Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls

Garvesh Raskutti[1] Martin J. Wainwright[1,2]
garveshr@stat.berkeley.edu wainwrig@stat.berkeley.edu

Bin Yu[1,2]
binyu@stat.berkeley.edu

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

Statistics Technical Report
October 1, 2009

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 design 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 studies the minimax rates of convergence for estimation of β^* for ℓ_p-losses and in the ℓ_2-prediction loss, assuming that β^* belongs to an 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 ℓ_2-loss and ℓ_2-prediction loss scales as $R_q\big(\frac{\log d}{n}\big)^{1-\frac{q}{2}}$. In addition, we provide lower bounds on minimax risks in ℓ_p-norms, 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 $\mathbb{B}_q(R_q)$, whereas our proofs of the upper bounds are direct and constructive, involving direct analysis of least-squares over ℓ_q-balls. For the special case $q = 0$, a comparison with ℓ_2-risks achieved by computationally efficient ℓ_1-relaxations reveals that although such methods can achieve the minimax rates up to constant factors, they require slightly stronger assumptions on the design matrix X than algorithms involving least-squares over the ℓ_0-ball.

1 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, including astrophysics, remote sensing and geophysics, 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. However, many applications require solving inference problems with $d \geq n$, so that consistency is not possible without imposing additional structure. Accordingly, an active line of research in high-dimensional inference is based on imposing various types of structural conditions, such as sparsity, manifold structure, or graphical model structure, and then studying the performance of different estimators. For instance, in the case of models with some type of sparsity constraint, a great deal of of work has studied the behavior of ℓ_1-based relaxations.

Complementary to the understanding of computationally efficient procedures are the fundamental or information-theoretic limitations of statistical inference, applicable to any algorithm regardless of its computational cost. There is a rich line of statistical work on such fundamental limits, an understanding of which can have two types of consequences. First, they can reveal gaps between the

performance of an optimal algorithm compared to known computationally efficient methods. Second, they can demonstrate regimes in which practical algorithms achieve the fundamental limits, which means that there is little point in searching for a more effective algorithm. As we shall see, the results in this paper lead to understanding of both types.

1.1 Problem set-up

The focus of this paper is a canonical instance of a high-dimensional inference problem, namely that of linear regression in d dimensions with sparsity constraints on the regression vector $\beta^* \in \mathbb{R}^d$. In this problem, we observe a pair $(Y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$, where X is the design matrix and Y is a vector of response variables. These quantities are linked by the standard linear model

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

where $w \sim N(0, \sigma^2 I_{n \times n})$ is observation noise. The goal is to estimate the unknown vector $\beta^* \in \mathbb{R}^d$ of regression coefficients. The sparse instance of this problem, in which β^* satisfies some type of sparsity constraint, has been investigated extensively over the past decade. Let X_i denote the i^{th} row of X and X_j denote the j^{th} column of X. A variety of practical algorithms have been proposed and studied, many based on ℓ_1-regularization, including basis pursuit [9], the Lasso [31], and the Dantzig selector [6]. Various authors have obtained convergence rates for different error metrics, including ℓ_2-error [6, 4, 37], prediction loss [4, 16], as well as model selection consistency [37, 25, 33, 38]. In addition, a range of sparsity assumptions have been analyzed, including the case of *hard sparsity* meaning that β^* has exactly $s \ll d$ non-zero entries, or *soft sparsity* assumptions, based on imposing a certain decay rate on the ordered entries of β^*.

Sparsity constraints These notions of sparsity can be defined more precisely in terms of the ℓ_q-balls[1] for $q \in [0, 1]$, defined as

$$\mathbb{B}_q(R_q) \quad := \quad \big\{\beta \in \mathbb{R}^d \;\mid\; \|\beta\|_q^q = \sum_{j=1}^{d} |\beta_j|^q \leq R_q\big\}, \tag{2}$$

where in the limiting case $q = 0$, we have the ℓ_0-ball

$$\mathbb{B}_0(s) \quad := \quad \big\{\beta \in \mathbb{R}^d \;\mid\; \sum_{j=1}^{d} \mathbb{I}[\beta_j \neq 0] \leq s\big\}, \tag{3}$$

corresponding to the set of vectors β with at most s non-zero elements.

Loss functions We consider estimators $\widehat{\beta} : \mathbb{R}^n \times \mathbb{R}^{n \times d} \to \mathbb{R}^d$ that are measurable functions of the data (y, X). Given any such estimator of the true parameter β^*, there are many criteria for determining the quality of the estimate. In a decision-theoretic framework, one introduces a loss function such that $\mathcal{L}(\widehat{\beta}, \beta^*)$ represents the loss incurred by estimating $\widehat{\beta}$ when $\beta^* \in \mathbb{B}_q(R_q)$ is the true parameter. The associated risk $\mathcal{R}$ is the expected value of the loss over distributions of (Y, X)—namely, the quantity $\mathcal{R}(\widehat{\beta}, \beta^*) = \mathbb{E}[\mathcal{L}(\widehat{\beta}, \beta^*)]$. Finally, in the minimax formalism, one seeks to choose an estimator that minimizes the worst-case risk given by

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \mathcal{R}(\widehat{\beta}, \beta^*). \tag{4}$$

[1] Strictly speaking, these sets are not "balls" when $q < 1$, since they fail to be convex.

Various choices of the loss function are possible, including (a) the *model selection loss*, which is zero if $\text{supp}(\widehat{\beta}) = \text{supp}(\beta^*)$ and one otherwise; (b) the *ℓ_p-losses*

$$\mathcal{L}_p(\widehat{\beta}, \beta^*) \quad := \quad \|\widehat{\beta} - \beta^*\|_p^p \; = \; \sum_{j=1}^{d} |\widehat{\beta}_j - \beta_j^*|^p, \tag{5}$$

and (c) the *ℓ_2-prediction loss* $\|X(\widehat{\beta} - \beta^*)\|_2^2/n$. In this paper, we study the ℓ_p-losses and the ℓ_2-prediction loss.

1.2 Our main contributions and related work

In this paper, we study minimax risks for the high-dimensional linear model (1), in which the regression vector β^* belongs to the ball $\mathbb{B}_q(R_q)$ for $0 \leq q \leq 1$. The core of the paper consists of four main theorems, corresponding to lower bounds on minimax rate for the cases of ℓ_p losses and the ℓ_2-prediction loss, and upper bounds for ℓ_2-norm loss and the ℓ_2-prediction loss. More specifically, in Theorem 1, we provide lower bounds for ℓ_p-losses that involve a maximum of two quantities: a term involving the diameter of the null-space restricted to the ℓ_q-ball, measuring the degree of non-identifiability of the model, and a term arising from the ℓ_p-metric entropy structure for ℓ_q-balls, measuring the massiveness of the parameter space. Theorem 2 is complementary in nature, devoted to upper bounds for ℓ_2-loss. For ℓ_2-loss, the upper and lower bounds match up to factors independent of the triple (n, d, R_q), and depend only on structural properties of the design matrix X (see Theorems 1 and 2). Finally, Theorems 3 and 4 provide upper and lower bounds for ℓ_2-prediction loss. For the ℓ_2-prediction loss, we provide upper and lower bounds on minimax risks that are again matching up to factors independent of (n, d, R_q), as summarized in Theorems 3 and 4. Structural properties of the design matrix X again play a role in minimax ℓ_2-prediction risks, but enter in a rather different way than in the case of ℓ_2-loss.

For the special case of the Gaussian sequence model (where $X = \sqrt{n} I_{n \times n}$), our work is closely related to the seminal work by of Donoho and Johnstone [14], who determined minimax rates for ℓ_p-losses over ℓ_q-balls. Our work applies to the case of general X, in which the sample size n need not be equal to the dimension d; however, we re-capture the same scaling as Donoho and Johnstone [14] when specialized to the case $X = \sqrt{n} I_{n \times n}$. In addition to our analysis of ℓ_p-loss, we also determine minimax rates for ℓ_2-prediction loss which, as mentioned above, can behave very differently from the ℓ_2-loss for general design matrices X. During the process of writing up our results, we became aware of concurrent work by Zhang (see the brief report [36]) that also studies the problem of determining minimax upper and lower bounds for ℓ_p-losses with ℓ_q-sparsity. We will be able to make a more thorough comparison once a more detailed version of their work is publicly available.

Naturally, our work also has some connections to the vast body of work on ℓ_1-based methods for sparse estimation, particularly for the case of hard sparsity ($q = 0$). Based on our results, the rates that are achieved by ℓ_1-methods, such as the Lasso and the Dantzig selector, are minimax optimal for ℓ_2-loss, but require somewhat stronger conditions on the design matrix than an "optimal" algorithm, which is based on searching the ℓ_0-ball. We compare the conditions that we impose in our minimax analysis to various conditions imposed in the analysis of ℓ_1-based methods, including the restricted isometry property of Candes and Tao [6], the restricted eigenvalue condition imposed in Menshausen and Yu [26], the partial Riesz condition in Zhang and Huang [37] and the restricted eigenvalue condition of Bickel et al. [4]. We find that "optimal" methods, which are based on minimizing least-squares directly over the ℓ_0-ball, can succeed for design matrices where ℓ_1-based methods are not known to work.

The remainder of this paper is organized as follows. In Section 2, we begin by specifying the assumptions on the design matrix that enter our analysis, and then state our main results. Section 3 is devoted to discussion of the consequences of our main results, including connections to the normal sequence model, Gaussian random designs, and related results on ℓ_1-based methods. In Section 4, we provide the proofs of our main results, with more technical aspects deferred to the appendices.

2 Main results

This section is devoted to the statement of our main results, and discussion of some of their consequences. We begin by specifying the conditions on the high-dimensional scaling and the design matrix X that enter different parts of our analysis, before giving precise statements of our main results.

In this paper, our primary interest is the high-dimensional regime in which $d \gg n$. For technical reasons, for $q \in (0, 1]$, we require the following condition on the scaling of (n, d, R_q):

$$\frac{d}{R_q n^{q/2}} = \Omega(d^{\kappa}) \quad \text{for some } \kappa > 0. \tag{6}$$

In the regime $d \geq n$, this assumption will be satisfied for all $q \in (0, 1]$ as long as $R_q = o(d^{\frac{1}{2}-\kappa'})$ for some $\kappa' \in (0, 1/2)$, which is a reasonable condition on the radius of the ℓ_q-ball for sparse models. In the work of Donoho and Johnstone [14] on the normal sequence model (special case of $X = I$), discussed at more length in the sequel, the effect of the scaling of the quantity $\frac{d}{R_q n^{q/2}}$ on the rate of convergence also requires careful treatment.

2.1 Assumptions on design matrices

Our first assumption, imposed throughout all of our analysis, is that the columns $\{X_j, j = 1, \ldots, d\}$ of the design matrix X are bounded in ℓ_2-norm:

Assumption 1 (Column normalization)**.** There exists a constant $0 < \kappa_c < +\infty$ such that

$$\frac{1}{\sqrt{n}} \max_{j=1,\ldots,d} \|X_j\|_2 \quad \leq \quad \kappa_c. \tag{7}$$

In addition, some of our results 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 the $\mathbb{B}_q(R_q)$-*kernel diameter* in the ℓ_p-norm

$$\mathrm{diam}_p(\mathcal{N}_q(X)) \quad := \quad \max_{\theta \in \mathcal{N}_q(X)} \|\theta\|_p \;=\; \max_{\|\theta\|_q^q \leq R_q,\, X\theta = 0} \|\theta\|_p. \tag{8}$$

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 $\beta^* = 0$ and $\beta^* = \Delta$. Consequently, this $\mathbb{B}_q(R_q)$-kernel diameter is a measure of the *lack of identifiability* of the linear model (1) over $\mathbb{B}_q(R_q)$.

Our second assumption, which is required only for achievable results for ℓ_2-error and lower bounds for ℓ_2-prediction error, imposes a lower bound on $\|X\theta\|_2/\sqrt{n}$ in terms of $\|\theta\|_2$ and a residual term:

Assumption 2 (Lower bound on restricted curvature)**.** There exists a constant $\kappa_\ell > 0$ and a function $f_\ell(R_q, n, d)$ such that

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \quad \geq \quad \kappa_\ell \|\theta\|_2 - f_\ell(R_q, n, d) \quad \text{for all} \quad \theta \in \mathbb{B}_q(2R_q). \tag{9}$$

Remarks: Conditions on the scaling for $f_\ell(R_q, n, d)$ are provided in Theorems 2 and 3. It is useful to recognize that the lower bound (9) is closely related to the diameter condition (8); in particular, Assumption 2 induces an upper bound on the $\mathbb{B}_q(R_q)$-kernel diameter in ℓ_2-norm, and hence the identifiability of the model:

Lemma 1. *If Assumption 2 holds for any $q \in (0, 1]$, then the $\mathbb{B}_q(R_q)$-kernel diameter in ℓ_2-norm is upper bounded as*

$$\operatorname{diam}_2(\mathcal{N}_q(X)) \quad \leq \quad \frac{f_\ell(R_q, n, d)}{\kappa_\ell}.$$

Proof. We prove the contrapositive statement. Note that if $\operatorname{diam}_2(\mathcal{N}_q(X)) > \frac{f_\ell(R_q,n,d)}{\kappa_\ell}$, then there must exist some $\theta \in \mathbb{B}_q(R_q)$ with $X\theta = 0$ and $\|\theta\|_2 > \frac{f_\ell(R_q,n,d)}{\kappa_\ell}$. We then conclude that

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

which implies there cannot exist any κ_ℓ for which the lower bound (9) holds. □

In Section 3.3, we discuss further connections between our assumptions, and the conditions imposed in analysis of the Lasso and other ℓ_1-based methods [6, 25, 4]. In the case $q = 0$, we find that Assumption 2 is weaker than any condition under which an ℓ_1-based method is known to succeed. Finally, in Section 3.2, we prove that versions of both Assumptions 1 and 2 hold with high probability for various classes of non-i.i.d. Gaussian random design matrices (see Proposition 1).

2.2 Risks in ℓ_p-norm

Having described our assumptions on the design matrix, we now turn to the main results that provide upper and lower bounds on minimax risks. In all of the statements to follow, we use the quantities $c_{q,p}$, $c'_{q,2}$, $\widetilde{c}_{q,2}$ etc. to denote numerical constants, independent of n, d, R_q, σ^2 and the design matrix X. We begin with lower bounds on the ℓ_p-risk.

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 is column-normalized (Assumption 1 with $\kappa_c < \infty$). For any $p \in [1, \infty)$, the minimax ℓ_p-risk over the ℓ_q ball is lower bounded as*

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \mathbb{E}\|\widehat{\beta} - \beta^*\|_p^p \quad \geq \quad c_{q,p} \; \max\left\{ \operatorname{diam}_p^p(\mathcal{N}_q(X)),\; R_q \left[\frac{\sigma^2}{\kappa_c^2}\frac{\log d}{n}\right]^{\frac{p-q}{2}} \right\}. \quad (10)$$

(b) **Conditions for $q = 0$:** *Suppose that $\frac{\|X\theta\|_2}{\sqrt{n}\|\theta\|_2} \leq \kappa_u$ for all $\theta \in \mathbb{B}_0(2s)$. Then for any $p \in [1, \infty)$, the minimax ℓ_p-risk over the ℓ_0-ball with radius $s = R_0$ is lower bounded as*

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_0(s)} \mathbb{E}\|\widehat{\beta} - \beta^*\|_p^p \quad \geq \quad c_{0,p} \; \max\left\{ \operatorname{diam}_p^p(\mathcal{N}_0(X)),\; s^{\frac{p}{2}}\left[\frac{\sigma^2}{\kappa_u^2}\frac{\log(d/s)}{n}\right]^{\frac{p}{2}} \right\}. \quad (11)$$

Note that both lower bounds consist of two terms. The first term is simply the diameter of the set $\mathcal{N}_q(X) = \mathrm{Ker}(X) \cap \mathbb{B}_q(R_q)$, which reflects the extent 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 ratios σ^2/κ_c^2 (or σ^2/κ_u^2) correspond to the inverse of the signal-to-noise ratio, comparing the noise variance σ^2 to the magnitude of the design matrix measured by κ_u. As the proof will clarify, the term $[\log d]^{\frac{p-q}{2}}$ in the lower bound (10), and similarly the term $\log(\frac{d}{s})$ in the bound (11), are reflections of the complexity 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 3.2 for an in-depth discussion of the case of random Gaussian designs.)

We now state upper bounds on the ℓ_2-norm minimax risk over ℓ_q balls. For these results, we require both the column normalization condition (Assumption 1) and the curvature condition (Assumption 2).

