# Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls

Garvesh Raskutti, Martin J. Wainwright, *Senior Member, IEEE* and Bin Yu, *Fellow, IEEE.*

***Abstract*—Consider the high-dimensional 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 with $d > n$, the quantity $\beta^* \in \mathbb{R}^d$ is an 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 estimating $\beta^*$ in either $\ell_2$-loss and $\ell_2$-prediction loss, assuming that $\beta^*$ belongs to an $\ell_q$-ball $\mathbb{B}_q(R_q)$ for some $q \in [0, 1]$. It is shown that under suitable regularity conditions on the design matrix $X$, the minimax optimal rate in $\ell_2$-loss and $\ell_2$-prediction loss scales as $R_q \big(\frac{\log d}{n}\big)^{1-\frac{q}{2}}$. The analysis in this paper reveals that conditions on the design matrix $X$ enter into the rates for $\ell_2$-error and $\ell_2$-prediction error in complementary ways in the upper and lower bounds. 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 constructive, involving direct analysis of least-squares over $\ell_q$-balls. For the special case $q = 0$, corresponding to models with an exact sparsity constraint, our results show that although computationally efficient $\ell_1$-based methods can achieve the minimax rates up to constant factors, they require slightly stronger assumptions on the design matrix $X$ than optimal algorithms involving least-squares over the $\ell_0$-ball.**

**Keywords:** compressed sensing; high-dimensional statistics; minimax rates; Lasso; sparse linear regression; $\ell_1$-relaxation. [1]

[1] G. Raskutti is with the Department of Statistics, University of California, Berkeley, CA, 94720 USA. e-mail: garveshr@stat.berkeley.edu. M. J. Wainwright and B. Yu are with the Department of Statistics and Department of Electrical Engineering and Computer Science, University of California, Berkeley. e-mail: wainwrig@stat.berkeley.edu and binyu@stat.berkeley.edu.
This work was partially supported by NSF grants DMS-0605165 and DMS-0907362 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. In addition, 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.

The initial version of this work was posted as arXiv:0910.2042 in September 2009. In addition, it was presented in part at the Allerton Conference on Control, Communication and Computer in October 2009.

## I. Introduction

The area of high-dimensional statistical inference concerns the estimation in the $d > n$ 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 general setting, it is impossible to obtain consistent estimators unless the ratio $d/n$ converges to zero. However, many applications require solving inference problems with $d > n$, in which setting the only hope is that the data has some type of lower-dimensional structure. Accordingly, an active line of research in high-dimensional inference is based on imposing various types of structural constraints, including sparsity, manifold structure, or Markov conditions, 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 work has studied the behavior of $\ell_1$-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.

### A. Problem set-up

The focus of this paper is a canonical instance of a high-dimensional inference problem, namely that estimating a high-dimensional regression vector $\beta^* \in \mathbb{R}^d$ with sparsity constraints based on observations from a linear model. 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

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

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 the regression vector $\beta^*$ satisfies some type of sparsity constraint, has been investigated extensively over the past decade. A variety of practical algorithms have been proposed and studied, many based on $\ell_1$-regularization, including basis pursuit [11], the Lasso [32, 11], and the Dantzig selector [8]. Various authors have obtained convergence rates for different error metrics, including $\ell_2$-norm error [4, 8, 25, 41], prediction loss [4, 16, 34], as well as model selection consistency [24, 36, 41, 43]. In addition, a range of sparsity assumptions have been analyzed, including the case of *hard sparsity* meaning that $\beta^*$ has exactly $s \ll d$ non-zero entries, or *soft sparsity* assumptions, based on imposing a certain decay rate on the ordered entries of $\beta^*$. Intuitively, soft sparsity means that while many of the co-efficients of the co-variates may be non-zero, many of the co-variates only make a small overall contribution to the model, which may be more applicable in some practical settings.

*a) Sparsity constraints:* One way in which to capture the notion of sparsity in a precise manner is in terms of the $\ell_q$-balls[2] for $q \in [0, 1]$, defined as

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

Note that in the limiting case $q = 0$, we have the $\ell_0$-ball

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

which corresponds to the set of vectors $\beta$ with at most $s$ non-zero elements. For $q \in (0, 1]$, membership of $\beta$ in $\mathbb{B}_q(R_q)$ enforces a "soft" form of sparsity, in that all of the coefficients of $\beta$ may be non-zero, but their absolute magnitude must decay at a relatively rapid rate. This type of soft sparsity is appropriate for various applications of high-dimensional linear regression, including image denoising, medical reconstruction and database updating, in which exact sparsity is not realistic.

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

*b) 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 $\beta^*$, 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. In the minimax formalism, one seeks to choose an estimator that minimizes the worst-case loss given by

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \mathcal{L}(\widehat{\beta}, \beta^*). \tag{2}$$

Note that the quantity (2) is random since $\widehat{\beta}$ depends on the noise $w$, and therefore, we must either provide bounds that hold with high probability or in expectation. In this paper, we provide results that hold with high probability, as shown in the statements our main results in results in Theorems 1 through 4.

Moreover, various choices of the loss function are possible, including (i) the *model selection loss*, which is zero if and only if the support $\text{supp}(\widehat{\beta})$ of the estimate agrees with the true support $\text{supp}(\beta^*)$, and one otherwise; (ii) the $\ell_2$*-loss*

$$\mathcal{L}_2(\widehat{\beta}, \beta^*) := \|\widehat{\beta} - \beta^*\|_2^2 \;=\; \sum_{j=1}^d |\widehat{\beta}_j - \beta_j^*|_2^2, \tag{3}$$

and (iii) the $\ell_2$*-prediction loss* $\|X(\widehat{\beta} - \beta^*)\|_2^2/n$. The information-theoretic limits of model selection have been studied extensively in past work (e.g., [37, 15, 1, 38]); in contrast, the analysis of this paper is focused on understanding the minimax rates associated with the $\ell_2$-loss and the $\ell_2$-prediction loss.

More precisely, the goal of this paper is to provide upper and lower bounds on the following four forms of minimax risk:

$$\mathcal{M}_2(\mathbb{B}_q(R_q), X) := \min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \|\widehat{\beta} - \beta^*\|_2^2,$$
$$\mathcal{M}_2(\mathbb{B}_0(s), X) := \min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_0(s)} \|\widehat{\beta} - \beta^*\|_2^2,$$
$$\mathcal{M}_n(\mathbb{B}_q(R_q), X) := \min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \frac{1}{n}\|X(\widehat{\beta} - \beta^*)\|_2^2,$$
$$\mathcal{M}_n(\mathbb{B}_0(s), X) := \min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_0(s)} \frac{1}{n}\|X(\widehat{\beta} - \beta^*)\|_2^2.$$

These quantities correspond to all possible combinations of minimax risk involving either the $\ell_2$-loss or the $\ell_2$-prediction loss, and with either hard sparsity ($q = 0$) or soft sparsity ($q \in (0, 1]$).

## B. Our contributions

The main contributions are derivations of optimal minimax rates both for $\ell_2$-norm and $\ell_2$-prediction losses, and perhaps more significantly, a thorough characterization of the conditions that are required on the design matrix $X$ in each case. The core of the paper consists of four main theorems, corresponding to upper and lower bounds on minimax rate for the $\ell_2$-norm loss (Theorems 1 and 2 respectively) and upper and lower bounds on $\ell_2$-prediction loss (Theorems 3 and Theorem 4) respectively. We note that for the linear model (1), the special case of orthogonal design $X = \sqrt{n} I_{n \times n}$ (so that $n = d$ necessarily holds) has been extensively studied in the statistics community (for example, see the papers [6, 14, 3] as well as references therein). In contrast, our emphasis is on the high-dimensional setting $d > n$, and our goal is to obtain results for general design matrices $X$.

More specifically, in Theorem 1, we provide lower bounds for the $\ell_2$-loss that involves a maximum of two quantities: a term involving the diameter of the null-space restricted to the $\ell_q$-ball, measuring the degree of non-identifiability of the model, and a term arising from the $\ell_2$-metric entropy structure for $\ell_q$-balls, measuring the complexity of the parameter space. Theorem 2 is complementary in nature, devoted to upper bounds that obtained by direct analysis of a specific estimator. We obtain upper and lower bounds that match up to factors that independent of the triple $(n, d, R_q)$, but depend on constants related to the structure of the design matrix $X$ (see Theorems 1 and 2). Finally, Theorems 3 and 4 are for $\ell_2$-prediction loss. For this loss, we provide upper and lower bounds on minimax rates that are again matching up to factors independent of $(n, d, R_q)$, but dependent again on the conditions of the design matrix.

A key part of our analysis is devoted to understanding the link between the prediction semi-norm—more precisely, the quantity $\|X\theta\|_2/\sqrt{n}$—and the $\ell_2$ norm $\|\theta\|_2$. In the high-dimensional setting (with $X \in \mathbb{R}^{n \times d}$ with $d \gg n$), these norms are in general incomparable, since the design matrix $X$ has a null-space of dimension at least $d - n$. However, for analyzing sparse linear regression models, it is sufficient to study the approximate equivalence of these norms only for elements $\theta$ lying in the $\ell_q$-ball, and this relationship between the two semi-norms plays an important role for the proofs of both the upper and the lower bounds. Indeed, for Gaussian noise models, the prediction semi-norm $\|X(\beta - \beta^*)\|_2/\sqrt{n}$ corresponds to the square-root Kullback-Leibler divergence between the distributions on $y$ indexed by $\beta$ and $\beta^*$, and so reflects the discriminability of these models. Our analysis shows that the conditions on $X$ enter in quite a different manner for $\ell_2$-norm and prediction losses. In particular, for the case $q > 0$, proving *upper bounds* on $\ell_2$-norm error and *lower bounds* on prediction error require relatively strong conditions on the design matrix $X$, whereas *lower bounds* on $\ell_2$-norm error and *upper bounds* on prediction error require only a very mild column normalization condition.

The proofs for the lower bounds in Theorems 1 and 3 involve a combination of a standard information-theoretic techniques (e.g. [5, 18, 39]) with results in the approximation theory literature (e.g., [17, 21]) on the metric entropy of $\ell_q$-balls. The proofs for the upper bounds in Theorems 2 and 4 involve direct analysis of the least-squares optimization over the $\ell_q$-ball. The basic idea involves concentration results for Gaussian random variables and properties of the $\ell_1$-norm over $\ell_q$-balls (see Lemma 5).

The remainder of this paper is organized as follows. In Section II, we state our main results and discuss their consequences. While we were writing up the results of this paper, we became aware of concurrent work by Zhang [42], and we provide a more detailed discussion and comparison in Section II-E, following the precise statement of our results. In addition, we also discuss a comparison between the conditions on $X$ imposed in our work, and related conditions imposed in the large body of work on $\ell_1$-relaxations. In Section III, we provide the proofs of our main results, with more technical aspects deferred to the appendices.

# II. Main results and their consequences

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.

## A. Assumptions on design matrices

Let $X^{(i)}$ denote the $i^{th}$ row of $X$ and $X_j$ denote the $j^{th}$ column of $X$. Our first assumption, which remains in force throughout most of our analysis, is that the columns $\{X_j, j = 1, \ldots, d\}$ of the design matrix $X$ are bounded in $\ell_2$-norm.

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

$$\frac{1}{\sqrt{n}} \max_{j=1,\ldots,d} \|X_j\|_2 \leq \kappa_c. \tag{4}$$

This is a fairly mild condition, since the problem can always be normalized to ensure that it is satisfied. Moreover, it would be satisfied with high probability for any random design matrix for which $\frac{1}{n}\|X_j\|_2^2 = \frac{1}{n}\sum_{i=1}^n X_{ij}^2$ satisfies a sub-exponential tail bound. This column normalization condition is required for all the theorems except for achievability bounds for $\ell_2$-prediction error when $q = 0$.

