Large-scale sparse regression models under weak assumptions

by

Garvesh Raskutti

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

and the Designated Emphasis

in

Communication, Computation and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bin Yu, Professor Martin Wainwright, Co-chairs
Professor Noureddine El Karoui
Professor Laurent El Ghaoui

Fall 2012

Large-scale sparse regression models under weak assumptions

Abstract

Large-scale sparse regression models under weak assumptions

by

Garvesh Raskutti

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Professor Martin Wainwright, Co-chairs

Many modern problems in science and other areas involve extraction of useful information from so-called 'big data.' However, many classical statistical techniques are not equipped to meet with the challenges posed by big data problems. Furthermore, existing statistical methods often result in intractable algorithms. Consequently the last $15-20$ years has seen a flurry of research on adapting existing methods and developing new methods that overcome some of the statistical and computational challenges posed by problems involving big data.

Regression is one of the oldest statistical techniques. For many modern regression problems involving big datasets, the number of predictors or covariates d is large compared the number of samples n, causing significant computational and statistical challenges. To overcome these challenges, many researchers have proposed imposing sparsity on the vector of regression co-efficients $\beta \in \mathbb{R}^d$. Furthermore, researchers have proposed using ℓ_1-based convex penalties for estimating β under the sparsity assumption since they yield implementable algorithms with desirable performance guarantees. While there was already an established body of work on developing procedures for sparse regression models, most existing results rely on very restrictive model assumptions. These assumptions are often not satisfied for many scientific problems. In this thesis, we relax 3 restrictive model assumptions that are commonly imposed in the literature for estimating sparse regression models.

The 3 assumptions are: (1) Strict sparsity, that is the vector of regression co-efficients β contains only a small number of non-zeros; (2) The covariates or predictors are independent; (3) Response depends linearly on covariates. Given that these 3 model assumptions are often not satisfied for many practical settings, it is important to understand whether existing theoretical results exhibit robustness to these assumptions.

In Chapter 2, we impose a weaker notion of sparsity known as ℓ_q-ball sparsity on β which ensures the vector of regression co-efficients lies in an ℓ_q ball, but need not have any non-zeros. We prove that under the weaker ℓ_q-ball sparsity assumption, it is possible to develop estimators with desirable mean-squared error behavior, even in the regime where $d \gg n$.

The weakest known condition under which the Lasso achieves optimal mean-squared error rate is the restricted eigenvalue condition [84, 11, 64]. Existing results prove that in

cases when the covariates are independent, the restricted eigenvalue condition is satisfied. However, the setting when predictors or covariates are correlated are also of interest and there was considerably less work dealing with this case. In Chapter 3, we prove that the restricted eigenvalue condition is satisfied for various correlated Gaussian designs, including time series models, spiked covariance models and others.

Finally, in Chapter 4 we analyze sparse additive models, a non-parametric analog of sparse linear models, in which each component function lies in an ellipsoid or more formally a Reproducing kernel Hilbert space $\mathcal{H}$. Hence we weaken the assumption that our response depends on the covariate via a linear function. A new ℓ_1-based polynomial-time method is developed and we prove that this method has desirable mean-squared error performance, even when $d \gg n$. Furthermore, we prove lower bounds on the mean-squared error for estimating sparse additive models that match the upper bounds for our method. Hence our algorithm is optimal in terms of mean-squared error rate.

Contents

Acknowledgments

First, I would like to express my deepest appreciation and gratitude to my advisors, Martin and Bin. They formed an excellent combination as advisors, providing endless support and guidance to me and their other students. I have great admiration for their incredible passion for all aspects of academia including research, mentoring students, and teaching. Both Bin and Martin have always gone above and beyond what is required of them especially in terms of giving time to their students, in spite of all their other commitments. All of the results presented in this thesis are based on joint projects with Martin and Bin, hence the use of 'we' throughout the thesis.

I would also like to thank Noureddine, the third member of my committee and Laurent, the outside member. Noureddine and Laurent have always been extremely generous with their time. I had a number of interesting discussions with Noureddine on random matrix theory and with Laurent relating to optimization, and analysis of text data. These discussions greatly helped with my dissertation, especially the material in Chapter 3.

In my first 2 years, I worked closely with Pradeep Ravikumar, a former post-doc with Bin and Martin and now an Assistant Professor. We wrote a paper together in which he was first author (not part of this thesis) and he set an outstanding example of how to be a top researcher. His (and co-authors) prior work on non-parametric sparse additive models inspired the work undertaken in Chapter 4 of this thesis. I learnt a lot working with him and hope to follow his example in the future.

During the summer at the end of my 3^{rd} year, I did an internship in the Natural Language Processing (NLP) group at Microsoft Research Asia (MSRA) under the supervision of Dr. Ming Zhou and Xiaohua Liu. I would like to thank Ming for giving me the opportunity to work under him and with other interns at MSRA, especially considering my background was not in NLP. While at MSRA, I learnt a lot about performing NLP tasks and large-scale news data and how some of the ideas from this thesis can be used in practice. I also found both Ming and Xiaohua gave me a great deal of freedom to pursue my own ideas, while I was learning from them.

Finally, my PhD would not have been as productive and enjoyable without the incredible support of my family and friends. I have met a number of great people (too many to mention), who have made the experience of doing a PhD at Berkeley very pleasant and straightforward. I would especially like to thank my parents and brothers for always being there on skype with plenty of love, support and advice.

Chapter 1

Introduction

With the rapid advancement of web-based technologies, we now have unprecedented access to massive amounts of data. With this data deluge, we as individuals, organizations, and researchers are constantly trying to find more efficient ways to extract relevant information from the mass of available data. Companies such as Netflix and Amazon are aiming to develop good recommender systems for individuals based on data from millions of users and products. Medical researchers are using data from the entire human genome to pinpoint causes for cancer and other genetic diseases. These are just some of the problems where we encounter the challenge of trying to automatically extract useful information from a massive dataset.

Given the sheer size and complexity of the datasets, automatically extracting useful information present a number of challenges including: (i) developing quantitative models that pinpoints the relevant information; and (ii) implementing these models given limited computational resources. Prior to the data deluge that begun in the 1990s, statistical methods were mainly developed for problems involving smaller datasets. Hence, these earlier methods are often not equipped to meet the challenges posed by modern large-scale data problems. Consequently the last 20 years have seen a large amount of research addressing some of the challenges posed by problems involving large datasets.

There are a number of approaches for addressing the first challenge of developing quantitative models for information extraction. A general framework that has been applied with some success is the principle of *parsimony*. The principle of parsimony arises from 'Occam's Razor,' which can be summarized as 'other things being equal, a simpler explanation is better than a more complex one.' Simple explanations are desirable in terms of human interpretability, since they are easier for individuals to parse and may often capture most of the relevant information within the data. If we consider the problem of developing movie recommendations for an individual based on the ratings of over millions of users for millions of films, one might expect that a small number of factors such as genre, actors, directors and others would capture most of the relevant information. Simple explanations are also attractive from a statistical and computational perspective. From a statistical point of view, simple models are less prone to overfit data than large complex models. The benefit from a computational perspective is that once a simple explanation is found, data storage and processing becomes much more straightforward. In this thesis, we apply the principle of parsimony to analyze and develop methods for regression problems involving large data.

A naiive approach for finding simple models for large datasets is to exhaustively search over all possible models. As you can imagine, for large complicated datasets, there are often millions or even trillions of different models. Consider the problem of predicting whether a patient has a type of cancer based on data consisting of gene expression measurements for approximately $50,000$ genes, which is a common scenario in medical research. Any subset of the $50,000$ genes is a reasonable model for prediction and if we wanted to find the best one, we would need to search over $2^{50,000}$ models which is several orders of magnitude greater than the number of atoms in the universe. Hence, an exhaustive search is clearly impossible. As a result, computation and algorithmic issues have started to play a more significant role in the development of statistical methods and ideas from convex optimization, approximation algorithms and computer science theory have been applied to large-scale statistical inference problems. This marriage of ideas between statistical methods and computation has seen the development of a number of computationally efficient algorithms for finding parsimonious models, even when datasets contain many more than $50,000$ predictors.

The focus of this thesis is on developing and analyzing parsimonious regression models for large-scale inference problems. Regression is one of the simplest and most widely used statistical methods. The goal of regression is to predict a response y based on predictors $[x_1, x_2, ..., x_d]$. In the next section we illustrate how Occam's Razor, combined with ideas from convex optimization may be used to develop methods for large-scale regression problems. First we provide a brief overview of past work. We then describe the main contributions of this thesis. A high-level summary of the challenge we address in this thesis is as follows: while a lot of progress has been made on the problem of estimating sparse or parsimonious regression models, most existing approaches rely on the data satisfying restrictive model assumptions including (1) the predictors being independent; (2) the response is dependent on only a small number of predictors; and (3) the relationship between the response and the predictors follows a parametric linear model. Many real-world scientific problems do not satisfy these assumptions and it is vital to develop reliable methods for these settings. In this thesis, we develop and analyze statistical models for reliably estimating sparse regression models when each of the 3 aforementioned assumptions is relaxed.

1.1 Sparse regression models

Regression is one of the oldest statistical methods, dating back to the early 1800's when Legendre and Gauss used the method of least-squares to determine the orbits of bodies around the sun. Since regression problems have arisen so frequently over the last two centuries, there is a large body of work on how and when to use different regression approaches. However prior to the 1990s, regression was generally applied to problems where the number of predictors d was quite small, on the order $5 - 100$. Now, with the vastly increasing size and complexity of datasets arising, the number of predictors d may be 1000's or even millions. When the number of predictors is so large, classical approaches require a large number of samples n which are often not available, and even if samples were available, fitting such a large regression model is computationally intensive.