Theorem 2 (Upper bounds on ℓ_2-risk)**.** *Consider the model (1) with a fixed design matrix $X \in \mathbb{R}^{n \times d}$ that is column-normalized (Assumption 1 with $\kappa_c < \infty$).*

(a) **Conditions for $q \in (0, 1]$:** *If X satisfies Assumption 2 with $f_\ell(R_q, n, d) = o(R_q{}^{1/2}(\frac{\log d}{n})^{1/2-q/4})$ and $\kappa_\ell > 0$, then there exist constants c_1 and c_2 such that the minimax ℓ_2-risk is upper bounded as*

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \|\widehat{\beta} - \beta^*\|_2^2 \;\; \leq \;\; 24 R_q \Big[\frac{\kappa_c^2}{\kappa_\ell^2} \frac{\sigma^2}{\kappa_\ell^2} \, \frac{\log d}{n}\Big]^{1-q/2}, \tag{12}$$

with probability greater than $1 - c_1 \exp(-c_2 n)$.

(b) **Conditions for $q = 0$:** *If X satisfies Assumption 2 with $f_\ell(s, n, d) = 0$ and $\kappa_\ell > 0$, then there exists constants c_1 and c_2 such that the minimax ℓ_2-risk is upper bounded as*

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_0(s)} \|\widehat{\beta} - \beta^*\|_2^2 \;\; \leq \;\; 6 \frac{\kappa_c^2}{\kappa_\ell^2} \, \frac{\sigma^2}{\kappa_\ell^2} \, \frac{s \, \log d}{n}, \tag{13}$$

with probability greater than $1 - c_1 \exp(-c_2 n)$. If, in addition, the design matrix satisfies $\frac{\|X\theta\|_2}{\sqrt{n}\|\theta\|_2} \leq \kappa_u$ for all $\theta \in \mathbb{B}_0(2s)$, then the minimax ℓ_2-risk is upper bounded as

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_0(s)} \|\widehat{\beta} - \beta^*\|_2^2 \;\; \leq \;\; 144 \, \frac{\kappa_u^2}{\kappa_\ell^2} \frac{\sigma^2}{\kappa_\ell^2} \, \frac{s \log(d/s)}{n}, \tag{14}$$

with probability greater than $1 - c_1 \exp(-c_2 s \log(d - s))$.

In the case of ℓ_2-risk and design matrices X that satisfy the assumptions of both Theorems 1 and 2, then these results identify the minimax risk up to constant factors. In particular, for $q \in (0, 1]$, the minimax ℓ_2-risk scales as

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \mathbb{E}\|\widehat{\beta} - \beta^*\|_2^2 \;\; = \;\; \Theta\Big(R_q \Big[\frac{\sigma^2 \log d}{n}\Big]^{1-q/2}\Big), \tag{15}$$

whereas for $q = 0$, the minimax ℓ_2-risk scales as

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_0(s)} \mathbb{E}\|\widehat{\beta} - \beta^*\|_2^2 \;\; = \;\; \Theta\Big(\frac{\sigma^2 \, s \log(d/s)}{n}\Big). \tag{16}$$

Note that the bounds with high probability can be converted to bound in expectation by a standard integration over the tail probability.

2.3 Risks in prediction norm

In this section, we investigate minimax risks in terms of the ℓ_2-prediction loss $\|X(\widehat{\beta} - \beta^*)\|_2^2/n$, and provide both lower and upper bounds on it.

Theorem 3 (Lower bounds on prediction risk)**.** *Consider the model* (1) *with a fixed design matrix* $X \in \mathbb{R}^{n \times d}$ *that is column-normalized (Assumption 1 with* $\kappa_c < \infty$*).*

(a) **Conditions for** $q \in (0, 1]$**:** *If the design matrix* X *satisfies Assumption 2 with* $\kappa_\ell > 0$ *and* $f_\ell(R_q, n, d) = o({R_q}^{1/2}(\frac{\log d}{n})^{1/2-q/4})$*, then the minimax prediction risk is lower bounded as*

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \mathbb{E}\frac{\|X\,(\widehat{\beta} - \beta)\|_2^2}{n} \quad \geq \quad c'_{2,q}\, R_q\, \kappa_\ell^2 \left[\frac{\sigma^2}{\kappa_c^2}\, \frac{\log d}{n}\right]^{1-q/2}. \tag{17}$$

(b) **Conditions for** $q = 0$**:** *Suppose that* X *satisfies Assumption 2 with* $\kappa_\ell > 0$ *and* $f_\ell(s, n, d) = 0$*, and moreover that* $\frac{\|X\theta\|_2}{\sqrt{n}\|\theta\|_2} \leq \kappa_u$ *for all* $\theta \in \mathbb{B}_0(2s)$*. Then the minimax prediction risk is lower bounded as*

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_0(s)} \mathbb{E}\frac{\|X(\widehat{\beta} - \beta)\|_2^2}{n} \quad \geq \quad c'_{0,q}\, \kappa_\ell^2\, \frac{\sigma^2}{\kappa_u^2}\, \frac{s \log(d/s)}{n}. \tag{18}$$

In the other direction, we have the following result:

Theorem 4 (Upper bounds on prediction risk)**.** *Consider the model* (1) *with a fixed design matrix* $X \in \mathbb{R}^{n \times d}$*.*

(a) **Conditions for** $q \in (0, 1]$**:** *If* X *satisfies the column normalization condition, then for some constant* $c_{2,q}$*, there exist* c_1 *and* c_2 *such that the minimax prediction risk is upper bounded as*

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \frac{1}{n}\|X(\widehat{\beta} - \beta^*)\|_2^2 \quad \leq \quad c_{2,q}\, \kappa_c^2\, R_q \left[\frac{\sigma^2}{\kappa_c^2}\, \frac{\log d}{n}\right]^{1-\frac{q}{2}}, \tag{19}$$

with probability greater than $1 - c_1 \exp\left(-c_2 R_q (\log d)^{1-q/2} n^{q/2}\right)$*.*

(b) **Conditions for** $q = 0$**:** *For any* X*, with probability greater than* $1 - \exp\left(-10 s \log(d/s)\right)$ *the minimax prediction risk is upper bounded as*

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_0(s)} \frac{1}{n}\|X(\widehat{\beta} - \beta^*)\|_2^2 \quad \leq \quad 81\frac{\sigma^2\, s \log(d/s)}{n}. \tag{20}$$

2.4 Some intuition

In order to provide the reader with some intuition, let us make some comments about the scalings that appear in our results.

First, as a basic check of our results, it can be verified that Lemma 1 ensures that the lower bounds on minimax rates stated in Theorem 1 for $p = 2$ are always less than or equal to the achievable rates stated in Theorem 2. In particular, since $f_\ell(R_q, n, d) = o({R_q}^{1/2}(\frac{\log d}{n})^{1/2-q/4})$ for $q \in (0, 1]$, Lemma 1 implies that $\operatorname{diam}_2^2(\mathcal{N}_q(X)) = o(R_q(\frac{\log d}{n})^{1-q/2})$, meaning that the achievable rates are always at least as large as the lower bounds in the case $q \in (0, 1]$. In the case of hard sparsity ($q = 0$), the upper and lower bounds are clearly consistent since $f_\ell(s, n, d) = 0$ implies the diameter of $\mathcal{N}_0(X)$ is 0.

Second, for the case $q = 0$, there is a concrete interpretation of the rate $\frac{s \log(d/s)}{n}$, which appears in Theorems 1(b), 2(b), 3(b) and 4(b)). Note that there are $\binom{d}{s}$ subsets of size s within $\{1, 2, \ldots, d\}$, and by standard bounds on binomial coefficients [11], we have $\log \binom{d}{s} = \Theta(s \log(d/s))$. Consquently, the rate $\frac{s \log(d/s)}{n}$ corresponds to the log number of models divided by the sample size n. Note that unless $s/d = \Theta(1)$, this rate is equivalent (up to constant factors) to $\frac{s \log d}{n}$.

Third, for $q \in (0, 1]$, the interpretation of the rate $R_q \big(\frac{\log d}{n}\big)^{1-q/2}$, appearing in parts (a) of Theorems 1 through 4, is less immediately obvious but can can understood as follows. Suppose that we choose a subset of size s_q of coefficients to estimate, and ignore the remaining $d - s_q$ coefficients. For instance, if we were to choose the top s_q coefficients of β^* in absolute value, then the fast decay imposed by the ℓ_q-ball condition on β^* would mean that the remaining $d - s_q$ coefficients would have relatively little impact. With this intuition, the rate for $q > 0$ can be interpreted as the rate that would be achieved by choosing $s_q = R_q \big(\frac{\log d}{n}\big)^{-q/2}$, and then acting as if the problem were an instance of a hard-sparse problem ($q = 0$) with $s = s_q$. For such a problem, we would expect to achieve the rate $\frac{s_q \log d}{n}$, which is exactly equal to $R_q \big(\frac{\log d}{n}\big)^{1-q/2}$. Of course, we have only made a very heuristic argument here; this truncation idea is made more precise in Lemma 2 to appear in the sequel.

Fourth, we note that the minimax rates for ℓ_2-prediction error and ℓ_2-norm error are essentially the same except that the design matrix structure enters minimax risks in *very different ways*. In particular, note that proving lower bounds on prediction risk requires imposing relatively strong conditions on the design X—namely, Assumptions 1 and 2 as stated in Theorem 3. In contrast, obtaining upper bounds on prediction risk requires very mild conditions. At the most extreme, the upper bound for $q = 0$ in Theorem 3 requires no assumptions on X while for $q > 0$ only the column normalization condition is required. All of these statements are reversed for ℓ_2-risks, where lower bounds can be proved with only Assumption 1 on X (see Theorem 1), whereas upper bounds require both Assumptions 1 and 2.

Lastly, in order to appreciate the difference between the conditions for ℓ_2-prediction error and ℓ_2 error, it is useful to consider a toy but illuminating example. Consider the linear regression problem defined by a design matrix $X = \begin{bmatrix} X_1 & X_2 & \cdots & X_d \end{bmatrix}$ with *identical columns*—that is, $X_j = \widetilde{X}_1$ for all $j = 1, \ldots, d$. We assume that vector $\widetilde{X}_1 \in \mathbb{R}^d$ is suitably scaled so that the column-normalization condition (Assumption 1) is satisfied. For this particular choice of design matrix, the linear observation model (1) reduces to $Y = (\sum_{j=1}^d \beta_j^*) \widetilde{X}_1 + w$. For the case of hard sparsity ($q = 0$), an elementary argument shows that the minimax risk in ℓ_2-prediction error scales as $\Theta(\frac{1}{n})$. This scaling implies that the upper bound (20) from Theorem 4 holds (but is not tight).[2] Consequently, this highly degenerate design matrix yields a very easy problem for ℓ_2-prediction, since the $1/n$ rate is essentially parametric. In sharp contrast, for the case of ℓ_2-norm error (still with hard sparsity $q = 0$), the model becomes unidentifiable. To see the lack of identifiability, let $e_i \in \mathbb{R}^d$ denote the unit-vector with 1 in position i, and consider the two regression vectors $\beta^* = c\, e_1$ and $\widetilde{\beta} = c\, e_2$, for some constant $c \in \mathbb{R}$. Both choices yield the same observation vector Y, and since the choice of c is arbitrary, the minimax ℓ_2-error is infinite. In this case, the lower bound (11) on ℓ_2-error from Theorem 1 holds (and is tight, since the kernel diameter is infinite). In contrast, the upper bound (13) on ℓ_2-error from Theorem 2 does not apply, because Assumption 2 is violated due to the extreme degeneracy of the design matrix.

3 Some consequences

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

[2]Note that the lower bound (18) on the ℓ_2-prediction error from Theorem 3 does not apply to this model, since this degenerate design matrix with identical columns does not satisfy any version of Assumption 2.

making explicit comparisons to the results of Donoho and Johnstone [14] on minimax risks over ℓ_q-balls.

3.1 Connections with the normal sequence model

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

$$y_i \quad = \quad \theta_i^* + \varepsilon_i, \qquad \text{for } i = 1, \ldots, n, \tag{21}$$

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 variates. Many non-parametric estimation problems, including regression and density estimation, are asymptotically equivalent to an instance of the Gaussian sequence model [28, 27, 5], 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 (21) 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, seminal work by Donoho and Johnstone [14] derived 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* (21) *with* $\Theta = \mathbb{B}_q(R_q)$ *for some* $q \in (0, 1]$. *Then there are constants* $c_q' \leq c_q$ *depending only on* q *such that*

$$c_q'\Big(\frac{2\tau^2 \log n}{n}\Big)^{1-\frac{q}{2}} \;\leq\; \min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \mathbb{E}\|\widehat{\beta} - \beta^*\|_2^2 \;\leq\; c_q\Big(\frac{2\tau^2 \log n}{n}\Big)^{1-\frac{q}{2}}. \tag{22}$$

These bounds follow from our main theorems, via the substitutions $n = d$, $\sigma^2 = \frac{\tau^2}{n}$, and $\kappa_u = \kappa_\ell = 1$. To be clear, Donoho and Johnstone [14] provide a far more careful analysis that yields sharper control of the constants than we have provided here.

3.2 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. $N(0, 1)$ variates. A variety of results are known for the singular values of random matrices X drawn from this ensemble (e.g., [2, 3, 12]); moreover, some past work [13, 6] has studied the behavior of different ℓ_1-based methods for the standard Gaussian ensemble, in which entries X_{ij} are i.i.d. $N(0, 1)$. In modeling terms, requiring that all entries of the design matrix X 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. To simplify notation, we define the shorthand $\rho(\Sigma) := \max_{j=1,\ldots,d} \Sigma_{jj}$, corresponding to the maximal variance of any element of X, and use $\Sigma^{1/2}$ to denote the symmetric square root of the covariance matrix.

In this model, each column X_j, $j = 1, \ldots, d$ has i.i.d. elements. Consequently, it is an immediate consequence of standard concentration results for χ^2_n variates (see Appendix I) that

$$\max_{j=1,\ldots,d} \frac{\|X_j\|_2}{\sqrt{n}} \leq \rho(\Sigma)\,\Big(1 + \sqrt{\frac{32 \log d}{n}}\Big). \tag{23}$$

Therefore, Assumption 1 holds as long as $n = \Omega(\log d)$ and $\rho(\Sigma)$ is bounded.

Showing that a version of Assumption 2 holds with high probability requires more work. We summarize our findings in the following result:

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 an $N(0, \Sigma)$ distribution. Then for some numerical constants $c_k \in (0, \infty)$, $k = 1, 2$, we have*

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{2}\|\Sigma^{1/2}v\|_2 - 6\left(\frac{\rho(\Sigma)\log d}{n}\right)^{1/2}\|v\|_1 \quad \text{for all } v \in \mathbb{R}^d \tag{24}$$

with probability $1 - c_1 \exp(-c_2 n)$.

Remarks: Past work by by Amini and Wainwright [1] in the analysis of sparse PCA has established an upper bound analogous to the lower bound (24) for the special case $\Sigma = I_{d \times d}$. We provide a proof of this matching upper bound for general Σ as part of the proof of Proposition 1 in Appendix E. The argument is based on Slepian's lemma [12] and its extension due to Gordon [15], combined with concentration of Gaussian measure results [22]. Note that we have made no effort to obtain sharp leading constants (i.e., the factors $1/2$ and 6 can easily be improved), but the basic result (24) suffices for our purposes.

Let us now discuss the implications of this result for Assumption 2. First, in the case $q = 0$, the bound (13) in Theorem 2 requires that Assumption 2 holds with $f_\ell(s, n, d) = 0$ for all $\theta \in \mathbb{B}_0(2s)$. To see the connection with Proposition 1, note that if $\theta \in \mathbb{B}_0(2s)$, then we have $\|\theta\|_1 \leq \sqrt{2s}\|\theta\|_2$, and hence

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \left\{\frac{\|\Sigma^{1/2}v\|_2}{2\|v\|_2} - 6\sqrt{2}\left(\frac{\rho(\Sigma)s\log d}{n}\right)^{1/2}\right\}\|v\|_2.$$

Therefore, as long as $\rho(\Sigma) < \infty$, $\min_{v \in \mathbb{B}_0(2s)} \frac{\|\Sigma^{1/2}v\|_2}{\|v\|_2} > 0$ and $\frac{s\log d}{n} = o(1)$, the condition needed for the bound (13) will be met.

