Estimation Error of the Constrained Lasso

Nissim Zerbib[1], Yen-Huan Li[2], Ya-Ping Hsieh[2], and Volkan Cevher[2]

***Abstract*— This paper presents a non-asymptotic upper bound for the estimation error of the constrained lasso, under the high-dimensional ($n \ll p$) setting. In contrast to existing results, the error bound in this paper is sharp, is valid when the parameter to be estimated is not exactly sparse (e.g., when it is weakly sparse), and shows explicitly the effect of over-estimating the ℓ_1-norm of the parameter to be estimated on the estimation performance. The results of this paper show that the constrained lasso is minimax optimal for estimating a parameter with bounded ℓ_1-norm, and also for estimating a weakly sparse parameter if its ℓ_1-norm is accessible.**

I. Introduction

A. Problem Formulation

Consider the linear regression problem. The goal is to estimate an unknown parameter $\beta^* \in \mathbb{R}^p$, given the *design matrix* $X \in \mathbb{R}^{n \times p}$, and the *sample*

$$y = X\beta^* + \sigma w \in \mathbb{R}^n,$$

for some $\sigma > 0$, where σw denotes the additive noise. We will mainly focus on the case when the parameter dimension p may scale with the sample size n and $n \ll p$, the so-called *high-dimensional setting*.

If the parameter β^* is known to be sparse, a widely-used estimator is the constrained lasso (which we will simply call as the lasso in this paper) [21], defined as

$$\hat{\beta}_n \in \arg\min_{\beta} \{f_n(\beta) : \beta \in c\mathcal{B}_1\}, \tag{1}$$

for some $c > 0$, where f_n is the normalized squared error function

$$f_n(\beta) := \frac{1}{2n}\|y - X\beta\|_2^2,$$

and $\mathcal{B}_1$ denotes the unit ℓ_1-norm ball in $\mathbb{R}^p$.

This paper studies the estimation error of the lasso in the linear regression model, under the high-dimensional setting.

*This work was supported in part by ERC Future Proof, SNF 200021-146750 and SNF CRSII2-147633.

[1]Nissim Zerbib was with the Laboratory for Information and Inference Systems (LIONS), École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland when this work was done. He is currently with the Computer Science Department, École Normale Supérieure, France. nissim.zerbib@ens.fr

[2]Yen-Huan Li, Ya-Ping Hsieh, and Volkan Cevher are with the Laboratory for Information and Inference Systems (LIONS), École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland. {yen-huan.li, ya-ping.hsieh, volkan.cevher}@epfl.ch

B. Related Work

If $c = \|\beta^*\|_1$, and the noise w has independent and identically distributed (i.i.d.) standard normal entries, the lasso is known to satisfy

$$\|\hat{\beta}_n - \beta^*\|_2 \leq L\sigma\sqrt{\frac{s \log p}{n}}, \tag{2}$$

with high probability for some constant $L > 0$, where s is the number of non-zero entries in β^* [5]. The bound (2) shows the lasso automatically adapts to β^*—the sparser β^* is, the smaller the estimation error bound.

This error bound (2), however, is not true in general when $c \neq \|\beta^*\|_1$. While (2) provides an $O((\sigma^2 n^{-1} \log p)^{\frac{1}{2}})$ error decaying rate, the minimax result in [18] shows that, with respect to the worst case of where β^* lies in $c\mathcal{B}_1$, *no estimator* can achieve an error decaying rate better than $O((\sigma^2 n^{-1} \log p)^{\frac{1}{4}})$. This gap is due to the possibility that β^* may lie strictly in $c\mathcal{B}_1$ or, in other words, $c > \|\beta^*\|_1$.

Therefore, a more general estimation error bound for the lasso is needed. Especially, a satisfactory estimation error bound for the lasso should be 1) *sharp* enough to recover (2) that varies with the sparsity of β^*, and 2) able to characterize the effect of the quantity $c - \|\beta^*\|_1$ on the estimation error.

Existing results, unfortunately, cannot provide such a satisfactory error bound. The proof in [5] for (2) fails when c is strictly larger than $\|\beta^*\|_1$. While the results in [16], [24] are valid as long as $c \geq \|\beta^*\|_1$, the derived bounds are independent of β^*, and hence not sharp enough to recover (2). The small-ball approach yields an estimation error bound that depends on β^* [10, Theorem 4.6], but the dependence is implicit, and even whether it can recover (2) is unclear. The results in [7], [15] recover (2) when $c = \|\beta^*\|_1$; the dependence on $c - \|\beta^*\|_1$, however, is also vague.

