (by (31)), and then

$$\|X\delta\|_2 = \|X\bar{\beta}_{q,\epsilon} - X\beta^*\|_2 \le \|X\bar{\beta}_{q,\epsilon} - y\|_2 + \|X\beta^* - y\|_2 \le 2\epsilon. \tag{44}$$

On the other hand, Proposition 3 is applicable to concluding that (37) holds, which shows $\delta \in C_q(s,1)$ (cf. (13)). Consequently, we obtain by the assumption of the q-REC$(s,t,1)$ that

$$\|\delta_{J_*}\|_2 \le \frac{\|X\delta\|_2}{\phi_q(s,t,1,X)}.$$

This, together with (43) and (44), implies that (42) holds under the event $\mathscr{A}$. Noting from Lemma 9 that $\mathbb{P}(\mathscr{A}) \ge 1 - \exp(-m)$, we obtain the conclusion. The proof is complete. □

In the special case when the underlying data is noise-free, Theorem 1 shows that (CP$_{q,\epsilon}$) can exactly predict the parameter for the deterministic linear regression with high probability under the lower-order REC. For the realistic scenario where the measurements are noisy-aware, Theorem 1 illustrates the stable recovery capability of (CP$_{q,\epsilon}$) in the sense that its solution approaches to the true sparse parameter within a tolerance proportional to the noise level with high probability. Moreover, Theorem 1 establishes the ℓ_2 recovery bound $\|\bar{\beta}_{q,\epsilon} - \beta^*\|_2 = O(\epsilon)$ under a weaker assumption than the RIP-type or MIP-type condition used in [16, 36], respectively.

As a special case of Theorem 1 when $q = 1$, the following corollary presents the ℓ_2 recovery bound of the ℓ_1 minimization problem (CP$_{1,\epsilon}$) as

$$\|\bar{\beta}_{1,\epsilon} - \beta^*\|_2 = O(\epsilon) \tag{45}$$

under the classical REC. This result improves the ones in [7, 9] under a weaker assumption, in which the ℓ_2 recovery bound (45) was obtained under the RIP-type conditions.

Corollary 1. *Let $\bar{\beta}_{1,\epsilon}$ be an optimal solution of* $(\mathrm{CP}_{1,\epsilon})$ *with ϵ given by* (29). *Suppose that X satisfies the* 1-REC$(s,t,1)$. *Then, with probability at least* $1-\exp(-m)$, *we have that*

$$\|\bar{\beta}_{1,\epsilon}-\beta^*\|_2^2 \le \frac{1+\frac{s}{t}}{\phi_1^2(s,t,1,X)}4\epsilon^2.$$

The other main theorem of this section is as follows, in which we exploit the statistical properties of the ℓ_q regularization problem $(\mathrm{RP}_{q,\lambda})$ under the q-REC. The results include the estimation of prediction loss and recovery bound of parameter approximation, and also the oracle property, which provides an upper bound on the prediction loss plus the violation of false parameter estimation.

Theorem 2. *Let $\hat{\beta}_{q,\lambda}$ be an optimal solution of* $(\mathrm{RP}_{q,\lambda})$ *with λ given by* (30). *Suppose that X satisfies the q*-REC(s,t,a) *and that* (26) *is satisfied. Then, with probability at least* $1-\exp(-m)-\left(n^b\sqrt{\pi\log n}\right)^{-1}$, *we have that*

$$\frac{1}{m}\|X\hat{\beta}_{q,\lambda}-X\beta^*\|_2^2 \le \left(\frac{2a\lambda}{(\phi_q(s,t,a,X)/\sqrt{m})^q}\right)^{\frac{2}{2-q}} s, \tag{46}$$

$$\frac{1}{2m}\|X\hat{\beta}_{q,\lambda}-X\beta^*\|_2^2+\lambda\|(\hat{\beta}_{q,\lambda})_{J^c}\|_q^q \le \left(\frac{2^{\frac{q}{2}}a\lambda}{(\phi_q(s,t,a,X)/\sqrt{m})^q}\right)^{\frac{2}{2-q}} s, \tag{47}$$

$$\|\hat{\beta}_{q,\lambda}-\beta^*\|_2^2 \le \left(1+a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)\left(\frac{2a\lambda}{(\phi_q(s,t,a,X)/\sqrt{m})^2}\right)^{\frac{2}{2-q}} s. \tag{48}$$

Proof. Write $\delta := \hat{\beta}_{q,\lambda}-\beta^*$ and fix $e\in\mathscr{A}\cap\mathscr{B}$. Note by (39) and (32) that

$$\frac{1}{m}\|\delta\|_1\|X^\top e\|_\infty \le \frac{a-1}{a+1}\lambda\|\delta\|_q^q.$$

This, together with Lemma 8, implies that

$$\begin{aligned}\frac{1}{2m}\|X\hat{\beta}_{q,\lambda}-X\beta^*\|_2^2 &\le \lambda\|\beta^*\|_q^q-\lambda\|\hat{\beta}_{q,\lambda}\|_q^q+\frac{a-1}{a+1}\lambda\|\delta\|_q^q\\ &\le \lambda\|\delta_J\|_q^q-\lambda\|(\hat{\beta}_{q,\lambda})_{J^c}\|_q^q+\frac{a-1}{a+1}\lambda\|\delta\|_q^q\end{aligned} \tag{49}$$

(noting that $\beta^*_{J^c}=0$ and by (8)). Let $J_* := J\cup J_0(\delta;t)$. One has by (40) and (7) that

$$\lambda\|\delta_J\|_q^q+\frac{a-1}{a+1}\lambda\|\delta\|_q^q \le a\lambda\|\delta_J\|_q^q \le a\lambda s^{1-\frac{q}{2}}\|\delta_J\|_2^q,$$

and by the assumption of the q-REC(s,t,a) that

$$\|\delta_J\|_2 \le \|\delta_{J^*}\|_2 \le \frac{\|X\delta\|_2}{\phi_q(s,t,a,X)}.$$

These two inequalities, together with (49), imply that

$$\frac{1}{2m}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^2 + \lambda\|(\hat{\beta}_{q,\lambda})_{J^c}\|_q^q \le \frac{a\lambda s^{1-\frac{q}{2}}}{\phi_q^q(s,t,a,X)}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^q.$$

This yields that

$$\text{(46) and (47) hold under the event } \mathscr{A}\cap\mathscr{B}. \tag{50}$$

Furthermore, it follows from [24, Lemma 7] that

$$\|\delta_{J_*^c}\|_2^2 \le t^{1-\frac{2}{q}}\|\delta_{J^c}\|_q^2 \le a^{\frac{2}{q}}t^{1-\frac{2}{q}}\|\delta_J\|_q^2 \le a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\|\delta_J\|_2^2.$$