To deal with these statistical and computational issues, for problems where the number of predictors d is large, a common statistical goal is to provide domain experts with simple, interpretable models which require fewer samples and less computation. This is where it is useful to combine the principle of parsimony with classical regression models. Applying this principle, one often aims to find the *sparsest* regression model that explains the data, that is the model containing the smallest numbers of predictors capturing the relevant information to predict y. Returning to the problem of determining whether a patient has a type of cancer based on the expression level of $50,000$ genes, the principle of parsimony would imply the response might only depend on a small handful of genes are useful predictors. The challenge then is to find a good small subset of genes. Classical approaches to finding a good small set of predictors involved using model selection methods such as Akaike Information Criterion [2], Bayesian Information Criterion [79] and minimum description length [73], which involve exhaustively searching over all 2^d subsets. As discussed earlier, such approaches are not feasible when d is large meaning new efficient approaches were required for finding sparse models.

Consequently, ideas from optimization theory have been successfully used to develop computationally efficient methods for estimating sparse regression models. In particular, instead of finding the best model over all 2^d models, many researchers have proposed optimizing over a convex set of models that encourages sparse solutions. There are a number of fast algorithms for optimization over convex sets [15] meaning convex methods can be applied to much larger problems than methods that involve exhaustive combinatorial searches. As a result methods that use convex algorithms are becoming increasingly popular in statistics. One of the best known convex method for estimating sparse regression models is the Lasso [82] (basis pursuit in the noiseless case [23]). The Dantzig selector [19] is another well-known convex approach that has many similar properties to the Lasso (see e.g. [11]). There has been a large body of work demonstrating that in addition to their computational benefits, the Lasso and Dantzig selectors have a number of desirable properties from an estimation and interpretation perspective (see e.g. [11, 60, 84, 90, 95, 97]).

While the use of convex methods for estimating sparse regression models has lead to reliable, computationally-efficient methods, there still remain a number of unresolved issues. One of the most significant issues is that existing analysis of convex methods rely on restrictive model assumptions including the 3 assumptions listed in the previous section. There are numerous important real-world regression problems that do not satisfy these aforementioned assumptions. Hence it is vital to develop methods that apply even when these assumptions do not hold. In this thesis, we provide analysis and methodology with provable guarantees in terms of ℓ_2 error and ℓ_2-prediction error for estimation of sparse regression models when we relax each of the aforementioned restrictive model assumptions. In the next section, we provide a summary of how we relax the 3 model assumptions and outline the main technical contributions of this thesis.

1.2 Contributions of this thesis

In this section, we provide a summary for Chapters 2, 3, and 4. First we describe each of the assumptions, explain how we weaken each assumption, and then summarize our results under the weakened assumptions.

1.2.1 Chapter 2: From strict to weak sparsity

Let us consider the standard linear model for regression:

$$y = \sum_{j=1}^{d} \beta_j x_j + w,$$

where w is Gaussian noise. Recall that we are most interested in the setting where d is large relative to the number of samples n. Strict sparsity requires that most of the β_j's are 0 or equivalently most of the x_j's have absolutely no effect on the response y. Requiring that most predictors have no effect on the response may be too restrictive for some problems. In image analysis for example, it is standard that co-efficients for images expressed in a wavelet basis exhibit sharp decay, but need not be exactly 0 (see e.g Hyvärinen at al. [46]). Hence a weaker form of sparsity that allows many of the predictors to be weekly correlated with the response is more appropriate.

In Chapter 2 in this thesis, we impose a weaker notion of sparsity by assuming the vector of regression co-efficients $\beta = [\beta_1, \beta_2, ..., \beta_d] \in \mathbb{R}^d$ lies in an ℓ_q-ball where $0 < q \leq 1$. ℓ_q-ball sparsity requires that $\|\beta\|_q^q := \sum_{j=1}^{d} |\beta_j|^q$ is bounded. Hence the regression co-efficients decay at a rate that is determined by q but none of the co-efficients need to be 0 exactly. We demonstrate in Chapter 2 that a number of desirable properties exhibited by strictly sparse models, also hold under weak ℓ_q-ball sparsity.

More concretely, we study the minimax error rate for estimating the regression parameter $\beta = [\beta_1, \beta_2, ...\beta_d] \in \mathbb{R}^d$ both in terms of ℓ_2-error and ℓ_2-prediction error, assuming that β belongs to an ℓ_q-ball $\mathbb{B}_q(R_q) = \{\beta \in \mathbb{R}^d \mid \|\beta\|_q^q \leq R_q\}$ for some $q \in [0, 1]$. We show that under suitable regularity conditions on the predictors $[x_1, x_2, ..., x_d]$, the optimal minimax rates in both ℓ_2 and ℓ_2-prediction error scale as $R_q \big(\frac{\log d}{n}\big)^{1-\frac{q}{2}}$.

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 the least-squares estimator over

ℓ_q-balls. Subsequent work by Negahban et al. [64] demonstrates that the Lasso estimator achieves the optimal mean-squared error rate under ℓ_q-ball sparsity. Hence it is possible to reliable estimate β under ℓ_q sparsity even when d is much larger than the sample size n.

One of the main contributions of this chapter is that we carefully characterize the conditions on the design matrix $X \in \mathbb{R}^{n \times d} = [x_1, x_2, ..., x_d]$ required for minimax optimal rates. Our analysis reveals that conditions on the design matrix X enter into the error rates for parameter estimation and prediction error in complementary ways in the upper and lower bounds. Our results also show that although computationally efficient convex methods can achieve the minimax rates up to constant factors, they require slightly stronger assumptions on the covariates $[x_1, x_2, ..., x_d]$ than optimal algorithms involving least-squares over the ℓ_q-ball.

1.2.2 Chapter 3: From independent to correlated predictors

Existing analysis on estimation of sparse regression models impose identifiability assumptions on the predictors $[x_1, x_2, ..., x_d]$ that are proven only to be satisfied if all d predictors are independent. In particular, well-known conditions such as restricted isometry property (RIP) [20], the irrepresentable condition [97], restricted eigenvalue (RE) [11, 84], restricted nullspace (RN) [25, 29, 33]. In particular, the RE condition is the weakest known condition for establishing optimal mean-squared error rate in the noisy setting, and RN is the weakest known condition for exact recovery in the noiseless setting. Previous results have shown that RE and RN conditions are satsified when all d predictors are independent. There has been considerably less work on understanding whether such identifiability conditions are satisfied when predictors are correlated.

In practice, it is rare that predictors are independent and they often exhibit a specific correlation structure (e.g. time series, spatial etc.). Hence it is important to understand whether existing identifiability assumptions hold when predictors may be correlated. In Chapter 3, we demonstrate that the RE and RN conditions are satisfied in many settings when predictors are correlated.

In particular, we prove directly that the RN and RE conditions hold with high probability for quite general classes of Gaussian matrices where each row has covariance $\Sigma \in \mathbb{R}^{d \times d}$ for which the predictors may be highly dependent. Examples not covered by previous results include when Σ is a Toeplitz matrix, a spiked covariance matrix, or any positive definite matrix. In this way, our results extend the attractive theoretical guarantees for ℓ_1-based methods to a much broader class of problems than the case of completely independent or unitary designs.

1.2.3 Chaoter 4: From linear to non-parametric models

There is a large body work on estimating sparse regression models under the assumption that the response variable y depends linearly on the predictors. While linear models are useful for many problems, non-parametric models are more suitable when the nature of the relation between the response and predictors is unknown. In a recent problem, Vu et al. [87] demonstrate that a sparse non-parametric model substantially outperforms existing parametric approaches in their problem involving image reconstruction using fMRI data.

While there has been some work on developing and analyzing sparse non-linear regression models (see e.g. [50, 54, 58, 71, 94]), the guarantees for practical problems that use high-dimensional non-parametric models and associated algorithms are considerably less understood compared to sparse linear models. In Chapter 4 we analyze and develop a new method for estimating sparse additive models, a sparse non-parametric model.

Sparse additive models are families of d-variate functions with the additive decomposition $f(x_1, x_2, ..., x_d) = \sum_{j \in S} f_j(x_j)$, where S is an unknown subset of cardinality $s \ll d$. Sparse additive models are a non-parametric analog of sparse linear models, where for linear models we would assume each univariate function f_j is linear. In Chapter 4, we consider the case where each f_j lies in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, and analyze a method for estimating the unknown function f based on kernels combined with a convex penalty. Working within a framework that allows both the dimension d and sparsity s to increase with n, we derive sharp bounds on both the mean-squared prediction error and population integrated mean-squared error over the class of sparse additive models.

The mean-sqaured error rate is $\frac{s \log d}{n} + s\delta_n^2$, where δ_n^2 is the mean-squared error rate for estimating a single univariate function in the RKHS $\mathcal{H}$. Our result captures the intuition that estimating sparse additive models decomposes into two low-dimensional sub-problems, a subset selection problem that has mean-squared error $\frac{s \log d}{n}$, and an s-variate estimation problem that has mean-squared error $s\delta_n^2$. We complement our upper bounds by deriving lower bounds, thereby showing the optimality of our method. Thus, we obtain optimal mean-sqaured error rates for many interesting classes of sparse additive models, including polynomials, splines, and Sobolev classes.

One of the main challenges in proving our upper bound was that many existing results for non-parametric problems rely on a global boundedness assumption on the function f. We prove that if instead of our conditions on each univariate function f_j, the d-variate function class $\mathcal{F}_{d,s,\mathcal{H}}$ is assumed to be globally bounded, then much faster mean-squared error rates are possible for any sparsity $s = \Omega(\sqrt{n})$. Hence we prove that global boundedness is a significant restriction in the setting where both s and d scale with n. Proving optimal upper bounds with only our univariate conditions requires new techniques that we present and discuss in Chapter 4.

Chapter 2

Minimax rates of estimation for weak ℓ_q-ball sparse high-dimensional linear regression

2.1 Outline

As was discussed in the introductory chapter, there has been 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 ℓ_1-based relaxations. Complementary to the understanding of computationally efficient procedures are the fundamental or information-theoretic limitations of statistical inference, applicable to any algorithm regardless of its computational cost. There is a rich line of statistical work on such fundamental limits, an understanding of which can have two types of consequences. First, they can reveal gaps between the performance of an optimal algorithm compared to known computationally efficient methods. Second, they can demonstrate regimes in which practical algorithms achieve the fundamental limits, which means that there is little point in searching for a more effective algorithm. As we shall see, the results in this chapter lead to understanding of both types.