The paper [18] assumed β^* lies in an ℓ_q-norm ball $\mathcal{B}_q$, $q \in [0, 1]$, and derived an estimation error bound for a *lasso-like* estimator, for which the ℓ_1-norm constraint in (1) is replaced by the corresponding ℓ_q-norm constraint. In contrast to [18], this paper will also consider the same assumption on β^*, but analyze the estimation performance of the lasso defined by (1) where an ℓ_1 norm constraint is used (cf. Corollary 2).

The authors are not aware of any existing work that discusses the estimation error of the lasso when $c < \|\beta^*\|_1$, though the analysis in [7] can be easily extended to this case, and yield an estimation error bound that is implicitly dependent on $\|\beta^*\|_1$. Note that in this case, the lasso cannot be consistent, i.e., the estimation error is always bounded away from zero no matter how large the sample size n is, because β^* is not a feasible solution of the optimization problem (1).

We note that while there are many well-studied estimators closely related to the constrained lasso, such as the penalized lasso, Dantzig selector, square-root lasso, and basis pursuit-type estimators [1], [2], [4], [12], [22], the analysis techniques in the cited works cannot be directly applied to study the constrained lasso when $c > \|\beta^*\|_1$. See Section III-B for a detailed discussion.

C. Contributions

The main result of this paper, Theorem IV.1, provides a *non-asymptotic* estimation error bound that is valid for any $c \geq \|\beta^*\|_1$, and for the case when β^* is not exactly sparse. It is sharp as it recovers (2) when $c = \|\beta^*\|_1$ (cf. Corollary 1). For the general case, it shows the following (cf. Corollary 2).

- For estimating any $\beta^* \in c\mathcal{B}_1$, the lasso is minimax optimal as long as $c \geq \|\beta^*\|_1$. The worst case (with respect to where β^* lies in $c\mathcal{B}_1$) error decaying rate is
$$\|\hat{\beta}_n - \beta^*\|_2 = O\left(\left(\frac{\sigma^2 \log p}{n}\right)^{\frac{1}{4}}\right).$$
- For estimating any *weakly sparse* $\beta^* \in c\mathcal{B}$ that is to mean it has bounded ℓ_q-norm for some $q \in (0, 1]$, the lasso is minimax optimal if $c = \|\beta^*\|_1$. The worst case error decaying rate is
$$\|\hat{\beta}_n - \beta^*\|_2 = O\left(\left(\frac{\sigma^2 \log p}{n}\right)^{\frac{1}{2}-\frac{1}{4}q}\right).$$

Formal statements can be found in Section IV.

The results in this paper are non-asymptotic, i.e., the error bounds (and the corresponding probability bounds) are valid for all finite values of the sample size n, parameter dimension p, sparsity level s, and other parameters that will be specified in Section IV.

II. Notation and Basic Definitions

Fix a vector $v \in \mathbb{R}^p$ for some $p \in \mathbb{N}$. Let $\mathcal{S} \subseteq \{1, \ldots, p\}$. The notation $v_{\mathcal{S}}$ denotes the sub-vector of v indexed by $\mathcal{S}$, and to lighten notation, v_i denotes $v_{\{i\}}$ for any $i \leq p$. Similarly, fix a matrix $X \in \mathbb{R}^{n\times p}$; $X_{i,j}$ denotes the (i, j)-th entry of X. Let $u \in \mathbb{R}^p$. The inner product $\langle u, v\rangle$ denotes $\sum_i u_i v_i$.

Fix $\mathcal{K} \subseteq \mathbb{R}^p$ and $\lambda \in \mathbb{R}$. The notations $\mathcal{K} - v$ and $\lambda\mathcal{K}$ denote the sets $\{u - v : u \in \mathcal{K}\}$ and $\{\lambda u : u \in \mathcal{K}\}$, respectively. The notation $\overline{\mathcal{K}}$ denotes the conic hull of $\mathcal{K}$, i.e.,
$$\overline{\mathcal{K}} := \{\rho v : v \in \mathcal{K}, \rho \geq 0\}.$$
The notation $|\mathcal{K}|$ denotes the cardinality of $\mathcal{K}$.

The ℓ_q-norm of v, denoted by $\|v\|_q$, is defined by $\|v\|_q^q := \sum_i |v_i|^q$, for any $q \in [0, \infty)$ (although rigorously speaking, $\|\cdot\|_q$ is a norm only when $q \geq 1$). The ℓ_0-norm is defined as $\|v\|_0 := |\{i : v_i \neq 0\}|$, and the ℓ_∞-norm is defined as $\|v\|_\infty := \max\{|v_i| : 1 \leq i \leq p\}$. The unit ℓ_q-norm ball is denoted by $\mathcal{B}_q$.

Some relevant notions about random variables (r.v.'s) and random vectors are provided below for completeness.

