Forty-Seventh Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 30 - October 2, 2009

Minimax rates of convergence for high-dimensional regression under ℓ_q-ball sparsity

Garvesh Raskutti[1] Martin J. Wainwright[1,2]
Bin Yu[1,2]

[1]Department of Statistics, and
[2]Department of EECS
UC Berkeley, Berkeley, CA 94720

***Abstract*— Consider the standard linear regression model $y = X\beta^* + w$, where $y \in \mathbb{R}^n$ is an observation vector, $X \in \mathbb{R}^{n \times d}$ is a measurement matrix, $\beta^* \in \mathbb{R}^d$ is the unknown regression vector, and $w \sim \mathcal{N}(0, \sigma^2 I)$ is additive Gaussian noise. This paper determines sharp minimax rates of convergence for estimation of β^* in ℓ_2 norm, assuming that β^* belongs to a weak ℓ_q-ball $\mathbb{B}_q(R_q)$ for some $q \in [0, 1]$. We show that under suitable regularity conditions on the design matrix X, the minimax error in squared ℓ_2-norm scales as $R_q\big(\frac{\log d}{n}\big)^{1-\frac{q}{2}}$. In addition, we provide lower bounds on rates of convergence for general ℓ_p norm (for all $p \in [1, +\infty], p \neq q$). Our proofs of the lower bounds are information-theoretic in nature, based on Fano's inequality and results on the metric entropy of the balls $\mathbb{B}_q(R_q)$. Matching upper bounds are derived by direct analysis of the solution to an optimization algorithm over $\mathbb{B}_q(R_q)$. We prove that the conditions on X required by optimal algorithms are satisfied with high probability by broad classes of non-i.i.d. Gaussian random matrices, for which RIP or other sparse eigenvalue conditions are violated. For $q = 0$, ℓ_1-based methods (Lasso and Dantzig selector) achieve the minimax optimal rates in ℓ_2 error, but require stronger regularity conditions on the design than the non-convex optimization algorithm used to determine the minimax upper bounds.**

I. Introduction

The area of high-dimensional statistical inference concerns the estimation in the "large d, small n" regime, where d refers to the ambient dimension of the problem and n refers to the sample size. Such high-dimensional inference problems arise in various areas of science and engineering, including communication and coding, imaging, natural language processing, database management, and computational biology, among others. In the absence of additional structure, it is frequently impossible to obtain consistent estimators unless the ratio d/n converges to zero. Accordingly, for applications in the high-dimensional regime with $d \gg n$, an active line of research is based on imposing various types of low-dimensional structural conditions, such as sparsity, manifold structure, or graphical model structure, and then studying the performance of various estimators.

In this paper, we consider one canonical family of high-dimensional inference problems—namely, high-dimensional linear regression. More concretely, suppose that we observe a vector $y \in \mathbb{R}^n$ of response variables and a design matrix $X \in \mathbb{R}^{n \times d}$ of covariates, linked by the standard linear model $y = X\beta^* + w$, where $w \in \mathbb{R}^n$ represents additive noise with $w \sim \mathcal{N}(0, \sigma^2 I_{n \times n})$, and the goal is to estimate the vector $\beta^* \in \mathbb{R}^d$ of regression coefficients. The sparse instance of this problem, in which β^* is assumed to either have exactly $s \ll d$ non-zero entries, has been studied extensively over the past decade. A variety of practical algorithms have been proposed and studied, many based on ℓ_1-regularization, such as basis pursuit [1], the Lasso [2], and the Dantzig selector [3]. Various authors have obtained convergence rates for prediction error [4], [5], ℓ_2-error [5], [3], [6], as well as model selection consistency [7], [8], [9], under either "hard sparsity" assumptions, meaning that β^* has exactly $s \ll d$ non-zero entries, or "weak sparsity" assumptions, based on imposing a certain decay rate on the ordered entries of β^*.

Of complementary interest to the achievable rates of practical algorithms are the fundamental or information-theoretic limits of performance, applicable to any algorithm regardless of computational complexity. An understanding of such fundamental limits has various consequences, including demonstrating when current algorithms are rate-optimal (up to constant factors), or conversely, when there are substantial gaps between the best-known practical algorithms and optimal algorithms. The information-theoretic limits of model selection for sparse high-dimensional regression have been investigated by various authors (e.g., [10], [11], [12], [13], [14]). In contrast for other error metrics, such as the ℓ_2-error or prediction error, there does not seem to be have been any analysis for general design matrices X.

The focus of this paper is the following question: given an instance of the sparse high-dimensional regression model, what is the optimal rate, in a minimax sense to be made precise, that it is possible for any algorithm to estimate the vector β^* in ℓ_p norm, where $p \in [1, \infty]$? For the case $p = 2$, we provide a sharp characterization of the minimax rate for general designs X, including both upper and lower bounds that are matching up to constant factors. Our analysis applies to signal or regression vectors β^* that belong to so-called "weak" ℓ_q-balls for $q \in [0, 1]$.