2.1.1 Problem set-up

The focus of this chapter 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

$$y = X\beta^* + w, \tag{2.1}$$

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 β^* 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 ℓ_1-regularization, including basis pursuit [23], the Lasso [82, 23], and the Dantzig selector [20]. Various authors have obtained convergence rates for different error metrics, including ℓ_2-norm error [11, 20, 60, 95], prediction loss [11, 36, 84], as well as model selection consistency [59, 90, 95, 97]. In addition, a range of sparsity assumptions have been analyzed, including the case of *hard sparsity* meaning that β^* has exactly $s \ll d$ non-zero entries, or *soft sparsity* assumptions, based on imposing a certain decay rate on the ordered entries of β^*. 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.

Sparsity constraints One way in which to capture the notion of sparsity in a precise manner is in terms of the ℓ_q-balls[1] 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 ℓ_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 β with at most s non-zero elements. For $q \in (0, 1]$, membership of β in $\mathbb{B}_q(R_q)$ enforces a "soft" form of sparsity, in that all of the coefficients of β 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.

Loss functions We consider estimators $\widehat{\beta} : \mathbb{R}^n \times \mathbb{R}^{n\times d} \to \mathbb{R}^d$ that are measurable functions of the data (y, X). Given any such estimator of the true parameter β^*, there are many criteria for determining the quality of the estimate. In a decision-theoretic framework, one

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

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

$$\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_q(R_q)} \mathcal{L}(\widehat{\beta}, \beta^*). \tag{2.2}$$

Note that the quantity (2.2) 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 chapter, we provide results that hold with high probability, as shown in the statements our main results in results in Theorems 7 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 ℓ_2*-loss*

$$\mathcal{L}_2(\widehat{\beta}, \beta^*) := \|\widehat{\beta} - \beta^*\|_2^2 \;=\; \sum_{j=1}^{d} |\widehat{\beta}_j - \beta^*_j|_2^2, \tag{2.3}$$

and (iii) the ℓ_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., [89, 3, 91]); in contrast, the analysis of this paper is focused on understanding the minimax rates associated with the ℓ_2-loss and the ℓ_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:

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

These quantities correspond to all possible combinations of minimax risk involving either the ℓ_2-loss or the ℓ_2-prediction loss, and with either hard sparsity ($q = 0$) or soft sparsity ($q \in (0, 1]$).

2.1.2 Our contributions

The main contributions are derivations of optimal minimax rates both for ℓ_2-norm and ℓ_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 ℓ_2-norm loss (Theorems 7 and 2 respectively) and upper and lower bounds on ℓ_2-prediction loss (Theorems 3 and Theorem 4) respectively. We note that for the linear model (3.1), 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 [13, 30, 8] 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 7, we provide lower bounds for the ℓ_2-loss that involves a maximum of two quantities: a term involving the diameter of the null-space restricted to the ℓ_q-ball, measuring the degree of non-identifiability of the model, and a term arising from the ℓ_2-metric entropy structure for ℓ_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 7 and 2). Finally, Theorems 3 and 4 are for ℓ_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 ℓ_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 θ lying in the ℓ_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 β and β^*, and so reflects the discriminability of these models. Our analysis shows that the conditions on X enter in quite a different manner for ℓ_2-norm and prediction losses. In particular, for the case $q > 0$, proving *upper bounds* on ℓ_2-norm error and *lower bounds* on prediction error require relatively strong conditions on the design matrix X, whereas *lower bounds* on ℓ_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 7 and 3 involve a combination of a standard information-theoretic techniques (e.g. [12, 41, 92]) with results in the approximation theory literature (e.g., [38, 51]) on the metric entropy of ℓ_q-balls. The proofs for the upper bounds in Theorems 2 and 4 involve direct analysis of the least-squares optimization over the ℓ_q-ball. The basic idea involves concentration results for Gaussian random variables and properties of the ℓ_1-norm over ℓ_q-balls (see Lemma 5).

The remainder of this paper is organized as follows. In Section 4.3, 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 [96], and we provide a more detailed discussion and comparison in Section 2.2.5, 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 ℓ_1-relaxations. In Section 3.4, we provide the proofs of our main results, with more technical aspects deferred to the appendices.

2.2 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.

2.2.1 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 ℓ_2-norm.

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

$$\frac{1}{\sqrt{n}} \max_{j=1,\ldots,d} \|X_j\|_2 \leq \kappa_c. \tag{2.4}$$

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 ℓ_2-prediction error when $q = 0$.

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

Asumption 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

$$\frac{\|X\theta\|_2}{\sqrt{n}} \geq \kappa_\ell \left(\|\theta\|_2 - f_\ell(R_q, n, d)\right) \tag{2.5}$$

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 ℓ_2-error and lower bounds for ℓ_2-prediction error. It is required in *upper bounding* ℓ_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 β is from β' in terms of the linear regression model. Indeed, note that for fixed X and β, 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 β and β' is given by $\frac{1}{2\sigma^2}\|X(\beta - \beta')\|_2^2$. Thus, the lower bound (2.5), when applied to the difference $\theta = \beta - \beta'$, ensures any pair (β, β') that are well-separated in ℓ_2-norm remain well-separated in the ℓ_2-prediction semi-norm. Interestingly, Assumption 2 is also essential in establishing *lower bounds* on the ℓ_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 ℓ_2-prediction semi-norm. As we show, it is this size that governs the difficulty of estimation in the prediction semi-norm.

The condition (2.5) 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 (2.5) 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 $\rho^2(\Sigma) = \max\limits_{j=1,\ldots,d} \Sigma_{jj}$. 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*

$$\frac{\|X\theta\|_2}{\sqrt{n}} \geq \frac{\lambda_{\min}(\Sigma^{1/2})}{4}\|\theta\|_2 - 18\,\rho(\Sigma)\; R_q \left(\frac{\log d}{n}\right)^{1-q/2}, \tag{2.6}$$

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 (2.6) is that Assumption 2 holds with

$$f_\ell(R_q, n, d) = \bar{c}\, \frac{\rho(\Sigma)}{\lambda_{\min}(\Sigma^{1/2})}\, R_q \left(\frac{\log d}{n}\right)^{1-q/2} \tag{2.7}$$

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 the Appendix in Raskutti et al. [67]. 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 ℓ_2-norm error, and lower bounds in ℓ_2-prediction error.

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

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

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \leq \kappa_u\, \|\theta\|_2 \text{ for all } \theta \in \mathbb{B}_0(2s). \tag{2.8}$$

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

$$\frac{1}{\sqrt{n}}\|X\theta\|_2 \geq \kappa_{0,\ell}\, \|\theta\|_2 \text{ for all } \theta \in \mathbb{B}_0(2s). \tag{2.9}$$

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 2.2.5, Assumption 3 is closely related to conditions imposed in analyses of ℓ_1-based relaxations, such as the restricted isometry property [20] as well as related but less restrictive sparse eigenvalue conditions [11, 60, 84]. Unlike the restricted isometry property, Assumption 3 does not require that the constants κ_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. [11], and we discuss these connections at more length in Section 2.2.5. 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 ℓ_2-error involve the set defined by intersecting the null space (or kernel) of X with the ℓ_q-ball, which we denote by $\mathcal{N}_q(X) := \text{Ker}(X) \cap \mathbb{B}_q(R_q)$. We define the $\mathbb{B}_q(R_q)$-*kernel diameter* in the ℓ_2-norm as

$$\operatorname{diam}_2(\mathcal{N}_q(X)) := \max_{\theta \in \mathcal{N}_q(X)} \|\theta\|_2 \;=\; \max_{\substack{\|\theta\|_q^q \leq R_q, \\ X\theta = 0}} \|\theta\|_2. \tag{2.10}$$

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 (3.1) 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 (3.1) 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 (2.10); in particular, these assumptions imply an upper bound bound on the $\mathbb{B}_q(R_q)$-kernel diameter in ℓ_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*

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

(b) Case $q = 0$: *If Assumption 3(b) is satisfied, then* $\operatorname{diam}_2(\mathcal{N}_0(X)) = 0$*. (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 2.2.5, we discuss further connections between our assumptions, and the conditions imposed in analysis of the Lasso and other ℓ_1-based methods [11, 20, 59, 64], for the case $q = 0$.

2.2.2 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 σ^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 σ^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 $d \geq 4s \geq c_2$. 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

$$\frac{d}{R_q n^{q/2}} \overset{(i)}{\geq} c_1 \, d^{\delta} \overset{(ii)}{\geq} c_2. \tag{2.11}$$

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 $R_q \leq c_3 d^{\frac{1}{2}-\delta'}$ for some $\delta' \in (0, 1/2)$ is one sufficient condition.)

2.2.3 Optimal minimax rates in ℓ_2-norm loss

We are now ready to state minimax bounds, and we begin with lower bounds on the ℓ_2-norm error:

Theorem 1 (Lower bounds on ℓ_2-norm error)**.** *Consider the linear model* (3.1) *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*

$$\mathcal{M}_2(\mathbb{B}_q(R_q), X) \geq \bar{c} \; \max \Big\{ \operatorname{diam}_2^2(\mathcal{N}_q(X)), \; R_q \left(\frac{\sigma^2}{\kappa_c^2} \frac{\log d}{n} \right)^{1-q/2} \Big\} \tag{2.12}$$

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*

$$\mathcal{M}_2(\mathbb{B}_0(s), X) \geq \bar{c} \; \max \Big\{ \operatorname{diam}_2^2(\mathcal{N}_0(X)), \; \frac{\sigma^2}{\kappa_u^2} \, \frac{s \; \log(d/s)}{n} \Big\} \tag{2.13}$$

with probability greater than 1/2.