(by (40) and (7)). By the assumption of the q-REC(s,t,a), one has by (46) that

$$\|\delta_{J_*}\|_2^2 \le \frac{\|X\delta\|_2^2}{\phi_q^2(s,t,a,X)} \le \left(\frac{2a\lambda}{(\phi_q(s,t,a,X)/\sqrt{m})^2}\right)^{\frac{2}{2-q}} s.$$

Hence we obtain that

$$\begin{aligned}\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 = \|\delta_{J_*}\|_2^2 + \|\delta_{J_*^c}\|_2^2 &\le \left(1 + a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)\|\delta_{J_*}\|_2^2 \\ &\le \left(1 + a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)\left(\frac{2a\lambda}{(\phi_q(s,t,a,X)/\sqrt{m})^2}\right)^{\frac{2}{2-q}} s.\end{aligned}$$

This shows that

$$\text{(48) holds under the event } \mathscr{A}\cap\mathscr{B}. \tag{51}$$

By assumption (26), Lemma 9 is applicable to concluding that

$$\mathbb{P}(\mathscr{A}\cap\mathscr{B}) \ge 1 - \exp(-m) - \left(n^b\sqrt{\pi\log n}\right)^{-1}.$$

This, together with (50) and (51), yields that (46)-(48) hold with probability at least $1 - \exp(-m) - \left(n^b\sqrt{\pi\log n}\right)^{-1}$. The proof is complete. □

Remark 4. (i) *It is worth noting that each of the estimations provided in Theorem 2 (cf.* (46)-(48)*) involves the term* $\phi_q(s,t,a,X)/\sqrt{m}$ *in the denominator, which scales as a constant if* X *has i.i.d. Gaussian entries; see Remark 1(ii).*

(ii) *Theorem 2 provides a unified framework of the statistical properties of the* ℓ_q *regularization problem under the weak* q-REC *that is one of the weakest regularity conditions in the literature, in which each of the obtained estimations depends on the noise amplitude and sample size. In particular, for the regularization parameter scaling as* $\lambda \asymp \max\left(\sigma\sqrt{\frac{\log n}{m}},\ \sigma^2\right)$ *(cf.* (30)*), Theorem 2 indicates the prediction loss and the* ℓ_2 *recovery bound for* $(\mathrm{RP}_{q,\lambda})$ *scale as*

$$\frac{1}{m}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^2 = O\left(\left(\sigma^2\frac{\log n}{m}\right)^{\frac{1}{2-q}} s\right),$$

and

$$\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 = O\left(\left(\sigma^2\frac{\log n}{m}\right)^{\frac{1}{2-q}} s\right). \tag{52}$$

(iii) *It was shown in [43] that the global solution of the FCP sparse linear regression, including the SCAD and MCP as special cases, has an* ℓ_2 *recovery bound* $O(\lambda^2 s)$ *under the SEC. Though the recovery bounds are slightly better than* (52)*, the condition required is substantially stronger than the q-REC. In [43], the authors also established the oracle property for the* ℓ_0 *regularization method under the SEC; while its* ℓ_2 *recovery bound cannot be guaranteed in their work. We shall see in section 5 that the* ℓ_q *regularization method performs better in parameter estimation than either the SCAD/MCP or the* ℓ_0 *regularization method via several numerical experiments.*

Remark 5. *Recently, some works concerned the statistical property for the local minimum of some nonconvex regularization problems; see [25, 27].*

(i) *Loh and Wainwright [27] studied the* ℓ_2 *recovery bound for the local minimum of a general regularization problem:*

$$\min \mathcal{L}_m(\beta; X) + \sum_{j=1}^{n} \rho_\lambda(\beta_j), \tag{53}$$

where $\mathcal{L}_m : \mathbb{R}^n \times \mathbb{R}^{m\times n} \to \mathbb{R}$ *is the loss function, and* $\rho_\lambda : \mathbb{R} \to \mathbb{R}$ *is the (possibly nonconvex) penalty function. In [27], the penalty function* ρ_λ *is assumed to satisfy the following assumptions:*

(a) $\rho_\lambda(0) = 0$ *and is symmetric around zero;*

(b) ρ_λ *is nondecreasing on* $\mathbb{R}_+$*;*

(c) *For* $t > 0$*, the function* $t \mapsto \frac{\rho_\lambda(t)}{t}$ *is nonincreasing in* t*;*

(d) ρ_λ *is differentiable for each* $t \neq 0$ *and subdifferentiable at* $t = 0$*, with* $\lim\limits_{t \to 0^+} \rho_\lambda^{'}(t) = \lambda L$*;*

(e) *There exists* $\mu > 0$ *such that* $\rho_{\lambda,\mu}(t) := \rho_\lambda(t) + \frac{\mu}{2}t^2$ *is convex.*

Loh and Wainwright established in [27, Theorem 1] the ℓ_2 *recovery bound for the critical point satisfying the first-order necessary condition of* (53) *under the restricted strong convex condition, which is a variant of the classical REC.*

The ℓ_q *norm can be reformulated as the penalty function* $\rho_\lambda(\beta_j) := \lambda|\beta_j|^q$*, however, it does not satisfy assumptions* (d) *or* (e)*; in particular, assumption* (e) *plays a key role in the establishment of oracle property and* ℓ_2 *recovery bound for the local minimum. Therefore, the result in [27] cannot be directly applied to the* ℓ_q *regularization problem, and the oracle property for the general local minimum of the* ℓ_q *regularization problem is still unsolved at this moment.*

(ii) *Liu et al. [25] studied the statistical property of the FCP sparse linear regression and presented the oracle property and* ℓ_2 *recovery bound for the certain local minimum, which satisfies a subspace second-order necessary condition and lies in the level set of the FCP regularized function at the true solution, under the SEC. Although the* ℓ_q *regularizer is beyond the FCP, our established Theorem 2 provides a theoretical result similar to [25] in the sense that the oracle property and* ℓ_2 *recovery bound are shown for the local minimum within the level set of the* ℓ_q *regularized function at the true solution.*

As an application of Theorem 2 to the case when $q = 1$, the following corollary presents the statistical properties of the ℓ_1 regularization problem under the classical REC, which covers [4, Theorem 7.2] as a special case when $a = 3$, $\theta = 0$ and $b = 0$. The same ℓ_2 recovery bound rate $O(\sigma^2 s \log n/m)$ was reported in [42] under the sparse Riesz condition, which is comparable with the classical REC; while the same oracle inequality rate $O(\sigma^2 s \log n/m)$ was established in [39] under the compatibility condition, which is slightly weaker than the classical REC but cannot guarantee the ℓ_2 recovery bound.