The main contribution of our work is to show that under suitable regularity conditions on design X, the ℓ_2 minimax rate scales as $R_q\big(\frac{\log d}{n}\big)^{1-\frac{q}{2}}$ (where $q \in [0, 1]$). Additionally we derive lower bounds on the ℓ_p ($p \geq 1$) minimax rate which scales as $R_q{}^{\frac{p-q}{2-q}}\big(\frac{\log d}{n}\big)^{\frac{p-q}{2}}$. We also demonstrate that for $q = 0$,

two ℓ_1-based methods achieve the minimax optimal rate but require stronger conditions on X than the conditions required for the non-convex optimization algorithm used to derive upper bounds on the minimax rate (see Section IV-C).

The remainder of this paper is organized as follows. We begin in Section II by setting up the model precisely, and specifying and discussing the assumptions on the design matrix X that underlie our analysis. Section III is devoted to stating our main results and discussing some of their consequences, including connections to the normal sequence model (Section IV-A), specialization to general Gaussian designs (Section IV-B), and comparison to ℓ_1-based rates and conditions (Section IV-C). We conclude with some discussion in Section V.

II. Problem set-up

We begin by setting up our model precisely, before describing the assumptions on the design matrix.

A. Observation model and sparsity constraint

Given a design matrix $X \in \mathbb{R}^{n \times d}$, this paper focuses on the sparse linear observation model

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

where $w \sim N(0, \sigma^2 I_{n\times n})$ is an n-vector of noise. The pair $(y, X) \in \mathbb{R}^n \times \mathbb{R}^{n\times d}$ are both observed, and the goal is to estimate the signal or regression vector β^*. In this paper, we assume that the regression vector β^* belongs to a weak ℓ_q-"ball", defined as follows. For a fixed parameter $q \in [0, 1]$, define the ℓ_q norm $\|\beta\|_q^q = \sum_{i=1}^d |\beta_i|^q$, and the ℓ_q-"ball" of radius R_q via

$$\mathbb{B}_q(R_q) \quad := \quad \big\{\beta \in \mathbb{R}^d \mid \sum_{i=1}^d |\beta_i|^q \leq R_q\big\}. \tag{2}$$

Note that in the special case $q = 0$, this set reduces to the ℓ_0 "ball" of radius $R_0 = s$, corresponding to the set of vectors with at most s non-zero entries. Therefore when $q = 0$, we have a a classical hard sparsity constraint.

Given a procedure that determines an estimate $\widehat{\beta}$ for the true parameter β^*, there are various criteria for determine the quality of the estimate. Typically a loss function $L(\widehat{\beta}, \beta^*)$ is used to assess performance of an estimator $\widehat{\beta}$. For this paper, we analyze the ℓ_p-loss function given by $L_p(\widehat{\beta}, \beta^*) = (\sum_{i=1}^d (\widehat{\beta}_i - \beta_i^*)^p)^{1/p}$ for $p \in [1, \infty]$. The ℓ_p risk is defined as $\mathbb{E}\|\widehat{\beta} - \beta^*\|_p^p$. We determine lower bounds on the minimax ℓ_p risk of convergence over the ℓ_q ball $\mathbb{B}_q(R_q)$ for $p > q$; this minimax risk is given by

$$\mathfrak{M}_p(\mathbb{B}_q(R_q); X) = \min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \mathbb{E}\|\widehat{\beta} - \beta^*\|_p^p, \tag{3}$$

assuming that we observe $y \in \mathbb{R}^n$ from the linear observation model (1) with design X, and the minimization is taken over all measurable functions $\widehat{\beta} = \widehat{\beta}(y)$. The main results presented in Section III highlight that the upper and lower bounds on minimax risk depend significantly on the assumptions imposed in X. In the next section, we define and briefly explain the conditions.

B. Assumptions on design matrices

We next specify and discuss the assumptions imposed on the design matrix; as will be clear, each of our results uses a *subset* of the following assumptions. Our first assumption provides an upper bound on $\|X\theta\|_2/\sqrt{n}$ in terms of the ordinary ℓ_2-norm $\|\theta\|_2$ and a residual term. This bound is required for proving *lower bounds* on the minimax ℓ_p-risk:

Assumption 1. There exists a constant $\psi_u(X) < \infty$ and function $f_u(n, d, R_q)$ such that

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \quad \leq \quad \psi_u(X)(\|\theta\|_2 + f_u(n, d, R_q)\, \|\theta\|_1)$$

for all $\theta \in \mathbb{B}_q(2R_q)$.

Our second assumption, which provides a lower bound on $\|X\theta\|_2/\sqrt{n}$ in terms of $\|\theta\|_2$ and a residual term, is required for proving achievable or *upper bounds* on the minimax ℓ_2- risk:

Assumption 2. There exists a constant $\psi_\ell(X) > 0$ and a function $f_\ell(n, d, R_q)$ such that

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \quad \geq \quad \psi_\ell(X)\,(\|\theta\|_2 - f_\ell(n, d, R_q)\, \|\theta\|_1)$$

for all $\theta \in \mathbb{B}_q(2R_q)$.