Definition II.1. A r.v. ξ is *subgaussian*, if there exists a constant $K > 0$ such that $(\mathbb{E}\,|\xi|^p)^{1/p} \leq K\sqrt{p}$ for all $p \geq 1$. The *subgaussian norm* of a subgaussian r.v. ξ is defined as the smallest K, i.e.,
$$\|\xi\|_{\psi_2} := \sup\{p^{-1/2}(\mathbb{E}\,|\xi|^p)^{1/p} : p \geq 1\}.$$

Definition II.2. A random vector $\eta \in \mathbb{R}^p$ is *isotropic*, if for any $v \in \mathbb{R}^p$,
$$\mathbb{E}\,\langle \eta, v\rangle^2 = \|v\|_2^2.$$

Definition II.3. A random vector $\eta \in \mathbb{R}^p$ is *subgaussian*, if the r.v. $\langle \eta, v\rangle$ is subgaussian for all $v \in \mathbb{R}^p$. The *subgaussian norm* of a subgaussian random vector η is defined as
$$\|\eta\|_{\psi_2} := \sup\{\|\langle \eta, v\rangle\|_{\psi_2} : v \in \mathbb{R}^p, \|v\|_2 = 1\}.$$

Remark. For example, both the standard normal r.v. and the Rademacher r.v. (random sign) are subgaussian, and a vector of either i.i.d. standard normal or i.i.d. Rademacher r.v.'s is a subgaussian random vector.

The Gaussian width is useful when studying a collection of subgaussian r.v.'s indexed by a subset in the metric space $(\mathbb{R}^p, \|\cdot\|_2)$ [20, Theorem 2.4.1].

Definition II.4 (Gaussian width)**.** The *Gaussian width* of a set $\mathcal{K} \subseteq \mathbb{R}^p$ is given by
$$w(\mathcal{K}) := \mathbb{E}\sup\{\langle g, v\rangle : v \in \mathcal{K}\},$$
where g is a vector of i.i.d. standard normal r.v.'s.

By Proposition III.2 below, the Gaussian width of a set of the form $\mathcal{C} \cap \mathcal{B}_2$, where $\mathcal{C} \subseteq \mathbb{R}^p$ is a closed convex cone, characterizes the sample size required for the lasso to have a small estimation error. We always have $w(\mathcal{C} \cap \mathcal{B}_2) \leq \sqrt{p}$. By Proposition III.2 and Theorem IV.1, this implies the possibility of doing estimation when $n < p$.

Proposition II.1. *We have the following:*

1) *If $\mathcal{K}_1 \subseteq \mathcal{K}_2$, then $w(\mathcal{K}_1) \leq w(\mathcal{K}_2)$.*
2) *If $\mathcal{K} = \mathbb{R}^p$, then $w(\mathcal{K} \cap \mathcal{B}_2) = \sqrt{p}$.*

Proof. The first assertion is obvious by definition. The second assertion is because
$$w(\mathbb{R}^p \cap \mathcal{B}_2) = w(\mathcal{B}_2) = (1/\sqrt{p})\mathbb{E}\,\|g\|_2^2 = \sqrt{p},$$
where g is a vector of i.i.d. standard normal r.v.'s. □

III. Relaxed Restricted Strong Convexity Condition

The key notion for deriving the results in this paper is the *relaxed restricted strong convexity (RSC) condition* introduced in the authors' unpublished work [7]. This section provides a brief discussion on the relaxed RSC condition, specialized for the lasso.

A. Definition of the Relaxed RSC Condition

Conventionally, linear regression is solved by the least-squares (LS) estimator, which works as long as the Hessian matrix $H_n := \nabla^2 f_n(\beta^*) \equiv n^{-1}X^TX$ is non-singular. Under the high-dimensional setting where $n < p$, however, the Hessian matrix H_n is always singular, and the LS approach fails, as illustrated by [3, Fig. 1].

The idea of the relaxed RSC condition is to require, *only in some directions*, that the Hessian matrix H_n behaves like a non-singular matrix.

Definition III.1 (Feasible Set)**.** The *feasible set* is defined as

$$\mathcal{F} := c\mathcal{B}_1 - \beta^* = \{\beta - \beta^* : \beta \in c\mathcal{B}_1\}.$$

That is, the feasible set is the set of all possible error vectors.

Definition III.2 (Relaxed RSC [7])**.** The (μ, t_n)*-relaxed RSC condition* holds for some $\mu > 0$ and $t_n \geq 0$, if and only if for all $v \in \mathcal{F} \setminus t_n\mathcal{B}_2$,

$$\langle \nabla f_n(\beta^* + v) - \nabla f_n(\beta^*), v\rangle \geq \mu\|v\|_2^2.$$

Remark. The parameter t_n in general can scale with the sample size n; therefore the subscript n is added.