We now turn to a more subtle condition on the design matrix $X$:

**Assumption 2** (Bound on restricted lower eigenvalue)**.** For $q \in (0, 1]$, there exists a constant $\kappa_\ell > 0$ and a function $f_\ell(R_q, n, d)$ such that

$$\frac{\|X\theta\|_2}{\sqrt{n}} \geq \kappa_\ell \left(\|\theta\|_2 - f_\ell(R_q, n, d)\right) \tag{5}$$

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

A few comments on this assumption are in order. For the case $q > 0$, this assumption is imposed when deriving upper bounds for the $\ell_2$-error and lower bounds for $\ell_2$-prediction error. It is required in *upper bounding* $\ell_2$-error because for any two distinct vectors $\beta, \beta' \in \mathbb{B}_q(R_q)$, the prediction semi-norm $\|X(\beta - \beta')\|_2/\sqrt{n}$ is closely related to the Kullback-Leibler divergence, which quantifies how distinguishable $\beta$ is from $\beta'$ in terms of the linear regression model. Indeed, note that for fixed $X$ and $\beta$, the vector $Y \sim \mathcal{N}(X\beta, \sigma^2 I_{n\times n})$, so that the Kullback-Leibler divergence between the distributions on $Y$ indexed by $\beta$ and $\beta'$ is given by $\frac{1}{2\sigma^2}\|X(\beta - \beta')\|_2^2$. Thus, the lower bound (5), when applied to the difference $\theta = \beta - \beta'$, ensures any pair $(\beta, \beta')$ that are well-separated in $\ell_2$-norm remain well-separated in the $\ell_2$-prediction semi-norm. Interestingly, Assumption 2 is also essential in establishing *lower bounds* on the $\ell_2$-prediction error. Here the reason is somewhat different—namely, it ensures that the set $\mathbb{B}_q(R_q)$ still suitably "large" when its diameter is measured in the $\ell_2$-prediction semi-norm. As we show, it is this size that governs the difficulty of estimation in the prediction semi-norm.

The condition (5) is almost equivalent to bounding the smallest singular value of $X/\sqrt{n}$ restricted to the set $\mathbb{B}_q(2R_q)$. Indeed, the only difference is the "slack" provided by $f_\ell(R_q, n, d)$. The reader might question why this slack term is actually needed. In fact, it is *essential* in the case $q \in (0, 1]$, since the set $\mathbb{B}_q(2R_q)$ spans all directions of the space $\mathbb{R}^d$. (This is not true in the limiting case $q = 0$.) Since $X$ must have a non-trivial null-space when $d > n$, the condition (5) can never be satisfied with $f_\ell(R_q, n, d) = 0$ whenever $d > n$ and $q \in (0, 1]$.

Interestingly, for appropriate choices of the slack term $f_\ell(R_q, n, d)$, the restricted eigenvalue condition is satisfied with high probability for many random matrices, as shown by the following result.

**Proposition 1.** *Consider a random matrix $X \in \mathbb{R}^{n\times d}$ formed by drawing each row i.i.d. from a $\mathcal{N}(0, \Sigma)$ distribution with maximal variance $\rho^2(\Sigma) = \max_{j=1,\ldots,d} \Sigma_{jj}$. If $\frac{\rho(\Sigma)}{\lambda_{\min}(\sqrt{\Sigma})} R_q \left(\frac{\log d}{n}\right)^{1/2-q/4} < c_1$ for a sufficiently small universal constant $c_1 > 0$, then*

$$\frac{\|X\theta\|_2}{\sqrt{n}} \geq \frac{\lambda_{\min}(\Sigma^{1/2})}{4}\|\theta\|_2 - 18\,\rho(\Sigma)\; R_q\,\left(\frac{\log d}{n}\right)^{1-q/2}, \tag{6}$$

*for all $\theta \in \mathbb{B}_q(2R_q)$ with probability at least $1 - c_2\exp(-c_3 n)$.*

An immediate consequence of the bound (6) is that Assumption 2 holds with

$$f_\ell(R_q, n, d) = \bar{c}\,\frac{\rho(\Sigma)}{\lambda_{\min}(\Sigma^{1/2})}\; R_q\,\left(\frac{\log d}{n}\right)^{1-q/2} \tag{7}$$

for some universal constant $\bar{c}$. We make use of this condition in Theorems 2(a) and 3(a) to follow. The proof of Proposition 1, provided in Appendix A, follows as a consequence of a random matrix result in Raskutti et al. [29]. In the same paper, on pp. $2248-49$ it is demonstrated that there are many interesting classes of non-identity covariance matrices, among them Toeplitz matrices, constant correlation matrices and spiked models, to which Proposition 1 can be applied.

For the special case $q = 0$, the following conditions are needed for upper and lower bounds in $\ell_2$-norm error, and lower bounds in $\ell_2$-prediction error.

**Assumption 3** (Sparse Eigenvalue Conditions)**.**

(a) There exists a constant $\kappa_u < +\infty$ such that

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \leq \kappa_u\,\|\theta\|_2 \text{ for all } \theta \in \mathbb{B}_0(2s). \tag{8}$$

(b) There exists a constant $\kappa_{0,\ell} > 0$ such that

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \geq \kappa_{0,\ell}\,\|\theta\|_2 \text{ for all } \theta \in \mathbb{B}_0(2s). \tag{9}$$

Assumption 2 adapted to the special case of $q = 0$ corresponding to exactly sparse models; however, in this case, no slack term $f_\ell(R_q, n, d)$ is involved. As we discuss at more length in Section II-E, Assumption 3 is closely related to conditions imposed in analyses of

$\ell_1$-based relaxations, such as the restricted isometry property [8] as well as related but less restrictive sparse eigenvalue conditions [4, 25, 34]. Unlike the restricted isometry property, Assumption 3 does not require that the constants $\kappa_u$ and $\kappa_{0,\ell}$ are close to one; indeed, they can be arbitrarily large (respectively small), as long as they are finite and non-zero. In this sense, it is most closely related to the sparse eigenvalue conditions introduced by Bickel et al. [4], and we discuss these connections at more length in Section II-E. The set $\mathbb{B}_0(2s)$ is a union of $2s$-dimensional subspaces, which does not span all direction of $\mathbb{R}^d$. Since the condition may be satisfied for $d > n$, no slack term $f_\ell(R_q, n, d)$ is required in the case $q = 0$.

In addition, our lower bounds on $\ell_2$-error involve the set defined by intersecting the null space (or kernel) of $X$ with the $\ell_q$-ball, which we denote by $\mathcal{N}_q(X) := \mathrm{Ker}(X) \cap \mathbb{B}_q(R_q)$. We define the $\mathbb{B}_q(R_q)$*-kernel diameter* in the $\ell_2$-norm as

$$\mathrm{diam}_2(\mathcal{N}_q(X)) := \max_{\theta \in \mathcal{N}_q(X)} \|\theta\|_2 = \max_{\substack{\|\theta\|_q^q \le R_q,\\ X\theta = 0}} \|\theta\|_2. \tag{10}$$

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 the set $\mathbb{B}_q(R_q)$.

It is useful to recognize that Assumptions 2 and 3 are closely related to the diameter condition (10); in particular, these assumptions imply an upper bound bound on the $\mathbb{B}_q(R_q)$-kernel diameter in $\ell_2$-norm, and hence limit the lack of identifiability of the model.

**Lemma 1** (Bounds on non-identifiability)**.**
*(a)* Case $q \in (0, 1]$: *If Assumption 2 holds, then the* $\mathbb{B}_q(R_q)$*-kernel diameter is upper bounded as*

$$\mathrm{diam}_2(\mathcal{N}_q(X)) = \mathcal{O}(f_\ell(R_q, n, d)).$$

*(b)* Case $q = 0$: *If Assumption 3(b) is satisfied, then* $\mathrm{diam}_2(\mathcal{N}_0(X)) = 0$. *(In words, the only element of* $\mathbb{B}_0(2s)$ *in the kernel of $X$ is the* $0$*-vector.)*

These claims follow in a straightforward way from the definitions given in the assumptions. In Section II-E, we discuss further connections between our assumptions, and the conditions imposed in analysis of the Lasso and other $\ell_1$-based methods [4, 8, 24, 26], for the case $q = 0$.

### *B. Universal constants and non-asymptotic statements*

Having described our assumptions on the design matrix, we now turn to the main results that provide upper and lower bounds on minimax rates. Before doing so, let us clarify our use of universal constants in our statements. Our main goal is to track the dependence of minimax rates on the triple $(n, d, R_q)$, as well as the noise variance $\sigma^2$ and the properties of the design matrix $X$. In our statement of the minimax rates themselves, we use $\bar{c}$ to denote a universal positive constant that is independent of $(n, d, R_q)$, the noise variance $\sigma^2$ and the parameters of the design matrix $X$. In this way, our minimax rates explicitly track the dependence of all of these quantities in a non-asymptotic manner. In setting up the results, we also state certain conditions that involve a separate set of universal constants denoted $c_1, c_2$ etc.; these constants are independent of $(n, d, R_q)$ but may depend on properties of the design matrix.

In this paper, our primary interest is the high-dimensional regime in which $d \gg n$. Our theory is non-asymptotic, applying to all finite choices of the triple $(n, d, R_q)$. Throughout the analysis, we impose the following conditions on this triple. In the case $q = 0$, we require that the sparsity index $s = R_0$ satisfies $d \ge 4s \ge c_2$. These bounds ensure that our probabilistic statements are all non-trivial (i.e., are violated with probability less than 1). For $q \in (0, 1]$, we require that for some choice of universal constants $c_1, c_2 > 0$ and $\delta \in (0, 1)$, the triple $(n, d, R_q)$ satisfies

$$\frac{d}{R_q n^{q/2}} \overset{(i)}{\ge} c_1 \, d^{\delta} \overset{(ii)}{\ge} c_2. \tag{11}$$

The condition (ii) ensures that the dimension $d$ is sufficiently large so that our probabilistic guarantees are all non-trivial (i.e., hold with probability strictly less than 1). In the regime $d > n$ that is of interest in this paper, the condition (i) on $(n, d, R_q)$ is satisfied as long as the radius $R_q$ does not grow too quickly in the dimension $d$. (As a concrete example, the bound $R_q \le c_3 d^{\frac{1}{2} - \delta'}$ for some $\delta' \in (0, 1/2)$ is one sufficient condition.)

### *C. Optimal minimax rates in $\ell_2$-norm loss*

We are now ready to state minimax bounds, and we begin with lower bounds on the $\ell_2$-norm error:

**Theorem 1** (Lower bounds on $\ell_2$-norm error)**.** *Consider the linear model (1) for a fixed design matrix $X \in \mathbb{R}^{n \times d}$.*

*(a)* Case $q \in (0, 1]$: *Suppose that $X$ is column-normalized (Assumption 1 holds with $\kappa_c < \infty$), and*

$R_q(\frac{\log d}{n})^{1-q/2} < c_1$ *for a universal constant* $c_1$*. Then*

$$\mathcal{M}_2(\mathbb{B}_q(R_q), X) \geq \bar{c}\,\max\Big\{\operatorname{diam}_2^2(\mathcal{N}_q(X)),\ R_q\Big(\frac{\sigma^2}{\kappa_c^2}\,\frac{\log d}{n}\Big)^{1-q/2}\Big\} \tag{12}$$

*with probability greater than* $1/2$*.*

*(b)* Case $q = 0$: *Suppose that Assumption 3(a) holds with* $\kappa_u > 0$*, and* $\frac{s\log(d/s)}{n} < c_1$ *for a universal constant* $c_1$*. Then*

$$\mathcal{M}_2(\mathbb{B}_0(s), X) \geq \bar{c}\,\max\Big\{\operatorname{diam}_2^2(\mathcal{N}_0(X)),\ \frac{\sigma^2}{\kappa_u^2}\,\frac{s\,\log(d/s)}{n}\Big\} \tag{13}$$

*with probability greater than* $1/2$*.*

The choice of probability $1/2$ is a standard convention for stating minimax lower bounds on rates.[3] Note that both lower bounds consist of two terms. The first term corresponds to the diameter of the set $\mathcal{N}_q(X) = \operatorname{Ker}(X) \cap \mathbb{B}_q(R_q)$, a quantity which reflects the extent which the linear model (1) is unidentifiable. Clearly, one cannot estimate $\beta^*$ any more accurately than the diameter of this set. In both lower bounds, the ratios $\sigma^2/\kappa_c^2$ (or $\sigma^2/\kappa_u^2$) correspond to the inverse of the signal-to-noise ratio, comparing the noise variance $\sigma^2$ to the magnitude of the design matrix measured by $\kappa_u$, since constants $c_q$ and $c_0$ do not depend on the design $X$. As the proof will clarify, the term $[\log d]^{1-\frac{q}{2}}$ in the lower bound (12), and similarly the term $\log(\frac{d}{s})$ in the bound (13), are reflections of the complexity of the $\ell_q$-ball, as measured by its metric entropy. For many classes of random Gaussian design matrices, the second term is of larger order than the diameter term, and hence determines the rate.