In addition, our lower bounds on the minimax risk involve the set defined by intersecting the kernel of X with the ℓ_q-ball, which we denote $\mathcal{N}_q(X) := \mathrm{Ker}(X) \cap \mathbb{B}_q(R_q)$. We define its diameter in the ℓ_p-norm

$$\mathrm{diam}_p(\mathcal{N}_q(X)) \quad := \quad \max_{\theta \in \mathbb{B}_q(R_q),\, X\theta = 0} \|\theta\|_p. \tag{4}$$

In the case of ℓ_2 risk, the significance of this diameter should be apparent: for any "perturbation" $\Delta \in \mathcal{N}_q(X)$, it follows immediately from the linear observation model (1) that no method could ever distinguish between β^* and $\beta^* + \Delta$.

It is worth observing that the lower bound (4) is closely related to the diameter condition (4): more specifically, in the case $p = 2$, the condition (4) implies that

$$\mathrm{diam}_2(\mathcal{N}_q(X)) \quad \leq \quad R_q^{1/q} f_\ell(n, d, R_q). \tag{5}$$

To see this fact, note that if $\mathrm{diam}_2(\mathcal{N}_q(X)) > R_q^{1/q} f_\ell(n,d,R_q)$, then there must exist some $\theta \in \mathbb{B}_q(R_q)$ with $X\theta = 0$ and

$$\|\theta\|_2 \;>\; f_\ell(n,d,R_q)\|\theta\|_q \;\geq\; f_\ell(n,d,R_q)\|\theta\|_1.$$

Consequently, we have

$$0 \;=\; \frac{1}{\sqrt{n}}\|X\theta\|_2 \quad < \quad \|\theta\|_2 - f_\ell(n,d,R_q)\|\theta\|_1,$$

which implies there cannot exist any $\psi_\ell(X)$ for which the lower bound (4) holds.

III. Statement of main results

We are now ready to state our main results, and discuss some of their consequences. In all of the statements to follow, we use the quantities $\kappa_{q,p}$, $\kappa'_{q,2}$, $\widetilde{\kappa}_{q,2}$ etc. to denote numerical constants, independent of n, d, R_q, σ^2 and the design matrix X.

A. Lower bounds on risks in ℓ_p-norm

We begin with a result on lower bounds:

Theorem 1 (Lower bounds on ℓ_p-risk)**.** *Consider the linear model* (1) *for a fixed design matrix $X \in \mathbb{R}^{n\times d}$.*

(a) **Conditions for $q \in (0,1]$:** *Suppose that X satisfies Assumption 1 with $f_u(n,d,R_q) = o\big({R_q}^{-1/2}(\frac{\log d}{n})^{q/4}\big)$. For any $p \in [1,\infty)$ with $p > q$, the minimax ℓ_p-risk $\mathfrak{M}_p(\mathbb{B}_q(R_q);X)$ is lower bounded by*

$$\kappa_{q,p}\max\left\{\mathrm{diam}_p^p(\mathcal{N}_q(X)), R_q^{\frac{p-q}{2-q}}\left[\frac{\sigma^2}{\psi_u^2(X)}\frac{\log d}{n}\right]^{\frac{p-q}{2}}\right\}. \tag{6}$$

(b) **Conditions for $q = 0$:** *Suppose that X satisfies Assumption 1 with $f_u(n,d,R_q) = 0$. For any $p \in [1,\infty)$, the minimax risk $\mathfrak{M}_p(\mathbb{B}_0(s);X)$ is lower bounded as*

$$\kappa_{0,p}\max\left\{\mathrm{diam}_p^p(\mathcal{N}_0(X)), \left[\frac{\sigma^2}{\psi_u^2(X)}\frac{s\log(\frac{d-s}{2s})}{n}\right]^{\frac{p}{2}}\right\}. \tag{7}$$

Note that both lower bounds consist of two terms. The first term is simply the diameter of the set $\mathcal{N}_q(X) = \mathrm{Ker}(X) \cap \mathbb{B}_q(R_q)$, reflecting the extent to which the linear model (1) is unidentifiable. Clearly, one cannot estimate β^* any more accurately than the diameter of this set. In both lower bounds, the ratio $\sigma^2/\psi_u^2(X)$ reflects a type of inverse signal-to-noise ratio, comparing the noise variance σ^2 to the parameter $\psi_u^2(X)$. As the proof will clarify, the term $[\log d]^{\frac{p-q}{2}}$ in the lower bound (6), and similarly the term $s\log(\frac{d-s}{2s})$ in the lower bound (7), are reflections of the "volume" of the ℓ_q-ball, as measured by its metric entropy. For many classes of design matrices, the second term is of larger order than the diameter term, and hence determines the rate. (In particular, see Section IV-B for an in-depth discussion of the case of random Gaussian designs.)