Proposition III.1. *The (μ, t_n)-relaxed RSC condition is equivalent to requiring*

$$\min\left\{\frac{v^TH_nv}{\|v\|_2^2} : v \in \mathcal{F} \setminus t_n\mathcal{B}_2\right\} \geq \mu,$$

i.e., it requires the smallest restricted eigenvalue *of H_n with respect to $\mathcal{F} \setminus t_n\mathcal{B}_2$ is bounded below by μ.*

Proof. By direct calculation, we obtain

$$\langle \nabla f_n(\beta^* + v) - \nabla f_n(\beta^*), v\rangle = v^TH_nv.$$

□

The validity of assuming the relaxed RSC condition is verified by the following proposition, which shows as long as the sample size n is sufficiently large (while it can be still less than p), the relaxed RSC condition can hold with high probability.

Proposition III.2. *Suppose that the rows of the design matrix X are i.i.d., isotropic, and subgaussian with subgaussian norm $\alpha > 0$. There exist constants $c_1, c_2 > 0$ such that for any $\delta \in (0, 1)$, if*

$$\sqrt{n} \geq c_1^2\alpha^2 w(\overline{\mathcal{F} \setminus t\mathcal{B}_2} \cap \mathcal{B}_2), \tag{3}$$

for some $t \geq 0$, the $(1 - \delta, t)$-relaxed RSC condition holds with probability at least $1 - \exp(-c_2\delta^2 n/\alpha^4)$.

Proof. Assume that (3) is satisfied. By [11, Theorem 2.3], with probability at least $1 - \exp(-c_2\delta^2 n)$, we have

$$\frac{\|Xv\|_2^2}{n} = \frac{v^TH_nv}{n} \geq (1 - \delta)\|v\|_2^2 \tag{4}$$

for any $v \in \mathcal{F} \setminus t\mathcal{B}_2$. The proposition follows by Proposition III.1. □

B. Discussions

One interesting special case of Proposition III.2 is when β^* has only $s < p$ non-zero entries and $c = \|\beta^*\|_1$. In this case, we can simply choose $t_n \equiv 0$; then $\overline{\mathcal{F} \setminus t_n\mathcal{B}_2}$ reduces to $\overline{\mathcal{F}}$, called the *tangent cone* in [4]. By [4, Proposition 3.10], the inequality (3) can be guaranteed, if

$$\sqrt{n} \geq c_1^2\alpha^2\sqrt{2s\log\left(\frac{p}{s}\right) + \frac{5}{4}s}.$$

Notice that the right-hand side can be much smaller than $\sqrt{p}$.

This observation is the main idea behind existing works on high-dimensional sparse parameter estimation in [1], [2], [4], [12], [22], to cite a few. Roughly speaking, the approach in the cited works can be summarized as follows.

1) Identify a convex cone $\mathcal{K}$ (possibly with a controlled small perturbation [12], [17]) in which the error vector $\tilde{\beta}_n - \beta^*$ lies, where $\tilde{\beta}_n$ denotes the estimator under consideration.
2) Derive a lower bound on the sample size n, such that the RSC (relaxed RSC with $t_n \equiv 0$, not necessary with respect to the ℓ_2-norm [22]) with respect to $\mathcal{K}$ holds with high probability.
3) Given that the RSC condition holds, the Hessian $H_n = n^{-1}X^TX$ behaves like a non-singular matrix with respect to the error vector, and classical approaches for analyzing the estimation error for the LS estimator applies.

While this existing approach is valid for analyzing the penalized lasso, Dantzig selector, square-root lasso, and basis pursuit-type estimators as shown in [1], [2], [4], [12], [22], it is not applicable to the constrained lasso. When $c > \|\beta^*\|_1$, the conic hull of all possible error vectors of the constrained lasso, $\overline{c\mathcal{B}_1 - \beta^*}$, is the whole space $\mathbb{R}^p$, and hence requiring the relaxed RSC condition with t_n=0 is equivalent to requiring the non-singularity of the Hessian H_n, which cannot hold when $n \ll p$.

The next section shows that the relaxed RSC condition with a non-zero t_n suffices for deriving minimax optimal estimation error bounds for the lasso.

IV. Main Result and Its Implications

The main theorem requires the following assumptions to be satisfied.

Assumption 1. *The noise w is a vector of i.i.d. mean-zero subgaussian r.v.'s of unit subgaussian norm.*

Assumption 2. *The design matrix X is normalized, i.e., $\sum_j X_{i,j}^2 \leq n$ for all $i \leq p$.*

Assumption 3. *The (μ, t_n)-relaxed RSC condition holds for some $\mu, t_n > 0$.*

The first assumption on the noise is valid in the standard Gaussian linear regression model, where w is a vector of i.i.d. standard normal r.v.'s, and the persistence framework in [10], where w is a vector of i.i.d. mean-zero bounded r.v.'s. The second assumption on the design matrix is standard as

in, e.g., [2] and [25]; without this assumption, the effect of noise can be arbitrarily small (when the entries of X are large compared to σ). Recall that we had discussed the validity of the third assumption in Section III.