We now state upper bounds on the $\ell_2$-norm minimax rate over $\ell_q$ balls. For these results, we require the column normalization condition (Assumption 1), and Assumptions 2 and 3. The upper bounds are proven by a careful analysis of constrained least-squares over the set $\mathbb{B}_q(R_q)$—namely, the estimator

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{B}_q(R_q)} \|y - X\beta\|_2^2. \tag{14}$$

**Theorem 2** (Upper bounds on $\ell_2$-norm loss)**.** *Consider the model (1) with a fixed design matrix* $X \in \mathbb{R}^{n\times d}$ *that is column-normalized (Assumption 1 with* $\kappa_c < \infty$*).*

[3] This probability may be made arbitrarily close to 1 by suitably modifying the constants in the statement.

*(a)* For $q \in (0,1]$: *Suppose that* $R_q(\frac{\log d}{n})^{1-q/2} < c_1$ *and* $X$ *satisfies Assumption 2 with* $\kappa_\ell > 0$ *and* $f_\ell(R_q, n, d) \leq c_2 R_q(\frac{\log d}{n})^{1-q/2}$*. Then*

$$\mathcal{M}_2(\mathbb{B}_q(R_q), X) \leq \bar{c}\,R_q\Big[\frac{\kappa_c^2}{\kappa_\ell^2}\,\frac{\sigma^2}{\kappa_\ell^2}\,\frac{\log d}{n}\Big]^{1-q/2} \tag{15}$$

*with probability greater than* $1 - c_3\,\exp(-c_4 \log d)$*.*

*(b)* For $q = 0$: *Suppose that* $X$ *satisfies Assumption 3(b) with* $\kappa_{0,\ell} > 0$*. Then*

$$\mathcal{M}_2(\mathbb{B}_0(s), X) \leq \bar{c}\,\frac{\kappa_c^2}{\kappa_{0,\ell}^2}\,\frac{\sigma^2}{\kappa_{0,\ell}^2}\,\frac{s\,\log d}{n} \tag{16}$$

*with probability greater than* $1 - c_1\exp(-c_2\log d)$*. If, in addition, the design matrix satisfies Assumption 3(a) with* $\kappa_u < \infty$*, then*

$$\mathcal{M}_2(\mathbb{B}_0(s), X) \leq \bar{c}\,\frac{\kappa_u^2}{\kappa_{0,\ell}^2}\,\frac{\sigma^2}{\kappa_{0,\ell}^2}\,\frac{s\log(d/s)}{n}, \tag{17}$$

*this bound holding with probability greater than* $1 - c_1\,\exp(-c_2 s\log(d/s))$*.*

In the case of $\ell_2$-error and design matrices $X$ that satisfy the assumptions of both Theorems 1 and 2, these results identify the minimax optimal rate up to constant factors. In particular, for $q \in (0,1]$, the minimax rate in $\ell_2$-norm scales as

$$\mathcal{M}_2(\mathbb{B}_q(R_q), X) \;=\; \Theta\Big(R_q\Big[\frac{\sigma^2\log d}{n}\Big]^{1-q/2}\Big), \tag{18}$$

whereas for $q = 0$, the minimax $\ell_2$-norm rate scales as

$$\mathcal{M}_2(\mathbb{B}_0(s), X) \;=\; \Theta\Big(\frac{\sigma^2\,s\log(d/s)}{n}\Big). \tag{19}$$

### *D. Optimal minimax rates in $\ell_2$-prediction norm*

In this section, we investigate minimax rates in terms of the $\ell_2$-prediction loss $\|X(\widehat{\beta} - \beta^*)\|_2^2/n$, and provide both lower and upper bounds on it. The rates match the rates for $\ell_2$, but the conditions on design matrix $X$ enter the upper and lower bounds in a different way, and we discuss these complementary roles in Section II-F.

**Theorem 3** (Lower bounds on prediction error)**.** *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)* For $q \in (0,1]$: *Suppose that* $R_q(\frac{\log d}{n})^{1-q/2} < c_1$*, and the design matrix* $X$ *satisfies Assumption 2 with* $\kappa_\ell > 0$ *and* $f_\ell(R_q, n, d) < c_2 R_q(\frac{\log d}{n})^{1-q/2}$*. Then*

$$\mathcal{M}_n(\mathbb{B}_q(R_q), X) \geq \bar{c}\,R_q\,\kappa_\ell^2\,\Big[\frac{\sigma^2}{\kappa_c^2}\,\frac{\log d}{n}\Big]^{1-q/2} \tag{20}$$

*with probability at least* $1/2$.

*(b)* For $q = 0$: *Suppose that* $X$ *satisfies Assumption 3(b) with* $\kappa_{0,\ell} > 0$ *and Assumption 3(a) with* $\kappa_u < \infty$*, and that* $\frac{s \log(d/s)}{n} < c_1$*, for some universal constant* $c_1$*. Then*

$$\mathcal{M}_n(\mathbb{B}_0(s), X) \geq \bar{c}\, \kappa_{0,\ell}^2 \, \frac{\sigma^2}{\kappa_u^2} \, \frac{s \log(d/s)}{n} \tag{21}$$

*with probability least* $1/2$.

In the other direction, we state upper bounds obtained via analysis of least-squares constrained to the ball $\mathbb{B}_q(R_q)$, a procedure previously defined (14).

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

*(a)* Case $q \in (0, 1]$: *If* $X$ *satisfies the column normalization condition, then with probability at least* $1 - c_1 \exp\left(-c_2 R_q (\log d)^{1-q/2} n^{q/2}\right)$*, we have*

$$\mathcal{M}_n(\mathbb{B}_q(R_q), X) \leq \bar{c}\, \kappa_c^2 \, R_q \left[\frac{\sigma^2}{\kappa_c^2} \, \frac{\log d}{n}\right]^{1-\frac{q}{2}}. \tag{22}$$

*(b)* Case $q = 0$: *For any* $X$*, with probability greater than* $1 - c_1 \exp\left(-c_2 s \log(d/s)\right)$*, we have*

$$\mathcal{M}_n(\mathbb{B}_0(s), X) \leq \bar{c}\, \frac{\sigma^2 \, s \log(d/s)}{n}. \tag{23}$$

We note that Theorem 4(b) was stated and proven in Bunea et. al [7] (see Theorem 3.1). However, we have included the statement here for completeness and so as to facilitate discussion.

### *E. Some remarks and comparisons*

In order to provide the reader with some intuition, let us make some comments about the scalings that appear in our results. We comment on the conditions we impose on $X$ in the next section.

- 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 [13], we have $\log \binom{d}{s} = \Theta(s \log(d/s))$. Consequently, the rate $\frac{s \log(d/s)}{n}$ corresponds to the log number of models divided by the sample size $n$. Note that in the regime where $d/s \sim d^\gamma$ for some $\gamma > 0$, this rate is equivalent (up to constant factors) to $\frac{s \log d}{n}$.
- For $q \in (0, 1]$, the interpretation of the rate $R_q \big(\frac{\log d}{n}\big)^{1-q/2}$, which appears in parts (a) of Theorems 1 through 4 can be 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 $\beta^*$ in absolute value, then the fast decay imposed by the $\ell_q$-ball condition on $\beta^*$ 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; we make this truncation idea and the optimality of the particular choice $s_q$ precise in Lemma 5 to follow in the sequel.
- It is also worthwhile considering the form of our results in the special case of the Gaussian sequence model, for which $X = \sqrt{n} I_{n \times n}$ and $d = n$. With these special settings, our results yields the same scaling (up to constant factors) as seminal work by Donoho and Johnstone [14], who determined minimax rates for $\ell_p$-losses over $\ell_q$-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 ($R_q(\frac{\log n}{n})^{1-q/2}$) as Donoho and Johnstone [14] when specialized to the case $X = \sqrt{n} I_{n \times n}$ and $\ell_p = \ell_2$. Other work by van de Geer and Loubes [35] derives bounds on prediction error for general thresholding estimators, again in the case $d = n$, and our results agree in this particular case as well.
- As noted in the introduction, during the process of writing up our results, we became aware of concurrent work by Zhang [42] on the problem of determining minimax upper and lower bounds for $\ell_p$-losses with $\ell_q$-sparsity for $q > 0$ and $p \geq 1$. There are notable differences between our and Zhang's results. First, we treat the $\ell_2$-prediction loss not covered by Zhang, and also show how assumptions on the design $X$ enter in complementary ways for $\ell_2$-loss versus prediction loss. We also have results for the important case of hard sparsity ($q = 0$), not treated in Zhang's paper. On the other hand, Zhang provides tight bounds for general $\ell_p$-losses ($p \geq 1$), not covered in this paper. It is also worth noting that the underlying proof techniques for the lower bounds are very different. We use

a direct information-theoretic approach based on Fano's method and metric entropy of $\ell_q$-balls. In contrast, Zhang makes use of an extension of the Bayesian least favorable prior approach used by Donoho and Johnstone [14]. Theorems 1 and 2 from his paper [42] (in the case $p = 2$) are similar to Theorems 1(a) and 2(a) in our paper, but the conditions on the design matrix $X$ imposed by Zhang are different from the ones imposed here. Furthermore, the conditions in Zhang are not directly comparable so it is difficult to say whether our conditions are stronger or weaker than his.

- Finally, in the special cases $q = 0$ and $q = 1$, subsequent work by Rigollet and Tsybakov [30] has yielded sharper results on the prediction error (compare our Theorems 3 and 4 to equations (5.24) and (5.25) in their paper). They explicitly take effects of the rank of $X$ into account, yielding tighter rates in the case $\text{rank}(X) \ll n$. In contrast, our results are based on the assumption $\text{rank}(X) = n$, which holds in many cases of interest.

### F. Role of conditions on $X$

In this subsection, we discuss the conditions on the design matrix $X$ involved in our analysis, and the different roles that they play in upper/lower bounds and different losses.

*1) Upper and lower bounds require complementary conditions:* It is worth noting that the minimax rates for $\ell_2$-prediction error and $\ell_2$-norm error are essentially the same except that the design matrix structure enters minimax rates in *very different ways*. In particular, note that proving lower bounds on prediction error for $q > 0$ 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 error 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 $\ell_2$-norm losses, where lower bounds for $q > 0$ can be proved with only Assumption 1 on $X$ (see Theorem 1), whereas upper bounds require both Assumptions 1 and 2.

In order to appreciate the difference between the conditions for $\ell_2$-prediction error and $\ell_2$ 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 rate in $\ell_2$-prediction error scales as $\Theta(\frac{1}{n})$. This scaling implies that the upper bound (23) from Theorem 4 holds (but is not tight). It is trivial to prove the correct upper bounds for prediction error using an alternative approach. [4] Consequently, this highly degenerate design matrix yields a very easy problem for $\ell_2$-prediction, since the $1/n$ rate is essentially low-dimensional parametric. In sharp contrast, for the case of $\ell_2$-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 $\ell_2$-error is infinite. In this case, the lower bound (13) on $\ell_2$-error from Theorem 1 holds (and is tight, since the kernel diameter is infinite). In contrast, the upper bound (16) on $\ell_2$-error from Theorem 2(b) does not apply, because Assumption 3(b) is violated due to the extreme degeneracy of the design matrix.

*2) Comparison to conditions required for $\ell_1$-based methods:* Naturally, our work also has some connections to the vast body of work on $\ell_1$-based methods for sparse estimation, particularly for the case of hard sparsity ($q = 0$). Based on our results, the rates that are achieved by $\ell_1$-methods, such as the Lasso and the Dantzig selector, are minimax optimal up to constant factors for $\ell_2$-norm loss, and $\ell_2$-prediction loss. However the bounds on $\ell_2$-error and $\ell_2$-prediction error for the Lasso and Dantzig selector require different conditions on the design matrix. We compare the conditions that we impose in our minimax analysis in Theorem 2(b) to various conditions imposed in the analysis of $\ell_1$-based methods, including the restricted isometry property of Candes and Tao [8], the restricted eigenvalue condition imposed in Meinshausen and Yu [25], the partial Riesz condition in Zhang and Huang [41] and the restricted eigenvalue condition of Bickel et al. [4]. We find that in the case where $s$ is known, "optimal" methods which are based on minimizing least-squares directly over the $\ell_0$-ball, can succeed for design matrices where $\ell_1$-based methods are not known to work for $q = 0$, as we discuss at more length in Section II-F2 to follow. As noted

[4]Note that the lower bound (21) on the $\ell_2$-prediction error from Theorem 3 does not apply to this model, since this degenerate design matrix with identical columns does not satisfy Assumption 3(b).

by a reviewer, unlike the direct methods that we have considered, $\ell_1$-based methods typically do not assume any prior knowledge of the sparsity index, but they do require knowledge or estimation of the noise variance.