Second, in the case $q \in (0, 1]$, Theorem 2(a) requires that Assumption 2 hold with the residual term $f_\ell(R_q, n, d) = o(R_q{}^{1/2}\frac{\log d}{n})^{1/2 - q/4}$. We claim that Proposition 1 guarantees this condition, as long as $\rho(\Sigma) < \infty$ and the minimum eigenvalue of Σ is bounded away from zero. In order to verify this claim, we require the following result:

Lemma 2. *For any vector $\theta \in \mathbb{B}_q(2R_q)$ and any positive number $\tau > 0$, we have*

$$\|\theta\|_1 \leq \sqrt{2R_q}\tau^{-q/2}\|\theta\|_2 + 2R_q\tau^{1-q}. \tag{25}$$

Although this type of result is standard (e.g, [14]), we provide a proof in Appendix A for completeness. In order to exploit Lemma 2, let us set $\tau = \sqrt{\frac{\log d}{n}}$. With this choice, we can substitute the resulting bound (25) into the lower bound (24), thereby obtaining that

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \left\{\frac{\|\Sigma^{1/2}v\|_2}{2\|v\|_2} - 6\sqrt{2\,\rho(\Sigma)}\,\sqrt{R_q}\left(\frac{\log d}{n}\right)^{1/2 - q/4}\right\}\|v\|_2 - 2R_q\rho(\Sigma)^{1/2}\left(\frac{\log d}{n}\right)^{1 - q/2}.$$

Recalling that the condition $\sqrt{R_q}\left(\frac{\log d}{n}\right)^{1/2 - q/4} = o(1)$ is required for consistency, we see that Assumption 2 holds as long as $\rho(\Sigma) < +\infty$ and the minimum eigenvalue of Σ is bounded away from zero.

Lastly, it is also worth noting that we can also obtain the following stronger result for the case $q = 0$, in the case that $\min_{v \in \mathbb{B}_0(2s)} \frac{\|\Sigma^{1/2} v\|_2}{\|v\|_2} > 0$ and $\max_{v \in \mathbb{B}_0(2s)} \frac{\|\Sigma^{1/2} v\|_2}{\|v\|_2} < \infty$. If the sparse eigenspectrum is bounded in this way, then as long as $n > c_3\, s \log(d/s)$, we have

$$3\|\Sigma^{1/2} v\|_2 \geq \frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{2}\|\Sigma^{1/2} v\|_2 \quad \text{for all } v \in \mathbb{B}_0(2s) \tag{26}$$

with probability greater than $1 - c_1 \exp(-c_2 n)$. This fact follows by applying the union bound over all $\binom{d}{2s}$ subsets of size $2s$, combined with standard concentration results for random matrices (e.g., see Davidson and Szarek [12] for $\Sigma = I$, and Wainwright [33] for the straightforward extensions to non-identity covariances).

3.3 Comparison to ℓ_1-based methods

In addition, it is interesting to compare our minimax rates of convergence for ℓ_2-error with known results for ℓ_1-based methods, including the Lasso [31] and the closely related Dantzig method [6]. Here we discuss only the case $q = 0$ since we are currently unaware of any ℓ_2-error bound for ℓ_1-based methods for $q \in (0, 1]$. For the Lasso, past work [37, 26] has shown that its ℓ_2-error is upper bounded by $\frac{s \log d}{n}$ under sparse eigenvalue conditions. Similarly, Candes and Tao [6] show the same scaling for the Dantzig selector, when applied to matrices that satisfy the more restrictive RIP conditions. More recent work by Bickel et. al [4] provides a simultaneous analysis of the Lasso and Dantzig selector under a common set of assumptions that are weaker than both the RIP condition and sparse eigenvalue conditions. Together with our results (in particular, Theorem 1(b)), this body of work shows that under appropriate conditions on the design X, the rates achieved by ℓ_1-methods in the case of hard sparsity ($q = 0$) are minimax-optimal.

Given that the rates are optimal, it is appropriate to compare the conditions needed by an "optimal" algorithm, such as that analyzed in Theorem 2, to those used in the analysis of ℓ_1-based methods. One set of conditions, known as the restricted isometry property [6] or RIP for short, is based on very strong constraints on the condition numbers of all submatrices of X up to size $2s$, requiring that they be near-isometries (i.e., with condition numbers extremely close to 1). Such conditions are satisfied by matrices with columns that are all very close to orthogonal (e.g., when X has i.i.d. $N(0, 1)$ entries and $n = \Omega(\log \binom{d}{2s})$), but are violated for many reasonable matrix classes (e.g., Toeplitz matrices) that arise in statistical practice. Zhang and Huang [37] imposed a weaker sparse Riesz condition, based on imposing constraints (different from those of RIP) on the condition numbers of all submatrices of X up to a size that grows as a function of s and n. Meinshausen and Yu [26] impose a bound in terms of the condition numbers or minimum and maximum restricted eigenvalues for submatrices of X up to size $s \log n$. It is unclear whether the conditions in Meinshausen and Yu [26] are weaker or stronger than the conditions in Zhang and Huang [37].

The weakest known sufficient conditions to date are due to Bickel et al. [4], who show that in addition to the column normalization condition (Assumption 1 in this paper), it suffices to impose a milder condition, namely a lower bound on a certain type of restricted eigenvalue (RE). They show that this RE condition is less restrictive than both the RIP condition [6] and the eigenvalue conditions imposed in Meinshausen and Yu [26]. For a given vector $\theta \in \mathbb{R}^d$, let $\theta_{(j)}$ refer to the j^{th} largest coefficient in absolute value, so that we have the ordering

$$\theta_{(1)} \geq \theta_{(2)} \geq \ldots \geq \theta_{(d-1)} \geq \theta_{(d)}.$$

For a given scalar c_0 and integer $s = 1, 2, \ldots, d$, let define the set

$$\Gamma(s, c_0) \quad := \quad \left\{ \theta \in \mathbb{R}^d \mid \sum_{j=s+1}^{d} |\theta_{(j)}| \leq c_0 \sum_{j=1}^{s} |\theta_{(j)}| \right\}.$$

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 $d = 3$, then the vector $(1, 1/2, 1/4) \in \Gamma(1, 1)$ whereas the vector $(1, 3/4, 3/4) \notin \Gamma(1, 1)$.

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

Assumption 3 (Restricted lower eigenvalues [4])**.** There exists a function $\kappa(X, c_0) > 0$ such that

$$\frac{1}{\sqrt{n}} \|X\theta\|_2 \quad \geq \quad \kappa(X, c_0) \|\theta\|_2 \quad \text{for all } \theta \in \Gamma(s, c_0).$$

Bickel et. al [4] require a slightly stronger condition for bounding the ℓ_2-loss in if s depends on n. However the conditions are equivalent for fixed s and Assumption 3 is much simpler to analyze and compare to Assumption 2. At this point, we have not seen conditions weaker than Assumption 3.

The following corollary of Proposition 1 shows that Assumption 3 is satisfied with high probability for broad classes of Gaussian random designs:

Corollary 2. *Suppose that* $\rho(\Sigma)$ *remains bounded,* $\min_{v \in \mathbb{B}_0(2s)} \frac{\|\Sigma^{1/2} v\|_2}{\|v\|_2} > 0$ *and that* $n > c_3 s \log d$ *for a sufficiently large constant. Then a randomly drawn design matrix* $X \in \mathbb{R}^{n \times d}$ *with i.i.d.* $N(0, \Sigma)$ *rows satisfies Assumption 3 with probability greater than* $1 - c_1 \exp(-c_2 n)$.

Proof. Note that for any vector $\theta \in \Gamma(s, c_0)$, we have

$$\|\theta\|_1 \leq (1 + c_0) \sum_{j=1}^{s} |\theta_{(j)}| \quad \leq \quad (1 + c_0) \sqrt{s} \|\theta\|_2.$$

Consequently, if the bound (24) holds, we have

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \left\{ \frac{\|\Sigma^{1/2} v\|_2}{2\|v\|_2} - 6(1 + c_0) \Big(\frac{\rho(\Sigma) s \log d}{n} \Big)^{1/2} \right\} \|v\|_2.$$

Since we have assumed that $n > c_3 s \log d$ for a sufficiently large constant, the claim follows. □

Combined with the discussion following Proposition 1, this result shows that both the conditions required by Theorem 2 of this paper and the analysis of Bickel et al. [4] (both in the case $q = 0$) hold with high probability for Gaussian random designs.

3.3.1 Comparison of RE assumption with Assumption 2

In the case $q = 0$, the condition required by the estimator that performs least-squares over the ℓ_0-ball—namely, the form of Assumption 2 used in Theorem 2(b)—is not stronger than Assumption 3. This fact was previously established by Bickel et al. (see p.7, [4]). We now provide a simple pedagogical example to show that the ℓ_1-based relaxation can fail to recover the true parameter while the optimal ℓ_0-based algorithm succeeds. In particular, let us assume that the noise vector $w = 0$, and consider the design matrix

$$X = \begin{bmatrix} 1 & -2 & -1 \\ 2 & -3 & -3 \end{bmatrix},$$

corresponding to a regression problem with $n = 2$ and $d = 3$. Say that the regression vector $\beta^* \in \mathbb{R}^3$ is hard sparse with one non-zero entry (i.e., $s = 1$). Observe that the vector $\Delta := \begin{bmatrix} 1 & 1/3 & 1/3 \end{bmatrix}$ belongs to the null-space of X, and moreover $\Delta \in \Gamma(1,1)$ but $\Delta \notin \mathbb{B}_0(2)$. All the 2×2 sub-matrices of X have rank two, we have $\mathbb{B}_0(2) \cap \ker(X) = \{0\}$, so that by known results from Cohen et. al. [10] (see, in particular, their Lemma 3.1), the condition $\mathbb{B}_0(2) \cap \ker(X) = \{0\}$ implies that the ℓ_0-based algorithm can exactly recover any 1-sparse vector. On the other hand, suppose that, for instance, the true regression vector is given by $\beta^* = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$, If applied to this problem with no noise, the Lasso would incorrectly recover the solution $\widehat{\beta} := \begin{bmatrix} 0 & -1/3 & -1/3 \end{bmatrix}$ since $\|\widehat{\beta}\|_1 = 2/3 \leq 1 = \|\beta^*\|_1$. Although this example is low-dimensional ($(s, d) = (1, 3)$), we suspect that higher dimensional examples of design matrices that satisfy the conditions required for the minimax rate but not satisfied for ℓ_1-based methods may be constructed using similar arguments. This construction highlights that there are instances of design matrices X for which ℓ_1-based methods fail to recover the true parameter β^* for $q = 0$ while the optimal ℓ_0-based algorithm succeeds.

In summary, for the hard sparsity case $q = 0$, methods based on ℓ_1-relaxation can achieve the minimax rate $\mathcal{O}\big(\frac{s \log d}{n}\big)$ for ℓ_2-error, but the current analyses of these ℓ_1-methods [6, 26, 4] are based on imposing stronger conditions on the design matrix X than those required by the "optimal" estimator that performs least-squares over the ℓ_0-ball.

4 Proofs of main results

In this section, we provide the proofs of our main theorems, with more technical lemmas and their proofs deferred to the appendices. To begin, we provide a high-level overview that outlines the main steps of the proofs.

Basic steps for lower bounds The proofs for the lower bounds follow an information-theoretic method based on Fano's inequality [11], as used in classical work on nonparametric estimation [19, 34, 35]. A key ingredient is a fine characterization of the metric entropy structure of ℓ_q balls [20, 8]. At a high-level, the proof of each lower bound follows the following three basic steps:

(1) Let $\|\cdot\|_*$ be the norm for which we wish to lower bound the minimax risk; for Theorem 1, the norm $\|\cdot\|_*$ corresponds to the ℓ_p norm, whereas for Theorem 3, it is the ℓ_2-prediction norm (the square root of the prediction loss). We first construct an δ_n-packing set for $\mathbb{B}_q(R_q)$ in the norm $\|\cdot\|_*$, where $\delta_n > 0$ is a free parameter to be determined in a later step. The packing set is constructed by deriving lower bounds on the packing numbers for $\mathbb{B}_q(R_q)$; we discuss the concepts of packing sets and packing numbers at more length in Section 4.1. For the case of ℓ_q-balls for $q > 0$, tight bounds on the packing numbers in ℓ_p norm have been developed in the approximation theory literature [20]. For $q = 0$, we use combinatorial to bound the packing numbers. We use Assumption 2 in order to relate the packing number in the ℓ_2-prediction norm to the packing number in ℓ_2-norm.

(2) The next step is to use a standard reduction to show that any estimator with minimax risk $\mathcal{O}(\delta_n^2)$ must be able to solve a hypothesis-testing problem over the packing set with vanishing error probability. More concretely, suppose that an adversary places a uniform distribution over the δ_n-packing set in $\mathbb{B}_q(R_q)$, and let this random variable be Θ. The problem of recovering Θ is a multi-way hypothesis testing problem, so that we may apply Fano's inequality to lower bound the probability of error. The Fano bound involves the log packing number and the mutual information $I(Y; \Theta)$ between the observation vector $y \in \mathbb{R}^n$ and the random parameter Θ chosen uniformly from the packing set.

(3) Finally, following a technique introduced by Yang and Barron [34], we derive an upper bound on the mutual information between Y and Θ by constructing an ϵ_n-covering set for $\mathbb{B}_q(R_q)$ with respect to the ℓ_2-prediction semi-norm. Using Lemma 4 in Section 4.1.2, we establish a link between covering numbers in ℓ_2-prediction semi-norm to covering numbers in ℓ_2-norm. Finally, we choose the free parameters $\delta_n > 0$ and and $\epsilon_n > 0$ so as to optimize the lower bound.

Basic steps for upper bounds The proofs for the upper bounds involve direct analysis of the natural estimator that performs least-squares over the ℓ_q-ball. The proof is constructive and involves two steps, the first of which is standard while the second step is more specific to the problem at hand:

(1) Since the estimator is based on minimizing the least-squares loss over the ball $\mathbb{B}_q(R_q)$, some straightforward algebra allows us to upper bound the ℓ_2-prediction error by a term that measures the supremum of a Gaussian empirical process over the ball $\mathbb{B}_q(2R_q)$. This step is completely generic and applies to any least-squares estimator involving a linear model.

(2) The second and more challenging step involves computing upper bounds on the supremum of the Gaussian process over $\mathbb{B}_q(2R_q)$. For each of the upper bounds, our approach is slightly different in the details. Common steps include upper bounds on the covering numbers of the ball $\mathbb{B}_q(2R_q)$, as well as on the image of these balls under the mapping $X : \mathbb{R}^d \to \mathbb{R}^n$. For the case $q = 1$, we make use of Lemma 2 in order to relate the ℓ_1-norm to the ℓ_2-norm for vectors that lie in an ℓ_q-ball. For $q \in (0, 1)$, we make use of some chaining and peeling results from empirical process theory (e.g., Van de Geer [32]).

4.1 Packing, covering, and metric entropy

The notion of packing and covering numbers play a crucial role in our analysis, so we begin with some background, with emphasis on the case of covering/packing for ℓ_q-balls.

Definition 1 (Covering and packing numbers)**.** Consider a metric space consisting of a set $\mathcal{S}$ and a metric $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$.

(a) An ϵ-covering of $\mathcal{S}$ in the metric ρ is a collection $\{\beta^1, \ldots, \beta^N\} \subset \mathcal{S}$ such that for all $\beta \in S$, there exists some $i \in \{1, \ldots, N\}$ with $\rho(\beta, \beta^i) \leq \epsilon$. The ϵ-covering number $N(\epsilon; \mathcal{S}, \rho)$ is the cardinality of the smallest ϵ-covering.

(b) A δ-packing of $\mathcal{S}$ in the metric ρ is a collection $\{\beta^1, \ldots, \beta^M\} \subset S$ such that $\rho(\beta^i, \beta^j) \geq \delta$ for all $i \neq j$. The δ-packing number $M(\delta; \mathcal{S}, \rho)$ is the cardinality of the largest δ-packing.

In simple terms, the covering number $N(\epsilon; \mathcal{S}, \rho)$ is the minimum number of balls with radius ϵ under the metric ρ required to completely cover the space, so that every point in $\mathcal{S}$ lies in some ball. The packing number $M(\delta; \mathcal{S}, \rho)$ is the maximum number of balls of radius δ under metric ρ that can be packed into the space so that there is no overlap between any of the balls. It is worth noting that the covering and packing numbers are (up to constant factors) essentially the same. In particular, the inequalities $M(\epsilon; \mathcal{S}, \rho) \leq N(\epsilon; \mathcal{S}, \rho) \leq M(\epsilon/2; \mathcal{S}, \rho)$ are standard (e.g., [29]). Consequently, given upper and lower bounds on the covering number, we can immediately infer similar upper and lower bounds on the packing number. Of interest in our results is the logarithm of the covering number $\log N(\epsilon; \mathcal{S}, \rho)$, a quantity known as the *metric entropy*.