The choice of probability 1/2 is a standard convention for stating minimax lower bounds on rates.[2] Note that both lower bounds consist of two terms. The first term corresponds to the diameter of the set $\mathcal{N}_q(X) = \mathrm{Ker}(X) \cap \mathbb{B}_q(R_q)$, a quantity which reflects the extent which the linear model (3.1) is unidentifiable. Clearly, one cannot estimate β^* any more accurately than the diameter of this set. In both lower bounds, the ratios σ^2/κ_c^2 (or σ^2/κ_u^2) correspond to the inverse of the signal-to-noise ratio, comparing the noise variance σ^2 to the magnitude of the design matrix measured by κ_u, 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 (2.12), and similarly the term $\log(\frac{d}{s})$ in the bound (2.13), are reflections of the complexity of the ℓ_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 ℓ_2-norm minimax rate over ℓ_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

$$\widehat{\beta} \in \arg \min_{\beta \in \mathbb{B}_q(R_q)} \|y - X\beta\|_2^2. \tag{2.14}$$

Theorem 2 (Upper bounds on ℓ_2-norm loss)**.** *Consider the model* (3.1) *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 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*

$$\mathcal{M}_2(\mathbb{B}_q(R_q), X) \leq \bar{c}\, R_q \Big[\frac{\kappa_c^2}{\kappa_\ell^2} \frac{\sigma^2}{\kappa_\ell^2}\, \frac{\log d}{n}\Big]^{1-q/2} \tag{2.15}$$

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

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

$$\mathcal{M}_2(\mathbb{B}_0(s), X) \leq \bar{c}\, \frac{\kappa_c^2}{\kappa_{0,\ell}^2}\, \frac{\sigma^2}{\kappa_{0,\ell}^2}\, \frac{s\, \log d}{n} \tag{2.16}$$

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

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

$$\mathcal{M}_2(\mathbb{B}_0(s),X)\leq\bar{c}\,\frac{\kappa_u^2}{\kappa_{0,\ell}^2}\frac{\sigma^2}{\kappa_{0,\ell}^2}\,\frac{s\log(d/s)}{n}, \tag{2.17}$$

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

In the case of ℓ_2-error and design matrices X that satisfy the assumptions of both Theorems 7 and 2, these results identify the minimax optimal rate up to constant factors. In particular, for $q\in(0,1]$, the minimax rate in ℓ_2-norm scales as

$$\mathcal{M}_2(\mathbb{B}_q(R_q),X)\;=\;\Theta\Big(R_q\Big[\frac{\sigma^2\log d}{n}\Big]^{1-q/2}\Big), \tag{2.18}$$

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

$$\mathcal{M}_2(\mathbb{B}_0(s),X)\;=\;\Theta\Big(\frac{\sigma^2\,s\log(d/s)}{n}\Big). \tag{2.19}$$

2.2.4 Optimal minimax rates in ℓ_2-prediction norm

In this section, we investigate minimax rates in terms of the ℓ_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 ℓ_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 2.2.6.

Theorem 3 (Lower bounds on prediction error)**.** *Consider the model* (3.1) *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* $f_\ell(R_q,n,d)<c_2R_q(\frac{\log d}{n})^{1-q/2}$*. Then*

$$\mathcal{M}_n(\mathbb{B}_q(R_q),X)\geq\bar{c}\,R_q\,\kappa_\ell^2\,\Big[\frac{\sigma^2}{\kappa_c^2}\,\frac{\log d}{n}\Big]^{1-q/2} \tag{2.20}$$

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*

$$\mathcal{M}_n(\mathbb{B}_0(s), X) \geq \; \bar{c}\, \kappa_{0,\ell}^2 \, \frac{\sigma^2}{\kappa_u^2} \, \frac{s\log(d/s)}{n} \tag{2.21}$$

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 (2.14).

Theorem 4 (Upper bounds on prediction error)**.** *Consider the model* (3.1) *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*

$$\mathcal{M}_n(\mathbb{B}_q(R_q), X) \leq \bar{c}\, \kappa_c^2 \, R_q \left[\frac{\sigma^2}{\kappa_c^2} \, \frac{\log d}{n}\right]^{1-\frac{q}{2}}. \tag{2.22}$$

(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*

$$\mathcal{M}_n(\mathbb{B}_0(s), X) \leq \bar{c}\, \frac{\sigma^2 \, s\log(d/s)}{n}. \tag{2.23}$$

We note that Theorem 4(b) was stated and proven in Bunea et. al [10] (see Theorem 3.1). However, we have included the statement here for completeness and so as to facilitate discussion.

2.2.5 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 7(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 [26], 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 7 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 β^* in absolute value, then the fast decay imposed by the ℓ_q-ball condition on β^* would mean that the remaining $d-s_q$ coefficients would have relatively little impact. With this intuition, the rate for $q > 0$ can be interpreted as the rate that would be achieved by choosing $s_q = R_q\big(\frac{\log d}{n}\big)^{-q/2}$, and then acting as if the problem were an instance of a hard-sparse problem ($q = 0$) with $s = s_q$. For such a problem, we would expect to achieve the rate $\frac{s_q \log d}{n}$, which is exactly equal to $R_q\big(\frac{\log d}{n}\big)^{1-q/2}$. Of course, we have only made a very heuristic argument here; 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 [30], who determined minimax rates for ℓ_p-losses over ℓ_q-balls. Our work applies to the case of general X, in which the sample size n need not be equal to the dimension d; however, we re-capture the same scaling ($R_q(\frac{\log n}{n})^{1-q/2}$)) as Donoho and Johnstone [30] 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 [55] 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 [96] on the problem of determining minimax upper and lower bounds for ℓ_p-losses with ℓ_q-sparsity for $q > 0$ and $p \geq 1$. There are notable differences between our and Zhang's results. First, we treat the ℓ_2-prediction loss not covered by Zhang, and also show how assumptions on the design X enter in complementary ways for ℓ_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 ℓ_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 ℓ_q-balls. In contrast, Zhang makes use of an extension of the Bayesian least favorable prior approach used by Donoho and Johnstone [30]. Theorems 1 and 2 from his paper [96] (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 [72] 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.

2.2.6 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.

Upper and lower bounds require complementary conditions

It is worth noting that the minimax rates for ℓ_2-prediction error and ℓ_2-norm error are essentially the same except that the design matrix structure enters minimax 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 ℓ_2-norm losses, where lower bounds for $q > 0$ can be proved with only Assumption 1 on X (see Theorem 7), whereas upper bounds require both Assumptions 1 and 2.

In order to appreciate the difference between the conditions for ℓ_2-prediction error and ℓ_2 error, it is useful to consider a toy but illuminating example. Consider the linear regression problem defined by a design matrix $X = \begin{bmatrix} X_1 & X_2 & \cdots & X_d \end{bmatrix}$ with *identical columns*—that is, $X_j = \widetilde{X}_1$ for all $j = 1, \ldots, d$. We assume that vector $\widetilde{X}_1 \in \mathbb{R}^d$ is suitably scaled so that the column-normalization condition (Assumption 1) is satisfied. For this particular choice of design matrix, the linear observation model (3.1) 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 ℓ_2-prediction error scales as $\Theta(\frac{1}{n})$. This scaling implies that the upper bound (2.23) from Theorem 4 holds (but is not tight). It is trivial to prove the correct upper bounds for prediction error using an alternative approach. [3] Consequently, this highly degenerate design matrix yields a very easy problem for ℓ_2-prediction, since the $1/n$ rate is essentially low-dimensional parametric. In sharp contrast, for the case of ℓ_2-norm error (still with hard sparsity $q = 0$), the model becomes unidentifiable. To see the lack of identifiability, let $e_i \in \mathbb{R}^d$ denote the unit-vector with 1 in position i, and consider the two regression vectors

[3] Note that the lower bound (2.21) on the ℓ_2-prediction error from Theorem 3 does not apply to this model, since this degenerate design matrix with identical columns does not satisfy Assumption 3(b).

$\beta^* = c\,e_1$ and $\widetilde{\beta} = c\,e_2$, for some constant $c \in \mathbb{R}$. Both choices yield the same observation vector y, and since the choice of c is arbitrary, the minimax ℓ_2-error is infinite. In this case, the lower bound (2.13) on ℓ_2-error from Theorem 7 holds (and is tight, since the kernel diameter is infinite). In contrast, the upper bound (2.16) on ℓ_2-error from Theorem 2(b) does not apply, because Assumption 3(b) is violated due to the extreme degeneracy of the design matrix.

Comparison to conditions required for ℓ_1-based methods

Naturally, our work also has some connections to the vast body of work on ℓ_1-based methods for sparse estimation, particularly for the case of hard sparsity ($q = 0$). Based on our results, the rates that are achieved by ℓ_1-methods, such as the Lasso and the Dantzig selector, are minimax optimal up to constant factors for ℓ_2-norm loss, and ℓ_2-prediction loss. However the bounds on ℓ_2-error and ℓ_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 ℓ_1-based methods, including the restricted isometry property of Candes and Tao [20], the restricted eigenvalue condition imposed in Meinshausen and Yu [60], the partial Riesz condition in Zhang and Huang [95] and the restricted eigenvalue condition of Bickel et al. [11]. We find that in the case where s is known, “optimal” methods which are based on minimizing least-squares directly over the ℓ_0-ball, can succeed for design matrices where ℓ_1-based methods are not known to work for $q = 0$, as we discuss at more length in Section 2.2.6 to follow. As noted by a reviewer, unlike the direct methods that we have considered, ℓ_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 [20] 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 [95] 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 [60] 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 [60] are weaker or stronger than the conditions in Zhang and Huang [95]. Bickel et al. [11] show that their restricted eigenvalue condition is less severe than both the RIP condition [20] and an earlier set of restricted eigenvalue conditions due to Meinshausen and Yu [60].

Here we state a restricted eigenvalue condition that is very closely related to the condition imposed in Bickel et. al [11], and as shown by Negahban et. al [64], and is sufficient for bounding the ℓ_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

$$\mathcal{C}(S; \alpha) := \big\{\theta \in \mathbb{R}^d \mid \|\theta_{S^c}\|_1 \leq \alpha \,\|\theta_S\|_1 + 4\|\beta^*_{S^c}\|_1\big\}, \tag{2.24}$$

where β^* 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. [11], while for $q \in (0, 1]$, this term is non-zero. With this notation, the restricted eigenvalue condition in Negahban et al. [64] 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 [64] show that under this restricted eigenvalue condition, the Lasso estimator has squared ℓ_2-error upper bounded by $\mathcal{O}\big(R_q(\frac{\log d}{n})^{1-q/2}\big)$. (To be clear, Negahban et al. [64] 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.