Theorem IV.1. *If Assumptions 1–3 are satisfied, then there exists a constant $c_3 > 0$ such that, for any $\tau > 0$ and $\mathcal{S} \subseteq \{1, \ldots, p\}$,*

$$\|\hat{\beta}_n - \beta^*\|_2 \leq \max\left\{t_n, \frac{c_3\sqrt{1+\tau}}{\mu} \cdot \sigma\sqrt{\frac{\log p}{n}}\gamma(t_n; \beta^*, \mathcal{S})\right\}$$

with probability at least $1 - ep^{-\tau}$, where

$$\gamma(t_n; \beta^*, \mathcal{S}) := 2\sqrt{|\mathcal{S}|} + \frac{2\|\beta^*_{\mathcal{S}^c}\|_1 + (c - \|\beta^*\|_1)}{t_n}. \quad (5)$$

Proof. See Section V-A. □

Theorem IV.1 immediately recovers the well-known result (2) up to a constant scaling.

Corollary 1. *Suppose that β^* has s non-zero entries, and $c = \|\beta^*\|_1$ in (1). Then if Assumptions 1–3 are satisfied, there exists a constant $c_3 > 0$ such that, for any $\tau > 0$, we have*

$$\|\hat{\beta}_n - \beta^*\|_2 \leq \frac{2c_3\sqrt{1+\tau}}{\mu} \cdot \sigma\sqrt{\frac{s\log p}{n}},$$

with probability at least $1 - ep^{-\tau}$.

Proof. Recall that in this case (cf. Section III), the relaxed RSC can hold with $t_n \equiv 0$, as discussed in Section III. Choosing $t_n \equiv 0$ and $\mathcal{S}$ as the support set of β^* in Theorem IV.1 completes the proof. □

In general, β^* may not be exactly sparse, and in practice, c can hardly be chosen as exactly $\|\beta^*\|_1$.

Definition IV.1 (Weak sparsity [12])**.** A vector $v \in \mathbb{R}^p$ is *q-weakly sparse* for some $q \in [0, 1]$, if and only if there exists some $C_q > 0$ such that $\|v\|_q^q := \sum_i |v_i|^q \leq C_q$.

Remark. A 0-weakly sparse parameter is exactly sparse.

Corollary 2. *Assume that β^* is q-weakly sparse for some $q \in [0, 1]$, $\log p \ll n$, and Assumptions 1–3 are satisfied with*

$$t_n = \begin{cases} \Theta\left(\sqrt{C_q}\left(\frac{(1+\tau)\sigma^2\log p}{\mu^2 n}\right)^{\frac{1}{2}-\frac{1}{4}q}\right) & \text{if } c = \|\beta^*\|_1 \\ \Theta\left(\sqrt{\delta + C_q}\left(\frac{(1+\tau)\sigma^2\log p}{\mu^2 n}\right)^{\frac{1}{4}}\right) & \text{if } c > \|\beta^*\|_1 \end{cases}, \quad (6)$$

where $\delta := c - \|\beta^*\|_1$ and $C_q := \|\beta^*\|_q^q$. Then we have, with probability at least $1 - ep^{-\tau}$,

$$\|\hat{\beta}_n - \beta^*\|_2 = O(t_n)$$

for any $\tau \in (0, 1)$.

Proof. See Section V-B. □

Remark. If t_n converges too fast to zero with respect to increasing n, the sample complexity bound (3) may not hold, and the validity of Assumption 3 in Corollary 2 would be in question. However, since

$$w(\overline{\mathcal{F} \setminus t_n\mathcal{B}_2} \cap \mathcal{B}_2) = \frac{w(\overline{\mathcal{F} \setminus \mathcal{B}_2} \cap \mathcal{B}_2)}{t_n} = \Theta\left(\frac{1}{t_n}\right),$$

the sample complexity bound (3) can hold as long as $t_n = \Omega(n^{-1/2})$, which is satisfied in Corollary 2.

The minimax error bound in [18, Theorem 3] shows that *no estimator* can achieve a better error decaying rate than

$$O\left(\sqrt{C_q}\left(\frac{\sigma^2\log p}{n}\right)^{\frac{1}{2}-\frac{1}{4}q}\right)$$

with probability larger than $1/2$ in the worst case, for estimating a q-weakly sparse parameter, $q \in (0, 1]$. According to Corollary 2, this implies:

- The lasso with $c \geq \|\beta^*\|_1$ is minimax optimal (up to a constant scaling) for estimating a parameter with bounded ℓ_1-norm.
- The lasso with $c = \|\beta^*\|_1$ is minimax optimal (up to a constant scaling) for estimating a q-weakly sparse parameter, $q \in (0, 1]$.