Corollary 2. *Let* $\hat{\beta}_{1,\lambda}$ *be an optimal solution of* $(\mathrm{RP}_{1,\lambda})$ *with*

$$\lambda = 2\sigma(1+\theta)\sqrt{\frac{2(1+b)\log n}{m}}.$$

Suppose that X satisfies the 1-REC$(s,t,3)$ *and that* (26) *is satisfied. Then, with probability at least* $1-\left(n^b\sqrt{\pi\log n}\right)^{-1}$, *we have that*

$$\frac{1}{m}\|X\hat{\beta}_{1,\lambda}-X\beta^*\|_2^2 \le \frac{288(1+b)(1+\theta)^2}{\phi_1^2(s,t,3,X)/m}\sigma^2 s\frac{\log n}{m},$$

$$\frac{1}{2m}\|X\hat{\beta}_{1,\lambda}-X\beta^*\|_2^2+\lambda\|(\hat{\beta}_{1,\lambda})_{J^c}\|_1 \le \frac{144(1+b)(1+\theta)^2}{\phi_1^2(s,t,3,X)/m}\sigma^2 s\frac{\log n}{m},$$

$$\|\hat{\beta}_{1,\lambda}-\beta^*\|_2^2 \le \frac{288(1+b)(1+\theta)^2\left(1+9\frac{s}{t}\right)}{\phi_1^4(s,t,3,X)/m^2}\sigma^2 s\frac{\log n}{m}.$$

4 Recovery Bounds for Random Design

In practical applications, it is a more realistic scenario that the design matrix X is random. In this section, we consider this situation and present the ℓ_2 recovery bounds for (CP$_{q,\epsilon}$) and (RP$_{q,\lambda}$) by virtue of the results obtained in the preceding section. In particular, throughout this section, we shall assume that the linear regression model (1) involves a Gaussian noise, i.e., $e\sim\mathscr{N}(0,\sigma^2\mathbb{I}_m)$, and

$$X\in\mathbb{R}^{m\times n} \text{ is a Gaussian random design with i.i.d. } \mathscr{N}(0,\Sigma) \text{ rows,}$$

that is, $X_{1\cdot},\ldots,X_{m\cdot}$ are i.i.d. random vectors with each $X_{i\cdot}\sim\mathscr{N}(0,\Sigma)$. Recall that a, θ, and b are given by (27), and let (s,t) be a pair of integers satisfying (5).

To study the statistical properties of (CP$_{q,\epsilon}$) and (RP$_{q,\lambda}$) with a random design X, we first provide some sufficient condition for the q-REC of X in terms of the population covariance matrix Σ. For this purpose, we use $\Sigma^{\frac{1}{2}}$ to denote the square root of Σ and $\zeta(\Sigma):=\max_{1\le j\le n}\Sigma_{j,j}$ to denote the maximal variance. Let $a>0$, and two random events related to the linear regression model (1) with X being a Gaussian random design are defined as follows

$$\mathscr{C}_a := \left\{\phi_q(s,t,a,X) > \frac{\sqrt{m}}{2}\phi_q(s,t,a,\Sigma^{\frac{1}{2}})\right\}, \tag{54}$$

and

$$\mathscr{D} := \left\{\max_{1\le j\le n}\|X_{\cdot j}\|_2 \le (1+\theta)\sqrt{m}\right\}. \tag{55}$$

The following lemma is taken from [1, Supplementary, Lemma 6], which is useful for providing a sufficient condition for the q-REC of X.

Lemma 10. *There exist universal positive constants* (c_1, c_2) *(independent of* m, n, Σ*) such that it holds with probability at least* $1 - \exp(-c_2 m)$ *that, for each* $\delta \in \mathbb{R}^n$

$$\frac{\|X\delta\|_2^2}{m} \geq \frac{1}{2}\|\Sigma^{\frac{1}{2}}\delta\|_2^2 - c_1\zeta(\Sigma)\frac{\log n}{m}\|\delta\|_1^2. \tag{56}$$

The following lemma calculates the probabilities of events $\mathscr{C}_c$ and $\mathscr{D}$, which is crucial for establishing the ℓ_2 recovery bounds of $(\mathrm{CP}_{q,\epsilon})$ and $(\mathrm{RP}_{q,\lambda})$ with a random design X. In particular, part (i) of this lemma shows that the Gaussian random design X satisfies the q-REC with high probability as long as the sample size m is sufficiently large and the square root of its population covariance matrix $\Sigma^{\frac{1}{2}}$ satisfies the q-REC; part (ii) of this lemma presents that each column of the Gaussian random design X has an Euclidean norm scaling as $\sqrt{m}$ with an overwhelming probability.

Lemma 11. (i) *Let* $a > 0$. *Suppose that* $\Sigma^{\frac{1}{2}}$ *satisfies the* q-REC(s, t, a). *Then, there exist universal positive constants* (c_1, c_2) *(independent of* m, n, q, s, t, a, Σ*) such that, if*

$$m > \frac{c_1\zeta(\Sigma)}{\phi_q^2(s, t, a, \Sigma^{\frac{1}{2}})}\left(\sqrt{s+t} + a\sqrt{s}\left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right)^2 \log n, \tag{57}$$

then

$$\mathbb{P}(\mathscr{C}_a) \geq 1 - \exp(-c_2 m). \tag{58}$$

(ii) *Suppose that* $\Sigma_{j,j} = 1$ *for all* $j = 1, \ldots, n$. *Then, there exist universal positive constants* (c_3, c_4) *and* $\tau \geq 1$ *(independent of* m, n, θ, Σ*) such that, if*

$$m > \frac{c_3\tau^4}{\theta^2}\log n, \tag{59}$$

then

$$\mathbb{P}(\mathscr{D}) \geq 1 - 2\exp(-c_4\theta^2 m/\tau^4). \tag{60}$$

Proof. (i) We first claim that

$$\phi_q(s, t, a, X) > \frac{\sqrt{m}}{2}\phi_q(s, t, a, \Sigma^{\frac{1}{2}}), \tag{61}$$

whenever (56) holds for each $\delta \in \mathbb{R}^n$. To this end, we suppose that (56) is satisfied for each $\delta \in \mathbb{R}^n$. Fix $\delta \in C_q(s, a)$, and let J, r, J_k (for each $k \in \mathbb{N}$)

and J_* be defined, respectively, as in the beginning of the proof of Lemma 5. Then (22) follows directly, and one has that

$$\begin{aligned}\|\delta\|_1 &= \|\delta_{J_*}\|_1 + \|\delta_{J_*^c}\|_1 \\ &\le \sqrt{s+t}\|\delta_{J_*}\|_2 + a\sqrt{s}\left(\frac{as}{t}\right)^{\frac{1}{q}-1}\|\delta_J\|_2 \\ &\le \left(\sqrt{s+t} + a\sqrt{s}\left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right)\|\delta_{J_*}\|_2.\end{aligned} \tag{62}$$