One set of conditions, known as the restricted isometry property [8] or RIP for short, is based on very strong constraints on the condition numbers of all sub-matrices of $X$ up to size $2s$, requiring that they be near-isometries (i.e., with condition numbers 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 [41] 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 [25] 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 [25] are weaker or stronger than the conditions in Zhang and Huang [41]. Bickel et al. [4] show that their restricted eigenvalue condition is less severe than both the RIP condition [8] and an earlier set of restricted eigenvalue conditions due to Meinshausen and Yu [25].

Here we state a restricted eigenvalue condition that is very closely related to the condition imposed in Bickel et. al [4], and as shown by Negahban et. al [26], and is sufficient for bounding the $\ell_2$-error in the Lasso algorithm. In particular, for a given subset $S \subset \{1, \ldots, d\}$ and constant $\alpha \geq 1$, let us define the set

$$\mathcal{C}(S;\alpha) := \big\{\theta \in \mathbb{R}^d \mid \|\theta_{S^c}\|_1 \leq \alpha \, \|\theta_S\|_1 + 4\|\beta^*_{S^c}\|_1\big\}, \tag{24}$$

where $\beta^*$ is the true parameter. Note that for $q = 0$, the term $\|\beta^*_{S^c}\|_1 = 0$ which is very closely related to the restricted eigenvalue condition in Bickel et al. [4], while for $q \in (0,1]$, this term is non-zero. With this notation, the restricted eigenvalue condition in Negahban et al. [26] can be stated as follows: there exists a constant $\kappa > 0$ such that

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \geq \kappa \|\theta\|_2 \quad \text{for all } \theta \in \mathcal{C}(S;3).$$

Negahban et. al [26] show that under this restricted eigenvalue condition, the Lasso estimator has squared $\ell_2$-error upper bounded by $\mathcal{O}\big(R_q(\frac{\log d}{n})^{1-q/2}\big)$. (To be clear, Negahban et al. [26] study a more general class of $M$-estimators, and impose a condition known as restricted strong convexity; however, it reduces to an RE condition in this special case.) For the case $q \in (0,1]$, the analogous restricted lower eigenvalue condition we impose is Assumption 2. Recall that this states that for $q \in (0,1]$, the eigenvalues restricted to the set

$$\{\theta \in \mathbb{R}^d \mid \theta \in \mathbb{B}_q(2R_q) \;\text{ and }\; \|\theta\|_2 \geq f_\ell(R_q,n,d)\}$$

remain bounded away from zero. Both conditions impose lower bounds on the restricted eigenvalues over sets of weakly sparse vectors.

*3) Comparison with restricted eigenvalue condition:* It is interesting to compare the restricted eigenvalue condition in Bickel et al. [4] with the condition underlying Theorem 2, namely Assumption 3(b). In the case $q = 0$, the condition required by the estimator that performs least-squares over the $\ell_0$-ball—namely, the form of Assumption 3(b) used in Theorem 2(b)—is not stronger than the restricted eigenvalue condition in Bickel et al. [4]. This fact was previously established by Bickel et al. (see p.7, [4]). We now provide a simple pedagogical example to show that the $\ell_1$-based relaxation can fail to recover the true parameter while the optimal $\ell_0$-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 $\Delta \in \mathcal{C}(S;3)$ but $\Delta \notin \mathbb{B}_0(2)$. All the $2 \times 2$ 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. [12] (see, in particular, their Lemma 3.1), the condition $\mathbb{B}_0(2) \cap \ker(X) = \{0\}$ implies that (in the noiseless setting $w = 0$), the $\ell_0$-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 the Lasso were applied to this problem with no noise, it would incorrectly recover the solution $\widehat{\beta} := \begin{bmatrix} 0 & -1/3 & -1/3 \end{bmatrix}$, since $\|\widehat{\beta}\|_1 = 2/3 < 1 = \|\beta^*\|_1$.

Although this example is low-dimensional with $(s,d) = (1,3)$, higher-dimensional examples of design matrices that satisfy the conditions required for the minimax rate but not satisfied for $\ell_1$-based methods may be constructed using similar arguments. This construction highlights that there are instances of design matrices $X$ for which $\ell_1$-based methods fail to recover the true parameter $\beta^*$ for $q = 0$ while the optimal $\ell_0$-based algorithm succeeds.

In summary, for the hard sparsity case $q = 0$, methods based on $\ell_1$-relaxation can achieve the minimax optimal rate $\mathcal{O}\big(\frac{s \log d}{n}\big)$ for $\ell_2$-error. However the current analyses of these $\ell_1$-methods [4, 8, 25, 34] are based on imposing stronger conditions on the design matrix $X$ than those required by the estimator that performs least-squares over the $\ell_0$-ball with $s$ known.

## III. 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.

### A. Basic steps for lower bounds

The proofs for the lower bounds follow an information-theoretic method based on Fano's inequality [13], as used in classical work on nonparametric estimation [20, 39, 40]. A key ingredient is a sharp characterization of the metric entropy structure of $\ell_q$ balls [10, 21]. At a high-level, the proof of each lower bound follows three basic steps. The first two steps are general and apply to all the lower bounds in this paper, while the third is different in each case:

(1) In order to lower bound the minimax risk in some norm $\|\cdot\|_*$, we let $M(\delta_n, \mathbb{B}_q(R_q))$ be the cardinality of a maximal packing of the ball $\mathbb{B}_q(R_q)$ in the norm $\|\cdot\|_*$, say with elements $\{\beta^1, \ldots, \beta^M\}$. A precise definition of a packing set is provided in the next section. A standard argument (e.g., [19, 39, 40]) yields a lower bound on the minimax rate in terms of the error in a multi-way hypothesis testing problem: in particular, the probability $\mathbb{P}\big[\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \|\widehat{\beta} - \beta\|_*^2 \geq \delta_n^2/4\big]$ is at most $\min_{\widetilde{\beta}} \mathbb{P}[\widetilde{\beta} \neq B]$, 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.

(2) The next step is to derive a lower bound on $\mathbb{P}[B \neq \widetilde{\beta}]$; in this paper, we make use of Fano's inequality [13]. Since $B$ is uniformly distributed over the packing set, we have

$$\mathbb{P}[B \neq \widetilde{\beta}] \geq 1 - \frac{I(y; B) + \log 2}{\log M(\delta_n, \mathbb{B}_q(R_q))},$$

where $I(y; B)$ is the mutual information between random parameter $B$ in the packing set and the observation vector $y \in \mathbb{R}^n$. (Recall that for two random variables $X$ and $Y$, the mutual information is given by $I(X, Y) = \mathbb{E}_Y[D(\mathbb{P}_{X|Y} \,\|\, \mathbb{P}_X)]$.) The distribution $\mathbb{P}_{Y|B}$ is the conditional distribution of $Y$ on $B$, where $B$ is the uniform distribution on $\beta$ over the packing set and $Y$ is the gaussian distribution induced by model (1).

(3) The final and most challenging step involves upper bounding $I(y; B)$ so that $\mathbb{P}[\widetilde{\beta} \neq B] \geq 1/2$. For each lower bound, the approach to upper bounding $I(y; B)$ is slightly different. Our proof for $q = 0$ is based on Generalized Fano method [18], whereas for the case $q \in (0, 1]$, we upper bound $I(y; B)$ by a more intricate technique introduced by Yang and Barron [39]. We derive an upper bound on the $\epsilon_n$-covering set for $\mathbb{B}_q(R_q)$ with respect to the $\ell_2$-prediction semi-norm. Using Lemma 3 in Section III-C2 and the column normalization condition (Assumption 1), we establish a link between covering numbers in $\ell_2$-prediction semi-norm to covering numbers in $\ell_2$-norm. Finally, we choose the free parameters $\delta_n > 0$ and $\epsilon_n > 0$ so as to optimize the lower bound.

### B. Basic steps for upper bounds

The proofs for the upper bounds involve direct analysis of the natural estimator that performs least-squares over the $\ell_q$-ball:

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

The proof is constructive and involves two steps, the first of which is standard while the second step is more specific to each problem:

(1) Since $\|\beta^*\|_q^q \leq R_q$ by assumption, it is feasible for the least-squares problem, meaning that we have $\|y - X\widehat{\beta}\|_2^2 \leq \|y - X\beta^*\|_2^2$. Defining the error vector $\widehat{\Delta} = \widehat{\beta} - \beta^*$ and performing some algebra, we obtain the inequality

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

(2) The second and more challenging step involves computing upper bounds on the supremum of the Gaussian process over $\mathbb{B}_q(2R_q)$, which allows us to upper bound $\frac{|w^T X\widehat{\Delta}|}{n}$. 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 $X : \mathbb{R}^d \to \mathbb{R}^n$. We also make use of some chaining and peeling results from empirical process theory (e.g., van de Geer [33]). For upper bounds in $\ell_2$-norm error (Theorem 2), Assumptions 2 for $q > 0$

and 3(b) for $q = 0$ are used to upper bound $\|\widehat{\Delta}\|_2^2$ in terms of $\frac{1}{n}\|X\widehat{\Delta}\|_2^2$.

### C. *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 $\ell_q$-balls in $\ell_2$ metric.

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

(a) An $\epsilon$-covering of $\mathcal{S}$ in the metric $\rho$ 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 $\epsilon$-covering number $N(\epsilon; \mathcal{S}, \rho)$ is the cardinality of the smallest $\epsilon$-covering.

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

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., [27]). 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 [10, 21, 31], are the (dyadic) entropy numbers $\epsilon_k(\mathcal{S}; \rho)$, defined as follows for $k = 1, 2, \ldots$

$$\epsilon_k(\mathcal{S}; \rho) := \inf\big\{\epsilon > 0 \;|\; N(\epsilon; \mathcal{S}, \rho) \leq 2^{k-1}\big\}. \quad (25)$$

By definition, note that we have $\epsilon_k(\mathcal{S}; \rho) \leq \delta$ if and only if $\log_2 N(\delta; \mathcal{S}, \rho) \leq k$. For the remainder of this paper, the only metric used will be $\rho = \ell_2$, so to simplify notation, we denote the $\ell_2$-packing and covering numbers by $M(\epsilon; \mathcal{S})$ and $N(\epsilon; \mathcal{S})$.

*1) Metric entropies of $\ell_q$-balls:* Central to our proofs is the metric entropy of the ball $\mathbb{B}_q(R_q)$ when the metric $\rho$ is the $\ell_2$-norm, a quantity which we denote by $\log N(\epsilon; \mathbb{B}_q(R_q))$. 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 [17, 21]; see Appendix B for the details. All bounds stated here apply to a dimension $d \geq 2$.

**Lemma 2.** *For $q \in (0, 1]$ there is a constant $U_q$, depending only on $q$, such that for all $\epsilon \in [U_q {R_q}^{1/q} \big(\frac{\log d}{d}\big)^{\frac{2-q}{2q}}, {R_q}^{1/q}]$, we have*

$$\log N(\epsilon; \mathbb{B}_q(R_q)) \;\leq\; U_q \; \Big({R_q}^{\frac{2}{2-q}} \big(\frac{1}{\epsilon}\big)^{\frac{2q}{2-q}} \log d\Big). \quad (26)$$

*Conversely, suppose in addition that $\epsilon < 1$ and $\epsilon^2 = \Omega\big(R_q^{2/(2-q)} \frac{\log d}{d^{\nu}}\big)^{1-\frac{q}{2}}$ for some fixed $\nu \in (0, 1)$, depending only on q. Then there is a constant $L_q \leq U_q$, depending only on q, such that*

$$\log N(\epsilon; \mathbb{B}_q(R_q)) \;\geq\; L_q \; \Big({R_q}^{\frac{2}{2-q}} \big(\frac{1}{\epsilon}\big)^{\frac{2q}{2-q}} \log d\Big). \quad (27)$$

**Remark:** In our application of the lower bound (27), our typical choice of $\epsilon^2$ will be of the order $\mathcal{O}\big(\frac{\log d}{n}\big)^{1-\frac{q}{2}}$. It can be verified that under the condition (11) from Section II-B, we are guaranteed that $\epsilon$ lies in the range required for the upper and lower bounds (26) and (27) to be valid.