Note that the error decaying rates in the two assertions are for the worst case. It is possible to have a better error decaying rate in special cases, as shown by Corollary 1.

V. Proofs

A. Proof of Theorem IV.1

Define $\Delta_n := \hat{\beta}_n - \beta^*$ for convenience.

By definition, Δ_n lies in either $t_n\mathcal{B}_2$ or $\mathcal{F} \setminus t_n\mathcal{B}_2$. In the former case, it holds trivially that $\|\Delta_n\|_2 \leq t_n$. We now consider the latter case.

Proposition V.1. *If the (μ, t_n)-relaxed RSC condition holds for some $\mu, t > 0$, and if $\Delta_n \in \mathcal{F} \setminus t\mathcal{B}_2$, then we have*

$$\|\Delta_n\|_2 \leq \frac{1}{\mu}\frac{\|\Delta_n\|_1}{\|\Delta_n\|_2}\frac{\langle -\nabla f_n(\beta^*), \Delta_n\rangle}{\|\Delta_n\|_1}. \quad (7)$$

Proof. By the relaxed RSC condition, we have

$$\langle \nabla f_n(\hat{\beta}_n) - \nabla f_n(\beta^*), \Delta_n\rangle \geq \mu\|\Delta_n\|_2. \quad (8)$$

Since (1) defines a convex optimization problem, we have, by the optimality condition of $\hat{\beta}_n$ [13],

$$\langle -\nabla f_n(\hat{\beta}_n), \Delta_n\rangle \geq 0. \quad (9)$$

Summing up (8) and (9), we obtain

$$\langle -\nabla f_n(\beta^*), \Delta_n\rangle \geq \mu\|\Delta_n\|_2^2,$$

which implies

$$\|\Delta_n\|_2 \leq \frac{1}{\mu}\frac{\|\Delta_n\|_1}{\|\Delta_n\|_2}\frac{\langle -\nabla f_n(\beta^*), \Delta_n\rangle}{\|\Delta_n\|_1}.$$

This completes the proof. □

The rest of this subsection is devoted to deriving an upper bound of the right-hand side of (7), which is independent of Δ_n.

We first derive a bound on $(\|\Delta_n\|_1/\|\Delta_n\|_2)$.

Proposition V.2. *The estimation error satisfies*

$$\|\Delta_n\|_1 \le 2(\|(\Delta_n)_{\mathcal{S}}\|_1 + \|\beta^*_{\mathcal{S}^c}\|_1) + (c - \|\beta^*\|_1),$$

for any $\mathcal{S} \subseteq \{1,\dots,p\}$*, where* $\mathcal{S}^c := \{1,\dots,p\} \setminus \mathcal{S}$.

Proof. By definition, we have $\hat{\beta}_n \in c\mathcal{B}_1$, and hence

$$\begin{aligned} c \ge \|\hat{\beta}_n\|_1 &= \|(\beta^* + \Delta_n)_{\mathcal{S}} + (\beta^* + \Delta_n)_{\mathcal{S}^c}\|_1 \\ &\ge \|\beta^*_{\mathcal{S}} + (\Delta_n)_{\mathcal{S}^c}\|_1 - \|\beta^*_{\mathcal{S}^c} + (\Delta_n)_{\mathcal{S}}\|_1 \\ &= \|\beta^*_{\mathcal{S}}\|_1 + \|(\Delta_n)_{\mathcal{S}^c}\|_1 - \|\beta^*_{\mathcal{S}^c}\|_1 - \|(\Delta_n)_{\mathcal{S}}\|_1 \\ &= \|\beta^*\|_1 - 2\|\beta^*_{\mathcal{S}^c}\|_1 + \|\Delta_n\|_1 - 2\|(\Delta_n)_{\mathcal{S}}\|_1, \end{aligned}$$

which proves the proposition. □

By Proposition V.2, we obtain

$$\begin{aligned} \frac{\|\Delta_n\|_1}{\|\Delta_n\|_2} &\le 2\frac{\|(\Delta_n)_{\mathcal{S}}\|_1}{\|\Delta_n\|_2} + \frac{2\|\beta^*_{\mathcal{S}^c}\|_1 + (c - \|\beta^*\|_1)}{\|\Delta_n\|_2} \\ &\le 2\frac{\|(\Delta_n)_{\mathcal{S}}\|_1}{\|(\Delta_n)_{\mathcal{S}}\|_2} + \frac{2\|\beta^*_{\mathcal{S}^c}\|_1 + (c - \|\beta^*\|_1)}{t_n} \\ &\le 2\sqrt{|\mathcal{S}|} + \frac{2\|\beta^*_{\mathcal{S}^c}\|_1 + (c - \|\beta^*\|_1)}{t_n}, \end{aligned} \tag{10}$$

if $\Delta_n \in \mathcal{F} \setminus t_n\mathcal{B}_2$.