B. Upper bounds on risks in ℓ_2-norm

We now state upper bounds on the ℓ_2-norm minimax risk over ℓ_q balls. For some of these results, we impose the following *column normalization condition* on the design matrix X:

$$\frac{\|X\|_j}{\sqrt{n}} := \big(\frac{1}{n}\sum_{i=1}^n X_{i,j}^2\big)^{\frac{1}{2}} \leq \gamma(X) \quad \text{for some finite } \gamma(X).$$

Theorem 2. *[Upper bounds on ℓ_2-risk] Consider the model* (1) *with a fixed design matrix $X \in \mathbb{R}^{n\times d}$.*

(a) **Conditions for $q \in (0,1]$:** *Suppose that X satisfies the column normalization condition* (8)*, and Assumption 2 holds with $f_\ell(n,d,R_q) = o\big({R_q}^{-1/2}(\frac{\log d}{n})^{q/4}\big)$. Then the minimax risk is upper bounded as*

$$\mathfrak{M}_2(\mathbb{B}_q(R_q);X) \leq 24 R_q\Big[\frac{\gamma^2(X)}{\psi_\ell^2(X)}\frac{\sigma^2}{\psi_\ell^2(X)}\frac{\log d}{n}\Big]^{1-q/2}, \tag{8}$$

(b) **Conditions for $q = 0$:** *Suppose that the design matrix X satisfies the column normalization condition* (8) *and Assumption 2 holds with $f_\ell(n,d,R_q) = 0$, then the minimax risk is upper bounded as*

$$\mathfrak{M}_2(\mathbb{B}_0(s);X) \leq 6\frac{\gamma^2(X)}{\psi_\ell^2(X)}\frac{\sigma^2}{\psi_\ell^2(X)}\frac{s\,\log d}{n}. \tag{9}$$

Alternatively, if the design matrix X satisfies both Assumptions 1 and 2 with $f_u(n,d,R_q) = 0$ and $f_\ell(n,d,R_q) = 0$, then the minimax risk $\mathfrak{M}_2(\mathbb{B}_0(s);X)$ is upper bounded as

$$\mathfrak{M}_2(\mathbb{B}_0(s);X) \;\leq\; 144\,\frac{\psi_u^2(X)}{\psi_\ell^2(X)}\frac{\sigma^2}{\psi_\ell^2(X)}\frac{s\log(d/s)}{n}. \tag{10}$$

For $q \in (0,1]$, note that if we substitute $p = 2$ into Theorem 1 (a) the upper bound from Theorem 2 (a) matches the lower bound up to constant, highlighting the optimality of the upper and lower bounds. For $q = 0$, if $\frac{d}{s} \to d^\alpha$ for some $\alpha \in (0,1)$, the lower bound from Theorem 1 (b) and the upper bound from Eq. (9) match. If such scaling on d and s does not hold, the upper bound from Eq. (10) which requires stronger conditions matches the lower bound.

IV. Some consequences

In this section, we discuss various consequences of our results. We begin by considering the classical Gaussian sequence model, which corresponds to a special case of our linear regression model. We make

explicit comparisons to the results of Donoho and Johnstone [15] on minimax risks over ℓ_q-balls for the Gaussian sequence model.

A. Connections with the normal sequence model

The normal (or Gaussian) sequence model is defined by the observation sequence

$$y_i = \theta_i^* + \varepsilon_i, \qquad \text{for } i = 1, \ldots, n, \quad (11)$$

where $\theta^* \in \Theta \subseteq \mathbb{R}^n$ is a fixed but unknown vector, and the noise variables $\varepsilon_i \sim \mathcal{N}(0, \frac{\tau^2}{n})$ are i.i.d. normal variables. Many non-parametric estimation problems, including non-parametric regression and density estimation, are asymptotically equivalent to an instance of the Gaussian sequence model [16], [17], [18], where the set Θ depends on the underlying "smoothness" conditions imposed on the functions. For instance, for functions that have an m^{th} derivative that is square-differentiable (a particular kind of Sobolev space), the set Θ corresponds to an ellipsoid; on the other hand, for certain choices of Besov spaces, it corresponds to an ℓ_q-ball.

In the case $\Theta = \mathbb{B}_q(R_q)$, our linear regression model (1) includes the normal sequence model (11) as a special case. In particular, it corresponds to setting $d = n$, the design matrix $X = I_{n\times n}$, and noise variance $\sigma^2 = \frac{\tau^2}{n}$. For this particular model, the seminal work by Donoho and Johnstone [15] determined sharp asymptotic results on the minimax error for general ℓ_p-norms over ℓ_q balls. Here we show that a corollary of our main theorems yields the same scaling in the case $p = 2$ and $q \in (0, 1]$.

Corollary 1. *Consider the normal sequence model* (11) *with* $\Theta = \mathbb{B}_q(R_q)$ *for some* $q \in (0, 1]$. *The the minimax risk in* ℓ_2*-norm is given by*

$$\mathfrak{M}_2(\mathbb{B}_q(R_q); I) \asymp R_q(\frac{\tau^2 \log n}{n})^{1-\frac{q}{2}}. \quad (12)$$

B. Random Gaussian Design

Another special case of particular interest is that of random Gaussian design matrices. A widely studied instance is the standard Gaussian ensemble, in which the entries of $X \in \mathbb{R}^{n\times d}$ are i.i.d. $N(0, 1)$ variables. A variety of results are known for the singular values of random matrices X drawn from this ensemble (e.g., [19], [20], [21]); moreover, some past work [22], [23] has studied the behavior of various ℓ_1-based methods for the standard Gaussian ensemble. In modeling terms, requiring that all entries are i.i.d. is an overly restrictive assumption, and not likely to be met in applications where the design matrix cannot be chosen. Accordingly, let us consider the more general class of Gaussian random design matrices $X \in \mathbb{R}^{n\times d}$, in which the rows are independent, but there can be arbitrary correlations between the columns of X. We show that random matrices drawn from such ensembles satisfy a version of Assumptions 1 and 2 with high probability.