By the assumption that $\Sigma^{\frac{1}{2}}$ satisfies the q-REC(s,t,a), it follows that

$$\|\Sigma^{\frac{1}{2}}\delta\|_2^2 \ge \phi_q^2(s,t,a,\Sigma^{\frac{1}{2}})\|\delta_{J_*}\|_2^2.$$

Substituting this inequality and (62) into (56) yields

$$\frac{\|X\delta\|_2^2}{m} \ge \left(\frac{1}{2}\phi_q^2(s,t,a,\Sigma^{\frac{1}{2}}) - c_1\zeta(\Sigma)\left(\sqrt{s+t} + a\sqrt{s}\left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right)^2\frac{\log n}{m}\right)\|\delta_{J_*}\|_2^2.$$

This, together with (57), shows that

$$\frac{\|X\delta\|_2^2}{m} \ge \frac{1}{4}\phi_q^2(s,t,a,\Sigma^{\frac{1}{2}})\|\delta_{J_*}\|_2^2.$$

Since δ and J satisfying (20) are arbitrary, we derive by (6) that (61) holds, as desired. Then, Lemma 10 is applicable to concluding (58).
(ii) Noting by the assumption that $\Sigma_{j,j} = 1$ for all $j = 1, \dots, n$, [46, Theorem 1.6] is applicable to showing that there exist universal positive constants (c_1, c_2) and $\tau \ge 1$ such that

$$\mathbb{P}\left(\cap_{j=1}^n \left\{(1-\theta)\sqrt{m} \le \|X_{\cdot j}\|_2 \le (1+\theta)\sqrt{m}\right\}\right) \ge 1 - 2\exp(-c_2\theta^2 m/\tau^4),$$

whenever m satisfies (59). Then it immediately follows from (55) that

$$\begin{aligned}\mathbb{P}(\mathscr{D}) &= \mathbb{P}(\cap_{j=1}^n\{\|X_{\cdot j}\|_2 \le (1+\theta)\sqrt{m}\}) \\ &\ge 1 - 2\exp(-c_2\theta^2 m/\tau^4),\end{aligned}$$

that is, (60) is proved. □

Remark 6. (i) *As a direct application of Lemma 11(i), the classical* REC *is satisfied by X with high probability if $\Sigma^{\frac{1}{2}}$ satisfies the classical* REC*(s,t,a) and*

$$m > \frac{c_1\zeta(\Sigma)}{\phi_1^2(s,t,a,\Sigma^{\frac{1}{2}})}\left(\sqrt{s+t} + a\sqrt{s}\right)^2\log n,$$

which covers [32, Corollary 1] as a special case when $t = 0$.

(ii) *Recall from Remark 1 that* $\phi_q(s,t,a,X)$ *given by* (6) *usually scales as* $\sqrt{m}$ *independent of* s *and* n *for the Gaussian random design* X. *Then Lemma 11(i) is applicable to indicating that* $\phi_q(s,t,a,\Sigma^{\frac{1}{2}})$ *usually scales as a constant independent of* s, m *and* n.

Below, we consider the dominant property (36) in the situation when X is a Gaussian random design. For the ℓ_q minimization problem ($\mathrm{CP}_{q,\epsilon}$), Proposition 3 is still applicable for the case when X is a Gaussian random design since it does not rely on the assumption of X, and thus, (37) holds with the same probability for the random design scenario; see Remark 3. In the following proposition, we show the dominant property (40) for the ℓ_q regularization problem ($\mathrm{RP}_{q,\lambda}$) with a random design by virtue of Proposition 4. Recall that ϵ, λ, ρ and the events $\mathscr{A}$ and $\mathscr{B}$ are given in the preceding section; see (29)-(32) for details.

Proposition 5. *Let* $\hat{\beta}_{q,\lambda}$ *be an optimal solution of* ($\mathrm{RP}_{q,\lambda}$) *with* λ *given by* (30). *Suppose that* $\Sigma_{j,j} = 1$ *for all* $j = 1, \ldots, n$. *Then, there exist universal positive constants* (c_1, c_2) *and* $\tau \geq 1$ *(independent of* $m, n, q, a, \theta, b, \epsilon, r, \lambda, \Sigma$*) such that, if*

$$m > \frac{c_1 \tau^4}{\theta^2} \log n, \tag{63}$$

then (40) *holds with probability at least* $(1 - \left(n^b\sqrt{\pi \log n}\right)^{-1})(1 - 2\exp(-c_2\theta^2 m/\tau^4)) - \exp(-m)$.

Proof. By (55), one sees by Proposition 4 that (40) holds under the event $\mathscr{A} \cap \mathscr{B} \cap \mathscr{D}$. Then it remains to estimate $\mathbb{P}(\mathscr{A} \cap \mathscr{B} \cap \mathscr{D})$. By Lemma 11(ii), there exist universal positive constants (c_1, c_2) and $\tau \geq 1$ such that

$$\mathbb{P}(\mathscr{D}) \geq 1 - 2\exp(-c_2\theta^2 m/\tau^4),$$

whenever m satisfies (63). From Lemma 9 (cf. (34)), we have also by (55) that

$$\mathbb{P}(\mathscr{B}|\mathscr{D}) \geq 1 - (n^b\sqrt{\pi \log n})^{-1}.$$

Then, it follows that

$$\begin{aligned}\mathbb{P}(\mathscr{B} \cap \mathscr{D}) &= \mathbb{P}(\mathscr{B}|\mathscr{D})\mathbb{P}(\mathscr{D}) \\ &\geq (1 - (n^b\sqrt{\pi \log n})^{-1})(1 - 2\exp(-c_2\theta^2 m/\tau^4)),\end{aligned}$$

and then by the elementary probability theory and (33) that,

$$\begin{aligned}\mathbb{P}(\mathscr{A}\cap\mathscr{B}\cap\mathscr{D}) &= \mathbb{P}(\mathscr{B}\cap\mathscr{D}) - \mathbb{P}(\mathscr{B}\cap\mathscr{D}\cap\mathscr{A}^c)\\ &\geq \mathbb{P}(\mathscr{B}\cap\mathscr{D}) + \mathbb{P}(\mathscr{A}) - 1\\ &\geq \left(1-\left(n^b\sqrt{\pi\log n}\right)^{-1}\right)(1-2\exp(-c_2\theta^2 m/\tau^4)) - \exp(-m),\end{aligned}$$

whenever m satisfies (63). The proof is complete. □

Now we are ready to present the main theorems of this section, in which we establish the ℓ_2 recovery bounds for $(\mathrm{CP}_{q,\epsilon})$ and $(\mathrm{RP}_{q,\lambda})$ when X is a Gaussian random design. The first theorem illustrates the stable recovery capability of the ℓ_q minimization problem $(\mathrm{CP}_{q,\epsilon})$ (within a tolerance proportional to the noise) with high probability when the design matrix is random as long as the vector β^* is sufficiently sparse and the sample size m is sufficiently large.