Comparison with restricted eigenvalue condition

It is interesting to compare the restricted eigenvalue condition in Bickel et al. [11] 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 ℓ_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. [11]. This fact was previously established by Bickel et al. (see p.7, [11]). We now provide a simple pedagogical example to show that the ℓ_1-based relaxation can fail to recover the true parameter while the optimal ℓ_0-based algorithm succeeds. In particular, let us assume that the noise vector $w = 0$, and consider the design matrix

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

corresponding to a regression problem with $n = 2$ and $d = 3$. Say that the regression vector $\beta^* \in \mathbb{R}^3$ is hard sparse with one non-zero entry (i.e., $s = 1$). Observe that the vector $\Delta := \begin{bmatrix} 1 & 1/3 & 1/3 \end{bmatrix}$ belongs to the null-space of X, and moreover $\Delta \in \mathcal{C}(S; 3)$ but $\Delta \notin \mathbb{B}_0(2)$. All the 2×2 sub-matrices of X have rank two, we have $\mathbb{B}_0(2) \cap \ker(X) = \{0\}$, so that by known results from Cohen et. al. [25] (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 ℓ_0-based algorithm can exactly recover any 1-sparse vector. On the other hand, suppose that, for instance, the true regression vector is given by $\beta^* = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$. If 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 ℓ_1-based methods may be constructed using similar arguments. This construction highlights that there are instances of design matrices X for which ℓ_1-based methods fail to recover the true parameter β^* for $q = 0$ while the optimal ℓ_0-based algorithm succeeds.

In summary, for the hard sparsity case $q = 0$, methods based on ℓ_1-relaxation can achieve the minimax optimal rate $\mathcal{O}\big(\frac{s \log d}{n}\big)$ for ℓ_2-error. However the current analyses of these ℓ_1-methods [11, 20, 60, 84] are based on imposing stronger conditions on the design matrix X than those required by the estimator that performs least-squares over the ℓ_0-ball with s known.

2.3 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.

2.3.1 Basic steps for lower bounds

The proofs for the lower bounds follow an information-theoretic method based on Fano's inequality [26], as used in classical work on nonparametric estimation [47, 92, 93]. A key ingredient is a sharp characterization of the metric entropy structure of ℓ_q balls [21, 51]. 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., [42, 92, 93]) 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 [26]. 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 β over the packing set and Y is the gaussian distribution induced by model (3.1).

(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 [41], whereas for the case $q \in (0, 1]$, we upper bound $I(y; B)$ by a more intricate technique introduced by Yang and Barron [92]. We derive an upper bound on the ϵ_n-covering set for $\mathbb{B}_q(R_q)$ with respect to the ℓ_2-prediction semi-norm. Using Lemma 3 in Section 2.3.3 and the column normalization condition (Assumption 1), we establish a link between covering numbers in ℓ_2-prediction semi-norm to covering numbers in ℓ_2-norm. Finally, we choose the free parameters $\delta_n > 0$ and $\epsilon_n > 0$ so as to optimize the lower bound.

2.3.2 Basic steps for upper bounds

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

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

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 $\|y - X\beta\|_2^2 \leq \|y - X\beta^*\|_2^2$. 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 [83]). For upper bounds in ℓ_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 $\frac{1}{n}\|X\widehat{\Delta}\|_2^2$.

2.3.3 Packing, covering, and metric entropy

The notion of packing and covering numbers play a crucial role in our analysis, so we begin with some background, with emphasis on the case of covering/packing for ℓ_q-balls in ℓ_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 ϵ-covering of $\mathcal{S}$ in the metric ρ is a collection $\{\beta^1, \ldots, \beta^N\} \subset \mathcal{S}$ such that for all $\beta \in S$, there exists some $i \in \{1, \ldots, N\}$ with $\rho(\beta, \beta^i) \leq \epsilon$. The ϵ-covering number $N(\epsilon; \mathcal{S}, \rho)$ is the cardinality of the smallest ϵ-covering.

(b) A δ-packing of $\mathcal{S}$ in the metric ρ is a collection $\{\beta^1, \ldots, \beta^M\} \subset S$ such that $\rho(\beta^i, \beta^j) > \delta$ for all $i \neq j$. The δ-packing number $M(\delta; \mathcal{S}, \rho)$ is the cardinality of the largest δ-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., [66]). 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 [21, 51, 78], are the (dyadic) entropy numbers $\epsilon_k(\mathcal{S}; \rho)$, defined as follows for $k = 1, 2, \ldots$

$$\epsilon_k(\mathcal{S}; \rho) := \inf \big\{ \epsilon > 0 \mid N(\epsilon; \mathcal{S}, \rho) \leq 2^{k-1} \big\}. \tag{2.25}$$

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 ℓ_2-packing and covering numbers by $M(\epsilon; \mathcal{S})$ and $N(\epsilon; \mathcal{S})$.

Metric entropies of ℓ_q-balls

Central to our proofs is the metric entropy of the ball $\mathbb{B}_q(R_q)$ when the metric ρ is the ℓ_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 [38, 51]; see the Appendix of Raskutti et al. [68] 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*

$$\log N(\epsilon; \mathbb{B}_q(R_q)) \;\leq\; U_q \, \Big({R_q}^{\frac{2}{2-q}} \big(\frac{1}{\epsilon}\big)^{\frac{2q}{2-q}} \log d \Big). \tag{2.26}$$

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

$$\log N(\epsilon; \mathbb{B}_q(R_q)) \;\geq\; L_q \, \Big({R_q}^{\frac{2}{2-q}} \big(\frac{1}{\epsilon}\big)^{\frac{2q}{2-q}} \log d \Big). \tag{2.27}$$

Remark: In our application of the lower bound (2.27), our typical choice of ϵ^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 (3.10) from Section 2.2.2, we are guaranteed that ϵ lies in the range required for the upper and lower bounds (2.26) and (2.27) to be valid.

Metric entropy of q-convex hulls

The proofs of the lower bounds all involve the Kullback-Leibler (KL) divergence between the distributions induced by different parameters β and β' in $\mathbb{B}_q(R_q)$. Here we show that for the linear observation model (3.1), these KL divergences can be represented as q-convex hulls of the columns of the design matrix, and provide some bounds on the associated metric entropy.

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

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

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

$$\text{absconv}_q(X/\sqrt{n}) := \Big\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{d} \theta_j X_j \;\mid\; \theta \in \mathbb{B}_q(R_q) \Big\}. \tag{2.28}$$

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 ℓ_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 κ_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*

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

The proof of this claim is provided in Appendix A in Raskutti et al. [68]. 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.

2.3.4 Proof of lower bounds

We begin by proving our main results that provide lower bounds on minimax rates, namely Theorems 7 and 3.

Proof of Theorem 7

Recall that for ℓ_2-norm error, the lower bounds in Theorem 7 are the maximum of two expressions, one corresponding to the diameter of the set $\mathcal{N}_q(X)$ intersected with the ℓ_q-ball, and the other correspond to the metric entropy of the ℓ_q-ball.

We begin by deriving the lower bound based on the diameter of $\mathcal{N}_q(X) = \mathbb{B}_q(R_q) \cap \ker(X)$. The minimax 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 $\|\widehat{\beta} - \beta\|_2 \geq \frac{1}{2} \operatorname{diam}_2(\mathcal{N}_q(X))$.

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

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

Overall, we conclude that

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

In the following subsections, we follow steps (1)–(3) outlined earlier so as to obtain the second term in our lower bounds on the ℓ_2-norm error and the ℓ_2-prediction error. As has already been mentioned, steps (1) and (2) are general, whereas step (3) is different in each case.

Proof of Theorem 7(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 ℓ_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

$$\mathbb{P}[B \neq \widetilde{\beta}] \geq 1 - \frac{I(y; B) + \log 2}{\log M(\delta_n, \mathbb{B}_q(R_q))}, \tag{2.29}$$

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 [92]. 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 ϵ_n-covering of $\mathbb{B}_q(R_q)$ in ℓ_2-norm. Using the upper bound on the metric entropy of $\operatorname{absconv}_q(X)$ provided by Lemma 3, we conclude that there exists a set $\{X\beta^1, \ldots, X\beta^N\}$ such that for all $X\beta \in \mathrm{absconv}_q(X)$, 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 [92], 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 (2.29) yields

$$\mathbb{P}[B \neq \widetilde{\beta}] \geq 1 - \frac{\log N(\epsilon_n; \mathbb{B}_q(R_q)) + \frac{c^2\,n}{\sigma^2}\kappa_c^2\,\epsilon_n^2 + \log 2}{\log M(\delta_n; \mathbb{B}_q(R_q))}. \tag{2.30}$$

The final step is to choose the packing and covering radii (δ_n and ϵ_n respectively) such that the lower bound (2.30) is greater than 1/2. In order to do so, suppose that we choose the pair (ϵ_n, δ_n) such that

$$\frac{c^2\,n}{\sigma^2}\kappa_c^2\,\epsilon_n^2 \leq \log N(\epsilon_n, \mathbb{B}_q(R_q)), \tag{2.31a}$$

$$\log M(\delta_n, \mathbb{B}_q(R_q)) \geq 4 \log N(\epsilon_n, \mathbb{B}_q(R_q)). \tag{2.31b}$$

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 ϵ_n and δ_n that satisfy the relations (2.31). From Lemma 2, relation (2.31a) is satisfied by choosing ϵ_n such that $\frac{c^2\,n}{2\sigma^2}\kappa_c^2\,\epsilon_n^2 \;=\; L_q\;\Big[R_q{}^{\frac{2}{2-q}}\big(\frac{1}{\epsilon_n}\big)^{\frac{2q}{2-q}}\log d\Big]$, 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 (2.31b), it suffices to choose δ_n such that

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

or equivalently such that

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

Substituting into equation (2.12), 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 7(a).