A related quantity, frequently used in the operator theory literature [20, 30, 8], are the (dyadic) entropy numbers $\epsilon_k(\mathcal{S}; \rho)$, defined as follows for $k = 1, 2, \ldots$

$$\epsilon_k(\mathcal{S}; \rho) \quad = \quad \inf \left\{ \epsilon > 0 \mid N(\epsilon; \mathcal{S}, \rho) \leq 2^{k-1} \right\}. \tag{27}$$

By definition, note that we have $\epsilon_k(\mathcal{S}; \rho) \leq \delta$ if and only if $\log N(\delta; \mathcal{S}, \rho) \leq k$.

4.1.1 Metric entropies of ℓ_q-balls

Central to our proofs is the metric entropy of the ball $\mathbb{B}_q(R_q)$ when the metric ρ is the ℓ_p-norm, a quantity which we denote by $\log N_{p,q}(\epsilon)$. The following result, which provides upper and lower bounds on this metric entropy that are tight up to constant factors, is an adaptation of results from the operator theory literature [20, 17]; see Appendix B for the details. All bounds stated here apply to a dimension $d \geq 2$.

Lemma 3. *Assume that $q \in (0, 1]$ and $p \in [1, \infty]$ with $p > q$. Then there is a constant $U_{q,p}$, depending only on q and p, such that*

$$\log N_{p,q}(\epsilon) \;\leq\; U_{q,p}\left[R_q{}^{\frac{p}{p-q}}\left(\frac{1}{\epsilon}\right)^{\frac{pq}{p-q}}\log d\right] \quad \text{for all } \epsilon \in (0, R_q{}^{1/q}). \tag{28}$$

Conversely, suppose in addition that $\epsilon < 1$ and $\epsilon^p = \Omega\big(\frac{\log d}{d^\nu}\big)^{\frac{p-q}{q}}$ for some fixed $\nu \in (0, 1)$, depending only on q and p. Then there is a constant $L_{q,p} \leq U_{q,p}$, depending only on q and p, such that

$$\log N_{p,q}(\epsilon) \;\geq\; L_{q,p}\left[R_q{}^{\frac{p}{p-q}}\left(\frac{1}{\epsilon}\right)^{\frac{pq}{p-q}}\log d\right]. \tag{29}$$

Remark: In our application of the lower bound (29), our typical choice of ϵ^p will be of the order $\mathcal{O}\big(\frac{\log d}{n}\big)^{\frac{p-q}{2}}$. It can be verified that as long as there exists a $\kappa \in (0, 1)$ such that $\frac{d}{R_q n^{q/2}} = \Omega(d^\kappa)$ (which is stated at the beginning of Section 2) and $p > q$, then there exists some fixed $\nu \in (0, 1)$, depending only on p and q, such that ϵ lies in the range required for the lower bound (29) to be valid.

4.1.2 Metric entropy of q-convex hulls

The proofs of the lower bounds all involve the Kullback-Leibler (KL) divergence between the distributions induced by different parameters β and β' in $\mathbb{B}_q(R_q)$. Here we show that for the linear observation model (1), these KL divergences can be represented as q-convex hulls of the columns of the design matrix, and provide some bounds on the associated metric entropy.

For two distributions $\mathbb{P}$ and $\mathbb{Q}$ that have densities $d\mathbb{P}$ and $d\mathbb{Q}$ with respect to some base measure μ, the Kullback-Leibler (KL) divergence is given by $D(\mathbb{P} \,\|\, \mathbb{Q}) = \int \log \frac{d\mathbb{P}}{d\mathbb{Q}}\, \mathbb{P}(d\mu)$. We use $\mathbb{P}_\beta$ to denote the distribution of $y \in \mathbb{R}$ under the linear regression model—in particular, it corresponds to the distribution of a $N(X\beta, \sigma^2 I_{n\times n})$ random vector. A straightforward computation then leads to

$$D(\mathbb{P}_\beta \,\|\, \mathbb{P}_{\beta'}) \;=\; \frac{1}{2\sigma^2}\|X\beta - X\beta'\|_2^2. \tag{30}$$

Therefore, control of KL-divergences requires understanding of the metric entropy of the q-convex hull of the rescaled columns of the design matrix X—in particular, the set

$$\operatorname{absconv}_q(X/\sqrt{n}) \;:=\; \Big\{\frac{1}{\sqrt{n}}\sum_{j=1}^d \theta_j X_j \;\mid\; \theta \in \mathbb{B}_q(1)\Big\}. \tag{31}$$

We have introduced the normalization by $1/\sqrt{n}$ for later technical convenience.

Under the column normalization condition, it turns out that the metric entropy of this set with respect to the ℓ_2-norm is essentially no larger than the metric entropy of $\mathbb{B}_q(R_q)$, as summarized in the following

Lemma 4. *Suppose that X satisfies the column normalization condition (Assumption 1 with constant κ_c). Then there is a constant $U'_{q,2}$ depending only on $q \in (0,1]$ such that*

$$\log N(\epsilon, \mathrm{absconv}_q(X/\sqrt{n}), \|\cdot\|_2) \quad \leq \quad U'_{q,2}\left[R_q^{\frac{2}{2-q}}\left(\frac{\kappa_c}{\epsilon}\right)^{\frac{2q}{2-q}} \log d\right].$$

The proof of this claim is provided in Appendix C. Note that apart from a different constant, this upper bound on the metric entropy is identical to that for $\log N_{2,q}(\epsilon/\kappa_c)$ from Lemma 3. Up to constant factors, this upper bound cannot be tightened in general (e.g., consider $n = d$ and $X = I$).

4.2 Proof of lower bounds

We begin by proving our main results that provide lower bounds on minimax risks, namely Theorems 1 and 3.

4.2.1 Proof of Theorem 1

Recall that the lower bounds in Theorem 1 are the maximum of two expressions, one corresponding to the diameter of the set $\mathcal{N}_q(X)$ intersected with the ℓ_q-ball, and the other correspond to the metric entropy of the ℓ_q-ball.

We begin by deriving the lower bound based on the diameter of $\mathcal{N}_q(X) = \mathbb{B}_q(R_q) \cap \ker(X)$. The minimax risk is lower bounded as

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \mathbb{E}\|\widehat{\beta} - \beta\|_p^p \quad \geq \quad \min_{\widehat{\beta}} \max_{\beta \in \mathcal{N}_q(X)} \mathbb{E}\|\widehat{\beta} - \beta\|_p^p,$$

where the inequality follows from the inclusion $\mathcal{N}_q(X) \subseteq \mathbb{B}_q(R_q)$. For any $\beta \in \mathcal{N}_q(X)$, we have $Y = X\beta + w = w$, so that Y contains no information about $\beta \in \mathcal{N}_q(X)$. Consequently, once $\widehat{\beta}$ is chosen, the adversary can always choose an element $\beta \in \mathcal{N}_q(X)$ such that $\|\widehat{\beta} - \beta\|_p \geq \frac{1}{2}\mathrm{diam}_p(\mathcal{N}_q(X))$. Indeed, if $\|\widehat{\beta}\|_p \geq \frac{1}{2}\mathrm{diam}_p(\mathcal{N}_q(X))$, then the adversary chooses $\beta = 0 \in \mathcal{N}_q(X)$. On the other hand, if $\|\widehat{\beta}\|_p \leq \frac{1}{2}\mathrm{diam}_p(\mathcal{N}_q(X))$, then the adversary can choose some $\beta \in \mathcal{N}_q(X)$ such that $\|\beta\|_p = \mathrm{diam}_p(\mathcal{N}_q(X))$. By triangle inequality, we then have $\|\beta - \widehat{\beta}\|_p \geq \|\beta\|_p - \|\widehat{\beta}\|_p \geq \frac{1}{2}\mathrm{diam}_p(\mathcal{N}_q(X))$. Overall, we conclude that

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \mathbb{E}\|\widehat{\beta} - \beta\|_p^p \quad \geq \quad \left(\frac{1}{2}\mathrm{diam}_p(\mathcal{N}_q(X))\right)^p.$$

In the following subsections, we establish the second terms in the lower bounds via the Fano method, a standard approach for minimax lower bounds. Our proofs of part (a) and (b) are based on slightly different arguments.

Proof of Theorem 1(a): Let $M = M_p(\delta_n)$ be the cardinality of a maximal packing of the ball $\mathbb{B}_q(R_q)$ in the ℓ_p metric, say with elements $\{\beta^1, \ldots, \beta^M\}$. A standard argument (e.g., [18, 34, 35]) yields a lower bound on the minimax ℓ_p-risk in terms of the error in a multi-way hypothesis testing problem: in particular, we have

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \mathbb{E}\|\widehat{\beta} - \beta\|_p^p \geq \frac{1}{2^p}\,\delta_n^p \min_{\widetilde{\beta}} \mathbb{P}[\widetilde{\beta} \neq B]$$

where the random vector $B \in \mathbb{R}^d$ is uniformly distributed over the packing set $\{\beta^1, \ldots, \beta^M\}$, and the estimator $\widetilde{\beta}$ takes values in the packing set. Applying Fano's inequality [11] yields the lower bound

$$\mathbb{P}[B \neq \widetilde{\beta}] \geq 1 - \frac{I(B;Y) + \log 2}{\log M_p(\delta_n)}, \tag{32}$$

where $I(B;Y)$ is the mutual information between random parameter B in the packing set and the observation vector $Y \in \mathbb{R}^n$.

It remains to upper bound the mutual information; we do so by following the procedure of Yang and Barron [34], which is based on covering the model space $\{\mathbb{P}_\beta, \ \beta \in \mathbb{B}_q(R_q)\}$ under the square-root Kullback-Leibler divergence. As noted prior to Lemma 4, for the Gaussian models given here, this square-root KL divergence takes the form $\sqrt{D(\mathbb{P}_\beta \,\|\, \mathbb{P}_{\beta'})} = \frac{1}{\sqrt{2\sigma^2}}\|X(\beta - \beta')\|_2$. Let $N = N_2(\epsilon_n)$ be the minimal cardinality of an ϵ_n-covering of $\mathbb{B}_q(R_q)$ in ℓ_2-norm. Using the upper bound on the dyadic entropy of $\operatorname{absconv}_q(X)$ provided by Lemma 4, we conclude that there exists a set $\{X\beta^1, \ldots, X\beta^N\}$ such that for all $X\beta \in \operatorname{absconv}_q(X)$, there exists some index i such that $\|X(\beta - \beta^i)\|_2/\sqrt{n} \leq c\,\kappa_c\,\epsilon_n$. Following the argument of Yang and Barron [34], we obtain that the mutual information is upper bounded as

$$I(B;Y) \quad \leq \quad \log N(\epsilon_n) + \frac{c^2\, n}{\sigma^2}\kappa_c^2 \epsilon_n^2.$$

Combining this upper bound with the Fano lower bound (32) yields

$$\mathbb{P}[B \neq \widetilde{\beta}] \quad \geq \quad 1 - \frac{\log N_2(\epsilon_n) + \frac{c^2\, n}{\sigma^2}\kappa_c^2\, \epsilon_n^2 + \log 2}{\log M_p(\delta_n)}. \tag{33}$$

The final step is to choose the packing and covering radii (δ_n and ϵ_n respectively) such that the lower bound (33) remains strictly above zero, say bounded below by $1/4$. In order to do so, suppose that we choose the pair (ϵ_n, δ_n) such that

$$\frac{c^2\, n}{\sigma^2}\kappa_c^2\, \epsilon_n^2 \quad \leq \quad \log N_2(\epsilon_n), \quad \text{and} \tag{34a}$$

$$\log M_p(\delta_n) \quad \leq \quad 4 \log N_2(\epsilon_n). \tag{34b}$$

As long as $N_2(\epsilon_n) \geq 2$, we are then guaranteed that

$$\mathbb{P}[B \neq \widetilde{\beta}] \quad \geq \quad 1 - \frac{2\log N_2(\epsilon_n) + \log 2}{4 \log N_2(\epsilon_n)} \quad \geq \quad 1/4, \tag{35}$$

as desired.

It remains to determine choices of ϵ_n and δ_n that satisfy the relations (34). From Lemma 3, relation (34a) is satisfied by choosing ϵ_n such that $\frac{c^2\, n}{\sigma^2}\kappa_c^2\, \epsilon_n^2 \;=\; L_{q,2}\left[R_q{}^{\frac{2}{2-q}}\left(\frac{1}{\epsilon_n}\right)^{\frac{2q}{2-q}} \log d\right]$, or equivalently such that

$$\left(\epsilon_n\right)^{\frac{4}{2-q}} \quad = \quad \Theta\Big(R_q{}^{\frac{2}{2-q}}\, \frac{\sigma^2}{\kappa_c^2}\, \frac{\log d}{n}\Big).$$

In order to satisfy the bound (34b), it suffices to choose δ_n such that

$$U_{q,p}\left[R_q{}^{\frac{p}{p-q}}\left(\frac{1}{\delta_n}\right)^{\frac{pq}{p-q}} \log d\right] \quad \leq \quad 4L_{q,2}\left[R_q{}^{\frac{2}{2-q}}\left(\frac{1}{\epsilon_n}\right)^{\frac{2q}{2-q}} \log d\right],$$

or equivalently such that

$$\begin{aligned}\delta_n^p &\geq \left[\frac{U_{q,p}}{4L_{q,2}}\right]^{\frac{p-q}{q}} \left\{(\epsilon_n)^{\frac{4}{2-q}}\right\}^{\frac{p-q}{2}} R_q^{\frac{2-p}{2-q}} \\ &= \left[\frac{U_{q,p}}{4L_{q,2}}\right]^{\frac{p-q}{q}} L_{q,2}^{\frac{p-q}{2}} R_q \left[\frac{\sigma^2}{\kappa_c^2}\frac{\log d}{n}\right]^{\frac{p-q}{2}}\end{aligned}$$

Combining this bound with the lower bound (35) on the hypothesis testing error probability and substituting into equation (10), we obtain

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \mathbb{E}\|\widehat{\beta} - \beta\|_p^p \geq c_{q,p} R_q \left[\frac{\sigma^2}{\kappa_c^2}\frac{\log d}{n}\right]^{\frac{p-q}{2}},$$

which completes the proof of Theorem 1(a).

Proof of Theorem 1(b): In order to prove Theorem 1(b), we require some definitions and an auxiliary lemma. For any integer $s \in \{1, \ldots, d\}$, we define the set

$$\mathcal{H}(s) := \{z \in \{-1, 0, +1\}^d \mid \|z\|_0 = s\}.$$

Although the set $\mathcal{H}$ depends on s, we frequently drop this dependence so as to simplify notation. We define the Hamming distance $\rho_H(z, z') = \sum_{j=1}^d \mathbb{I}[z_j \neq z'_j]$ between the vectors z and z'. We prove the following result in Appendix D:

Lemma 5. *There exists a subset $\widetilde{\mathcal{H}} \subset \mathcal{H}$ with cardinality $|\widetilde{\mathcal{H}}| \geq \exp(\frac{s}{2}\log\frac{d-s}{s/2})$ such that $\rho_H(z, z') \geq \frac{s}{2}$ for all $z, z' \in \widetilde{\mathcal{H}}$.*

Now consider a rescaled version of the set $\widetilde{\mathcal{H}}$, say $\sqrt{\frac{2}{s}}\delta_n\widetilde{\mathcal{H}}$ for some $\delta_n > 0$ to be chosen. For any elements $\beta, \beta' \in \frac{\delta_n}{\sqrt{s}}\widetilde{\mathcal{H}}$, we have the following bounds on the ℓ_2-norm of their difference:

$$\|\beta - \beta'\|_2^2 \geq \delta_n^2, \quad \text{and} \tag{36a}$$
$$\|\beta - \beta'\|_2^2 \leq 8\delta_n^2. \tag{36b}$$

Consequently, the rescaled set $\sqrt{\frac{2}{s}}\delta_n\widetilde{\mathcal{H}}$ is an δ_n-packing set in ℓ_2 norm with $M_2(\delta_n) = |\widetilde{\mathcal{H}}|$ elements, say $\{\beta^1, \ldots, \beta^M\}$. Using this packing set, we now follow the same classical steps as in the proof of Theorem 1(a), up until the Fano lower bound (32).