Theorem 3. *Let $\bar\beta_{q,\epsilon}$ be an optimal solution of $(\mathrm{CP}_{q,\epsilon})$ with ϵ given by (29). Suppose that $\Sigma^{\frac{1}{2}}$ satisfies the q-$\mathrm{REC}(s,t,1)$. Then, there exist universal positive constants (c_1,c_2) (independent of $m,n,q,s,t,\epsilon,\Sigma$) such that, if (57) is satisfied, then it holds with probability at least $(1-\exp(-m))(1-\exp(-c_2m))$ that*

$$\|\bar\beta_{q,\epsilon}-\beta^*\|_2^2 \leq \frac{16(1+\left(\frac{s}{t}\right)^{\frac{2}{q}-1})}{m\phi_q^2(s,t,1,\Sigma^{\frac{1}{2}})}\epsilon^2. \tag{64}$$

Proof. To simplify the proof, corresponding to inequalities (42) and (64), we define the following two events

$$\begin{aligned}\mathscr{E}_1 &:= \left\{\|\bar\beta_{q,\epsilon}-\beta^*\|_2^2 \leq \frac{4(1+\left(\frac{s}{t}\right)^{\frac{2}{q}-1})}{\phi_q^2(s,t,1,X)}\epsilon^2\right\},\\ \mathscr{E}_2 &:= \left\{\|\bar\beta_{q,\epsilon}-\beta^*\|_2^2 \leq \frac{16(1+\left(\frac{s}{t}\right)^{\frac{2}{q}-1})}{m\phi_q^2(s,t,1,\Sigma^{\frac{1}{2}})}\epsilon^2\right\}.\end{aligned}$$

Then, by the definition of $\mathscr{C}_1$ (54), we have that $\mathscr{C}_1\cap\mathscr{E}_1\subseteq\mathscr{E}_2$ and thus

$$\mathbb{P}(\mathscr{E}_2)\geq\mathbb{P}(\mathscr{E}_1\cap\mathscr{C}_1)=\mathbb{P}(\mathscr{E}_1|\mathscr{C}_1)\mathbb{P}(\mathscr{C}_1). \tag{65}$$

Note by Theorem 1 that

$$\mathbb{P}(\mathscr{E}_1|\mathscr{C}_1)\geq 1-\exp(-m). \tag{66}$$

By Lemma 11(i) (with $a = 1$), there exist universal positive constants (c_1, c_2) such that (57) ensures (58). Then we obtain by (65) and (66) that

$$\mathbb{P}(\mathscr{E}_2) \geq (1 - \exp(-m))(1 - \exp(-c_2 m)),$$

whenever m satisfies (57). The proof is complete. □

As a direct application of Theorem 3 to the special case when $q = 1$, the following corollary presents the ℓ_2 recovery bound of the ℓ_1 minimization problem $(\text{CP}_{1,\epsilon})$ with a Gaussian random design as

$$\|\bar{\beta}_{1,\epsilon} - \beta^*\|_2 = O(\epsilon)$$

under the classical REC.

Corollary 3. *Let $\bar{\beta}_{1,\epsilon}$ be an optimal solution of* $(\text{CP}_{1,\epsilon})$ *with ϵ given by* (29)*. Suppose that $\Sigma^{\frac{1}{2}}$ satisfies the* 1-REC$(s, t, 1)$*. Then, there exist universal positive constants (c_1, c_2) (independent of $m, n, q, s, t, \epsilon, \Sigma$) such that, if*

$$m > \frac{c_1 \zeta(\Sigma)}{\phi_1^2(s, t, 1, \Sigma^{\frac{1}{2}})}(\sqrt{s+t} + \sqrt{s})^2 \log n,$$

then it holds with probability at least $(1 - \exp(-m))(1 - \exp(-c_2 m))$ *that*

$$\|\bar{\beta}_{1,\epsilon} - \beta^*\|_2^2 \leq \frac{16(1 + \frac{s}{t})}{m\phi_1^2(s, t, 1, \Sigma^{\frac{1}{2}})}\epsilon^2.$$

The other main theorem of this section is as follows, in which we exploit the estimation of prediction loss, the oracle property and the ℓ_2 recovery bound of parameter approximation of the ℓ_q regularization problem $(\text{RP}_{q,\lambda})$ with a Gaussian random design under the q-REC of the square root of its population covariance matrix.

Theorem 4. *Let $\hat{\beta}_{q,\lambda}$ be an optimal solution of* $(\text{RP}_{q,\lambda})$ *with λ given by* (30)*. Suppose that $\Sigma_{j,j} = 1$ for all $j = 1, \ldots, n$ and $\Sigma^{\frac{1}{2}}$ satisfies the q-*REC(s, t, a)*. Then, there exist universal positive constants (c_1, c_2, c_3, c_4) and $\tau \geq 1$ (independent of $m, n, q, s, t, a, \theta, b, \epsilon, r, \lambda, \Sigma$) such that, if*

$$m > \max\left\{ \frac{c_1(\sqrt{s+t} + a^{\frac{1}{q}}\sqrt{s}\left(\frac{s}{t}\right)^{\frac{1}{q}-1})^2}{\phi_q^2(s, t, a, \Sigma^{\frac{1}{2}})} \log n, \ \frac{c_3 \tau^4}{\theta^2} \log n \right\}, \tag{67}$$

then it holds with probability at least $\left(1-\exp(-m)-\left(n^b\sqrt{\pi\log n}\right)^{-1}\right)(1-\exp(-c_2m)-2\exp(-c_4\theta^2m/\tau^4))$ *that*

$$\frac{1}{m}\|X\hat{\beta}_{q,\lambda}-X\beta^*\|_2^2\leq\left(\frac{2^{q+1}a\lambda}{\phi_q^q(s,t,a,\Sigma^{\frac{1}{2}})}\right)^{\frac{2}{2-q}}s, \tag{68}$$

$$\frac{1}{2m}\|X\hat{\beta}_{q,\lambda}-X\beta^*\|_2^2+\lambda\|(\hat{\beta}_{q,\lambda})_{J^c}\|_q^q\leq\left(\frac{8^{\frac{q}{2}}a\lambda}{\phi_q^q(s,t,a,\Sigma^{\frac{1}{2}})}\right)^{\frac{2}{2-q}}s, \tag{69}$$

$$\|\hat{\beta}_{q,\lambda}-\beta^*\|_2^2\leq\left(1+a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)\left(\frac{8a\lambda}{\phi_q^2(s,t,a,\Sigma^{\frac{1}{2}})}\right)^{\frac{2}{2-q}}s. \tag{70}$$