*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 $\beta$ and $\beta'$ 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 $\mu$, 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

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

Note that the KL-divergence is proportional to the squared prediction semi-norm. Hence control of KL-divergences are equivalent up to constant to control of the prediction semi-norm. 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, we define the set

$$\text{absconv}_q(X/\sqrt{n}) := \Big\{\frac{1}{\sqrt{n}}\sum_{j=1}^{d} \theta_j X_j \;|\; \theta \in \mathbb{B}_q(R_q)\Big\}. \quad (28)$$

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 $\ell_2$-norm is essentially no larger than the metric entropy of $\mathbb{B}_q(R_q)$, as summarized in the following

**Lemma 3.** *Suppose that $X$ satisfies the column normalization condition (Assumption 1 with constant $\kappa_c$) and $\epsilon \in [U_q {R_q}^{1/q} \big(\frac{\log d}{d}\big)^{\frac{2-q}{2q}}, {R_q}^{1/q}]$. Then there is a constant $U'_q$ depending only on $q \in (0,1]$ such that*

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

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 $\log N(\epsilon; \mathbb{B}_q(R_q))$ from Lemma 2.

### D. *Proof of lower bounds*

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

*1) Proof of Theorem 1:* Recall that for $\ell_2$-norm error, 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 $\ell_q$-ball, and the other correspond to the metric entropy of the $\ell_q$-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 rate is lower bounded as

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \|\widehat{\beta} - \beta\|_2^2 \geq \min_{\widehat{\beta}} \max_{\beta \in \mathcal{N}_q(X)} \|\widehat{\beta} - \beta\|_2^2,$$

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, there always exists an element $\beta \in \mathcal{N}_q(X)$ such that $\|\widehat{\beta} - \beta\|_2 \geq \frac{1}{2} \text{diam}_2(\mathcal{N}_q(X))$. Indeed, if $\|\widehat{\beta}\|_2 \geq \frac{1}{2}\text{diam}_2(\mathcal{N}_q(X))$, then the adversary chooses $\beta = 0 \in \mathcal{N}_q(X)$. On the other hand, if $\|\widehat{\beta}\|_2 \leq \frac{1}{2}\text{diam}_2(\mathcal{N}_q(X))$, then there exists $\beta \in \mathcal{N}_q(X)$ such that $\|\beta\|_2 = \text{diam}_2(\mathcal{N}_q(X))$. By triangle inequality, we then have

$$\|\beta - \widehat{\beta}\|_2 \geq \|\beta\|_2 - \|\widehat{\beta}\|_2 \geq \frac{1}{2}\text{diam}_2(\mathcal{N}_q(X)).$$

Overall, we conclude that

$$\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \|\widehat{\beta} - \beta\|_2^2 \geq \big\{\frac{1}{2}\text{diam}_2(\mathcal{N}_q(X))\big\}^2.$$

In the following subsections, we follow steps (1)–(3) outlined earlier so as to obtain the second term in our lower bounds on the $\ell_2$-norm error and the $\ell_2$-prediction error. As has already been mentioned, steps (1) and (2) are general, whereas step (3) is different in each case.

*a) Proof of Theorem 1(a):* Let $M(\delta_n, \mathbb{B}_q(R_q))$ be the cardinality of a maximal packing of the ball $\mathbb{B}_q(R_q)$ in the $\ell_2$ metric, say with elements $\{\beta^1, \ldots, \beta^M\}$. Then, by the standard arguments referred to earlier in step (1), we have

$$\mathbb{P}\big[\min_{\widehat{\beta}} \max_{\beta \in \mathbb{B}_q(R_q)} \|\widehat{\beta} - \beta\|_2^2 \geq \ \delta_n^2/4\big] \geq \min_{\widetilde{\beta}} \mathbb{P}[\widetilde{\beta} \neq B],$$

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 (step (2)) yields the lower bound

$$\mathbb{P}[B \neq \widetilde{\beta}] \geq 1 - \frac{I(y; B) + \log 2}{\log M(\delta_n, \mathbb{B}_q(R_q))}, \tag{29}$$

where $I(y; B)$ 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 (step (3)); we do so using a procedure due to Yang and Barron [39]. It 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 3, 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 $N(\epsilon_n; \mathbb{B}_q(R_q))$ be the minimal cardinality of an $\epsilon_n$-covering of $\mathbb{B}_q(R_q)$ in $\ell_2$-norm. Using the upper bound on the metric entropy of $\text{absconv}_q(X)$ provided by Lemma 3, we conclude that there exists a set $\{X\beta^1, \ldots, X\beta^N\}$ such that for all $X\beta \in \text{absconv}_q(X)$, there exists some index $i$ such that $\|X(\beta - \beta^i)\|_2/\sqrt{n} \leq c\,\kappa_c\,\epsilon_n$ for some $c > 0$. Following the argument of Yang and Barron [39], we obtain that the mutual information is upper bounded as

$$I(y; B) \leq \log N(\epsilon_n; \mathbb{B}_q(R_q)) + \frac{c^2\, n}{\sigma^2} \kappa_c^2 \epsilon_n^2.$$

Combining this upper bound with the Fano lower bound (29) yields

$$\mathbb{P}[B \neq \widetilde{\beta}] \geq 1 - \frac{\log N(\epsilon_n; \mathbb{B}_q(R_q)) + \frac{c^2\, n}{\sigma^2} \kappa_c^2\, \epsilon_n^2 + \log 2}{\log M(\delta_n; \mathbb{B}_q(R_q))}. \tag{30}$$

The final step is to choose the packing and covering radii ($\delta_n$ and $\epsilon_n$ respectively) such that the lower bound (30) is greater than $1/2$. In order to do so, suppose that we choose the pair $(\epsilon_n, \delta_n)$ such that

$$\frac{c^2\, n}{\sigma^2} \kappa_c^2\, \epsilon_n^2 \leq \log N(\epsilon_n, \mathbb{B}_q(R_q)), \tag{31a}$$

$$\log M(\delta_n, \mathbb{B}_q(R_q)) \geq 4 \log N(\epsilon_n, \mathbb{B}_q(R_q)). \tag{31b}$$

As long as $N(\epsilon_n, \mathbb{B}_q(R_q)) \geq 2$, we are then guaranteed that

$$\mathbb{P}[B \neq \widetilde{\beta}] \geq 1 - \frac{\log N(\epsilon_n, \mathbb{B}_q(R_q)) + \log 2}{4 \log N(\epsilon_n, \mathbb{B}_q(R_q))} \;\geq\; 1/2,$$

as desired.

It remains to determine choices of $\epsilon_n$ and $\delta_n$ that satisfy the relations (31). From Lemma 2, relation (31a) is satisfied by choosing $\epsilon_n$ such that $\frac{c^2\, n}{2\sigma^2} \kappa_c^2\, \epsilon_n^2 = L_q \left[ {R_q}^{\frac{2}{2-q}} \left(\frac{1}{\epsilon_n}\right)^{\frac{2q}{2-q}} \log d \right]$, or equivalently such that

$$\big(\epsilon_n\big)^{\frac{4}{2-q}} = \Theta\big( {R_q}^{\frac{2}{2-q}} \frac{\sigma^2}{\kappa_c^2} \frac{\log d}{n} \big).$$

In order to satisfy the bound (31b), it suffices to choose $\delta_n$ such that

$$U_q \left[ {R_q}^{\frac{2}{2-q}} \big(\frac{1}{\delta_n}\big)^{\frac{2q}{2-q}} \log d \right] \geq 4 L_q \left[ {R_q}^{\frac{2}{2-q}} \big(\frac{1}{\epsilon_n}\big)^{\frac{2q}{2-q}} \log d \right],$$

or equivalently such that

$$\begin{aligned} \delta_n^2 &\leq \big[\frac{U_q}{4L_q}\big]^{\frac{2-q}{q}} \left\{ \big(\epsilon_n\big)^{\frac{4}{2-q}} \right\}^{\frac{2-q}{2}} \\ &= \big[\frac{U_q}{4L_q}\big]^{\frac{2-q}{q}}\; {L_q}^{\frac{2-q}{2}}\; R_q \left[ \frac{\sigma^2}{\kappa_c^2} \frac{\log d}{n} \right]^{\frac{2-q}{2}} \end{aligned}$$

Substituting into equation (12), we obtain

$$\mathbb{P}\Big[ \mathcal{M}_2(\mathbb{B}_q(R_q), X) \geq c_q\; R_q\; \big(\frac{\sigma^2}{\kappa_c^2} \frac{\log d}{n}\big)^{1-\frac{q}{2}} \Big] \geq \frac{1}{2},$$

for some absolute constant $c_q$. This completes the proof of Theorem 1(a).

*b) 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

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

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 $z'$. Next we require the following result:

**Lemma 4.** *For $d, s$ even and $s < 2d/3$, 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 $z, z' \in \widetilde{\mathcal{H}}$.*

Note that if $d$ and/or $s$ is odd, we can embed $\widetilde{\mathcal{H}}$ into a $d-1$ and/or $s-1$-dimensional hypercube and the result holds. Although results of this type are known (e.g., see Lemma 4, [6]), for completeness, we provide a proof of Lemma 4 in Appendix D. 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 $\beta, \beta' \in \sqrt{\frac{2}{s}} \delta_n \widetilde{\mathcal{H}}$, we have

$$\frac{2}{s} \delta_n^2 \times \rho_H(\beta, \beta') \leq \|\beta - \beta'\|_2^2 \leq \frac{8}{s} \delta_n^2 \times \rho_H(\beta, \beta').$$

Therefore by applying Lemma 4 and noting that $\rho_H(\beta, \beta') \leq s$ for all $\beta, \beta' \in \widetilde{\mathcal{H}}$, we have the following bounds on the $\ell_2$-norm of their difference for all elements $\beta, \beta' \in \sqrt{\frac{2}{s}} \delta_n \widetilde{\mathcal{H}}$:

$$\|\beta - \beta'\|_2^2 \geq \delta_n^2, \quad \text{and} \tag{32a}$$

$$\|\beta - \beta'\|_2^2 \leq 8 \delta_n^2. \tag{32b}$$

Consequently, the rescaled set $\sqrt{\frac{2}{s}} \delta_n \widetilde{\mathcal{H}}$ is an $\delta_n$-packing set of $\mathbb{B}_0(s)$ in $\ell_2$ norm with $M(\delta_n, \mathbb{B}_0(s)) = |\widetilde{\mathcal{H}}|$ 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 (29) (steps (1) and (2)).

At this point, we use an alternative upper bound on the mutual information (step (3)), 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 [13]. 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 (32b), we conclude that

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

Substituting this upper bound into the Fano lower bound (29), we obtain

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

Setting $\delta_n^2 = \frac{1}{16} \frac{\sigma^2}{\kappa_u^2} \frac{s}{2n} \log \frac{d-s}{s/2}$ ensures that this probability is at least $1/2$. Consequently, combined with the lower bound (12), we conclude that

$$\mathbb{P}\Big[ \mathcal{M}_2(\mathbb{B}_0(s), X) \geq \frac{1}{16}\; \big( \frac{\sigma^2}{\kappa_u^2} \frac{s}{2n} \log \frac{d-s}{s/2} \big) \Big] \geq 1/2.$$

As long as $d/s \geq 3/2$, we are guaranteed that $\log(d/s - 1) \geq c \log(d/s)$ for some constant $c > 0$, from which the result follows.

*2) Proof of Theorem 3:* We use arguments similar to the proof of Theorem 1 in order to establish lower bounds on prediction error $\|X(\widehat{\beta} - \beta^*)\|_2/\sqrt{n}$.