Proof of Theorem 7(b) In order to prove Theorem 7(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, [13]), for completeness, we provide a proof of Lemma 4 in Appendix D of Raskutti et al. [68]. 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 ℓ_2-norm of their difference for all elements $\beta, \beta' \in \sqrt{\frac{2}{s}}\delta_n \widetilde{\mathcal{H}}$:

$$\|\beta - \beta'\|_2^2 \geq \delta_n^2, \quad \text{and} \tag{2.32a}$$

$$\|\beta - \beta'\|_2^2 \leq 8\delta_n^2. \tag{2.32b}$$

Consequently, the rescaled set $\sqrt{\frac{2}{s}}\delta_n\widetilde{\mathcal{H}}$ is an δ_n-packing set of $\mathbb{B}_0(s)$ in ℓ_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 7(a), up until the Fano lower bound (2.29) (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 [26]. For the linear observation model (3.1), 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 (2.32b), we conclude that

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

Substituting this upper bound into the Fano lower bound (2.29), 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 (2.12), 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.

Proof of Theorem 3

We use arguments similar to the proof of Theorem 7 in order to establish lower bounds on prediction error $\|X(\widehat{\beta} - \beta^*)\|_2/\sqrt{n}$.

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

$$\delta_n^2 := \bar{c}\,R_q\,\big(\frac{\sigma^2}{\kappa_c^2}\,\frac{\log d}{n}\big)^{1-q/2}, \tag{2.33}$$

and let $\{\beta^1, \ldots, \beta^M\}$ be an δ_n packing of the ball $\mathbb{B}_q(R_q)$ in the ℓ_2 metric, say with a total of $M(\delta_n; \mathbb{B}_q(R_q))$ 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 $f_\ell(R_q, n, d) \leq c_2 R_q \left(\frac{\log d}{n}\right)^{1-q/2}$ and $R_q \left(\frac{\log d}{n}\right)^{1-q/2} < c_1$. From Assumption 2, for each $i \neq j$, we are guaranteed that

$$\frac{\|X(\beta^i - \beta^j)\|_2}{\sqrt{n}} \geq \kappa_\ell \|\beta^i - \beta^j\|_2, \tag{2.34}$$

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

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

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

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

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 $f_\ell(R_q, n, d) \leq c_2 R_q \left(\frac{\log d}{n}\right)^{1-q/2}$ by assumption, the lower bound (2.34) 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 7(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[4] 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 $\text{absconv}_q(X)$ is upper bounded (up to a constant factor) by the ϵ covering number of $\mathbb{B}_q(R_q)$ in ℓ_2-norm, which we denote by $N(\epsilon_n; \mathbb{B}_q(R_q))$. Following the same reasoning as in Theorem 2(a), the mutual information is upper bounded as

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

[4] 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.

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

$$1 - \frac{\log N(\epsilon_n; \mathbb{B}_q(R_q)) + \frac{n}{\sigma^2}\,\kappa_c^2\,\epsilon_n^2 + \log 2}{\log M(\delta_n; \mathbb{B}_q(R_q))}. \tag{2.35}$$

Lastly, we choose the packing and covering radii (δ_n and ϵ_n respectively) such that the lower bound (2.35) remains bounded below by 1/2. As in the proof of Theorem 7(a), it suffices to choose the pair (ϵ_n, δ_n) to satisfy the relations (2.31a) and (2.31b). The same choice of ϵ_n ensures that relation (2.31a) holds; moreover, by making a sufficiently small choice of the universal constant $\bar{c}$ in the definition (2.33) of δ_n, we may ensure that the relation (2.31b) 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.

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 δ_n-packing set in ℓ_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 (2.32a) and (2.32b) 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

$$\frac{\|X(\beta^i - \beta^j)\|}{\sqrt{n}} \;\leq\; \kappa_u \|\beta^i - \beta^j\|_2 \;\leq\; \kappa_u \sqrt{8}\,\delta_n. \tag{2.36}$$

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

$$\frac{\|X(\beta^i - \beta^j)\|_2}{\sqrt{n}} \geq \kappa_{0,\ell}\delta_n. \tag{2.37}$$

We can follow the same steps as in the proof of Theorem 7(b), thereby obtaining an upper bound the mutual 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 κ_ℓ from the lower bound (2.37), 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 7(b) allows us to further lower bound this quantity in terms of $\log(d/s)$, leading to the claimed form of the bound.

2.3.5 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

$$\widehat{\beta} \in \arg \min_{\|\beta\|_q^q \leq R_q} \|y - X\beta\|_2^2. \tag{2.38}$$

Note that (2.38) 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

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \frac{2|w^T X\widehat{\Delta}|}{n}. \tag{2.39}$$

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

Proof of Theorem 2

We begin with upper bounds on the minimax rate in squared ℓ_2-norm.

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 $R_q\,\big(\frac{\log d}{n}\big)^{1-q/2} < c_2$. 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

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

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 (4.24) 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 (2.39), 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

$$\kappa_\ell^2 \|\widehat{\Delta}\|_2^2 \leq 2\sigma\kappa_c \sqrt{\frac{3\log d}{n}}\, \|\widehat{\Delta}\|_1 \tag{2.40}$$

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

It remains to upper bound the ℓ_1-norm in terms of the ℓ_2-norm and a residual term. Since both $\widehat{\beta}$ and β^* belong to $\mathbb{B}_q(R_q)$, we have $\|\widehat{\Delta}\|_q^q = \sum_{j=1}^d |\widehat{\Delta}_j|^q \leq 2R_q$. 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*

$$\|\theta\|_1 \leq \sqrt{2R_q}\tau^{-q/2}\|\theta\|_2 + 2R_q\tau^{1-q}. \tag{2.41}$$

Although this type of result is standard (e.g, [30]), we provide a proof in Appendix A of Raskutti et al. [68].

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).

Proof of Theorem 2(b) In order to establish the bound (2.16), we follow the same steps with $f_\ell(s, n, d) = 0$, thereby obtaining the following simplified form of the bound (2.40):

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

By definition of the estimator, we have $\|\widehat{\Delta}\|_0 \le 2s$, from which we obtain $\|\widehat{\Delta}\|_1 \le \sqrt{2s}\|\widehat{\Delta}\|_2$. Canceling out a factor of $\|\widehat{\Delta}\|_2$ from both sides yields the claim (2.16).

Establishing the sharper upper bound (2.17) requires more precise control on the right-hand side of the inequality (2.39). The following lemma, proved in Appendix A of Raskutti et al. [68], provides this control:

Lemma 6. *Suppose that $\frac{\|X\theta\|_2}{\sqrt{n}\|\theta\|_2} \le \kappa_u$ 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*

$$\sup_{\|\theta\|_0 \le 2s, \|\theta\|_2 \le r} \frac{1}{n} \big|w^T X \theta\big| \le 6 \, \sigma \, r \, \kappa_u \sqrt{\frac{s \log(d/s)}{n}} \tag{2.42}$$

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 in Appendix H of Raskutti et al. [68]. In particular, consider the event $\mathcal{E}$ that there exists some $\theta \in \mathbb{B}_0(2s)$ such that

$$\frac{1}{n} \left|w^T X \theta\right| \ge 6 \sigma \kappa_u \|\theta\|_2 \sqrt{\frac{s \log(d/s)}{n}}. \tag{2.43}$$

Then we claim that

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

for some $c > 0$. This claim follows from Lemma 9 in Appendix H of Raskutti et al. [68] 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 $r \ge \sigma \kappa_u \sqrt{\frac{s \log(d/s)}{n}}$, we are guaranteed that $g(r) \ge \sigma^2 \kappa_u^2 \frac{s \log(d/s)}{n}$, and $\mu = \sigma^2 \kappa_u^2 \frac{s \log(d/s)}{n}$, so that Lemma 9 in Appendix H 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.

Proof of Theorem 4

We again make use of the elementary inequality (2.39) to establish upper bounds on the prediction error.

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

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

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

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 G. 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 in Appendix H 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 (2.39), when the event $\mathcal{E}$ from equation (2.43) 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 σ^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.

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

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

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

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

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

We now apply Lemma 9 in Appendix H with the sequence $a_n = n$, the function $f(v; X) = \frac{1}{n} |w^T X v|$, 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 in Appendix H. 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 (2.39), 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.

2.4 Conclusion

The main contribution of this paper is to obtain optimal minimax rates of convergence for the linear model (3.1) 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 ℓ_2-norm and ℓ_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 ℓ_2-prediction error for general design matrices X and general $q \in [0, 1]$. We also derive optimal minimax rates in ℓ_2-error, with similar rates obtained in concurrent work by Zhang [96] 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 ℓ_2-norm and ℓ_2-prediction loss functions. On one hand, it is possible to obtain lower bounds on ℓ_2-norm error (see Theorem 7) or upper bounds on ℓ_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 ℓ_2-norms (see Assumption 1). On the other hand, in order to obtain upper bounds on ℓ_2-norm error (Theorem 2) or lower bound on ℓ_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 ℓ_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 $X^T X$ measured over the ℓ_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 ℓ_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 [69].

Chapter 3

Restricted eigenvalue properties for correlated Gaussian designs

3.1 Introduction

Using the ℓ_1-norm to enforce sparsity has been very successful, as evidenced by the widespread use of methods such as basis pursuit [23], the Lasso [82] and the Dantzig selector [20]. There is now a well-developed theory on what conditions are required on the design matrix $X \in \mathbb{R}^{n \times d}$ for such ℓ_1-based relaxations to reliably estimate β^*. In the case of noiseless observation models, it is known that imposing a *restricted nullspace property* on the design matrix $X \in \mathbb{R}^{n \times d}$ is both necessary and sufficient for the basis pursuit linear program to recover β^* exactly. The nullspace property and its link to the basis pursuit linear program has been discussed in various papers [25, 29, 33]. In the case of noisy observations, exact recovery of β^* is no longer possible, and one goal is to obtain an estimate $\widehat{\beta}$ such that the ℓ_2-error $\|\widehat{\beta} - \beta^*\|_2$ is well-controlled. To this end, various sufficient conditions for the success of ℓ_1-relaxations have been proposed, including restricted eigenvalue conditions [11, 60] and the restricted Riesz property [95]. Of the conditions mentioned, one of weakest known sufficient conditions for bounding ℓ_2-error are the restricted eigenvalue (RE) conditions due to [11] and [84]. In this chapter, we consider a restricted eigenvalue condition that is essentially equivalent to the RE condition in [11]. As shown by [70], a related restriction is actually necessary for obtaining good control on the ℓ_2-error in the minimax setting.