Proof. To simplify the proof, we define the following six events

$$\mathscr{F}_1=\{(46)\text{ happens}\},\quad \mathscr{F}_2=\{(47)\text{ happens}\},\quad \mathscr{F}_3=\{(48)\text{ happens}\},$$
$$\mathscr{G}_1=\{(68)\text{ happens}\},\quad \mathscr{G}_2=\{(69)\text{ happens}\},\quad \mathscr{G}_3=\{(70)\text{ happens}\}.$$

Fix $i\in\{1,2,3\}$. Then, we have by (54) that $\mathscr{C}_a\cap\mathscr{F}_i\subseteq\mathscr{G}_i$ and thus

$$\mathbb{P}(\mathscr{G}_i)\geq\mathbb{P}(\mathscr{C}_a\cap\mathscr{F}_i). \tag{71}$$

By Lemma 11, there exist universal positive constants (c_1,c_2,c_3,c_4) and $\tau\geq 1$ such that, (67) ensures (58) and (60). Then it follows from (58) and (60) that

$$\mathbb{P}(\mathscr{C}_a\cap\mathscr{D})\geq\mathbb{P}(\mathscr{C}_a)+P(\mathscr{D})-1\geq 1-\exp(-c_2m)-2\exp(-c_4\theta^2m/\tau^4), \tag{72}$$

whenever m satisfies (67). Recall from Theorem 2 that

$$\mathbb{P}(\mathscr{F}_i|\mathscr{C}_a\cap\mathscr{D})\geq 1-\exp(-m)-\left(n^b\sqrt{\pi\log n}\right)^{-1}.$$

This, together with (72), implies that

$$\begin{aligned}\mathbb{P}(\mathscr{C}_a\cap\mathscr{F}_i)&\geq\mathbb{P}(\mathscr{F}_i|\mathscr{C}_a\cap\mathscr{D})\,\mathbb{P}(\mathscr{C}_a\cap\mathscr{D})\\&\geq\left(1-\exp(-m)-\left(n^b\sqrt{\pi\log n}\right)^{-1}\right)(1-\exp(-c_2m)-2\exp(-c_4\theta^2m/\tau^4)).\end{aligned}$$

Then, one has by (71) that

$$\mathbb{P}(\mathscr{G}_i)\geq\left(1-\exp(-m)-\left(n^b\sqrt{\pi\log n}\right)^{-1}\right)(1-\exp(-c_2m)-2\exp(-c_4\theta^2m/\tau^4)),$$

whenever m satisfies (67). The proof is complete. □

As an application of Theorem 4 to the special case when $q = 1$ and $a = 3$, the following corollary presents the statistical properties of the ℓ_1 regularization problem with a Gaussian random design under the classical REC. A similar ℓ_2 recovery bound was shown in [46, Theorem 3.1] by using a different analytic technique.

Corollary 4. *Let $\hat{\beta}_{1,\lambda}$ be an optimal solution of* $(\mathrm{RP}_{1,\lambda})$ *with*

$$\lambda = 2\sigma(1+\theta)\sqrt{\frac{2(1+b)\log n}{m}}.$$

Suppose that $\Sigma_{j,j} = 1$ for all $j = 1, \ldots, n$ and $\Sigma^{\frac{1}{2}}$ satisfies the $1\text{-REC}(s,t,3)$*. Then, there exist universal positive constants (c_1, c_2, c_3, c_4) and $\tau \geq 1$ (independent of $m, n, s, t, \theta, b, \Sigma$) such that, if*

$$m > \max\left\{ \frac{c_1(\sqrt{s+t} + 3\sqrt{s})^2}{\phi_1^2(s,t,3,\Sigma^{\frac{1}{2}})} \log n,\ \frac{c_3\tau^4}{\theta^2}\log n \right\},$$

then it holds with probability at least $(1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1})(1 - \exp(-c_2 m) - 2\exp(-c_4\theta^2 m/\tau^4))$ *that*

$$\frac{1}{m}\|X\hat{\beta}_{1,\lambda} - X\beta^*\|_2^2 \leq \frac{1152(1+b)(1+\theta)^2}{\phi_1^2(s,t,3,\Sigma^{\frac{1}{2}})} \frac{s\log n}{m}\sigma^2,$$

$$\frac{1}{2m}\|X\hat{\beta}_{1,\lambda} - X\beta^*\|_2^2 + \lambda\|(\hat{\beta}_{1,\lambda})_{J^c}\|_1 \leq \frac{576(1+b)(1+\theta)^2}{\phi_1^2(s,t,3,\Sigma^{\frac{1}{2}})} \frac{s\log n}{m}\sigma^2,$$

$$\|\hat{\beta}_{1,\lambda} - \beta^*\|_2^2 \leq \frac{4608(1+b)(1+\theta)^2(1+9\frac{s}{t})}{\phi_1^4(s,t,3,\Sigma^{\frac{1}{2}})} \frac{s\log n}{m}\sigma^2.$$

5 Numerical Experiments

The purpose of this section is to carry out the numerical experiments to illustrate the stability of the ℓ_q optimization methods, verify the established theory of the ℓ_2 recovery bounds in the preceding sections and compare the numerical performance of the ℓ_q regularization methods with another two widely used nonconvex regularization methods, namely the SCAD and MCP. In particular, we are concerned with the cases when $q = 0$, $q = 1/2$, $2/3$ and 1. To solve the ℓ_q minimization problems, we will apply the iterative reweighted algorithm [11, 13] . To solve the ℓ_q regularization problems, we will apply the iterative hard thresholding algorithm [5] for $q = 0$, the proximal gradient algorithm [24] for $q = 1/2$ and $2/3$, and FISTA [3] for $q = 1$, respectively. The proximal gradient algorithm proposed in [27] will be used to solve the SCAD and MCP. All numerical experiments are performed in MATLAB R2014b and executed on a personal desktop (Intel Core i7-4790, 3.60 GHz, 8.00 GB of RAM).

The simulated data are generated via a standard process; see, e.g., [1, 24]. Specifically, we randomly generate an i.i.d. Gaussian ensemble $X \in \mathbb{R}^{m\times n}$ and a sparse vector $\beta^* \in \mathbb{R}^n$ with the sparsity being equal to s. The observation y is then generated by the MATLAB script

$$y = X * \beta^* + sigma * randn(m, 1),$$

where $sigma$ is the noise level, that is the standard deviation of Gaussian noise. In the numerical experiments, the dimension of variables and the noise level are set as $n = 1024$ and $sigma = 0.01$, respectively.