Proposition 1. *Consider a random design matrix* $X \in \mathbb{R}^{n\times d}$ *formed by drawing each row* $X_i \in \mathbb{R}^d$ *i.i.d. from a* $N(0, \Sigma)$ *distribution. Then for some numerical constants* $c_i > 0$, $i = 1, 2$ *with probability* $1 - c_1 \exp(-c_2 n)$, *we have for all* $v \in \mathbb{R}^d$,

$$\frac{\|Xv\|_2}{\sqrt{n}} \leq 3\|\Sigma^{1/2}v\|_2 + 9\left(\sqrt{\frac{[\max_i \Sigma_{ii}]\log d}{n}}\right)\|v\|_1,$$

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{2}\|\Sigma^{1/2}v\|_2 - 9\left(\sqrt{\frac{[\max_i \Sigma_{ii}]\log d}{n}}\right)\|v\|_1.$$

The proof of Proposition 1 uses Slepian's lemma [21] and Gordon's inequality [24] combined with concentration of Gaussian measure results [25].

For $q = 0$, we note that if $v \in \mathbb{B}_0(2s)$, then $\|v\|_1 \leq \sqrt{2s}\|v\|_2$, which implies that

$$\sqrt{\frac{\log d}{n}}\,\|v\|_1 \leq \sqrt{\frac{2s\log d}{n}}\,\|v\|_2;$$

Consequently, once n is sufficiently large such that $\frac{\max_i \Sigma_{ii} s\log d}{n} < 1$, then Assumptions 1 and 2 both hold with $f_u(n, d, R_q) = f_\ell(n, d, R_q) = 0$.

For $q > 0$, Theorem 2 require that Assumption 2 holds with $f_\ell(n, d, R_q) = f_\ell(n, d, R_q) = o\big({R_q}^{-1/2}(\frac{\log d}{n})^{q/4}\big)$. But since $\log d/n = o(1)$, we have $(\frac{\log d}{n})^{q/4} \geq \big(\frac{\log d}{n}\big)^{1/2}$ for n sufficiently large. Therefore Proposition 1 implies that Gaussian random designs satisfy Assumption 2 holds in the case $q > 0$.

Based on (5) and Proposition 1,

$$\operatorname{diam}_2^2(\mathcal{N}_q(X)) \leq 81{R_q}^{2/q}\frac{[\max_i \Sigma_{ii}]\log d}{n}.$$

Since the minimax rate for ℓ_2 error is lower bounded by a maximum of $\operatorname{diam}_2^2(\mathcal{N}_q(X))$ and $\mathcal{O}((\frac{\log d}{n})^{1-q/2})$, the diameter term is clearly of lower order.

Let us now discuss the implications of this result for Assumptions 1 and 2. We note that past work by Amini and Wainwright [26], in the analysis of sparse PCA, established the upper bound (13a) in the very special case $\Sigma = I_{d\times d}$. More generally, consider an ensemble with any covariance matrix Σ whose eigenspectrum is bounded from below and above. This class of matrices includes, for instance, any Toeplitz matrix. For such matrices, Proposition 1 guarantees that

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{\lambda_{\min}(\Sigma)}{2}\|v\|_2 - 9\sqrt{\frac{\log d}{n}}\|v\|_1$$

with a similarly simplified upper bound. Consequently, for any vector $v \in \mathbb{R}^d$ such that $\|v\|_1/\|v\|_2 =$

$o(\sqrt{n/\log d})$, we are guaranteed (for n large enough) that $\frac{\|Xv\|_2}{\sqrt{n}} \geq c'\|v\|_2$ for some constant $c' > 0$.

C. Comparison to ℓ_1-based methods

We compare the optimal minimax rates of convergence for ℓ_2 error with ℓ_1-based methods, the Lasso selector [2] and the closely related Dantzig selector [3]. Here we focus discuss only the case $q = 0$ since we are currently unaware any ℓ_2-error bound for ℓ_1-based methods for $q \in (0, 1]$. For the Lasso, past work [9], [27] has shown that its ℓ_2-error is upper bounded by $\frac{s \log d}{n}$ which is the minimax rate from Theorem 2 (b) Eq. (9). Candes and Tao [3] showed that the Dantzig selector also achieves the minimax optimal rate under restricted isometry properties (RIP) to be discussed below. More recently, Bickel et. al [5] showed (amongst other results) that both the Lasso and Dantzig selector achieve the rate $\frac{s \log d}{n}$ under some restricted eigenvalue (RE) conditions. This shows that both the Lasso and Dantzig achieve the minimax rate derived in Theorem 2 (b), Eq. (9) for $q = 0$.