At this point, we use an alternative upper bound on the mutual information, namely the bound $I(Y; B) \leq \frac{1}{\binom{M}{2}}\sum_{i \neq j} D(\beta^i \,\|\, \beta^j)$, which follows from the convexity of mutual information [11]. For the linear observation model (1), we have $D(\beta^i \,\|\, \beta^j) = \frac{1}{2\sigma^2}\|X(\beta^i - \beta^j)\|_2^2$. Since $(\beta - \beta') \in \mathbb{B}_0(2s)$ by construction, from the assumptions on X and the upper bound bound (36b), we conclude that

$$I(Y; B) \leq \frac{8n\kappa_u^2\delta_n^2}{2\sigma^2}.$$

Substituting this upper bound into the Fano lower bound (32), we obtain

$$\mathbb{P}[B \neq \widetilde{\beta}] \geq 1 - \frac{\frac{8n\kappa_u^2}{2\sigma^2}\delta_n^2 + \log(2)}{\frac{s}{2}\log\frac{d-s}{s/2}}.$$

Setting $\delta_n^2 = \frac{1}{32}\frac{\sigma^2}{\kappa_u^2}\frac{s}{2n}\log\frac{d-s}{s/2}$ ensures that this probability is at least $1/4$. Consequently, combined with the lower bound (10), we conclude that

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \mathbb{E}\|\widehat{\beta} - \beta\|_p^p \quad \geq \quad \frac{1}{2^p}\frac{1}{4}(\frac{1}{32})^{p/2}\left[\frac{\sigma^2}{\kappa_u^2}\frac{s}{2n}\log\frac{d-s}{s/2}\right]^{\frac{p}{2}}.$$

As long as the ratio $d/s \geq 1+\delta$ for some $\delta > 0$ we have $\log(d/s - 1) \geq c\log(d/s)$ for some constant $c > 0$, from which the result follows.

4.2.2 Proof of Theorem 3

We use arguments similar to the proof of Theorem 1 in order to establish lower bounds on prediction error $\|X(\widehat{\beta} - \beta^*)\|_2/\sqrt{n}$.

Proof of Theorem 3(a): For some $\delta_n^2 = \Omega(R_q(\frac{\log d}{n})^{1-q/2})$, let $\{\beta^1, \ldots, \beta^M\}$ be an δ_n packing of the ball $\mathbb{B}_q(R_q)$ in the ℓ_2 metric, say with a total of $M = M(\delta_n/\kappa_c)$ elements. We first show that if n is sufficiently large, then this set is also a $\kappa_\ell\delta_n/2$-packing set in the prediction (semi)-norm. From Assumption 2, for each $i \neq j$,

$$\frac{\|X(\beta^i - \beta^j)\|_2}{\sqrt{n}} \quad \geq \quad \kappa_\ell \|\beta^i - \beta^j\|_2 - f_\ell(R_q, n, d). \tag{37}$$

Using the assumed lower bound on δ_n^2—namely, $\delta_n^2 = \Omega\big(R_q(\frac{\log d}{n})^{1-\frac{q}{2}}\big)$—and the initial lower bound (37), we conclude that $\frac{\|X(\beta^i-\beta^j)\|_2}{\sqrt{n}} \geq \kappa_\ell\delta_n/2$ once n is larger than some finite number.

We have thus constructed a $\kappa_\ell\delta_n/2$-packing set in the (semi)-norm $\|X(\beta^i - \beta^j)\|_2$. As in the proof of Theorem 2(a), we follow a standard approach to reduce the problem of lower bounding the minimax error to the error probability of a multi-way hypothesis testing problem. After this step, we apply the Fano inequality to lower bound this error probability via

$$\mathbb{P}[XB \neq X\widetilde{\beta}] \geq 1 - \frac{I(XB^i; Y) + \log 2}{\log M_2(\delta_n)},$$

where $I(XB^i; Y)$ now represents the mutual information[3] between random parameter XB (uniformly distributed over the packing set) and the observation vector $Y \in \mathbb{R}^n$.

From Lemma 4, the $\kappa_c\,\epsilon$-covering number of the set $\mathrm{absconv}_q(X)$ is upper bounded (up to a constant factor) by the ϵ covering number of $\mathbb{B}_q(R_q)$ in ℓ_2-norm, which we denote by $N_2(\epsilon_n)$. Following the same reasoning as in Theorem 2(a), the mutual information is upper bounded as

$$I(XB; Y) \quad \leq \quad \log N_2(\epsilon_n) + \frac{n}{2\sigma^2}\kappa_c^2\,\epsilon_n^2.$$

Combined with the Fano lower bound, we obtain

$$\mathbb{P}[XB \neq X\widetilde{\beta}] \quad \geq \quad 1 - \frac{\log N_2(\epsilon_n) + \frac{n}{\sigma^2}\kappa_c^2\,\epsilon_n^2 + \log 2}{\log M_p(\delta_n)}. \tag{38}$$

Lastly, we choose the packing and covering radii (δ_n and ϵ_n respectively) such that the lower bound (38) remains strictly above zero, say bounded below by $1/4$. It suffices to choose the pair (ϵ_n, δ_n) to satisfy the relations (34a) and (34b). As long as $\epsilon_n^2 > \frac{\log d}{n}$ and $N_2(\epsilon_n) \geq 2$, we are then guaranteed that

$$\mathbb{P}[XB \neq X\widetilde{\beta}] \quad \geq \quad 1 - \frac{2\log N_2(\epsilon_n) + \log 2}{4\log N_2(\epsilon_n)} \quad \geq \quad 1/4,$$

[3]Despite the difference in notation, this mutual information is the same as $I(B; Y)$, since it measures the information between the observation vector y and the discrete index i.

as desired. Recalling that we have constructed a $\delta_n \kappa_\ell/2$ covering in the prediction (semi)-norm, we obtain

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \mathbb{E}\|X(\widehat{\beta} - \beta)\|_2^2/n \;\; \geq \;\; c'_{2,q} \, R_q \, \kappa_\ell^2 \left[\frac{\sigma^2}{\kappa_c^2} \frac{\log d}{n}\right]^{1-q/2},$$

for some constant $c'_{2,q} > 0$. This completes the proof of Theorem 3(a).

Proof of Theorem 3(b): Recall the assertion of Lemma 5, which guarantees the existence of a set $\frac{\delta_n^2}{2s}\widetilde{\mathcal{H}}$ is an δ_n-packing set in ℓ_2-norm with $M_p(\delta_n) = |\widetilde{\mathcal{H}}|$ elements, say $\{\beta^1, \ldots, \beta^M\}$, such that the bounds (36a) and (36b) hold, and such that $\log |\widetilde{\mathcal{H}}| \geq \frac{s}{2} \log \frac{d-s}{s/2}$. By construction, the difference vectors $(\beta^i - \beta^j) \in \mathbb{B}_0(2s)$, so that by assumption, we have

$$\|X(\beta^i - \beta^j)\|/\sqrt{n} \;\leq\; \kappa_u \|\beta^i - \beta^j\|_2 \;\leq\; \kappa_u \sqrt{8}\, \delta_n. \tag{39}$$

In the reverse direction, since Assumption 2 holds with $f_\ell(R_q, n, d) = 0$, we have

$$\|X(\beta^i - \beta^j)\|_2/\sqrt{n} \;\; \geq \;\; \kappa_\ell \delta_n. \tag{40}$$

We can follow the same steps as in the proof of Theorem 1(b), thereby obtaining an upper bound the mutual information of the form $I(XB; y) \leq 8\kappa_u^2 n \delta_n^2$. Combined with the Fano lower bound, we have

$$\mathbb{P}[XB \neq X\widetilde{\beta}] \;\; \geq \;\; 1 - \frac{\frac{8\, n \kappa_u^2}{2\sigma^2}\delta_n^2 + \log(2)}{\frac{s}{2n} \log \frac{d-s}{s/2}}.$$

Remembering the extra factor of κ_ℓ from the lower bound (40), we obtain the lower bound

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_0(s)} \mathbb{E}\frac{1}{n}\|X(\widehat{\beta} - \beta)\|_2^2 \;\; \geq \;\; c'_{0,q} \, \kappa_\ell^2 \, \frac{\sigma^2}{\kappa_u^2} \, s \log \frac{d-s}{s/2}.$$

Repeating the argument from the proof of Theorem 1(b) allows us to further lower bound this quantity in terms of $\log(d/s)$, leading to the claimed form of the bound.

4.3 Proof of achievability results

We now turn to the proofs of our main achievability results, namely Theorems 2 and 4, that provide upper bounds on minimax risks. We prove all parts of these theorems by analyzing the family of M-estimators

$$\widehat{\beta} \;\; \in \;\; \arg \min_{\|\beta\|_q^q \leq R_q} \|Y - X\beta\|_2^2.$$

We begin by deriving an elementary inequality that is useful throughout the analysis. Since the vector β^* satisfies the constraint $\|\beta^*\|_q^q \leq R_q$ meaning β^* is a feasible point, we have $\|Y - X\beta\|_2^2 \leq \|Y - X\beta^*\|_2^2$. Defining $\widehat{\Delta} = \widehat{\beta} - \beta^*$ and performing some algebra, we obtain the inequality

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \;\; \leq \;\; \frac{2|w^T X \widehat{\Delta}|}{n}. \tag{41}$$

4.3.1 Proof of Theorem 2

We begin with the proof of Theorem 2, in which we upper bound the minimax risk in squared ℓ_2-norm.

Proof of Theorem 2(a): To begin, we may apply Assumption 2 to the inequality (41) to obtain

$$\begin{aligned}\big[\max(0,\ \kappa_\ell\|\widehat{\Delta}\|_2 - f_\ell(R_q,n,d))\big]^2 &\leq 2|w^T X\widehat{\Delta}|/n \\ &\leq \frac{2}{n}\|w^T X\|_\infty\|\widehat{\Delta}\|_1.\end{aligned}$$

Since $w_i \sim N(0,\sigma^2)$ and the columns of X are normalized, each entry of $\frac{2}{n}w^T X$ is zero-mean Gaussian with variance at most $4\sigma^2\kappa_c^2/n$. Therefore, by union bound and standard Gaussian tail bounds, we obtain that the inequality

$$\big[\max(0,\ \kappa_\ell\|\widehat{\Delta}\|_2 - f_\ell(R_q,n,d))\big]^2 \leq 2\sigma\kappa_c\sqrt{\frac{3\log d}{n}}\,\|\widehat{\Delta}\|_1 \tag{42}$$

holds with probability greater than $1 - c_1\exp(-c_2 n)$. Consequently, we may conclude that at least one of the two following alternatives must hold

$$\|\widehat{\Delta}\|_2 \leq \frac{2f_\ell(R_q,n,d)}{\kappa_\ell}, \quad \text{or} \tag{43a}$$

$$\|\widehat{\Delta}\|_2^2 \leq \frac{2\sigma\kappa_c}{\kappa_\ell^2}\sqrt{\frac{3\log d}{n}}\|\widehat{\Delta}\|_1. \tag{43b}$$

Suppose first that alternative (43a) holds. Consequently for we have

$$\|\widehat{\Delta}\|_2^2 \leq o\left(R_q\big(\frac{\log d}{n}\big)^{1-q/2}\right),$$

which is the same up to constant rate than claimed in Theorem 2(a).

On the other hand, suppose that alternative (43b) holds. Since both $\widehat{\beta}$ and β^* belong to $\mathbb{B}_q(R_q)$, we have $\|\widehat{\Delta}\|_q^q = \sum_{j=1}^d |\widehat{\Delta}_j|^q \leq 2R_q$. Therefore we can exploit Lemma 2 by setting $\tau = \frac{2\sigma\kappa_c}{\kappa_\ell^2}\sqrt{\frac{3\log d}{n}}$, thereby obtaining the bound $\|\widehat{\Delta}\|_2^2 \leq \tau\|\widehat{\Delta}\|_1$, and hence

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

Viewed as a quadratic in the indeterminate $x = \|\widehat{\Delta}\|_2$, this inequality is equivalent to the constraint $f(x) = ax^2 + bx + c \leq 0$, with $a = 1$,

$$b = -\sqrt{2R_q}\tau^{1-q/2}, \quad \text{and} \quad c = -2R_q\tau^{2-q}.$$

Since $f(0) = c < 0$ and the positive root of $f(x)$ occurs at $x^* = (-b + \sqrt{b^2 - 4ac})/(2a)$, some algebra shows that we must have

$$\|\widehat{\Delta}\|_2^2 \ \leq\ 4\max\{b^2,\ |c|\} \ \leq\ 24R_q\Big[\frac{\kappa_c^2}{\kappa_\ell^2}\frac{\sigma^2}{\kappa_\ell^2}\ \frac{\log d}{n}\Big]^{1-q/2},$$

with high probability (stated in Theorem 2(a) which completes the proof of Theorem 2(a).

Proof of Theorem 2(b): In order to establish the bound (13), we follow the same steps with $f_\ell(s,n,d) = 0$, thereby obtaining the following simplified form of the bound (42):

$$\|\widehat{\Delta}\|_2^2 \leq \frac{\kappa_c}{\kappa_\ell}\,\frac{\sigma}{\kappa_\ell}\sqrt{\frac{3\log d}{n}}\,\|\widehat{\Delta}\|_1.$$

By definition of the estimator, we have $\|\widehat{\Delta}\|_0 \leq 2s$, from which we obtain $\|\widehat{\Delta}\|_1 \leq \sqrt{2s}\|\widehat{\Delta}\|_2$. Canceling out a factor of $\|\widehat{\Delta}\|_2$ from both sides yields the claim (13).

Establishing the sharper upper bound (14) requires more precise control on the right-hand side of the inequality (41). The following lemma, proved in Appendix F, provides this control:

Lemma 6. *If* $\frac{\|X\theta\|_2}{\sqrt{n}\|\theta\|_2} \leq \kappa_u$ *for all* $\theta \in \mathbb{B}_0(2s)$, *then for any* $r > 0$, *we have*

$$\sup_{\|\theta\|_0 \leq 2s, \|\theta\|_2 \leq r} \frac{1}{n}\left|w^T X\theta\right| \quad \leq \quad 6\,\sigma\, r\, \kappa_u \sqrt{\frac{s\log(d/s)}{n}} \tag{44}$$

with probability greater than $1 - c_1\exp(-c_2\min\{n, s\log(d-s)\})$.

Let us apply this lemma to the basic inequality (41). We may upper bound the right-hand side as

$$\left|\frac{w^T X\Delta}{n}\right| \quad \leq \quad \|\Delta\|_2 \sup_{\|\theta\|_0 \leq 2s, \|\theta\|_2 \leq 1} \frac{1}{n}\left|w^T X\theta\right| \quad \leq \quad 6\,\|\Delta\|_2\,\sigma\,\kappa_u\sqrt{\frac{s\log(d/s)}{n}}.$$

Consequently, we have

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \quad \leq \quad 12\,\sigma\,\|\widehat{\Delta}\|_2\,\kappa_u\sqrt{\frac{s\log(d/s)}{n}},$$

with high probability. By Assumption 2, we have $\|X\widehat{\Delta}\|_2^2/n \geq \kappa_\ell^2\|\widehat{\Delta}\|_2^2$. Cancelling out a factor of $\|\widehat{\Delta}\|_2$ and re-arranging yields $\|\widehat{\Delta}\|_2 \leq 12\,\frac{\kappa_u\sigma}{\kappa_\ell^2}\sqrt{\frac{s\log(d/s)}{n}}$ with high probability as claimed.

4.3.2 Proof of Theorem 4

We again make use of the elementary inequality (41) to establish upper bounds on the prediction risk.

Proof of Theorem 4(a): So as to facilitate tracking of constants in this part of the proof, we consider the rescaled observation model, in which $\widetilde{w} \sim N(0, I_n)$ and $\widetilde{X} := \sigma^{-1}X$. Note that if X satisfies Assumption 1 with constant κ_c, then $\widetilde{X}$ satisfies it with constant $\widetilde{\kappa_c} = \kappa_c/\sigma$. Moreover, if we establish a bound on $\|\widetilde{X}(\widehat{\beta} - \beta^*)\|_2^2/n$, then multiplying by σ^2 recovers a bound on the original prediction loss.

We first deal with the case $q = 1$. In particular, we have

$$\left|\frac{1}{n}\widetilde{w}^T\widetilde{X}\theta\right| \quad \leq \quad \|\frac{\widetilde{w}^T\widetilde{X}}{n}\|_\infty\|\theta\|_1 \quad \leq \quad \sqrt{\frac{3\widetilde{\kappa_c}^2\sigma^2\log d}{n}}\,(2\,R_1),$$

where the second inequality holds with probability $1 - c_1\exp(-c_2\log d)$, using standard Gaussian tail bounds. (In particular, since $\|\widetilde{X}_i\|_2/\sqrt{n} \leq \widetilde{\kappa_c}$, the variate $\widetilde{w}^T\widetilde{X}_i/n$ is zero-mean Gaussian with variance at most $\widetilde{\kappa_c}^2/n$.) This completes the proof for $q = 1$.

Turning to the case $q \in (0, 1)$, in order to establish upper bounds over $\mathbb{B}_q(2R_q)$, we require the following analog of Lemma 6, proved in Appendix G.1. So as to lighten notation, let us introduce the shorthand $g(R_q, n, d) := \sqrt{R_q}\,(\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}}$.