*a) Proof of Theorem 3(a):* For some universal constant $\bar{c} > 0$ to be chosen, define

$$\delta_n^2 := \bar{c}\, R_q\, \Big(\frac{\sigma^2}{\kappa_c^2}\,\frac{\log d}{n}\Big)^{1-q/2}, \tag{33}$$

and let $\{\beta^1, \ldots, \beta^M\}$ be an $\delta_n$ packing of the ball $\mathbb{B}_q(R_q)$ in the $\ell_2$ metric, say with a total of $M(\delta_n; \mathbb{B}_q(R_q))$ elements. We first show that if $n$ is sufficiently large, then this set is also a $\kappa_\ell \delta_n$-packing set in the prediction (semi)-norm. From the theorem assumptions, we may choose universal constants $c_1, c_2$ such that $f_\ell(R_q, n, d) \leq c_2 R_q\, \big(\frac{\log d}{n}\big)^{1-q/2}$ and $R_q\, \big(\frac{\log d}{n}\big)^{1-q/2} < c_1$. From Assumption 2, for each $i \neq j$, we are guaranteed that

$$\frac{\|X(\beta^i - \beta^j)\|_2}{\sqrt{n}} \geq \kappa_\ell \|\beta^i - \beta^j\|_2, \tag{34}$$

as long as $\|\beta^i - \beta^j\|_2 \geq f_\ell(R_q, n, d)$. Consequently, for any fixed $\bar{c} > 0$, we are guaranteed that

$$\|\beta^i - \beta^j\|_2 \overset{(i)}{\geq} \delta_n \overset{(ii)}{\geq} c_2 R_q\, \Big(\frac{\log d}{n}\Big)^{1-q/2}.$$

where inequality (i) follows since $\{\beta^j\}_{j=1}^M$ is a $\delta_n$-packing set. Here step (ii) follows because the theorem conditions imply that

$$R_q\Big(\frac{\log d}{n}\Big)^{1-q/2} \leq \sqrt{c_1}\Big[R_q\Big(\frac{\log d}{n}\Big)^{1-q/2}\Big]^{1/2},$$

and we may choose $c_1$ as as small as we please. (Note that all of these statements hold for an arbitrarily small choice of $\bar{c} > 0$, which we will choose later in the argument.) Since $f_\ell(R_q, n, d) \leq c_2 R_q\, \big(\frac{\log d}{n}\big)^{1-q/2}$ by assumption, the lower bound (34) guarantees that $\{\beta^1, \beta^2, \ldots, \beta^M\}$ form a $\kappa_\ell \delta_n$-packing set in the prediction (semi)-norm $\|X(\beta^i - \beta^j)\|_2$.

Given this packing set, we now follow a standard approach, as in the proof of Theorem 1(a), 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

$$\mathbb{P}[XB \neq X\widetilde{\beta}] \geq 1 - \frac{I(y; XB) + \log 2}{\log M(\delta_n; \mathbb{B}_q(R_q))},$$

where $I(y; XB)$ now represents the mutual information[5] between random parameter $XB$ (uniformly distributed over the packing set) and the observation vector $y \in \mathbb{R}^n$.

From Lemma 3, the $\kappa_c\, \epsilon$-covering number of the set $\operatorname{absconv}_q(X)$ is upper bounded (up to a constant factor) by the $\epsilon$ covering number of $\mathbb{B}_q(R_q)$ in $\ell_2$-norm, which we denote by $N(\epsilon_n; \mathbb{B}_q(R_q))$. Following the same reasoning as in Theorem 2(a), the mutual information is upper bounded as

$$I(y; XB) \leq \log N(\epsilon_n; \mathbb{B}_q(R_q)) + \frac{n}{2\sigma^2}\kappa_c^2\, \epsilon_n^2.$$

Combined with the Fano lower bound, $\mathbb{P}[XB \neq X\widetilde{\beta}]$ is lower bounded by

$$1 - \frac{\log N(\epsilon_n; \mathbb{B}_q(R_q)) + \frac{n}{\sigma^2}\,\kappa_c^2\,\epsilon_n^2 + \log 2}{\log M(\delta_n; \mathbb{B}_q(R_q))}. \tag{35}$$

Lastly, we choose the packing and covering radii ($\delta_n$ and $\epsilon_n$ respectively) such that the lower bound (35) remains bounded below by $1/2$. As in the proof of Theorem 1(a), it suffices to choose the pair $(\epsilon_n, \delta_n)$ to satisfy the relations (31a) and (31b). The same choice of $\epsilon_n$ ensures that relation (31a) holds; moreover, by making a sufficiently small choice of the universal constant $\bar{c}$ in the definition (33) of $\delta_n$, we may ensure that the relation (31b) also holds. Thus, as long as $N_2(\epsilon_n) \geq 2$, we are then guaranteed that

$$\begin{aligned}\mathbb{P}[XB \neq X\widetilde{\beta}] &\geq 1 - \frac{\log N(\delta_n; \mathbb{B}_q(R_q)) + \log 2}{4 \log N(\delta_n; \mathbb{B}_q(R_q))}\\ &\geq 1/2,\end{aligned}$$

as desired.

*b) Proof of Theorem 3(b):* Recall the assertion of Lemma 4, which guarantees the existence of a set $\frac{\delta_n^2}{2s}\widetilde{\mathcal{H}}$ is an $\delta_n$-packing set in $\ell_2$-norm with $M(\delta_n; \mathbb{B}_q(R_q)) = |\widetilde{\mathcal{H}}|$ elements, say $\{\beta^1, \ldots, \beta^M\}$, such that the bounds (32a) and (32b) 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 3(a), we have

$$\frac{\|X(\beta^i - \beta^j)\|}{\sqrt{n}} \leq \kappa_u \|\beta^i - \beta^j\|_2 \leq \kappa_u \sqrt{8}\, \delta_n. \tag{36}$$

In the reverse direction, since Assumption 3(b) holds, we have

$$\frac{\|X(\beta^i - \beta^j)\|_2}{\sqrt{n}} \geq \kappa_{0,\ell}\delta_n. \tag{37}$$

We can follow the same steps as in the proof of Theorem 1(b), thereby obtaining an upper bound the mutual

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

information of the form $I(y; XB) \leq 8\kappa_u^2 n\delta_n^2$. 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 $\kappa_\ell$ from the lower bound (37), we obtain the lower bound

$$\mathbb{P}\Big[\mathcal{M}_n(\mathbb{B}_0(s), X) \geq c'_{0,q}\,\kappa_\ell^2\,\frac{\sigma^2}{\kappa_u^2}\,s\log\frac{d-s}{s/2}\Big] \geq \frac{1}{2}.$$

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.

### *E. 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 rates. We prove all parts of these theorems by analyzing the family of $M$-estimators

$$\widehat{\beta} \in \arg\min_{\|\beta\|_q^q \leq R_q} \|y - X\beta\|_2^2. \tag{38}$$

Note that (38) is a non-convex optimization problem for $q \in [0, 1)$, so it is not an algorithm that would be implemented in practice. Step (1) for upper bounds provided above implies that if $\widehat{\Delta} = \widehat{\beta} - \beta^*$, then

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \frac{2|w^T X\widehat{\Delta}|}{n}. \tag{39}$$

The remaining sections are devoted to step (2), which involves controlling $\frac{|w^T X\widehat{\Delta}|}{n}$ for each of the upper bounds.

*1) Proof of Theorem 2:* We begin with upper bounds on the minimax rate in squared $\ell_2$-norm.

*a) Proof of Theorem 2(a):* Recall that this part of the theorem deals with the case $q \in (0, 1]$. We split our analysis into two cases, depending on whether the error $\|\widehat{\Delta}\|_2$ is smaller or larger than $f_\ell(R_q, n, d)$.

**Case 1:** First, suppose that $\|\widehat{\Delta}\|_2 < f_\ell(R_q, n, d)$. Recall that the theorem is based on the assumption $R_q\left(\frac{\log d}{n}\right)^{1-q/2} < c_2$. As long as the constant $c_2 \ll 1$ is sufficiently small (but still independent of the triple $(n, d, R_q)$), we can assume that

$$c_1 R_q\left(\frac{\log d}{n}\right)^{1-q/2} \leq \sqrt{R_q}\Big[\frac{\kappa_c^2}{\kappa_\ell^2}\frac{\sigma^2}{\kappa_\ell^2}\frac{\log d}{n}\Big]^{1/2-q/4}.$$

This inequality, combined with the assumption $f_\ell(R_q, n, d) \leq c_1 R_q\left(\frac{\log d}{n}\right)^{1-q/2}$ imply that the error $\|\widehat{\Delta}\|_2$ satisfies the bound (15) for all $\bar{c} \geq 1$.

**Case 2:** Otherwise, we may assume that $\|\widehat{\Delta}\|_2 > f_\ell(R_q, n, d)$. In this case, Assumption 2 implies that $\frac{\|X\widehat{\Delta}\|_2^2}{n} \geq \kappa_\ell^2\|\widehat{\Delta}\|_2^2$, and hence, in conjunction with the inequality (39), we obtain

$$\kappa_\ell^2\|\widehat{\Delta}\|_2^2 \;\leq\; 2|w^T X\widehat{\Delta}|/n \;\leq;\frac{2}{n}\|w^T X\|_\infty\|\widehat{\Delta}\|_1.$$

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 vector 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

$$\kappa_\ell^2\|\widehat{\Delta}\|_2^2 \leq 2\sigma\kappa_c\sqrt{\frac{3\log d}{n}}\,\|\widehat{\Delta}\|_1 \tag{40}$$

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

It remains to upper bound the $\ell_1$-norm in terms of the $\ell_2$-norm and a residual term. Since both $\widehat{\beta}$ and $\beta^*$ belong to $\mathbb{B}_q(R_q)$, we have $\|\widehat{\Delta}\|_q^q = \sum_{j=1}^d |\widehat{\Delta}_j|^q \leq 2R_q$. We exploit the following lemma:

**Lemma 5.** *For any vector $\theta \in \mathbb{B}_q(2R_q)$ and any positive number $\tau > 0$, we have*

$$\|\theta\|_1 \leq \sqrt{2R_q}\tau^{-q/2}\|\theta\|_2 + 2R_q\tau^{1-q}. \tag{41}$$

Although this type of result is standard (e.g, [14]), we provide a proof in Appendix E.

We can exploit Lemma 5 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 $g(x) = ax^2 + bx + c \leq 0$, with $a = 1$,

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

Since $g(0) = c < 0$ and the positive root of $g(x)$ 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).

*b) Proof of Theorem 2(b):* In order to establish the bound (16), we follow the same steps with $f_\ell(s, n, d) = 0$, thereby obtaining the following simplified form of the bound (40):

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

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 (16).

Establishing the sharper upper bound (17) requires more precise control on the right-hand side of the inequality (39). The following lemma, proved in Appendix F, provides this control:

**Lemma 6.** *Suppose that* $\frac{\|X\theta\|_2}{\sqrt{n}\|\theta\|_2} \leq \kappa_u$ *for all* $\theta \in \mathbb{B}_0(2s)$. *Then there are universal positive constants* $c_1, c_2$ *such that for any* $r > 0$*, we have*

$$\sup_{\|\theta\|_0 \leq 2s, \|\theta\|_2 \leq r} \frac{1}{n}|w^T X\theta| \leq 6\,\sigma\, r\, \kappa_u \sqrt{\frac{s\log(d/s)}{n}} \tag{42}$$

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

Lemma 6 holds for a fixed radius $r$, whereas we would like to choose $r = \|\widehat{\Delta}\|_2$, which is a random quantity. To extend Lemma 6 so that it also applies uniformly over an interval of radii (and hence also to a random radius within this interval), we use a "peeling" result, stated as Lemma 9 in Appendix H. In particular, consider the event $\mathcal{E}$ that there exists some $\theta \in \mathbb{B}_0(2s)$ such that

$$\frac{1}{n}\left|w^T X\theta\right| \geq 6\sigma\kappa_u\|\theta\|_2\sqrt{\frac{s\log(d/s)}{n}}. \tag{43}$$

Then we claim that

$$\mathbb{P}[\mathcal{E}] \leq \frac{2\exp(-c\,s\log(d/s))}{1 - \exp(-c\,s\log(d/s))}$$

for some $c > 0$. This claim follows from Lemma 9 in Appendix H by choosing the function $f(v; X) = \frac{1}{n}|w^T X v|$, the set $A = \mathbb{B}_0(2s)$, the sequence $a_n = n$, and the functions $\rho(v) = \|v\|_2$, and $g(r) = 6\sigma r\kappa_u\sqrt{\frac{s\log(d/s)}{n}}$. For any $r \geq \sigma\kappa_u\sqrt{\frac{s\log(d/s)}{n}}$, we are guaranteed that $g(r) \geq \sigma^2\kappa_u^2\frac{s\log(d/s)}{n}$, and $\mu = \sigma^2\kappa_u^2\frac{s\log(d/s)}{n}$, so that Lemma 9 may be applied. We use a similar peeling argument for two of our other achievability results.