Finally, we compare the conditions needed for upper bounding the ℓ_2-error (Assumption 2) using an optimal method (Theorem 2) to those imposed in previous analyses of ℓ_1-based methods for $q = 0$. One set of conditions, known as the restricted isometry property or RIP for short [23] is based on constraining the condition numbers of various submatrices of X. In particular, for a given subset $T \subseteq \{1, 2, \ldots, d\}$, let us define the condition number of the sub-matrix $X_T \in \mathbb{R}^{n \times |T|}$ via $\delta(T) = \sigma_{\max}(X_T)/\sigma_{\min}(X_T)$, where $\sigma_{\max}$ and $\sigma_{\min}$ refer (respectively) to the maximum and minimum singular values of X_T. In order to guarantee good recovery in ℓ_2-norm for hard-sparse signals ($q = 0$), the RIP condition requires that the worst-case condition numbers $\delta_t = \sup_{|T|=t} \delta(T)$ be extremely close to 1 for subsets up to size $2s$ [3]. These restrictive conditions are satisfied only by matrices that are extremely close to orthogonal (e.g., when X is drawn from the standard i.i.d. Gaussian design with n sufficiently large), but fail to be satisfied for any matrix with sub-blocks that are not close to orthogonal (e.g., Toeplitz matrices).

Bickel et al. [5] show that it suffices to impose a much weaker lower bound on the restricted eigenvalues (RE). Before stating the assumption, we define the following notation. For any subset $S \subset \{1, \ldots d\}$, define $\|\theta_S\|_1 = \sum_{i \in S} |\theta_i|$. For each integer $s = 1, 2, \ldots d$, define the set $\Gamma(s, c_0)$ as

$$\Gamma(s, c_0) = \big\{\theta \in \mathbb{R}^d;\ |S| \leq s \mid \|\theta_S\|_1 \geq c_0 \|\theta_{S^c}\|_1\big\}.$$

In words, the set $\Gamma(s, c_0)$ contains all vectors in $\mathbb{R}^d$ where the ℓ_1-norm of the largest s co-ordinates provides an upper bound (up to constant c_0) to the ℓ_1 norm over the smallest $d - s$ co-ordinates. For example if $d = 3$, $(1, 1/2, 1/4) \in \Gamma(1, 1)$ while $(1, 3/4, 3/4) \notin \Gamma(1, 1)$.

With this notation, the restricted eigenvalue (RE) assumption can be stated as follows:

Assumption 3. There exists a function $\kappa(X, c_0) > 0$ such that

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \quad \geq \quad \kappa(X, c_0)\|\theta\|_2,$$

for all $\theta \in \Gamma(s, c_0)$.

We now claim that the condition required by the optimal method—namely, Assumption 2—is weaker than Assumption 3 for $q = 0$. In the case $q = 0$, Theorem 2 requires that there exists a $\psi_\ell(X) > 0$ such that

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \quad \geq \quad \psi_\ell(X)\|\theta\|_2,$$

for all $\theta \in \mathbb{B}_0(2s)$. (This corresponds to Assumption 2 with $f_\ell(n, d, R_q) = 0$.) To prove that Assumption 2 is weaker than Assumption 3, it suffices to show that $\mathbb{B}_0(2s) \subset \Gamma(s, c_0)$ for all $c_0 \geq 1$. Note that if $\theta \in \mathbb{B}_0(2s)$, then there exist disjoint subsets S_1 and S_2, $|S_i| \leq s$ (i=1,2) and $\|\theta_{S_1}\|_1 \geq \|\theta_{S_2}\|_1$. Since $\|\theta_{S_1^c}\|_1 = \|\theta_{S_2}\|_1$, $\|\theta_{S_1}\|_1 \geq \|\theta_{S_1^c}\|_1$. Hence $\theta \in \Gamma(s, c_0)$.

In summary, for the hard sparsity case $q = 0$, the Lasso and Dantzig selectors achieve the minimax ℓ_2-error in Eq. (9) of $\mathcal{O}(\frac{s \log d}{n})$. However current analyses of ℓ_1-methods are based on imposing stronger conditions on the design matrix X than those required by the analysis of the NP-hard combinatoric optimization algorithm used to derive the minimax rate. We believe the analysis and conclusion generalizes to $q \in (0, 1]$ and are currently in the process of analyzing this case.

V. Conclusion

In this paper, we determine sharp minimax rates of convergence for Model (1) for ℓ_2 error and lower bounds for ℓ_p error $p \in [1, \infty]$ under the assumption that the true parameter lies in an ℓ_q ($q \in [0, 1]$). It was shown that the ℓ_2 rates of convergence scale as $R_q\big(\frac{\log d}{n}\big)^{1 - \frac{q}{2}}$ under suitable conditions on the design matrix X. Gaussian random design matrices with suitably structured covariance matrices satisfy the conditions with high probability. It was also shown that the Lasso and Dantzig selectors achieve the minimax risk for $q = 0$. However, the conditions on the design matrix X required for the Lasso and Dantzig selectors to achieve the minimax risk are stronger than the conditions required for the non-convex algorithm used to derive upper bounds for the minimax risk.

There are a variety of open questions and extensions to our analysis. In particular, it would be interesting to see whether the assumptions on design matrix hold for other matrices (e.q Bernoulli or general sub-Gaussian matrices). Further, it would be interesting to determine whether analysis of the model under noise distribution with heavier tails lead to different rates.