Lemma 7. *For* $q \in (0, 1)$, *suppose that* $g(R_q, n, d) = o(1)$ *and* $d = \Omega(n)$. *Then for any fixed radius* r *such that* $r \geq c_3\widetilde{\kappa_c}^{\frac{q}{2}}\,g(R_q, n, d)$ *for some numerical constant* $c_3 > 0$, *we have*

$$\sup_{\theta \in \mathbb{B}_q(2R_q),\ \frac{\|\widetilde{X}\theta\|_2}{\sqrt{n}} \leq r} \frac{1}{n}\left|\widetilde{w}^T\widetilde{X}\theta\right| \quad \leq \quad c_4 r\,\widetilde{\kappa_c}^{\frac{q}{2}}\sqrt{R_q}\,(\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}},$$

with probability greater than $1 - c_1\exp(-c_2\,n\,g^2(R_q, n, d))$.

Note that Lemma 7 above holds for any fixed radius $r \geq c_3 \widetilde{\kappa_c}^{\frac{q}{2}} g(R_q, n, d)$. We would like the apply the result of Lemma 7 to $r = \frac{\|X\Delta\|_2}{\sqrt{n}}$, which is a random quantity. In Appendix H, we state and prove a "peeling" result that allows us to strengthen Lemma 7 in a way suitable for our needs. In particular, if we define the event

$$\mathcal{E} \quad := \quad \big\{\exists\, \theta \in \mathbb{B}_q(2R_q) \text{ such that } \frac{1}{n}\big|\widetilde{w}^T \widetilde{X}\theta\big| \geq c_4 \frac{\|\widetilde{X}\theta\|_2}{\sqrt{n}}\, \widetilde{\kappa_c}^{\frac{q}{2}} \sqrt{R_q}\, (\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}}\big\}, \tag{45}$$

then we claim that

$$\mathbb{P}[\mathcal{E}] \quad \leq \quad \frac{2\exp(-c\, n\, g^2(R_q, n, d))}{1 - \exp(-c\, n\, g^2(R_q, n, d))}.$$

This claim follows from Lemma 9 in Appendix H by making the choices $f_n(v; X_n) = \frac{1}{n}|w^T X v|$, $\rho(v) = \frac{\|Xv\|_2}{\sqrt{n}}$, and $g(r) = c_3\, r\, \widetilde{\kappa_c}^{\frac{q}{2}} \sqrt{R_q}\, (\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}}$.

Returning to the main thread, from the basic inequality (41), when the event $\mathcal{E}$ from equation (45) holds, we have

$$\frac{\|\widetilde{X}\Delta\|_2^2}{n} \quad \leq \quad \frac{\|\widetilde{X}\Delta\|_2}{\sqrt{n}} \sqrt{\widetilde{\kappa_c}^q R_q (\frac{\log d}{n})^{1-q/2}}.$$

Canceling out a factor of $\frac{\|X\Delta\|_2}{\sqrt{n}}$, squaring both sides, multiplying by σ^2 and simplifying yields

$$\frac{\|X\Delta\|_2^2}{n} \quad \leq \quad c^2\, \sigma^2 (\frac{\kappa_c}{\sigma})^q\, R_q (\frac{\log d}{n})^{1-q/2} \quad = \quad c^2\, \kappa_c^2\, R_q\, (\frac{\sigma^2}{\kappa_c^2}\, \frac{\log d}{n})^{1-q/2},$$

as claimed.

Proof of Theorem 4(b): For this part, we require the following lemma, proven in Appendix G.2:

Lemma 8. *Suppose that* $\frac{d}{2s} \geq 2$. *Then for any* $r > 0$, *we have*

$$\sup_{\theta \in \mathbb{B}_0(2s), \frac{\|X\theta\|_2}{\sqrt{n}} \leq r} \frac{1}{n}\big|w^T X\theta\big| \quad \leq \quad 9\, r\, \sigma \sqrt{\frac{s \log(\frac{d}{s})}{n}}$$

with probability greater than $1 - \exp\big(-10 s \log(\frac{d}{2s})\big)$.

Consequently, combining this result with the basic inequality (41), we conclude that

$$\frac{\|X\Delta\|_2^2}{n} \leq 9\, \frac{\|X\Delta\|_2}{\sqrt{n}}\, \sigma \sqrt{\frac{s \log(\frac{d}{s})}{n}},$$

with high probability, from which the result follows.

5 Discussion

The main contribution of this paper was to analyze minimax rates of convergence for the linear model (1) under high-dimensional scaling, in which the sample size n and problem dimension d tend to infinity. We provided lower bounds for the ℓ_p-norm for all $p \in [1, \infty]$ with $p \neq q$, as well

as for the ℓ_2-prediction loss. In addition, for both the ℓ_2-loss and ℓ_2-prediction loss, we derived a set of upper bounds that match our lower bounds up to constant factors, so that the minimax rates are exactly determined in these cases. The rates may be viewed as an extension of the rates for the case of ℓ_2-loss from Donoho and Johnstone [14] on the Gaussian sequence model to more general design matrices X. In particular substituting $X = I$ and $d = n$ into Theorems 1 and 2, yields the same rates as those expressed in Donoho and Johnstone [14] (see Corollary 1), although they provided much sharper control of the constant pre-factors than the analysis given here.

Apart from the rates themselves, our analysis highlights how conditions on the design matrix X enter in complementary manners for different loss functions. On one hand, it is possible to obtain lower bounds on ℓ_2-risk (see Theorem 1) or upper bounds on ℓ_2-prediction risk (see Theorem 4) under very mild assumptions on X—in particular, our analysis requires only that the columns of $X/\sqrt{n}$ have bounded ℓ_2-norms (see, in particular, Assumption 1). On the other hand, in order to obtain upper bounds on ℓ_2 risk (Theorem 2) or lower bound on ℓ_2-norm prediction risk (Theorem 3), the design matrix X must satisfy, in addition to column normalization, other more restrictive conditions. In particular, our analysis was based on imposed on a certain type of lower bound on the curvature of $X^T X$ measured over the ℓ_q-ball (see Assumption 2). As shown in Lemma 1, this lower bound is intimately related to the *degree of non-identifiability* over the ℓ_q-ball of the high-dimensional linear regression model .

In addition, we showed that Assumption 2 is not unreasonable—in particular, it is satisfied with high probability for broad classes of Gaussian random matrices, in which each row is drawn in an i.i.d. manner from a $N(0, \Sigma)$ distribution (see Proposition 1). This result applies to Gaussian ensembles with much richer structure than the standard Gaussian case ($\Sigma = I_{d \times d}$). Finally, we compared to the weakest known sufficient conditions for ℓ_1-based relaxations to be consistent in ℓ_2-norm for $q = 0$—namely, the restricted eigenvalue (RE) condition, of Bickel et al. [4] and showed that the oracle least-squares over the ℓ_0-ball method can succeed with even milder conditions on the design. In addition, we also proved that the RE condition holds with high probability for broad classes for Gaussian random matrices, as long as the covariance matrix Σ is not degenerate. The analysis highlights how the structure of X determines whether ℓ_1-based relaxations achieve the minimax optimal rate.

The results and analysis from our paper can be extended in a number of ways. First, the assumption of independent Gaussian noise is somewhat restrictive and it would be interesting to analyze the model under different noise assumption, either noise with heavier tails or some degree of dependency. In addition, we are currently working on extending our analysis to non-parametric sparse additive models.

Acknowledgements

We thank Arash Amini for useful discussion, particularly regarding the proofs of Proposition 1 and Lemma 9. This work was partially supported by NSF grant DMS-0605165 to MJW and BY. In addition, BY was partially supported by the NSF grant SES-0835531 (CDI), the NSFC grant 60628102 and a grant from the MSRA. MJW was supported by an Sloan Foundation Fellowship and AFOSR Grant FA9550-09-1-0466. During this work, GR was financially supported by a Berkeley Graduate Fellowship.

A Proof of Lemma 2

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

$$\|\theta\|_1 = \|\theta_S\|_1 + \sum_{j \notin S} |\theta_j| \leq \sqrt{|S|}\|\theta\|_2 + \tau \sum_{j \notin S} \frac{|\theta_j|}{\tau}.$$

Since $|\theta_j|/\tau < 1$ for all $i \notin S$, we obtain

$$\begin{aligned} \|\theta\|_1 &\leq \sqrt{|S|}\|\theta\|_2 + \tau \sum_{j \notin S} \big(|\theta_i|/\tau\big)^q \\ &\leq \sqrt{|S|}\|\theta\|_2 + 2R_q \tau^{1-q}. \end{aligned}$$

Finally, we observe $2R_q \geq \sum_{j \in S} |\theta_j|^q \geq |S|\tau^q$, from which the result follows.

B Proof of Lemma 3

The result is obtained by inverting known results on (dyadic) entropy numbers of ℓ_q-balls; there are some minor technical subtleties in performing the inversion. For a d-dimensional ℓ_q ball with $q \in (0, p)$, it is known [30, 20, 17] that for all integers $k \in [\log d, d]$, the dyadic entropy numbers ϵ_k of the ball $\mathbb{B}_q(1)$ with respect to the ℓ_p-norm scale as

$$\epsilon_k(\ell_q, \|\cdot\|_p) = C_{q,p} \left[\frac{\log(1 + \frac{d}{k})}{k}\right]^{1/q - 1/p}. \tag{46}$$

Moreover, for $k \in [1, \log d]$, we have $\epsilon_k(\ell_q) \leq C_{q,p}$.

We first establish the upper bound on the metric entropy. Since $d \geq 2$, we have

$$e_k(\ell_q) \leq C_{q,p}\left[\frac{\log(1 + \frac{d}{2})}{k}\right]^{1/q - 1/p} \leq C_{q,p}\left[\frac{\log d}{k}\right]^{1/q - 1/p}.$$

Inverting this inequality for $k = \log N_{p,q}(\epsilon)$ and allowing for a ball radius R_q yields

$$\log N_{p,q}(\epsilon) \leq \big(C_{q,p} \frac{{R_q}^{1/q}}{\epsilon}\big)^{\frac{pq}{p-q}} \log d, \tag{47}$$

as claimed.

We now turn to proving the lower bound on the metric entropy, for which we require the existence of some fixed $\nu \in (0, 1)$ such that $k \leq d^{1-\nu}$. Under this assumption, we have $1 + \frac{d}{k} \geq \frac{d}{k} \geq d^{\nu}$, and hence

$$C_{q,p}\left[\frac{\log(1 + \frac{d}{k})}{k}\right]^{1/q - 1/p} \geq C_{q,p}\left[\frac{\nu \log d}{k}\right]^{1/q - 1/p}$$

Accounting for the radius R_q as was done for the upper bound yields

$$\log N_{p,q}(\epsilon) \geq \nu\big(\frac{C_{q,p} {R_q}^{1/q}}{\epsilon}\big)^{\frac{pq}{p-q}} \log d,$$

as claimed.

Finally, let us check that our assumptions on k needed to perform the inversion are ensured by the conditions that we have imposed on ϵ. The condition $k \geq \log d$ is ensured by setting $\epsilon < 1$. Turning to the condition $k \leq d^{1-\nu}$, from the bound (47) on k, it suffices to choose ϵ such that $\big(\frac{C_{q,p}}{\epsilon}\big)^{\frac{pq}{p-q}} \log d \leq d^{1-\nu}$. This condition is ensured by enforcing the lower bound $\epsilon^p = \Omega\big(\frac{\log d}{d^{1-\nu}}\big)^{\frac{p-q}{q}}$ for some $\nu \in (0, 1)$.

C Proof of Lemma 4

We deal first with (dyadic) entropy numbers, as previously defined (27), and show that

$$\epsilon_{2k-1}(\text{absconv}_q(X/\sqrt{n}),\ \|\cdot\|_2) \ \leq \ c\,\kappa_c\ \min\Big\{1, \big(\frac{\log(1+\frac{d}{k})}{k}\big)^{\frac{1}{q}-\frac{1}{2}}\Big\}. \tag{48}$$

We prove this intermediate claim by combining a number of known results on the behavior of dyadic entropy numbers. First, using Corollary 9 from Guédon and Litvak [17], for all $k = 1, 2, \ldots$, we have

$$\epsilon_{2k-1}(\text{absconv}_q(X/\sqrt{n}),\ \|\cdot\|_2) \ \leq \ c\,\epsilon_k(\text{absconv}_1(X),\ \|\cdot\|_2)\ \min\Big\{1, \big(\frac{\log(1+\frac{d}{k})}{k}\big)^{\frac{1}{q}-1}\Big\}.$$

Using Corollary 2.4 from Carl and Pajor [7], we obtain

$$\epsilon_k(\text{absconv}_1(X/\sqrt{n}),\ \|\cdot\|_2) \ \leq \ \frac{c}{\sqrt{n}}\,|\!|\!|X|\!|\!|_{1\to 2}\ \min\Big\{1, \big(\frac{\log(1+\frac{d}{k})}{k}\big)^{1/2}\Big\},$$

where $|\!|\!|X|\!|\!|_{1\to 2}$ denotes the norm of X viewed as an operator from $\ell_1^d \to \ell_2^n$. More specifically, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}|\!|\!|X|\!|\!|_{1\to 2} &= \frac{1}{\sqrt{n}}\sup_{\|u\|_1=1}\|Xu\|_2 \\
&= \frac{1}{\sqrt{n}}\sup_{\|v\|_2=1}\sup_{\|u\|_1=1} v^T X u \\
&= \max_{i=1,\ldots,d}\|X_i\|_2/\sqrt{n} \ \leq \ \kappa_c.
\end{aligned}$$

Overall, we have shown that $\epsilon_{2k-1}(\text{absconv}_q(X/\sqrt{n}),\ \|\cdot\|_2) \ \leq \ c\,\kappa_c\ \min\Big\{1, \big(\frac{\log(1+\frac{d}{k})}{k}\big)^{\frac{1}{q}-\frac{1}{2}}\Big\}$, as claimed. Finally, under the stated assumptions, we may invert the upper bound (48) by the same procedure as in the proof of Lemma 3 (see Appendix B), thereby obtaining the claim.

D Proof of Lemma 5

In this appendix, we prove Lemma 5. Our proof is inspired by related results from the approximation theory literature (see, e.g., Kühn [20]). For each even integer $s = 2, 4, 6, \ldots, d$, let us define the set

$$\mathcal{H} := \big\{z \in \{-1, 0, +1\}^d \ \mid \ \|z\|_0 = s\big\}. \tag{49}$$

Note that the cardinality of this set is $|\mathcal{H}| = \binom{d}{s}2^s$, and moreover, we have $\|z - z'\|_0 \leq 2s$ for all pairs $z, z' \in \mathcal{H}$. We now define the Hamming distance ρ_H on $\mathcal{H} \times \mathcal{H}$ via $\rho_H(z, z') = \sum_{j=1}^d \mathbb{I}[z_j \neq z'_j]$. For some fixed element $z \in \mathcal{H}$, consider the set $\{z' \in \mathcal{H} \ \mid \ \rho_H(z, z') \leq s/2\}$. Note that its cardinality is upper bounded as

$$\big|\{z' \in \mathcal{H} \ \mid \ \rho_H(z, z') \leq s/2\}\big| \ \leq \ \binom{d}{s/2}3^{s/2}.$$

To see this, note that we simply choose a subset of size $s/2$ where z and z' agree and then choose the other $s/2$ co-ordinates arbitrarily.

Now consider a set $\mathcal{A} \subset \mathcal{H}$ with cardinality at most $|\mathcal{A}| \leq m := \frac{\binom{d}{s}}{\binom{d}{s/2}}$. The set of elements $z \in \mathcal{H}$ that are within Hamming distance $s/2$ of some element of $\mathcal{A}$ has cardinality at most

$$|\{z \in \mathcal{H} \mid : \rho_H(z, z') \leq s/2 \text{ for some } z' \in \mathcal{A}\}| \quad \leq \quad |\mathcal{A}| \binom{d}{s/2} 3^{s/2} \quad < \quad |\mathcal{H}|,$$

where the final inequality holds since $m\binom{d}{s/2}3^{s/2} \; < \; |\mathcal{H}|$. Consequently, for any such set with cardinality $|\mathcal{A}| \leq m$, there exists a $z \in \mathcal{H}$ such that $\rho_H(z, z') > s/2$ for all $z' \in \mathcal{A}$. By inductively adding this element at each round, we then create a set with $\mathcal{A} \subset \mathcal{H}$ with $|\mathcal{A}| > m$ such that $\rho_H(z, z') > s/2$ for all $z, z' \in \mathcal{A}$.