For each sparsity level (e.g., $s/n = 5\%, 10\%, 15\%, 20\%$), we randomly generate the data X, β^*, y for 100 times and run the algorithms mentioned above to solve the ℓ_q optimization problems for $q = 0$, $1/2$, $2/3$ and 1 as well as the SCAD and MCP. The parameter ϵ in the ℓ_q minimization methods $(\mathrm{CP})_{q,\epsilon}$ is set as $\epsilon = sigma * \sqrt{m + 2\sqrt{2m}}$ in order to guarantee that $\|e\|_2^2$ is no more than ϵ^2 with overwhelming probability [9, 11]. The parameter λ in the ℓ_q regularization methods $(\mathrm{RP}_{q,\lambda})$ is chosen by 10-fold cross validation. To simplify the notations, the solution of different problems will all be denoted as $\hat{\beta}$. In order to reveal the dependence of ℓ_2 recovery bounds on sample size and inspired by the established theorems in the preceding sections (e.g., (67)), we report the numerical results for a range of sample sizes of the form $m = \Omega(s \log n)$.

The first experiment is conduct to show the performance on parameter estimation of the ℓ_q minimization methods. Figure 1 plots the estimated error $\|\hat{\beta} - \beta^*\|_2$ along with different sample size m. From Figure 1, we can see that the estimated error of each minimization method decreases consistently as the sample size increases. In addition, we find that the lower the q, the better the corresponding minimization method to achieve a more accurate solution.

The second experiment is carried out to show the performance on parameter estimation of the ℓ_q regularization methods and compare the performance with the SCAD and MCP. The corresponding result is displayed in Figure 2, which plots the estimated error $\|\hat{\beta} - \beta^*\|_2$ along with the sample size m. As shown by Figure 2, the estimated error of each regularization method

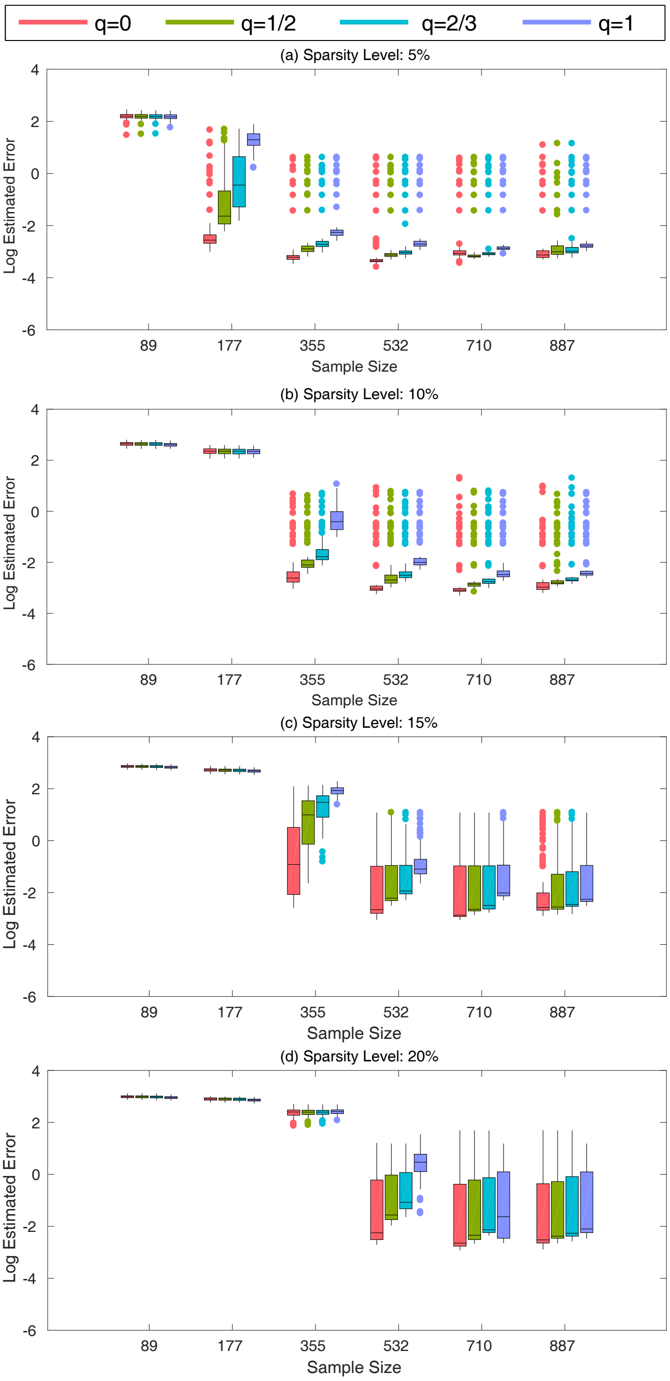

Figure 1: Boxplots of estimated error versus sample size for different constrained optimization methods.

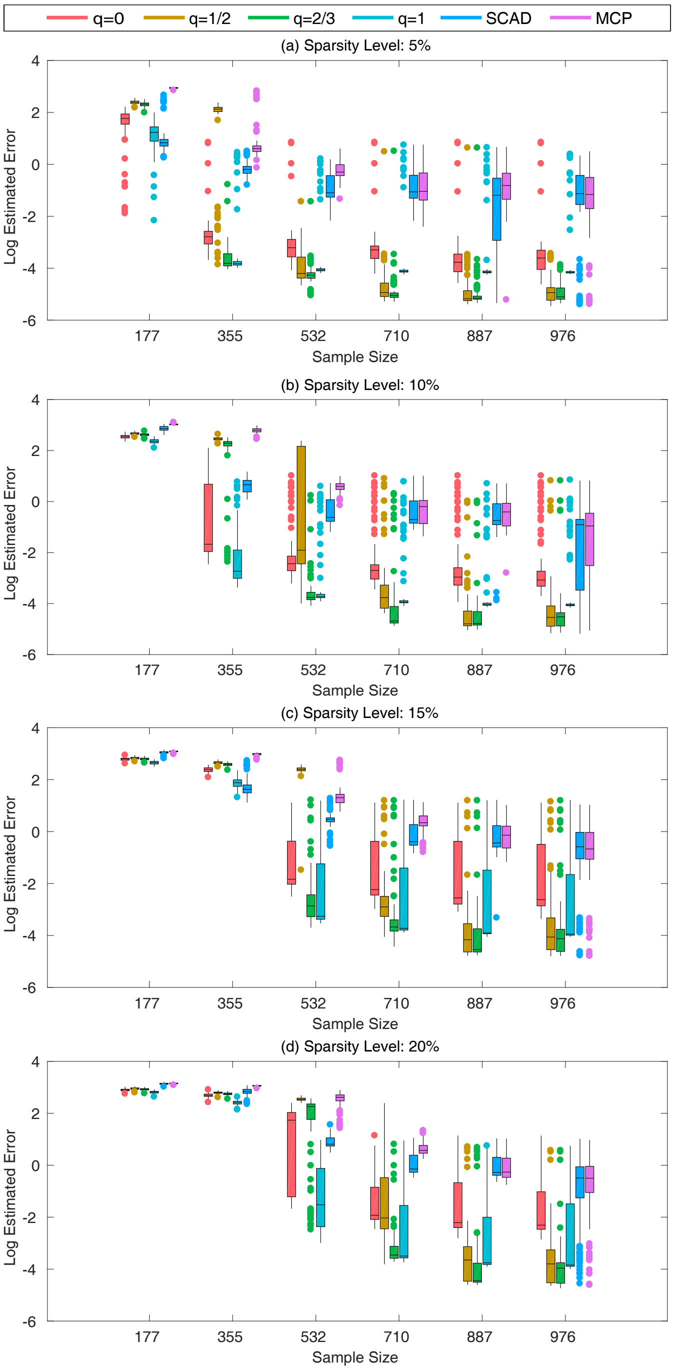

Figure 2: Boxplots of estimated error versus sample size for different regularization methods.