Thus, in the setting of linear regression with random design, the interesting question is the following: for what ensembles of design matrices do the restricted nullspace and eigenvalue conditions hold with high probability? To date, the main routes to establishing these properties have been via either incoherence conditions [29, 33] or via the restricted isometry property [19], both of which are sufficient but not necessary conditions [25, 85]. The restricted isometry property (RIP) holds with high probability for various classes of random matrices with i.i.d. entries, including sub-Gaussian matrices [62] with sample size

$n = \Omega(s \log(d/s))$, and for i.i.d. subexponential random matrices [1] provided that $n = \Omega(s \log^2(d/s))$. It has also been demonstrated that RIP is satisfied for matrices from unitary ensembles (e.g., [39, 40, 74, 75]), for which the rows are generated based on independent draws from a set of uncorrelated basis functions.

Design matrices based on i.i.d. or unitary ensembles are well-suited to the task of compressed sensing [19, 28], where the matrix X can be chosen by the user. However, in most of machine learning and statistics, the design matrix is not under control of the statistician, but rather is specified by nature. As a concrete example, suppose that we are fitting a linear regression model to predict heart disease on the basis of a set of d covariates (e.g., diet, exercise, smoking etc.). In this setting, it is not reasonable to assume that the different covariates are i.i.d. or unitary—for instance, one would expect a strong positive correlation between amount of exercise and healthiness of diet. Nonetheless, at least in practice, ℓ_1-methods still work very well in settings where the covariates are correlated and non-unitary, but currently lacking is the corresponding theory that guarantees the performance of ℓ_1-relaxations for dependent designs.

The main contribution of this chapter is a direct proof that both the restricted nullspace and eigenvalue conditions hold with high probability for a broad class of dependent Gaussian design matrices. In conjunction with known results on ℓ_1-relaxation, our main result implies as corollaries that the basis pursuit algorithm reliably recovers β^* exactly in the noiseless setting, and that in the case of observations contaminated by Gaussian noise, the Lasso and Dantzig selectors produces a solution $\widehat{\beta}$ such that $\|\widehat{\beta} - \beta^*\|_2 = \mathcal{O}(\sqrt{\frac{s \log d}{n}})$. Our theory requires that the sample size n scale as $n = \Omega(s \log d)$, where s is the sparsity index of the regression vector β^* and d is its dimensions. For sub-linear sparsity ($s/d \to 0$), this scaling matches known optimal rates in a minimax sense for the sparse regression problem [70], and hence cannot be improved upon by any algorithm. The class of matrices covered by our result allows for correlation among different covariates, and hence covers many matrices for which restricted isometry or incoherence conditions fail to hold but the restricted eigenvalue condition holds. Interestingly, one can even sample the rows of the design matrix X from a multivariate Gaussian with a degenerate covariance matrix Σ, and nonetheless, our results still guarantee that the restricted nullspace and eigenvalue conditions will hold with high probability (see Section 3.3.3). Consequently, our results extend theoretical guarantees on ℓ_1-relaxations with optimal rates of convergence to a much broader class of random designs.

The remainder of this chapter is organized as follows. We begin in Section 4.2 with background on sparse linear models, the basis pursuit and Lasso ℓ_1-relaxations, and sufficient conditions for their success. In Section 4.3, we state our main result, discuss its consequences for ℓ_1-relaxations, and illustrate it with some examples. Section 3.4 contains the proof of our main result, which exploits Gaussian comparison inequalities and concentration of measure for Lipschitz functions.

3.2 Background

We begin with background on sparse linear models and sufficient conditions for the success of ℓ_1-relaxations.

3.2.1 High-dimensional sparse models and ℓ_1-relaxation

In the classical linear model, a scalar output $y_i \in \mathbb{R}$ is linked to a d-dimensional vector $X_i \in \mathbb{R}^d$ of covariates via the relation $y_i = X_i^T \beta^* + w_i$, where w_i is a scalar observation noise. If we make a set of n such observations, then they can be written in the matrix-vector form

$$y \;=\; X\beta^* + w, \tag{3.1}$$

where $y \in \mathbb{R}^n$ is the vector of outputs, the matrix $X \in \mathbb{R}^{n\times d}$ is the set of covariates (in which row $X_i \in \mathbb{R}^d$ represents the covariates for i^{th} observation), and $w \in \mathbb{R}^n$ is a noise vector where $w \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$. Given the pair (y, X), the goal is to estimate the unknown regression vector $\beta^* \in \mathbb{R}^d$.

In many applications, the linear regression model is high-dimensional in nature, meaning that the number of observations n may be substantially smaller than the number of covariates d. In this $d \gg n$ regime, it is easy to see that without further constraints on β^*, the statistical model (3.1) is not identifiable, since (even when $w = 0$), there are many vectors β^* that are consistent[1] with the observations y and X. This identifiability concern may be eliminated by imposing some type of sparsity assumption on the regression vector $\beta^* \in \mathbb{R}^d$. The simplest assumption is that of *exact sparsity*: in particular, we say that $\beta^* \in \mathbb{R}^d$ is s-sparse if its support set

$$S(\beta^*) := \big\{ j \in \{1, \ldots, d\} \;\mid\; \beta^*_j \neq 0 \big\} \tag{3.2}$$

has cardinality at most s.

Disregarding computational cost, the most direct approach to estimating an s-sparse β^* in the linear regression model would be solving a quadratic optimization problem with an ℓ_0-constraint, say

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^d} \|y - X\beta\|_2^2 \qquad \text{such that } \|\beta\|_0 \leq s, \tag{3.3}$$

where $\|\beta\|_0$ simply counts the number of non-zero entries in β. Of course, this problem is non-convex and combinatorial in nature, since it involves searching over all $\binom{d}{s}$ subsets of size s. A natural relaxation is to replace the non-convex ℓ_0 constraint with the ℓ_1-norm,

[1] Indeed, any vector β^* in the nullspace of X, which has dimension at least $d - n$, leads to $y = 0$ when $w = 0$.

which leads to the *constrained form of the Lasso* [23, 82], given by

$$\widehat{\beta} \in \arg\min_{\beta\in\mathbb{R}^d} \|y - X\beta\|_2^2 \qquad \text{such that } \|\beta\|_1 \leq R, \tag{3.4}$$

where R is a radius to be chosen by the user. Equivalently, by Lagrangian duality, this program can also be written in the penalized form

$$\widehat{\beta} \in \arg\min_{\beta\in\mathbb{R}^d} \big\{\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\big\}, \tag{3.5}$$

where $\lambda > 0$ is a regularization parameter. In the case of noiseless observations, obtained by setting $w = 0$ in the observation model (3.1), a closely related convex program is the *basis pursuit linear program* [23], given by

$$\widehat{\beta} \in \arg\min_{\widehat{\beta}\in\mathbb{R}^d} \|\beta\|_1 \qquad \text{such that } X\beta = y. \tag{3.6}$$

[23] also study the constrained Lasso (3.4), which they refer to as relaxed basis pursuit. Another closely related estimator based on ℓ_1-relaxation is the Dantzig selector [20].

3.2.2 Sufficient conditions for success

The high-dimensional linear model under the exact sparsity constraint has been extensively analyzed. Accordingly, as we discuss here, there is a good understanding of the necessary and sufficient conditions for the success of ℓ_1-based relaxations such as basis pursuit and the Lasso.

Restricted nullspace in noiseless setting: In the noiseless setting ($w = 0$), it is known that the basis pursuit linear program (LP) (3.6) recovers β^* exactly if and only if the design matrix X satisfies a restricted nullspace condition. In particular, for a given subset $S \subset \{1, \ldots, d\}$ and constant $\alpha \geq 1$, let us define the set

$$\mathcal{C}(S;\alpha) := \big\{\theta \in \mathbb{R}^d \mid \|\theta_{S^c}\|_1 \leq \alpha\, \|\theta_S\|_1\big\}. \tag{3.7}$$

For a given sparsity index $s \leq d$, we say that the matrix X satisfies the *restricted nullspace (RN) condition* of order s if $\text{null}(X) \cap \mathcal{C}(S;1) = \{0\}$ for all subsets S of cardinality s. Although this definition appeared in earlier work [29, 33], the terminology of restricted nullspace is due to [25].

This restricted nullspace property is important, because the basis pursuit LP recovers any vector s-sparse vector β^* exactly if and only if X satisfies the restricted nullspace property of order s. See the chapters [25, 29, 32, 33] for more discussion of restricted nullspaces and equivalence to exact recovery of basis pursuit.

Restricted eigenvalue condition for ℓ_2 error: In the noisy setting, it is impossible to recover β^* exactly, and a more natural criterion is to bound the ℓ_2-error between β^* and an estimate $\widehat{\beta}$. Various conditions have been used to analyze the ℓ_2-norm convergence rate of ℓ_1-based methods, including the restricted isometry property [20], various types of restricted eigenvalue conditions [84, 11, 60], and a partial Riesz condition [95]. Of all these conditions, the least restrictive are the restricted eigenvalue conditions due to [11] and [84]. As shown by [11], their restricted eigenvalue (RE) condition is less severe than both the RIP condition [20] and an earlier set of restricted eigenvalue conditions due to [60]. All of these conditions involve lower bounds on $\|X\theta\|_2$ that hold uniformly over the previously defined set $\mathcal{C}(S;\alpha)$,

Here we state a condition that is essentially equivalent to the restricted eigenvalue condition due to [11]. In particular, we say that the $d\times d$ sample covariance matrix X^TX/n satisfies the *restricted eigenvalue (RE) condition* over S with parameters $(\alpha,\gamma) \in [1,\infty) \times (0,\infty)$ if

$$\frac{1}{n}\theta^T X^T X\theta \;=\; \frac{1}{n}\|X\theta\|_2^2 \;\geq\; \gamma^2\,\|\theta\|_2^2 \qquad \text{for all } \theta \in \mathcal{C}(S;\alpha). \tag{3.8}$$

If this condition holds uniformly for all subsets S with cardinality s, we say that X^TX/n *satisfies a restricted eigenvalue condition of order s with parameters* (α,γ). On occasion, we will also say that a deterministic $d \times d$ covariance matrix Σ satisfies an RE condition, by which we mean that $\|\Sigma^{1/2}\theta\|_2 \geq \gamma\|\theta\|_2$ for all $\theta \in \mathcal{C}(S;\alpha)$. It is straightforward to show that the RE condition for some α implies the restricted nullspace condition for the same α, so that the RE condition is slightly stronger than the RN property.