To conclude, let us lower bound the cardinality m. We have

$$m = \frac{\binom{d}{s}}{\binom{d}{s/2}} \; = \; \frac{(d - s/2)!\,(s/2)!}{(d-s)!\;s!} = \prod_{j=1}^{s/2} \frac{d-s+j}{s/2+j} \; \geq \; (\frac{d-s}{s/2})^{s/2},$$

where the final inequality uses the fact that the ratio $\frac{d-s+j}{s/2+j}$ is decreasing as a function of j.

E Proof of Proposition 1

In this appendix, we prove both parts of Proposition 1. In addition to proving the lower bound (24), we also prove the analogous upper bound

$$\frac{\|Xv\|_2}{\sqrt{n}} \leq 3\|\Sigma^{1/2}v\|_2 + 6\left[\frac{\rho(\Sigma)\log d}{n}\right]^{1/2} \|v\|_1 \qquad \text{for all } v \in \mathbb{R}^d. \tag{50}$$

Our approach to proving the bounds (24) and (50) is based on Slepian's lemma [23, 12] as well as an extension thereof due to Gordon [15]. For the reader's convenience, we re-state versions of this lemma here. Given some index set $U \times V$, let $\{Y_{u,v},\ (u,v) \in U \times V\}$ and $\{Z_{u,v},\ (u,v) \in U \times V\}$ be a pair of zero-mean Gaussian processes. Given the semi-norm on these processes defined via $\sigma(X) = \mathbb{E}[X^2]^{1/2}$, Slepian's lemma asserts that if

$$\sigma(Y_{u,v} - Y_{u',v'}) \quad \leq \quad \sigma(Z_{u,v} - Z_{u',v'}) \qquad \text{for all } (u,v) \text{ and } (u',v') \text{ in } U \times V, \tag{51}$$

then

$$\mathbb{E}[\sup_{(u,v)\in U\times V} Y_{u,v}] \leq \mathbb{E}[\sup_{(u,v)\in U\times V} Z_{u,v}]. \tag{52}$$

One version of Gordon's extension [15, 23] asserts that if the inequality (51) holds for (u,v) and (u',v') in $U \times V$, and holds with *equality* when $v = v'$, then

$$\mathbb{E}[\sup_{u\in U} \inf_{v\in V} Y_{u,v}] \quad \leq \quad \mathbb{E}[\sup_{u\in U} \inf_{v\in V} Z_{u,v}]. \tag{53}$$

Turning to the problem at hand, any random matrix X from the given ensemble can be written as $W\Sigma^{1/2}$, where $W \in \mathbb{R}^{n\times d}$ is a matrix with i.i.d. $N(0,1)$ entries, and $\Sigma^{1/2}$ is the symmetric matrix square root. We choose the set U as the unit ball $S^{n-1} = \{u \in \mathbb{R}^n \mid \|u\|_2 = 1\}$, and for some radius r, we choose V as the set

$$V(r) := \{v \in \mathbb{R}^d \mid \|\Sigma^{1/2}v\|_2 = 1, \|v\|_q^q \leq r\}.$$

(Although this set may be empty for certain choices of r, our analysis only concerns those choices for which it is non-empty.) For a matrix M, we define the associated Frobenius norm $|||M|||_F = [\sum_{i,j} M_{ij}^2]^{1/2}$, and for any $v \in V(r)$, we introduce the convenient shorthand $\tilde{v} = \Sigma^{1/2}\, v$.

With these definition, consider the centered Gaussian process $Y_{u,v} = u^T W v$ indexed by $S^{n-1} \times V(r)$. Given two pairs (u, v) and (u', v') in $S^{n-1} \times V(r)$, we have

$$\begin{aligned} \sigma^2(Y_{u,v} - Y_{u',v'}) &= |||u\,\tilde{v}^T - u'(\tilde{v}')^T|||_F^2 \\ &= |||u\tilde{v}^T - u'\,\tilde{v}^T + u'\,\tilde{v}^T - u'(\tilde{v}')^T|||_F^2 \\ &= \|\tilde{v}\|_2^2\, \|u - u'\|_2^2 + \|u'\|_2^2 \|\tilde{v} - \tilde{v}'\|_2^2 + 2(u^T u' - \|u'\|_2^2)(\|\tilde{v}\|_2^2 - \tilde{v}^T\tilde{v}') \quad (54) \end{aligned}$$

Now by the Cauchy-Schwarz inequality and the equalities $\|u\|_2 = \|u'\|_2 = 1$ and $\|\tilde{v}\|_2 = \|\tilde{v}'\|_2$, we have $u^T u' - \|u\|_2^2 \leq 0$, and $\|\tilde{v}\|_2^2 - \tilde{v}^T\tilde{v}' \geq 0$. Consequently, we may conclude that

$$\sigma^2(Y_{u,v} - Y_{u',v'}) \;\leq\; \|u - u'\|_2^2 + \|\tilde{v} - \tilde{v}'\|_2^2. \quad (55)$$

We claim that the Gaussian process $Y_{u,v}$ satisfies the conditions Gordon's lemma in terms of the zero-mean Gaussian process $Z_{u,v}$ given by

$$Z_{u,v} \;=\; g^T u + h^T(\Sigma^{1/2}\, v), \quad (56)$$

where $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^d$ are both standard Gaussian vectors (i.e., with i.i.d. $N(0,1)$ entries). To establish this claim, we compute

$$\begin{aligned} \sigma^2(Z_{u,v} - Z_{u',v'}) &= \|u - u'\|_2^2 + \|\Sigma^{1/2}\,(v - v')\|_2^2 \\ &= \|u - u'\|_2^2 + \|\tilde{v} - \tilde{v}'\|_2^2. \end{aligned}$$

Thus, from equation (55), we see that Slepian's condition (51) holds. On the other hand, when $v = v'$, we see from equation (54) that

$$\sigma^2(Y_{u,v} - Y_{u',v}) \;=\; \|u - u'\|_2^2 \;=\; \sigma^2(Z_{u,v} - Z_{u,v'}),$$

so that the equality required for Gordon's inequality is also satisfied.

Establishing an upper bound: We begin by exploiting Slepian's inequality (52) to establish the upper bound (50). We have

$$\begin{aligned} \mathbb{E}\big[\sup_{v \in V(r)} \|Xv\|_2\big] &= \mathbb{E}\big[\sup_{(u,v) \in S^{n-1} \times V(r)} u^T X v\big] \\ &\leq \mathbb{E}\big[\sup_{(u,v) \in S^{n-1} \times V(r)} Z_{u,v}\big] \\ &= \mathbb{E}[\sup_{\|u\|_2 = 1} g^T u] + \mathbb{E}[\sup_{v \in V(r)} h^T(\Sigma^{1/2} v)] \\ &\leq \mathbb{E}[\|g\|_2] + \mathbb{E}[\sup_{v \in V(r)} h^T(\Sigma^{1/2} v)]. \end{aligned}$$

By convexity, we have $\mathbb{E}[\|g\|_2] \leq \sqrt{\mathbb{E}[\|g\|_2^2]} = \sqrt{n}$, from which we can conclude that

$$\mathbb{E}\big[\sup_{v \in V(r)} \|Xv\|_2\big] \;\leq\; \sqrt{n} + \mathbb{E}[\sup_{v \in V(r)} h^T(\Sigma^{1/2} v)]. \quad (57)$$

Turning to the remaining term, we have

$$\sup_{v \in V(r)} |h^T(\Sigma^{1/2} v)| \leq \sup_{v \in V(r)} \|v\|_1 \, \|\Sigma^{1/2} h\|_\infty \; \leq \; r \|\Sigma^{1/2} h\|_\infty.$$

Since each element $(\Sigma^{1/2} h)_i$ is zero-mean Gaussian with variance at most $\rho(\Sigma) = \max_i \Sigma_{ii}$, standard results on Gaussian maxima (e.g., [23]) imply that $\mathbb{E}[\|\Sigma^{1/2} h\|_\infty] \leq \sqrt{3\rho(\Sigma) \, \log d}$. Putting together the pieces, we conclude that for $q = 1$

$$\mathbb{E}\big[\sup_{v \in V(r)} \|Xv\|_2 / \sqrt{n} \big] \quad \leq \quad \underbrace{1 + \big[3\rho(\Sigma) \, \frac{\log d}{n}\big]^{1/2} \, r}_{t_u(r)}. \tag{58}$$

Having controlled the expectation, it remains to establish sharp concentration. Let $f : \mathbb{R}^D \to \mathbb{R}$ be Lipschitz function with constant L with respect to the ℓ_2-norm. Then if $w \sim N(0, I_{D \times D})$ is standard normal, we are guaranteed [22] that for all $t > 0$,

$$\mathbb{P}\big[|f(w) - \mathbb{E}[f(w)]| \geq t\big] \quad \leq \quad 2 \exp(-\frac{t^2}{2L^2}). \tag{59}$$

Note the dimension-independent nature of this inequality. We apply this result to the random matrix $W \in \mathbb{R}^{n \times d}$, viewed as a standard normal random vector in $D = n \, d$ dimensions. First, letting $f(W) = \sup_{v \in V(r)} \|W \Sigma^{1/2} v\|_2 / \sqrt{n}$, we find that

$$\begin{aligned} \sqrt{n}\big[f(W) - f(W')\big] \quad &= \quad \sup_{v \in V(r)} \|W \Sigma^{1/2} v\|_2 - \sup_{v \in V(r)} \|W' \Sigma^{1/2} v\|_2 \\ &\leq \quad \sup_{v \in V(r)} \|\Sigma^{1/2} v\|_2 |\!|\!| (W - W') |\!|\!|_F \\ &= \quad |\!|\!| W - W' |\!|\!|_F \end{aligned}$$

since $\|\Sigma^{1/2} v\|_2 = 1$ for all $v \in V(r)$. We have thus shown that the Lipschitz constant $L \leq 1/\sqrt{n}$. Recalling the definition of $t_u(r)$ from the upper bound (58), we set $t = t_u(r)/2$ in the tail bound (59), thereby obtaining

$$\mathbb{P}\big[\sup_{v \in V(r)} \|Xv\|_2 \geq \frac{3}{2} t_u(r; q) \big] \quad \leq \quad 2 \exp(-n \frac{t_u(r)^2}{8}). \tag{60}$$

We now exploit this family of tail bounds to upper bound the probability of the event

$$\mathcal{T} \quad := \quad \big\{ \exists \, v \in \mathbb{R}^d \text{ s.t. } \|\Sigma^{1/2} v\|_2 = 1 \text{ and } \|Xv\|_2 \geq 3 t_u(\|v\|_1) \big\}.$$

We do so using Lemma 9 from Appendix H. In particular, for the case $\mathcal{E} = \mathcal{T}$, we may apply this lemma with the objective functions $f(v; X) = \|Xv\|_2$, sequence $a_n = n$, the constraint $\rho(\cdot) = \|\cdot\|_1$, the set $S = \{v \in \mathbb{R}^d \mid \|\Sigma^{1/2} v\|_2 = 1\}$, and $g(r) = 3 t_u(r)/2$. Note that the bound (60) means that the tail bound (65) holds with $c = 4/72$. Therefore, by applying Lemma 9, we conclude that $\mathbb{P}[\mathcal{T}] \leq c_1 \exp(-c_2 n)$ for some numerical constants c_i.

Finally, in order to extend the inequality to arbitrary $v \in \mathbb{R}^d$, we note that the rescaled vector $\check{v} = v / \|\Sigma^{1/2} v\|_2$ satisfies $\|\Sigma^{1/2} \check{v}\|_2 = 1$. Consequently, conditional on the event $\mathcal{T}^c$, we have

$$\|X \check{v}\|_2 / \sqrt{n} \quad \leq \quad 3 + 3\big[\sqrt{(3\rho(\Sigma) \, \log d)/n}\big] \, \|\check{v}\|_1,$$

or equivalently, after multiplying through by $\|\Sigma^{1/2} v\|_2$, the inequality

$$\|Xv\|_2 / \sqrt{n} \quad \leq \quad 3 \|\Sigma^{1/2} v\|_2 + 3(\sqrt{(3\rho(\Sigma) \, \log d)/n}) \|v\|_1,$$

thereby establishing the claim (50).

Establishing the lower bound (24): We now exploit Gordon's inequality in order to establish the lower bound (24). We have

$$-\inf_{v \in V(r)} \|Xv\|_2 \;=\; \sup_{v \in V} -\|Xv\|_2 \;=\; \sup_{v \in V(r)} \inf_{u \in U} u^T X v.$$

Applying Gordon's inequality, we obtain

$$\begin{aligned}
\mathbb{E}[\sup_{v \in V(r)} -\|Xv\|_2] &\leq \mathbb{E}\big[\sup_{v \in V(r)} \inf_{u \in S^{n-1}} Z_{u,v}\big] \\
&= \mathbb{E}[\inf_{u \in S^{n-1}} g^T u] + \mathbb{E}[\sup_{v \in V(r)} h^T \Sigma^{1/2} v] \\
&\leq -\mathbb{E}[\|g\|_2] + \big[3\rho(\Sigma)\,\log d\big]^{1/2}\, r.
\end{aligned}$$

where we have used our previous derivation to upper bound $\mathbb{E}[\sup_{v \in V(r)} h^T \Sigma^{1/2} v]$. Noting[4] that $\mathbb{E}[\|g\|_2] \geq \sqrt{n}/2$ for all $n \geq 1$, we divide by $\sqrt{n}$ and add 1 to both sides so as to obtain

$$\mathbb{E}\big[\sup_{v \in V(r)} \big(1 - \|Xv\|_2/\sqrt{n}\big)\big] \;\leq\; \underbrace{1/2 + \big[3\rho(\Sigma)\,\log d\big]^{1/2}\, r}_{t_\ell(r)} \tag{61}$$

Next define the function $f(W) = \sup_{v \in V(r)} \big(1 - \|W\Sigma^{1/2} v\|_2/\sqrt{n}\big)$. The same argument as before shows that its Lipschitz constant is at most $1/\sqrt{n}$. Setting $t = t_\ell(r)/2$ in the concentration statement (59) and combining with the lower bound (61), we conclude that

$$\mathbb{P}\big[\sup_{v \in V(r)} \big(1 - \|Xv\|_2\big) \geq \frac{3}{2} t_\ell(r)\big] \;\leq\; 2\exp\big(-n\frac{t_\ell^2(r)}{8}\big). \tag{62}$$

Define the event

$$\widetilde{\mathcal{T}} \;:=\; \big\{\exists\, v \in \mathbb{R}^d \text{ s.t. } \|\Sigma^{1/2} v\|_2 = 1 \text{ and } \big(1 - \|Xv\|_2\big) \geq 3t_\ell(\|v\|_1)\big\}.$$

We can now apply Lemma 9 with $a_n = n$, $g(r) = 3t_\ell(r)/2$ and $\mu = 1/2$ to conclude that there exist constants c_i such that $\mathbb{P}[\widetilde{\mathcal{T}}] \leq c_1 \exp(-c_2 n)$.

Finally, to extend the claim to all vectors v, we consider the rescaled vector $\check{v} = v/\|\Sigma^{1/2} v\|_2$. Conditioned on the event $\widetilde{\mathcal{T}}^c$, we have for all $v \in \mathbb{R}^d$,

$$1 - \|X\check{v}\|_2/\sqrt{n} \;\leq\; \frac{3}{2} + 3\,(\sqrt{(3\rho(\Sigma)\,\log d)/n})\,\|\check{v}\|_1,$$

or equivalently, after multiplying through by $\|\Sigma^{1/2} v\|_2$ and re-arranging,

$$\|Xv\|_2/\sqrt{n} \;\geq\; \frac{1}{2}\|\Sigma^{1/2} v\|_2 - 3\,(\sqrt{(3\rho(\Sigma)\,\log d)/n})\,\|v\|_1,$$

as claimed.

[4] In fact, $|\mathbb{E}[\|g\|_2] - \sqrt{n}| = o(\sqrt{n})$, but this simple bound is sufficient for our purposes.

F Proof of Lemma 6

For a given radius $r > 0$, define the set

$$\mathbb{S}(s, r) \quad := \quad \big\{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq 2s, \quad \|\theta\|_2 \leq r\big\},$$

and the random variables $Z_n = Z_n(s, r)$ given by

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

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

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

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 an $(r\epsilon)$-covering in ℓ_2-norm of the ball $\mathbb{B}_2(r)$ with at most $2^{2s \log(1/\epsilon)}$ elements (e.g., [24]).