APPENDIX

A. Proof of Lemma ??

Defining the set $S = \{i \mid |\Delta_i| > \tau\}$, we have

$$\begin{aligned}
\|\Delta\|_1 &= \|\Delta_S\|_1 + \sum_{i \notin S} |\Delta_i| \\
&\leq \sqrt{|S|}\|\Delta\|_2 + \tau \sum_{i \notin S} \frac{|\Delta_i|}{\tau} \\
&\leq \sqrt{|S|}\|\Delta\|_2 + \tau \sum_{i \notin S} \big(\frac{|\Delta_i|}{\tau}\big)^q \\
&\leq \sqrt{|S|}\|\Delta\|_2 + 2R_q \tau^{1-q}.
\end{aligned}$$

Lastly, we have

$$2R_q \geq \sum_{i \in S} |\Delta_i|^q \geq |S| \tau^q,$$

so that we conclude that

$$\|\Delta\|_1 \leq \tau^{-q/2} \sqrt{2R_q}\, \|\Delta\|_2 + 2R_q \tau^{1-q},$$

as claimed.

B. Proof of Lemma ??

For a given radius $r > 0$, define the set $\mathbb{S}(s, r) := \mathbb{B}_0(2s) \cap \mathbb{B}_2(r)$, and the random variables $Z_n = Z_n(s, r)$ given by

$$Z_n := \sup_{\theta \in \mathbb{S}(s,r)} \frac{1}{n} |w^T X \theta|.$$

For a given $\epsilon \in (0, 1)$ to be chosen, let us upper bound the minimal cardinality of a set that covers $\mathbb{S}(s, r)$ up to $(r\epsilon)$-accuracy in ℓ_2-norm. We claim that we may find such a covering set $\{\theta^1, \ldots, \theta^N\} \subset \mathbb{S}(s, r)$ with cardinality $N = N(s, r, \epsilon)$ that is upper bounded as

$$\log N(s, r, \epsilon) \leq \log \binom{d}{2s} + 2s \log(1/\epsilon).$$

To establish this claim, note that here are $\binom{d}{2s}$ subsets of size $2s$ within $\{1, 2, \ldots, d\}$. Moreover, for any $2s$-sized subset, there is a $r\epsilon$ covering in ℓ_2-norm of the ball $\mathbb{B}_2(r)$ with at most $2^{2s \log(1/\epsilon)}$ elements (e.g., [35]).

Consequently, for each $\theta \in \mathbb{S}(s, r)$, we may find some θ^i such that $\|\theta - \theta^i\|_2 \leq r\epsilon$. By triangle inequality, we then have

$$\begin{aligned}
\frac{1}{n}|w^T X \theta| &\leq \frac{1}{n}|w^T X \theta^i| + \frac{1}{n}|w^T X(\theta - \theta^i)| \\
&\leq \frac{1}{n}|w^T X \theta^i| + \frac{\|w\|_2}{\sqrt{n}} \frac{\|X(\theta - \theta^i)\|_2}{\sqrt{n}}.
\end{aligned}$$

Given the assumptions on X, we have $\|X(\theta - \theta^i)\|_2/\sqrt{n} \leq \psi_u(X)\|\theta - \theta^i\|_2 \leq \psi_u(X)\, r\epsilon$. Moreover, since the variate $\|w\|_2^2/\sigma^2$ is χ^2 with n degrees of freedom, we have $\frac{\|w\|_2}{\sqrt{n}} \leq 2\sigma$ with probability $1 - c_1 \exp(-c_2 n)$. Putting together the pieces, we conclude that

$$\frac{1}{n}|w^T X \theta| \leq \frac{1}{n}|w^T X \theta^i| + 2\psi_u(X)\, \sigma\, r\, \epsilon$$

with high probability. Taking the supremum over θ on both sides yields

$$Z_n \leq \max_{i=1,2,\ldots,N} \frac{1}{n}|w^T X \theta^i| + 2\psi_u(X)\, \sigma\, r\, \epsilon.$$

It remains to bound the finite maximum over the covering set. We begin by observing that each variate $w^T X \theta^i/n$ is zero-mean Gaussian with variance $\sigma^2 \|X\theta^i\|_2^2/n^2$. Under the given conditions on θ^i and X, this variance is at most $\sigma^2 \psi_u^2(X) r^2/n$, so that by standard Gaussian tail bounds, we conclude that

$$\begin{aligned}
Z_n &\leq \sigma\, r\, \psi_u(X) \sqrt{\frac{3 \log N(s, r, \epsilon)}{n}} + 2\psi_u(X)\, \sigma r\, \epsilon \\
&= \sigma\, r\, \psi_u(X) \Big\{ \sqrt{\frac{3 \log N(s, r, \epsilon)}{n}} + 2\epsilon \Big\}. \quad (14)
\end{aligned}$$