decreases consistently as the sample size increases, and that the lower-order regularization method (e.g., when $q = 1/2, 2/3$) outperforms the ℓ_0/ℓ_1 regularization method in the sense that its estimated error decreases faster when the sample size increases and achieves a more accurate solution than the ℓ_0/ℓ_1 regularization method. This is due to the fact that the q-REC is satisfied when the sample size is larger than a certain level (see Lemma 11(i)) and the lower-order regularization method requires a weaker q-REC to guarantee its nice statistical property (see Theorems 2 and 4). This result is consistent with the existing empirical studies on the ℓ_q regularization methods as in [40, 24]. In addition, it is obvious that the lower-order regularization methods perform much better than the SCAD and MCP to achieve an accurate solution no matter whether the sparsity level is high or low.

The third experiment is implemented to study the performance on variable selection of the ℓ_q regularization methods as well as the SCAD and MCP. We use two criteria to characterize the capability of variable selection, namely the sensitivity $= \frac{\text{true positive}}{\text{true positive+false negative}}$ and the specificity $= \frac{\text{true negative}}{\text{true negative+false positive}}$, which respectively measures the proportion of positives and negatives that are correctly identified. The larger values of both sensitivity and specificity mean the higher capability of variable selection. The results are illustrated by averaging over the 100 random trials. Tables 1 and 2 respectively chart the sensitivity and specificity of these methods at a sparsity level 10% corresponding to Figure 2(b). It is illustrated that the sensitivity and specificity of all these methods increase as the sample size grows, except for the specificity of Lasso, which is resulted from the fact that there are many small nonzero coefficients estimated by Lasso. We also note that the lower-order regularization method (e.g., when $q = 1/2, 2/3$) outperforms the other regularization methods in the sense that it can almost completely select the true model from a small number of samples.

Table 1: Sensitivity of different regularization methods.

Method	Sample size					
	177	355	532	710	887	976
q=0	0.3029	0.8931	0.9824	0.9873	0.9902	0.9912
q=1/2	0.2902	0.5108	0.9412	0.9873	0.9892	0.9941
q=2/3	0.3108	0.9333	0.9922	0.9931	0.9941	0.9971
q=1	0.5088	0.9980	1.0000	1.0000	1.0000	1.0000
SCAD	0.2882	0.8471	0.9157	0.9363	0.9324	0.9422
MCP	0.1373	0.4539	0.8461	0.9088	0.9353	0.9382

Table 2: Specificity of different regularization methods.

Method	Sample size					
	177	355	532	710	887	976
q=0	0.9229	0.9882	0.9980	0.9986	0.9989	0.9990
q=1/2	0.8810	0.8119	1.0000	1.0000	1.0000	1.0000
q=2/3	0.8782	0.9999	1.0000	1.0000	1.0000	1.0000
q=1	0.8088	0.7454	0.7680	0.7473	0.7357	0.6120
SCAD	0.9653	0.9900	0.9906	0.9908	0.9919	0.9925
MCP	0.9466	0.9757	0.9909	0.9919	0.9925	0.9925

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40(5):2452–2482, 2012.

[2] S. Aronoff. *Remote Sensing for GIS Managers*. Environmental Systems Research, Redlands, 2004.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[5] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

[6] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 64(3):330–2, 2007.

[7] T. T. Cai, G. W. Xu, and J. Zhang. On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 55(7):3388–3397, 2009.

[8] E. J. Candès, J. K. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[9] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):410–412, 2006.

[10] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[11] E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

[12] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

[13] R. Chartrand and W. T. Yin. Iteratively reweighted algorithms for compressive sensing. In *IEEE International Conference on Acoustics*, pages 3869–3872, 2008.

[14] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[15] I. Daubechies, R. Devore, and M. Fornasier. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[16] Z. L. Dong, X. Q. Yang, and Y. H. Dai. A unified recovery bound estimation for noise-aware ℓ_q optimization model in compressed sensing. *arXiv preprint arXiv:1609.01531*, 2016.

[17] D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[18] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

[19] D. L Donoho and X. M. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[20] J. Q. Fan and R. Z. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[21] M. A T Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007.

[22] S. Foucart and M.-J. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q < 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

[23] J. Herman, R. Kucera, and J. Simsa. *Equations and Inequalities: Elementary Problems and Theorems in Algebra and Number Theory.* Springer, Berlin, 2000.

[24] Y. H. Hu, C. Li, K. W. Meng, J. Qin, and X. Q. Yang. Group sparse optimization via $\ell_{p,q}$ regularization. *Journal of Machine Learning Research*, 18(30):1–52, 2017.

[25] H. C. Liu, T. Yao, R. Z. Li, and Y. Y. Ye. Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory for local solutions. *Mathematical Programming*, 166(1-2):207–240, 2017.

[26] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Annals of Statistics*, 40(3):1637–1664, 2012.

[27] P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(1):559–616, 2015.

[28] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[29] S. Negahban, P. Ravikumar, M. J Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[30] J. Qin, Y. H. Hu, F. Xu, H. K. Yalamanchili, and J. W. Wang. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*, 67(3):294–303, 2014.

[31] C. R. Rao and M. Statistiker. *Linear Statistical Inference and Its Applications.* Wiley New York, New York, 1973.

[32] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(2):2241–2259, 2010.

[33] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

[34] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[35] S. Ross. *A First Course in Probability.* Pearson, London, 2009.

[36] C. B. Song and S. T. Xia. Sparse signal recovery by ℓ_q minimization under restricted isometry property. *IEEE Signal Processing Letters*, 21(9):1154–1158, 2014.

[37] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[38] S. A. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[39] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:2009, 2009.

[40] Z. B. Xu, X. Y. Chang, F. M. Xu, and H. Zhang. $\mathrm{L}_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1013–1027, 2012.

[41] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010.

[42] C. H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[43] C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

[44] T. Zhang. Some sharp performance bounds for least squares regression with ℓ_1 regularization. *Annals of Statistics*, 37:2109–2144, 2009.

[45] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010.

[46] S. H. Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.