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

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

Given the assumptions on X, we have $\|X(\theta - \theta^k)\|_2/\sqrt{n} \leq \kappa_u r \|\theta - \theta^k\|_2 \quad \leq \quad \kappa_u \, \epsilon$. 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 $1 - c_1 \exp(-c_2 n)$, using standard tail bounds (see Appendix I). Putting together the pieces, we conclude that

$$\frac{1}{n}|w^T X\theta| \quad \leq \quad \frac{1}{n}|w^T X\theta^k| + 2\kappa_u \, \sigma \, r \, \epsilon$$

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

$$Z_n \quad \leq \quad \max_{k=1,2,\ldots,N} \frac{1}{n}|w^T X\theta^k| + 2\kappa_u \, \sigma \, r \, \epsilon.$$

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

$$\begin{aligned} Z_n \quad &\leq \quad \sigma \; r \; \kappa_u \sqrt{\frac{3 \log N(s, r, \epsilon)}{n}} + 2\kappa_u \, \sigma r \, \epsilon \\ &= \quad \sigma \; r \; \kappa_u \Big\{ \sqrt{\frac{3 \log N(s, r, \epsilon)}{n}} + 2\epsilon \Big\}. \end{aligned} \tag{63}$$

with probability greater than $1 - c_1 \exp(-c_2 \log N(s, r, \epsilon))$.

Finally, suppose that $\epsilon = \sqrt{\frac{s \log(d/2s)}{n}}$. 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 (63), we conclude that

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

as claimed. Since $\log N(s, r, \epsilon) \geq s \log(d - 2s)$, this event occurs with probability at least $1 - c_1 \exp(-c_2 \min\{n, s \log(d - s)\})$, as claimed.

G Proofs for Theorem 4

This appendix is devoted to the proofs of technical lemmas used in Theorem 4.

G.1 Proof of Lemma 7

For $q \in (0, 1)$, let us define the set

$$\mathbb{S}_q(R_q, r) := \mathbb{B}_q(2R_q) \cap \big\{ \theta \in \mathbb{R}^d \mid \|\widetilde{X}\theta\|_2/\sqrt{n} \leq r \big\}.$$

We seek to bound the random variable $Z(R_q, r) := \sup_{\theta \in \mathbb{S}_q(R_q, r)} \frac{1}{n} |\widetilde{w}^T \widetilde{X}\theta|$, which we do by a chaining result—in particular, Lemma 3.2 in van de Geer [32]). Adopting the notation from this lemma, we seek to apply it with $\epsilon = \delta/2$, and $K = 4$. Suppose that $\frac{\|X\theta\|_2}{\sqrt{n}} \leq r$, and

$$\sqrt{n}\delta \geq c_1 r \tag{64a}$$

$$\sqrt{n}\delta \geq c_1 \int_{\frac{\delta}{16}}^{r} \sqrt{\log N(t; \mathbb{S}_q)}\, dt =: J(r, \delta). \tag{64b}$$

where $N(t; \mathbb{S}_q)$ is the covering number for $\mathbb{S}_q$ in the ℓ_2-prediction norm (defined by $\|X\theta\|/\sqrt{n}$). As long as $\frac{\|\widetilde{w}\|_2^2}{n} \leq 16$, Lemma 3.2 guarantees that

$$\mathbb{P}\big[Z(R_q, r) \geq \delta,\ \frac{\|\widetilde{w}\|_2^2}{n} \leq 16\big] \leq c_1 \exp\big(-c_2 \frac{n\delta^2}{r^2}\big).$$

By tail bounds on χ^2 random variables (see Appendix I), we have $\mathbb{P}[\|\widetilde{w}\|_2^2 \geq 16n] \leq c_4 \exp(-c_5 n)$. Consequently, we conclude that

$$\mathbb{P}\big[Z(R_q, r) \geq \delta] \leq c_1 \exp\big(-c_2 \frac{n\delta^2}{r^2}\big) + c_4 \exp(-c_5 n)$$

For some $c_3 > 0$, let us set

$$\delta \;\; = \;\; c_3 \, r \, \widetilde{\kappa_c}^{\frac{q}{2}} \sqrt{R_q} \, (\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}},$$

and let us verify that the conditions (64a) and (64b) hold. Given our choice of δ, we find that

$$\frac{\delta}{r}\sqrt{n} = \Omega(n^{q/4}(\log d)^{1/2-q/4}),$$

and since $d, n \to \infty$, we see that condition (64a) holds. Turning to verification of the inequality (64b), we first provide an upper bound for $\log N(\mathbb{S}_q, t)$. Setting $\gamma = \frac{\widetilde{X}\theta}{\sqrt{n}}$ and from the definition (31) of $\mathrm{absconv}_q(X/\sqrt{n})$, we have

$$\sup_{\theta \in \mathbb{S}_q(R_q, r)} \frac{1}{n}|\widetilde{w}^T \widetilde{X}\theta| \leq \sup_{\gamma \in \mathrm{absconv}_q(X/\sqrt{n}), \|\gamma\|_2 \leq r} \frac{1}{\sqrt{n}}|\widetilde{w}^T \gamma|.$$

We may apply the bound in Lemma 4 to conclude that $\log N(\epsilon; \mathbb{S}_q)$ is upper bounded by $c' \; R_q^{\frac{2}{2-q}} \; \left(\frac{\widetilde{\kappa_c}}{\epsilon}\right)^{\frac{2q}{2-q}} \log d$.

Using this upper bound, we have

$$\begin{aligned} J(r, \delta) := \int_{\delta/16}^{r} \sqrt{\log N(\mathbb{S}_q, t)} dt \;\; &\leq \;\; \int_0^r \sqrt{\log N(\mathbb{S}_q, t)} dt \\ &\leq \;\; c \;\; R_q^{\frac{1}{2-q}} \; \widetilde{\kappa_c}^{\frac{q}{2-q}} \; \sqrt{\log d} \int_0^r t^{-q/(2-q)} dt \\ &= \;\; c' R_q^{\frac{1}{2-q}} \; \widetilde{\kappa_c}^{\frac{q}{2-q}} \; \sqrt{\log d} \, r^{1-\frac{q}{2-q}}. \end{aligned}$$

Using this upper bound, let us verify that the inequality (64b) holds as long as $r = \Omega(\widetilde{\kappa_c}^{\frac{q}{2}} \sqrt{R_q} \, (\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}})$, as assumed in the statement of Lemma 7. With our choice of δ, we have

$$\begin{aligned} \frac{J}{\sqrt{n}\,\delta} &\leq \frac{c' R_q^{\frac{1}{2-q}} \; \widetilde{\kappa_c}^{\frac{q}{2-q}} \; \sqrt{\frac{\log d}{n}} \; r^{1-\frac{q}{2-q}}}{c_3 \, r \, \widetilde{\kappa_c}^{\frac{q}{2}} \sqrt{R_q} \, (\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}}} \\ &= \frac{c' R_q^{\frac{1}{2-q}-\frac{1}{2}-\frac{q}{2(2-q)}} \; \widetilde{\kappa_c}^{\frac{q}{2-q}-\frac{q}{2}\frac{q}{2-q}-\frac{q}{2}} \; \left(\frac{\log d}{n}\right)^{\frac{q}{4}-\frac{q}{2-q}\left(\frac{1}{2}-\frac{q}{4}\right)}}{c_3} \\ &= \frac{c'}{c_3}, \end{aligned}$$

so that condition (64b) will hold as long as we choose $c_3 > 0$ large enough. Overall, we conclude that $\mathbb{P}[Z(R_q, r) \geq c_3 \, r \, \widetilde{\kappa_c}^{\frac{q}{2}} \sqrt{R_q} \, (\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}}] \leq c_1 \exp(-R_q (\log d)^{1-\frac{q}{2}} n^{\frac{q}{2}})$, which concludes the proof.

G.2 Proof of Lemma 8

First, consider a fixed subset $S \subset \{1, 2, \ldots, d\}$ of cardinality $|S| = s$. Applying the SVD to the sub-matrix $X_S \in \mathbb{R}^{n \times s}$, we have $X_S = VDU$, where $V \in \mathbb{R}^{n \times s}$ has orthonormal columns, and $DU \in \mathbb{R}^{s \times s}$. By construction, for any $\Delta_S \in \mathbb{R}^s$, we have $\|X_S \Delta_S\|_2 = \|DU\Delta_S\|_2$. Since V has orthonormal columns, the vector $\widetilde{w}_S = V^T w \in \mathbb{R}^s$ has i.i.d. $N(0, \sigma^2)$ entries. Consequently, for any

Δ_S such that $\frac{\|X_S\Delta_S\|_2}{\sqrt{n}} \leq r$, we have

$$\begin{aligned}\Big|\frac{w^T X_S \Delta_S}{n}\Big| &= \Big|\frac{\widetilde{w}_S^T}{\sqrt{n}}\frac{DU\Delta_S}{\sqrt{n}}\Big| \\ &\leq \frac{\|\widetilde{w}_S\|_2}{\sqrt{n}}\frac{\|DU\Delta_S\|_2}{\sqrt{n}} \\ &\leq \frac{\|\widetilde{w}_S\|_2}{\sqrt{n}}\, r.\end{aligned}$$

Now the variate $\sigma^{-2}\|\widetilde{w}_S\|_2^2$ is χ^2 with s degrees of freedom, so that by standard χ^2 tail bounds (see Appendix I), we have

$$\mathbb{P}\big[\frac{\|\widetilde{w}_S\|_2^2}{\sigma^2 s} \geq 1 + 4\delta\big] \leq \exp(-s\delta), \text{ valid for all } \delta \geq 1.$$

Setting $\delta = 20\log(\frac{d}{2s})$ and noting that $\log(\frac{d}{2s}) \geq \log 2$ by assumption, we have (after some algebra)

$$\mathbb{P}\left[\frac{\|\widetilde{w}_S\|_2^2}{n} \geq \frac{\sigma^2 s}{n}\big(81\log(d/s)\big)\right] \leq \exp(-20s\log(\frac{d}{2s})).$$

We have thus shown that for each fixed subset, we have the bound

$$\Big|\frac{w^T X_S \Delta_S}{n}\Big| \leq r\sqrt{\frac{81\sigma^2 s\log(\frac{d}{2s})}{n}},$$

with probability at least $1 - \exp(-20s\log(\frac{d}{2s}))$.

Since there are $\binom{d}{2s} \leq \big(\frac{de}{2s}\big)^{2s}$ subsets of size s, applying a union bound yields that

$$\begin{aligned}\mathbb{P}\Big[\sup_{\theta\in\mathbb{B}_0(2s),\ \frac{\|X\theta\|_2}{\sqrt{n}}\leq r} |\frac{w^T X\theta}{n}| \geq r\sqrt{\frac{81\sigma^2 s\log(\frac{d}{2s})}{n}}\Big] &\leq \exp\Big(-20s\log(\frac{d}{2s}) + 2s\log\frac{de}{2s}\Big) \\ &\leq \exp\big(-10s\log(\frac{d}{2s})\big),\end{aligned}$$

as claimed.

H Large deviations for random objectives

In this appendix, we state a result on large deviations of the constrained optimum of random objective functions of the form $f(v; X)$, where $v \in \mathbb{R}^d$ is the optimization vector, and X is some random vector. Of interest is the optimization problem $\sup_{\rho(v)\leq r,\ v\in S} f(v; X_n)$, where $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is some non-negative and increasing constraint function, and S is a non-empty set. With this set-up, our goal is to bound the probability of the event defined by

$$\mathcal{E} := \big\{\exists\, v \in S \text{ such that } f(v; X) \geq 2g(\rho(v)))\big\},$$

where $g : \mathbb{R} \to \mathbb{R}$ is non-negative and strictly increasing.

Lemma 9. *Suppose that $g(r) \geq \mu$ for all $r \geq 0$, and that there exists some constant $c > 0$ such that for all $r > 0$, we have the tail bound*

$$\mathbb{P}\big[\sup_{v \in S,\, \rho(v) \leq r} f_n(v; X_n) \geq g(r)\big] \quad \leq \quad 2\exp(-c\, a_n\, g^2(r)), \tag{65}$$

for some $a_n > 0$. Then we have

$$\mathbb{P}[\mathcal{E}_n] \quad \leq \quad \frac{2\exp(-ca_n\mu^2)}{1 - \exp(-ca_n\mu^2)}. \tag{66}$$

Proof. Our proof is based on a standard peeling technique (e.g., see van de Geer [32] pp. 82). By assumption, as v varies over S, we have $g(r) \in [\mu, \infty)$. Accordingly, for $m = 1, 2, \ldots$, defining the sets

$$S_m \quad := \quad \big\{v \in S \mid 2^{m-1}\mu \leq g(\rho(v)) \leq 2^m\mu\big\},$$

we may conclude that if there exists $v \in S$ such that $f(v, X) \geq 2h(\rho(v))$, then this must occur for some m and $v \in S_m$. By union bound, we have

$$\mathbb{P}[\mathcal{E}] \quad \leq \quad \sum_{m=1}^{\infty} \mathbb{P}\big[\exists\, v \in S_m \text{ such that } f(v, X) \geq 2g(\rho(v))\big].$$

If $v \in S_m$ and $f(v, X) \geq 2g(\rho(v))$, then by definition of S_m, we have $f(v, X) \geq 2\,(2^{m-1})\,\mu = 2^m\mu$. Since for any $v \in S_m$, we have $g(\rho(v)) \leq 2^m\mu$, we combine these inequalities to obtain

$$\begin{aligned}
\mathbb{P}[\mathcal{E}] \quad &\leq \quad \sum_{m=1}^{\infty} \mathbb{P}\big[\sup_{\rho(v) \leq g^{-1}(2^m\mu)} f(v, X) \geq 2^m\mu\big] \\
&\leq \quad \sum_{m=1}^{\infty} 2\exp\big(-ca_n\,[g(g^{-1}(2^m\mu))]^2\big) \\
&= \quad 2\sum_{m=1}^{\infty} \exp\big(-ca_n\,2^{2m}\mu^2\big),
\end{aligned}$$

from which the stated claim follows by upper bounding this geometric sum. □

I Some tail bounds for χ^2-variates

The following large-deviations bounds for centralized χ^2 are taken from Laurent and Massart [21]. Given a centralized χ^2-variate Z with m degrees of freedom, then for all $x \geq 0$,

$$\mathbb{P}\left[Z - m \geq 2\sqrt{mx} + 2x\right] \quad \leq \quad \exp(-x), \qquad \text{and} \tag{67a}$$

$$\mathbb{P}\left[Z - m \leq -2\sqrt{mx}\right] \quad \leq \quad \exp(-x). \tag{67b}$$

The following consequence of this bound is useful: for $t \geq 1$, we have

$$\mathbb{P}\big[\frac{Z - m}{m} \geq 4t\big] \quad \leq \quad \exp(-mt). \tag{68}$$

Starting with the bound (67a), setting $x = tm$ yields $\mathbb{P}\big[\frac{Z-m}{m} \geq 2\sqrt{t} + 2t\big] \leq \exp(-tm)$, Since $4t \geq 2\sqrt{t} + 2t$ for $t \geq 1$, we have $\mathbb{P}[\frac{Z-m}{m} \geq 4t] \leq \exp(-tm)$ for all $t \geq 1$.

References

[1] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semdefinite relaxations for sparse principal component analysis. *Annals of Statistics*, 5B:2877–2921, 2009.

[2] Z. D. Bai and Y. Q. Yin. Convergence to the semicircle law. *Annals of Probability*, 16(2): 863–875, April 2001.

[3] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked populations models. *Journal of Multivariate Analysis*, 97(6):1382–1408, July 2006.

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

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

[6] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[7] B. Carl and A. Pajor. Gelfand numbers of operators with values in a Hilbert space. *Invent. math.*, 94:479–504, 1988.

[8] B. Carl and I. Stephani. *Entropy, compactness and the approximation of operators*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, UK, 1990.

[9] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.

[10] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. of. American Mathematical Society*, 22(1):211–231, July 2008.

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

[12] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–336. Elsevier, Amsterdan, NL, 2001.

[13] D. Donoho, M. Elad, and V. M. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Info Theory*, 52(1):6–18, January 2006.

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

[15] Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[16] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.

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

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

[19] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.

[20] T. Kühn. A lower estimate for entropy numbers. *Journal of Approximation Theory*, 110:120–124, 2001.

[21] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1303–1338, 1998.

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

[23] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

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

[25] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[26] N. Meinshausen and B.Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

[27] M. Nussbaum. Asympototically equivalence of density estimation and gaussian white noise. *Annals of Statistics*, 24(6):2399–2430, 1996.

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

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

[30] C. Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. *Journal of Approximation Theory*, 40:121–128, 1984.

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

[32] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

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

[34] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

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

[36] C. H. Zhang. Least squares estimation and variable selection under minimax concave penalty. In *Mathematisches Forschungsinstitut Oberwolfach: Sparse Recovery Problems in High Dimensions*, pages 908–911, March 2009.

[37] 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.

[38] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.