Now we bound the term $\langle -\nabla f_n(\beta^*), \Delta_n\rangle / \|\Delta_n\|_1$.

Proposition V.3. *If the design matrix* X *is normalized, i.e.,* $\sum_j X^2_{i,j} \le n$ *for all* $i \le p$*, there exists a universal constant* $c_3 > 0$ *such that for any* $\tau > 0$*, we have*

$$\frac{\langle -\nabla f_n(\beta^*), \Delta_n\rangle}{\|\Delta_n\|_1} \le c_3\sigma\sqrt{\frac{(1+\tau)\log p}{n}},$$

with probability at least $1 - ep^{-\tau}$.

Proof. We note that

$$\begin{aligned} \frac{\langle -\nabla f_n(\beta^*), \Delta_n\rangle}{\|\Delta_n\|_1} &\le \sup\{\langle -\nabla f_n(\beta^*), v\rangle : \|v\|_1 = 1\} \\ &= \| - \nabla f_n(\beta^*)\|_\infty. \end{aligned}$$

By direct calculation, we obtain

$$(\nabla f_n(\beta^*))_i = \frac{1}{n}\sum_{j=1}^n X_{i,j} w_j$$

for all $i \le p$, and hence, by a Hoeffding-type inequality [23, Proposition 5.10], there exists a universal constant $L > 0$ such that for any $\varepsilon > 0$,

$$\mathbb{P}\{|(\nabla f_n(\beta^*))_i| \ge \varepsilon\} \le e \cdot \exp\left(-\frac{L\varepsilon^2 n}{\sigma^2}\right).$$

By the union bound, this implies

$$\begin{aligned} \mathbb{P}\{\|\nabla f_n(\beta^*)\|_\infty \ge \varepsilon\} &\le \sum_{i=1}^p \mathbb{P}\{|(\nabla f_n(\beta^*))_i| \ge \varepsilon\} \\ &\le e \cdot \exp\left(-\frac{L\varepsilon^2 n}{\sigma^2} + \log p\right). \end{aligned}$$

Choosing

$$\varepsilon = \sigma\sqrt{\frac{(1+\tau)\log p}{Ln}}$$

completes the proof. □

Theorem IV.1 follows by combining (10) and Proposition V.3.

B. Proof of Corollary 2

Define $\mathcal{S}_n := \{i : |\beta^*_i| \ge \rho_n\}$ for some $\rho_n > 0$. Then we have $|\mathcal{S}_n| \le C_q\rho_n^{-q}$, as

$$C_q \ge \sum_{i\in\mathcal{S}_n} |\beta^*_i|^q \ge |\mathcal{S}_n|\rho_n^q.$$

Moreover, we have

$$\|\beta^*_{\mathcal{S}_n^c}\|_1 = \sum_{i\in\mathcal{S}_n^c} |\beta^*_i|^q |\beta^*_i|^{1-q} \le \sum_{i\in\mathcal{S}_n^c} |\beta^*_i|^q \rho_n^{1-q} \le C_q\rho_n^{1-q}.$$

Applying Theorem IV.1 with $\mathcal{S} = \mathcal{S}_n$, we obtain

$$\begin{aligned} \|\hat{\beta}_n - \beta^*\|_2 &\le \max\left\{t, \frac{c_3\sqrt{1+\tau}\sigma}{\mu}\sqrt{\frac{\log p}{n}}\gamma_n\right\} \\ &\le t_n + \frac{c_3\sqrt{1+\tau}\sigma}{\mu}\sqrt{\frac{\log p}{n}}\gamma_n, \end{aligned} \tag{11}$$

with probability at least $1 - ep^{-\tau}$, where

$$\gamma_n := 2\sqrt{C_q\rho_n^{-q}} + \frac{2C_q\rho_n^{1-q} + (c - \|\beta^*\|_1)}{t_n}.$$

The corollary follows by optimizing over t_n and ρ_n by the inequality for arithmetic and geometric means on (11). Specifically, the best possible error decaying rate can be achieved when

$$\rho_n = \Theta\left(\left(\frac{(1+\tau)\sigma^2\log p}{\mu^2 n}\right)^{\frac{1}{2}}\right),$$

and t_n is chosen as in (6).

VI. Discussions

This paper focuses on the case where the design matrix X has subgaussian rows and the noise w has subgaussian entries. This is simply for convenience of presentation, and the analysis framework can be easily extended to more general cases.

Proposition III.2, which shows the validity of the relaxed RSC condition, can be easily extended for design matrices whose rows are not necessarily subgaussian, with a possibly worse sample complexity bound compared to (3). The interested reader is referred to [6], [14], [19] for the details.