Finally, suppose that we $\epsilon = \sqrt{\frac{s \log(d/2s)}{n}}$. With this choice and recalling that $n \leq d$ by assumption, we obtain

$$\begin{aligned}
\frac{\log N(s, r, \epsilon)}{n} &\leq \frac{\log \binom{d}{2s}}{n} + \frac{s \log \frac{n}{s \log(d/2s)}}{n} \\
&\leq \frac{\log \binom{d}{2s}}{n} + \frac{s \log(d/s)}{n} \\
&\leq \frac{2s + 2s \log(d/s)}{n} + \frac{s \log(d/s)}{n},
\end{aligned}$$

where the final line uses standard bounds on binomial coefficients. Since $d/s \geq 2$ by assumption, we conclude that our choice of ϵ guarantees that $\frac{\log N(s,r,\epsilon)}{n} \leq 5\, s \log(d/s)$. Substituting these relations into the inequality (14), we conclude that

$$Z_n \leq \sigma r \psi_u(X) \Big\{ 4\sqrt{\frac{s \log(d/s)}{n}} + 2\sqrt{\frac{s \log(d/s)}{n}} \Big\},$$

as claimed.

REFERENCES

[1] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[3] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[4] E. Greenshtein and Y. Ritov, "Persistency in high dimensional linear predictor-selection and the virtue of overparametrization," *Bernoulli*, vol. 10, pp. 971–988, 2004.

[5] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *Annals of Statistics*, 2008, to appear.

[6] N. Meinshausen and P. Buhlmann, "High-dimensional graphs and variable selection with the lasso," *Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[7] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE Trans. Information Theory*, vol. 55, pp. 2183–2202, May 2009.

[8] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[9] C. H. Zhang and J. Huang, "The sparsity and bias of the lasso selection in high-dimensional linear regression," *Annals of Statistics*, vol. 36, no. 4, pp. 1567–1594, 2008.

[10] M. Akcakaya and V. Tarokh, "Shannon theoretic limits on noisy compressive sampling," Harvard University, Tech. Rep. arXiv:cs.IT:0711.0366, November 2007.

[11] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions on sparsity pattern recovery," UC Berkeley, Tech. Rep. arXiv:cs.IT:0804.1839, April 2008.

[12] G. Reeves, "Sparse signal sampling using noisy linear projections," Master's thesis, UC Berkeley, December 2007.

[13] M. J. Wainwright, "Information-theoretic bounds for sparsity recovery in the high-dimensional and noisy setting," Department of Statistics, UC Berkeley, Tech. Rep. 725, January 2007, presented at International Symposium on Information Theory, June 2007; To appear in IEEE Trans. Info Theory.

[14] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," UC Berkeley, Tech. Rep. arXiv:0806.0604, June 2008, presented at ISIT 2008, Toronto, Canada.

[15] D. L. Donoho and I. M. Johnstone, "Minimax risk over ℓ_p-balls for ℓ_q-error," *Prob. Theory and Related Fields*, vol. 99, pp. 277–303, 1994.

[16] M. S. Pinsker, "Optimal filtering of square integrable signals in gaussian white noise," *Probl. Pered. Inform. (Probl. Inf. Tranmission)*, vol. 16, pp. 120–133, 1980.

[17] M. Nussbaum, "Asympototically equivalence of density estimation and gaussian white noise," *Annals of Statistics*, vol. 24, no. 6, pp. 2399–2430, 1996.

[18] L. Brown and M. Low, "Asymptotic equivalence of nonparametric regression and white noise," *Annals of Statistics*, vol. 24, pp. 2384–2398, 1996.

[19] Z. D. Bai and Y. Q. Yin, "Convergence to the semicircle law," *Annals of Probability*, vol. 16, no. 2, pp. 863–875, April 2001.

[20] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked populations models," *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, July 2006.

[21] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices, and Banach spaces," in *Handbook of Banach Spaces*. Amsterdan, NL: Elsevier, 2001, vol. 1, pp. 317–336.

[22] D. Donoho, M. Elad, and V. M. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Info Theory*, vol. 52, no. 1, pp. 6–18, January 2006.

[23] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Info Theory*, vol. 51, no. 12, pp. 4203–4215, December 2005.

[24] Y. Gordon, "Some inequalities for guassian processes and applications," *Israel Journal of Mathematics*, vol. 50, no. 4, pp. 265–289, 1985.

[25] M. Ledoux, *The Concentration of Measure Phenomenon*, ser. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society, 2001.

[26] A. A. Amini and M. J. Wainwright, "High-dimensional analysis of semidefinite programming relaxations for sparse principal component analysis," *Annals of Statistics*, 2009, to appear.

[27] N. Meinshausen and B.Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2009.

[28] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.

[29] T. Kühn, "A lower estimate for entropy numbers," *Journal of Approximation Theory*, vol. 110, pp. 120–124, 2001.

[30] O. Guedon and A. E. Litvak, "Euclidean projections of p-convex body," in *Geometric aspects of functional analysis*. Springer-Verlag, 2000, pp. 95–108.

[31] R. Z. Has'minskii, "A lower bound on the risks of nonparametric estimates of densities in the uniform metric," *Theory Prob. Appl.*, vol. 23, pp. 794–798, 1978.

[32] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, vol. 27, no. 5, pp. 1564–1599, 1999.

[33] B. Yu, "Assouad, Fano and Le Cam," *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam*, pp. 423–435, 1996.

[34] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[35] J. Matousek, *Lectures on discrete geometry*. New York: Springer-Verlag, 2002.