Returning to the main thread, we have

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 6\,\sigma\,\|\widehat{\Delta}\|_2\,\kappa_u\sqrt{\frac{s\log(d/s)}{n}},$$

with high probability. By Assumption 3(b), we have $\|X\widehat{\Delta}\|_2^2/n \geq \kappa_\ell^2\|\widehat{\Delta}\|_2^2$. Canceling 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.

*2) Proof of Theorem 4:* We again make use of the elementary inequality (39) to establish upper bounds on the prediction error.

*a) 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 $\kappa_c$, 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 $\sigma^2$ recovers a bound on the original prediction loss.

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

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

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 G1. So as to lighten notation, let us introduce the shorthand $h(R_q, n, d) := \sqrt{R_q}\,(\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}}$.

**Lemma 7.** *For* $q \in (0, 1)$*, suppose that there is a universal constant* $c_1$ *such that* $h(R_q, n, d) < c_1 < 1$*. Then there are universal constants* $c_i$*,* $i = 2, \ldots, 5$ *such that for any fixed radius* $r$ *with* $r \geq c_2\widetilde{\kappa_c}^{\frac{q}{2}}\,h(R_q, n, d)$*, we have*

$$\sup_{\substack{\theta\in\mathbb{B}_q(2R_q)\\ \frac{\|\widetilde{X}\theta\|_2}{\sqrt{n}} \leq r}} \frac{1}{n}\left|\widetilde{w}^T\widetilde{X}\theta\right| \leq c_3 r\,\widetilde{\kappa_c}^{\frac{q}{2}}\,\sqrt{R_q}\,(\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}},$$

*with probability greater than* $1 - c_4\exp(-c_5\,n\,h^2(R_q, n, d))$.

Once again, we require the peeling result (Lemma 9 from Appendix H) to extend Lemma 7 to hold for random radii. In this case, we define the event $\mathcal{E}$ as there exists some $\theta \in \mathbb{B}_q(2R_q)$ such that

$$\frac{1}{n}\left|\widetilde{w}^T\widetilde{X}\theta\right| \geq c_3\frac{\|\widetilde{X}\theta\|_2}{\sqrt{n}}\,\widetilde{\kappa_c}^{\frac{q}{2}}\,\sqrt{R_q}\,(\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}}.$$

By Lemma 9 with the choices $f(v; X) = |w^T X v|/n$, $A = \mathbb{B}_q(2R_q)$, $a_n = n$, $\rho(v) = \frac{\|Xv\|_2}{\sqrt{n}}$, and

$g(r) = c_3\, r\, \widetilde{\kappa_c}^{\frac{q}{2}} h(R_q, n, d)$, we have

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

Returning to the main thread, from the basic inequality (39), when the event $\mathcal{E}$ from equation (43) holds, we have

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

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

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

as claimed.

*b) Proof of Theorem 4(b):* For this part, we require the following lemma, proven in Appendix G2:

**Lemma 8.** *As long as $\frac{d}{2s} \geq 2$, then for any $r > 0$, we have*

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

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

By using a peeling technique (see Lemma 9 in Appendix H), we now extend the result to hold uniformly over all radii. Consider the event $\mathcal{E}$ that there exists some $\theta \in \mathbb{B}_0(2s)$ such that

$$\frac{1}{n}\, |w^T X\theta| \geq 9\sigma \frac{\|\widetilde{X}\theta\|_2}{\sqrt{n}} \sqrt{\frac{s\log(d/s)}{n}}\}.$$

We now apply Lemma 9 with the sequence $a_n = n$, the function $f(v;X) = \frac{1}{n}|w^T Xv|$, the set $A = \mathbb{B}_0(2s)$, and the functions

$$\rho(v) = \frac{\|Xv\|_2}{\sqrt{n}}, \text{ and } g(r) = 9\, r\, \widetilde{\kappa_c}^{\frac{q}{2}} \sqrt{\frac{s\log(d/s)}{n}}.$$

We take $r \geq \sigma\kappa_u \sqrt{\frac{s\log(d/s)}{n}}$, which implies that $g(r) \geq \sigma^2 \kappa_u^2 \frac{s\log(d/s)}{n}$, and $\mu = \sigma^2\kappa_u^2 \frac{s\log(d/s)}{n}$ in Lemma 9. Consequently, we are guaranteed that

$$\mathbb{P}[\mathcal{E}] \leq \frac{2\exp(-c\, s\log(d/s))}{1 - \exp(-c\, s\log(d/s))}.$$

Combining this tail bound with the basic inequality (39), 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.

## IV. Conclusion

The main contribution of this paper is to obtain optimal minimax rates of convergence for the linear model (1) under high-dimensional scaling, in which the sample size $n$ and problem dimension $d$ are allowed to scale, for general design matrices $X$. We provided matching upper and lower bounds for the $\ell_2$-norm and $\ell_2$-prediction loss, so that the optimal minimax rates are determined in these cases. To our knowledge, this is the first paper to present minimax optimal rates in $\ell_2$-prediction error for general design matrices $X$ and general $q \in [0,1]$. We also derive optimal minimax rates in $\ell_2$-error, with similar rates obtained in concurrent work by Zhang [42] under different conditions on $X$.

Apart from the rates themselves, our analysis highlights how conditions on the design matrix $X$ enter in complementary manners for the $\ell_2$-norm and $\ell_2$-prediction loss functions. On one hand, it is possible to obtain lower bounds on $\ell_2$-norm error (see Theorem 1) or upper bounds on $\ell_2$-prediction error (see Theorem 4) under very mild assumptions on $X$—in particular, our analysis requires only that the columns of $X/\sqrt{n}$ have bounded $\ell_2$-norms (see Assumption 1). On the other hand, in order to obtain upper bounds on $\ell_2$-norm error (Theorem 2) or lower bound on $\ell_2$-norm prediction error (Theorem 3), the design matrix $X$ must satisfy, in addition to column normalization, other more restrictive conditions. Indeed both lower bounds in prediction error and upper bounds in $\ell_2$-norm error require that elements of $\mathbb{B}_q(R_q)$ are well separated in prediction semi-norm $\|X(\cdot)\|_2/\sqrt{n}$. In particular, our analysis was based on imposed on a certain type of restricted lower eigenvalue condition on $X^TX$ measured over the $\ell_q$-ball, as formalized in Assumption 2. As shown by our results, this lower bound is intimately related to the degree of non-identifiability over the $\ell_q$-ball of the high-dimensional linear regression model. Finally, we note that similar techniques can be used to obtain minimax-optimal rates for more general problems of sparse non-parametric regression [28].

## Appendix

### A. *Proof of Proposition 1*

Under the stated conditions, Theorem 1 from Raskutti et al. [29] guarantees that

$$\frac{\|X\theta\|_2}{\sqrt{n}} \geq \frac{\lambda_{\min}(\Sigma^{1/2})}{4} \|\theta\|_2 - 9\,\big(\frac{\rho^2(\Sigma)\log d}{n}\big)^{1/2}\, \|\theta\|_1 \tag{44}$$

for all $\theta \in \mathbb{R}^d$, with probability greater than $1 - c_1 \exp(-c_2 n)$. When $\theta \in \mathbb{B}_q(2R_q)$, Lemma 5 guarantees that

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

for all $\tau > 0$. We now set $\tau = \sqrt{\frac{\log d}{n}}$ and substitute the result into the lower bound (44). Following some algebra, we find that $\frac{\|X\theta\|_2}{\sqrt{n}}$ is lower bounded by

$$\Big\{ \frac{\lambda_{\min}(\Sigma^{1/2})}{4} - 18 \rho(\Sigma)\, \sqrt{R_q}\, \big(\frac{\log d}{n}\big)^{1/2 - q/4} \Big\} \|\theta\|_2 - 18\, R_q \rho(\Sigma)\, \big(\frac{\log d}{n}\big)^{1-q/2}.$$

Consequently, as long as

$$\frac{\lambda_{\min}(\Sigma^{1/2})}{8} > 18 \rho(\Sigma)\, \sqrt{R_q}\, \big(\frac{\log d}{n}\big)^{1/2 - q/4},$$

then we are guaranteed that

$$\frac{\|X\theta\|_2}{\sqrt{n}} \geq \frac{\lambda_{\min}(\Sigma^{1/2})}{4} \|\theta\|_2 - 18\, R_q \rho(\Sigma)\, \big(\frac{\log d}{n}\big)^{1-q/2}$$

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

### B. Proof of Lemma 2

The result is obtained by inverting known results on (dyadic) entropy numbers of $\ell_q$-balls; there are some minor technical subtleties in performing the inversion. For a $d$-dimensional $\ell_q$ ball with $q \in (0,1]$, it is known [17, 21, 31] that for all integers $k \in [\log d, d]$, the dyadic entropy numbers $\epsilon_k$ of the ball $\mathbb{B}_q(1)$ with respect to the $\ell_2$-norm scale as

$$\epsilon_k(\mathbb{B}_q(1)) = C_q\, \Big(\frac{\log(1+\frac{d}{k})}{k}\Big)^{1/q - 1/2}. \tag{45}$$

Moreover, for $k \in [1, \log d]$, we have $\epsilon_k(\mathbb{B}_q(1)) \leq C_q$.

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

$$\epsilon_k(\mathbb{B}_q(1)) \leq C_q \big(\frac{\log(1+\frac{d}{2})}{k}\big)^{1/q-1/2} \leq C_q \big(\frac{\log d}{k}\big)^{1/q-1/2}.$$

Inverting this inequality for $k = \log N(\epsilon; \mathbb{B}_q(1))$ and allowing for a ball radius $R_q$ yields

$$\log N(\epsilon; \mathbb{B}_q(R_q)) \leq \big(C_q \, \frac{{R_q}^{1/q}}{\epsilon}\big)^{\frac{2q}{2-q}} \log d, \tag{46}$$

as claimed. The conditions $\epsilon \leq {R_q}^{1/q}$ and $\epsilon \geq C_q {R_q}^{1/q} \big(\frac{\log d}{d}\big)^{\frac{2-q}{2q}}$ ensure that $k \in [\log d, d]$.

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

$$C_q \big(\frac{\log(1+\frac{d}{k})}{k}\big)^{1/q-1/2} \geq C_q \big(\frac{\nu \log d}{k}\big)^{1/q-1/2}.$$

Accounting for the radius $R_q$ (as was done for the upper bound) yields

$$\log N(\epsilon; \mathbb{B}_q(R_q)) \geq \nu \big(\frac{C_q {R_q}^{1/q}}{\epsilon}\big)^{\frac{2q}{2-q}} \log d,$$

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 $\epsilon$. The condition $k \geq \log d$ is ensured by setting $\epsilon < 1$. Turning to the condition $k \leq d^{1-\nu}$, from the bound (46) on $k$, it suffices to choose $\epsilon$ such that $\big(\frac{C_q {R_q}^{1/q}}{\epsilon}\big)^{\frac{2q}{2-q}} \log d \leq d^{1-\nu}$. This condition is ensured by enforcing the lower bound $\epsilon^2 = \Omega\big({R_q}^{2/(2-q)} \frac{\log d}{d^{1-\nu}}\big)^{\frac{2-q}{q}}$ for some $\nu \in (0,1)$.

### C. Proof of Lemma 3

We deal first with (dyadic) entropy numbers, as previously defined (25), and show that $\epsilon_{2k-1}(\text{absconv}_q(X/\sqrt{n}),\ \|\cdot\|_2)$ is upper bounded by

$$c\,\kappa_c\ \min\Big\{1, \big(\frac{\log(1+\frac{d}{k})}{k}\big)^{\frac{1}{q}-\frac{1}{2}}\Big\}. \tag{47}$$

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 $k = 1, 2, \ldots$, $\epsilon_{2k-1}(\text{absconv}_q(X/\sqrt{n}),\ \|\cdot\|_2)$ is upper bounded as follows:

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

Using Corollary 2.4 from Carl and Pajor [9], $\epsilon_k(\text{absconv}_1(X/\sqrt{n}),\ \|\cdot\|_2)$ is upper bounded as follows:

$$\frac{c}{\sqrt{n}}\ |||X|||_{1\to 2}\ \min\Big\{1, \big(\frac{\log(1+\frac{d}{k})}{k}\big)^{1/2}\Big\},$$

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 $\epsilon_{2k-1}(\text{absconv}_q(\frac{X}{\sqrt{n}}), \|\cdot\|_2)$ is upper bounded by

$$c\,\kappa_c\;\min\Big\{1, \big(\frac{\log(1+\frac{d}{k})}{k}\big)^{\frac{1}{q}-\frac{1}{2}}\Big\},$$

as claimed. Finally, under the stated assumptions, we may invert the upper bound (47) by the same procedure as in the proof of Lemma 2 (see Appendix B), thereby obtaining the claim.