Theorem IV.1 can be easily extended for possibly non-subgaussian noise. One only needs to replace the Hoeffding-type inequality in the proof of Proposition V.3 by Bernstein's inequality [9] or other appropriate concentration inequalities for sums of independent r.v.'s. Note that the obtained estimation error bound may be worse, as shown in [8].

Finally, we remark that by Proposition III.2 and the union bound, Theorem IV.1 also implies an estimation error bound for the *random design* case, where the design matrix X is a random matrix independent of the noise w. Such an error bound can be useful for compressive sensing, where

the design matrix is not given, but can be chosen by the practitioner.

Corollary 3. *Suppose the rows of the design matrix X are i.i.d., isotropic, and subgaussian with subgaussian norm $\alpha > 0$, and X is independent of the noise w. Then there exist constants $c_1, c_2, c_3 > 0$ such that, if* (3) *and Assumptions 2 and 3 are satisfied, for any $\tau > 0$ and $\mathcal{S} \subseteq \{1, \dots, p\}$, we have*

$$\|\hat{\beta}_n - \beta^*\|_2 \leq \max\left\{ t_n, \frac{c_3\sqrt{1+\tau}\sigma}{1-\delta}\sqrt{\frac{\log p}{n}}\gamma(t_n; \beta^*, \mathcal{S}) \right\}$$

with probability at least $1 - ep^{-\tau} - \exp(-c_2\delta^2 n/\alpha^4)$ (with respect to the design matrix X and the noise w), where $\gamma(t_n; \beta^*, \mathcal{S})$ is defined as in (5)*.*

Corollary 2 can be extended for the random design case in the same manner.

REFERENCES

[1] A. Belloni, V. Chernozhukov, and L. Wang, "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, vol. 98, no. 4, pp. 791–806, 2011.

[2] P. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Stat.*, vol. 37, no. 4, pp. 1705–1732, 2009.

[3] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[4] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, pp. 805–849, 2012.

[5] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and generalizations*. Boca Raton, FL: CRC Press, 2015.

[6] V. Koltchinskii and S. Mendelson, "Bounding the smallest singular value of a random matrix without concentration," *Int. Math. Res. Not.*, 2015.

[7] Y.-H. Li, Y.-P. Hsieh, N. Zerbib, and V. Cevher, "A geometric view on constrained M-estimators," 2015, arXiv:1506.08163 [math.ST].

[8] Y.-H. Li, J. Scarlett, P. Ravikumar, and V. Cevher, "Sparsistency of ℓ_1-regularized M-estimators," in *Proc. 18th Inf. Conf. Artificial Intelligence and Statistics*, 2015, pp. 644–652.

[9] P. Massart, *Concentration Inequalities and Model Selection*. Berlin: Springer-Verl., 2007.

[10] S. Mendelson, "Learning without concentration," *J. ACM*, vol. 62, no. 3, 2015.

[11] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Reconstruction and subgaussian operators in asymptotic geometric analysis," *Geom. Funct. Anal.*, vol. 17, pp. 1248–1282, 2007.

[12] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

[13] Y. Nesterov, *Introductory Lectures on Convex Optimization*. Boston, MA: Kluwer, 2004.

[14] S. Oymak and J. A. Tropp, "Universality laws for randomized dimension reduction, with applications," 2015, arXiv:1511.09433v1 [math.PR].

[15] Y. Plan and R. Vershynin, "The generalized Lasso with non-linear observations," 2015, arXiv:1502.0407v1 [cs.IT].

[16] Y. Plan, R. Vershynin, and E. Yudovina, "High-dimensional estimation with geometric constraints," 2014, arXiv:1404.3749v1 [math.PR].

[17] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated Gaussian designs," *J. Mach. Learn. Res.*, vol. 11, pp. 2241–2259, 2010.

[18] ——, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6976–6994, Oct. 2011.

[19] V. Sivakumar, A. Banerjee, and P. Ravikumar, "Beyond sub-Gaussian measurements: High-dimensional estimation with sub-exponential designs," in *Adv. Neural Information Processing Systems 28*, 2015.

[20] M. Talagrand, *Upper and Lower Bounds for Stochastic Processes*. Berlin: Springer, 2014.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[22] S. van de Geer, "The deterministic Lasso," Seminar für Statistik, Eidgenössische Technische Hochschule, Research Report No. 140, 2007.

[23] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*, Y. Eldar and G. Kutyniok, Eds. Cambridge Univ. Press, 2012, ch. 5, pp. 210–268.

[24] ——, "Estimation in high dimensions: A geometric perspective," in *Sampling Theory, a Renaissance*, G. E. Pfander, Ed. Cham: Birkhäuser, 2015, ch. 1, pp. 3–66.

[25] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.