### D. Proof of Lemma 4

Our proof is inspired by related results from the approximation theory literature (see, e.g., Kühn [21] and Birgé and Massart [6]). For each even integer $s = 2, 4, 6, \ldots, d$, let us define the set

$$\mathcal{H} := \big\{z \in \{-1, 0, +1\}^d \;\mid\; \|z\|_0 = s\big\}. \tag{48}$$

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 $\rho_H$ 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

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

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

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

where the final inequality uses the fact that the ratio $\frac{d-s+j}{s/2+j}$ is decreasing as a function of $j$ (see Kühn [21] pp. 122–123 and Birgé and Massart [6], Lemma 4 for details).

### E. Proof of Lemma 5

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 $|\theta_j|/\tau \leq 1$ 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.

### F. Proof of Lemma 6

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

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

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

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

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 $\ell_2$-norm. We claim that we may find such a covering set $\{\theta^1, \ldots, \theta^N\} \subset \mathbb{S}(s,r)$ with cardinality $N = N(\epsilon, \mathbb{S}(s,r))$ that is upper bounded as

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

To establish this claim, note that here are $\binom{d}{2s}$ subsets of size $2s$ within $\{1, 2, \ldots, d\}$. Moreover, for any $2s$-sized subset, there is an $(r\epsilon)$-covering in $\ell_2$-norm of the ball $\mathbb{B}_2(r)$ with at most $2^{2s\log(1/\epsilon)}$ elements (e.g., [23]).

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

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

Given the assumptions on $X$, we have $\|X(\theta - \theta^k)\|_2/\sqrt{n} \leq \kappa_u\|\theta - \theta^k\|_2 \;\leq\; \kappa_u\, r\epsilon$. Moreover, since the variate $\|w\|_2^2/\sigma^2$ is $\chi^2$ 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

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

with high probability. Taking the supremum over $\theta$ on both sides yields

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

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 $\theta^k$ 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

$$\begin{aligned} Z_n &\leq \sigma \, r \, \kappa_u \sqrt{\frac{3\log N(\epsilon; \mathbb{S}(s,r))}{n}} + 2\kappa_u \, \sigma r \, \epsilon \\ &= \sigma \, r \, \kappa_u \Big\{\sqrt{\frac{3\log N(\epsilon; \mathbb{S}(s,r))}{n}} + 2\epsilon\Big\}. \end{aligned} \tag{49}$$

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

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

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

where the final line uses standard bounds on binomial coefficients. Since $d/s \geq 2$ by assumption, we conclude that our choice of $\epsilon$ guarantees that $\frac{\log N(\epsilon;\mathbb{S}(s,r))}{n} \leq 5\, s\log(d/s)$. Substituting these relations into the inequality (49), 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 $\log N(\epsilon; \mathbb{S}(s,r)) \geq cs\log(d/s)$, 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.

*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 [33]). 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

$$\sqrt{n}\delta \geq c_1 r \tag{50a}$$

$$\sqrt{n}\delta \geq c_1 \int_{\frac{\delta}{16}}^{r} \sqrt{\log N(t; \mathbb{S}_q)}dt \;=: J(r,\delta). \tag{50b}$$

where $N(t; \mathbb{S}_q; \|.\|_2/\sqrt{n})$ is the covering number for $\mathbb{S}_q$ in the $\ell_2$-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 $\chi^2$ 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

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

For some $c_3 > 0$, let us set

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

and let us verify that the conditions (50a) and (50b) hold. Given our choice of $\delta$, we find that

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

By the condition (11), the dimension $d$ is lower bounded so that condition (50a) holds. Turning to verification of the inequality (50b), 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 (28) of $\operatorname{absconv}_q(X/\sqrt{n})$, we have

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

We may apply the bound in Lemma 3 to conclude that $\log N(\epsilon; \mathbb{S}_q)$ is upper bounded by $c' \, R_q^{\frac{2}{2-q}} \, \big(\frac{\widetilde{\kappa}_c}{\epsilon}\big)^{\frac{2q}{2-q}} \log d$.

Using this upper bound, we have

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

Using the definition of $\delta$, the condition (11), and the condition $r = \Omega(\widetilde{\kappa_c}^{\frac{q}{2}} \sqrt{R_q}\, (\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}})$, it is straightforward to verify that $(\delta/16, r)$ lies in the range of $\epsilon$ specified in Lemma 3.

Using this upper bound, let us verify that the inequality (50b) 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 $\delta$, we have

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

so that condition (50b) will hold as long as we choose $c_3 > 0$ large enough. Overall, we conclude that

$$\mathbb{P}[Z(R_q, r) \geq c_3\, r\, \widetilde{\kappa_c}^{\frac{q}{2}} \sqrt{R_q}\, (\frac{\log d}{n})^{\frac{1}{2}-\frac{q}{4}}]$$

is at most $c_1 \exp(-R_q (\log d)^{1-\frac{q}{2}} n^{\frac{q}{2}})$, which concludes the proof.

*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 $\Delta_S$ 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 $\chi^2$ with $s$ degrees of freedom, so that by standard $\chi^2$ 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)

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

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

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

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

as claimed.

### H. Peeling argument

In this appendix, we state a tail bound for the constrained optimum of a random objective function of the form $f(v; X)$, where $v \in \mathbb{R}^d$ is the vector to be optimized over, and $X$ is some random vector. More precisely, consider the problem $\sup_{\rho(v)\leq r,\ v \in A} f(v; X)$, where $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is some non-negative and increasing constraint function, and $A$ is a non-empty set. Our goal is to bound the probability of the event defined by

$$\mathcal{E} := \big\{X \in \mathbb{R}^{n\times d} \mid \exists\, v \in A \text{ s. t. } f(v; X) \geq 2g(\rho(v))\big\},$$

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*

$$\mathbb{P}\big[\sup_{v\in A,\ \rho(v)\leq r} f(v; X) \geq g(r)\big] \leq 2\exp(-c\, a_n\, g(r)), \tag{51}$$

*for some $a_n > 0$. Then we have*

$$\mathbb{P}[\mathcal{E}] \leq \frac{2\exp(-4ca_n\mu)}{1 - \exp(-4ca_n\mu)}.$$

*Proof:* Our proof is based on a standard peeling technique (e.g., [2], [33], p. 82). By assumption, as $v$ varies over $A$, we have $g(r) \in [\mu, \infty)$. Accordingly, for $m = 1, 2, \ldots$, defining the sets

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

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

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

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

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

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

### *I. Some tail bounds for $\chi^2$-variates*

The following large-deviations bounds for centralized $\chi^2$ are taken from Laurent and Massart [22]. Given a centralized $\chi^2$-variate $Z$ with $m$ degrees of freedom, then for all $x \geq 0$,

$$\mathbb{P}\big[Z - m \geq 2\sqrt{mx} + 2x\big] \leq \exp(-x), \text{ and} \tag{52a}$$

$$\mathbb{P}\big[Z - m \leq -2\sqrt{mx}\big] \leq \exp(-x). \tag{52b}$$

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

$$\mathbb{P}\big[\frac{Z - m}{m} \geq 4t\big] \leq \exp(-mt). \tag{53}$$

Starting with the bound (52a), 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] M. Akcakaya and V. Tarokh. Shannon theoretic limits on noisy compressive sampling. *IEEE Trans. Info Theory*, 56(1):492–504, January 2010.

[2] K. S. Alexander. Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 475—493. UC Press, Berkeley, 1985.

[3] A. R. Barron, L. Birge, and P. Massart. Risk bounds for model selection by penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

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

[5] L. Birgé. Approximation dans les espaces metriques et theorie de l'estimation. *Z. Wahrsch. verw. Gebiete*, 65:181–327, 1983.

[6] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

[7] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35(4):1674–1697, 2007.

[8] 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.

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

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

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

[12] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. of. American Mathematical Society*, 22(1):211–231, January 2009.

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

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

[15] A. K. Fletcher, S. Rangan, and V. K. Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772, 2009.

[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 a $p$-convex body. In *Geometric aspects of functional analysis*, pages 95–108. Springer-Verlag, 2000.

[18] T. S. Han and S. Verdu. Generalizing the Fano inequality. *IEEE Transactions on Information Theory*, 40:1247–1251, 1994.

[19] 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.

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

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

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

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

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

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

[26] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Proceedings of NIPS*, 2009. Full length version: arxiv.org/abs/1010.2731v1.

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

[28] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Technical Report http://arxiv.org/abs/1008.3654, UC Berkeley, Department of Statistics, August 2010.

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

[30] P. Rigollet and A. B. Tsybakov. Exponential screening and optimal rates of sparse estimation. Technical report, Princeton, 2010.

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

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

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

[34] S. van de Geer. The deterministic Lasso. In *Proc. of Joint Statistical Meeting*, 2007.

[35] S. van de Geer and J.-M. Loubes. Adaptive estimation in regression, using soft thresholding type penalties. *Statistica Neerlandica*, 56:453–478, 2002.

[36] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009.

[37] M. J. Wainwright. Information-theoretic bounds for sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Information Theory*, 55(12): 5728–5741, 2009.

[38] W. Wang, M. J. Wainwright, and K. Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Info Theory*, 56(6):2967–2979, June 2010.

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

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

[41] 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.

[42] C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 2010. To appear.

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

*Biographies*

- Garvesh Raskutti is currently a PhD candidate in the Department of Statistics at the University of California, Berkeley. He received his Bachelor of Science (majors in Mathematics and Statistics) and Bachelor of Engineering (major in Electrical Engineering), and Masters of Engineering Science at the University of Melbourne. His research interests include statistical machine learning, mathematical statistics, convex optimization, information theory, and high-dimensional statistics. He was awarded a Berkeley Graduate Fellowship from 2007-09.
- Martin Wainwright is currently a professor at University of California at Berkeley, with a joint appointment between the Department of Statistics and the Department of Electrical Engineering and Computer Sciences. He received a Bachelor's degree in Mathematics from University of Waterloo, Canada, and Ph.D. degree in Electrical Engineering and Computer Science (EECS) from Massachusetts

Institute of Technology (MIT). His research interests include statistical signal processing, coding and information theory, statistical machine learning, and high-dimensional statistics. He has been awarded an Alfred P. Sloan Foundation Fellowship, an NSF CAREER Award, the George M. Sprowls Prize for his dissertation research (EECS department, MIT), a Natural Sciences and Engineering Research Council of Canada 1967 Fellowship, an IEEE Signal Processing Society Best Paper Award in 2008, and several outstanding conference paper awards.

- Bin Yu is Chancellor's Professor in the Departments of Statistics and Electrical Engineering  Computer Science at UC Berkeley. She is currently the chair of department of Statistics, and a founding co-director of the Microsoft Lab on Statistics and Information Technology at Peking University, China. Her Ph.D. is in Statistics from UC Berkeley, and she has published over 80 papers in a wide range of research areas including empirical process theory, information theory (MDL), MCMC methods, signal processing, machine learning, high dimensional data inference (boosting and Lasso and sparse modeling in general), bioinformatics, and remotes sensing. Her current research interests include statistical machine learning for high dimensional data and solving data problems from remote sensing, neuroscience, and text documents.

  She was a 2006 Guggenheim Fellow, and is a Fellow of AAAS, IEEE, IMS (Institute of Mathematical Statistics) and ASA (American Statistical Association). She is a co-chair of National Scientific Committee of SAMSI and on the Board of Mathematical Sciences and Applications of the National Academy of Sciences in the US. She has been invited to deliver the 2012 Tukey Memorial Lecture in Statistics at the 8th World Congress in Probability and Statistics, the quadrennial joint conference of the Bernoulli Society and IMS in Istanbul, Turkey.