Again, the RE condition is important because it yields guarantees on the ℓ_2-error of any Lasso estimate $\widehat{\beta}$. For instance, if X satisfies the RE condition of order s with parameters $\alpha \geq 3$ and $\gamma > 0$, then it can be shown that (with appropriate choice of the regularization parameter) any Lasso estimate $\widehat{\beta}$ satisfies the error bound $\|\widehat{\beta} - \beta^*\|_2 = \mathcal{O}(\sqrt{\frac{s\log d}{n}})$ with high probability over the Gaussian noise vector w. A similar result holds for the Dantzig selector provided the RE condition is satisfied for $\alpha \geq 1$. Bounds with this scaling have appeared in various chapters on sparse linear models [18, 11, 20, 60, 84, 85]. Moreover, this ℓ_2-convergence rate is known to be minimax optimal [68] in the regime $s/d \to 0$.

3.3 Main result and its consequences

Thus, in order to provide performance guarantees (either exact recovery or ℓ_2-error bounds) for ℓ_1-relaxations applied to sparse linear models, it is sufficient to show that the RE or RN conditions hold. Given that our interest is in providing sufficient conditions for these properties, the remainder of the chapter focuses on providing conditions for the RE condition to hold for random designs, which implies that the RN condition is satisfied.

3.3.1 Statement of main result

Our main result guarantees that the restricted eigenvalue (and hence restricted nullspace) conditions hold for a broad class of random Gaussian designs. In particular, we consider the linear model $y_i = X_i^T \beta^* + w_i$, in which each row $X_i \sim \mathcal{N}(0, \Sigma)$. We define $\rho^2(\Sigma) = \max_{j=1,\ldots,d} \Sigma_{jj}$ to be the maximal variance, and let $\Sigma^{1/2}$ denote the square root of Σ.

Theorem 5. *For any Gaussian random design $X \in \mathbb{R}^{n \times d}$ with i.i.d. $\mathcal{N}(0, \Sigma)$ rows, there are universal positive constants c, c' such that*

$$\frac{\|Xv\|_2}{\sqrt{n}} \;\geq\; \frac{1}{4}\|\Sigma^{1/2}v\|_2 - 9\,\rho(\Sigma)\,\sqrt{\frac{\log d}{n}}\,\|v\|_1 \qquad \textit{for all } v \in \mathbb{R}^d, \tag{3.9}$$

with probability at least $1 - c' \exp(-cn)$.

The proof of this claim is given later in Section 3.4. Note that we have not tried to obtain sharpest possible leading constants (i.e., the factors of 1/4 and 9 can easily be improved).

In intuitive terms, Theorem 5 provides some insight into the eigenstructure of the sample covariance matrix $\widehat{\Sigma} = X^T X/n$. One implication of the lower bound (3.9) is that the nullspace of X cannot contain any vectors that are "overly" sparse. In particular, for any vector $v \in \mathbb{R}^d$ such that $\|v\|_1/\|\Sigma^{1/2}v\|_2 = o(\sqrt{\frac{n}{\log d}})$, the right-hand side of the lower bound (3.9) will be strictly positive, showing that v cannot belong to the nullspace of X. In the following corollary, we formalize this intuition by showing how Theorem 5 guarantees that whenever the population covariance Σ satisfies the RE condition of order s, then the sample covariance $\widehat{\Sigma} = X^T X/n$ satisfies the same property as long as the sample size is sufficiently large.

Corollary 1 (Restricted eigenvalue property)**.** *Suppose that Σ satisfies the RE condition of order s with parameters (α, γ). Then for universal positive constants c, c', c'', if the sample size satisfies*

$$n \;>\; c''\,\frac{\rho^2(\Sigma)\,(1+\alpha)^2}{\gamma^2}\,s\log d, \tag{3.10}$$

then the matrix $\widehat{\Sigma} = X^T X/n$ satisfies the RE condition with parameters $(\alpha, \frac{\gamma}{8})$ with probability at least $1 - c' \exp(-cn)$.

Proof. Let S be an arbitrary subset of cardinality s, and suppose that $v \in \mathcal{C}(S; \alpha)$. By definition, we have

$$\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1 \;\leq\; (1+\alpha)\|v_S\|_1,$$

and consequently $\|v\|_1 \leq (1+\alpha)\sqrt{s}\|v\|_2$. By assumption, we also have $\|\Sigma^{1/2}v\|_2 \geq \gamma\|v\|_2$ for all $v \in \mathcal{C}(S;\alpha)$. Substituting these two inequalities into the bound (3.9) yields

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \left\{\frac{\gamma}{4} - 9\,(1+\alpha)\,\rho(\Sigma)\,\sqrt{\frac{s\log d}{n}}\right\}\|v\|_2.$$

Under the assumed scaling (3.10) of the sample size, we have

$$9\,(1+\alpha)\,\rho(\Sigma)\,\sqrt{\frac{s\log d}{n}} \leq \gamma/8,$$

which shows that the RE condition holds for X^TX/n with parameter $(\alpha, \gamma/8)$ as claimed. □

Remarks:

(a) From the definitions, it is easy to see that if the RE condition holds with $\alpha = 1$ and any $\gamma > 0$ (even arbitrarily small), then the RN condition also holds. Indeed, if the matrix X^TX/n satisfies the $(1,\gamma)$-RE condition, then for any $v \in \mathcal{C}(S;1)\backslash\{0\}$, we have $\frac{\|Xv\|_2}{\sqrt{n}} \geq \gamma\|v\|_2 > 0$, which implies that $\mathcal{C}(S,1) \cap (X) = \{0\}$.

(b) As previously discussed, it is known [11, 83, 85] that if X^TX/n satisfies the RE condition, then the ℓ_2 error of the Lasso under the sparse linear model with Gaussian noise satisfies the bound

$$\|\widehat{\beta} - \beta^*\|_2 = \mathcal{O}(\sqrt{\frac{s\log d}{n}}) \qquad \text{with high probability.}$$

Consequently, in order to ensure that the ℓ_2-error is bounded, the sample size must scale as $n = \Omega(s\log d)$, which matches the scaling (3.10) required in Corollary 1, as long as the sequence of covariance matrices Σ have diagonal entries that stay bounded.

(c) Finally, we note that Theorem 5 guarantees that the sample covariance X^TX/n satisfies a property that is slightly stronger than the RE condition. As shown by [64], this strengthening also leads to error bounds for the Lasso when β^* is not exactly s-sparse, but belongs to an ℓ_q-ball. The resulting rates are known to be minimax-optimal for these ℓ_q-balls [70].

3.3.2 Comparison to related work

At this point, we provide a brief comparison of our results with some related results in the literature beyond the chapters discussed in the introduction. [44] showed that a certain class of random Toeplitz matrices, where the entries in the first row and first column are Bernoulli random variables and the rest fill out the Toeplitz structure satisfy RIP (and hence the weaker RE condition) provided that $n = \Omega(s^3 \log(d/s))$. In Section 3.3.3, we demonstrate that Gausssian design matrices where the covariance matrix is a Toeplitz matrix satisfies the RE condition under the milder scaling requirement $n = \Omega(s \log(d))$. It would be of interest to determine such scaling can be established for the random Toeplitz matrices considered by [44].

It is worth comparing the scaling (3.10) to a related result due to [85]. In particular, their Lemma 10.1 also provides sufficient conditions for a restricted eigenvalue condition to hold for design matrices with dependent columns. They show that if the true covariance matrix satisfies an RE condition, and if the elementwise maximum $\|\widehat{\Sigma} - \Sigma\|_\infty$ between the sample covariance $\widehat{\Sigma} = X^T X/n$ and true covariance Σ is suitably bounded, then the sample covariance also satisfies the RE condition. Their result applied to the case of Gaussian random matrices guarantees that $\widehat{\Sigma}$ satisfies the RE property as long as $n = \Omega(s^2 \log d)$ and Σ satisfies the RE condition. By contrast, Corollary 1 guarantees the RE condition with the less restrictive scaling $n = \Omega(s \log d)$. Note that if $s = \mathcal{O}(\sqrt{n})$, our scaling condition is satisfied while their condition fails. This quadratic-linear gap in sparsity between s^2 and s arises from the difference between a local analysis (looking at individual entries of $\widehat{\Sigma}$) versus the global analysis of this chapter, which studies the full random matrix. On the other hand, the result of [85] applies more generally, including the case of sub-Gaussian random matrices (e.g., those with bounded entries) in addition to the Gaussian matrices considered in Theorem 5.

Finally, in work that followed up on the initial posting of this work [70], a chapter by [98] provides an extension of Theorem 5 to the case of correlated random matrices with sub-Gaussian entries. Theorem 1.6 in her chapter establishes that certain families of sub-Gaussian matrices satisfy the RE condition w.h.p. with sample size $n = \Omega(s \log(d/s)$. This extension is based on techniques developed by [62], while we use Gaussian comparison inequalities and simple concentration results for the case of Gaussian design.

3.3.3 Some illustrative examples

Let us illustrate some classes of matrices to which our theory applies. We will see that Corollary 1 applies to many sequences of covariance matrices $\Sigma = \Sigma_{d \times d}$ that have much more structure than the identity matrix. Our theory allows for the maximal eigenvalue of Σ to be arbitrarily large, or for Σ to be rank-degenerate, or for both of these degeneracies to occur at the same time. In all cases, we consider sequences of matrices for which the maximum variance $\rho^2(\Sigma) = \max_{j=1,\ldots,d} \Sigma_{jj}$ stays bounded. Under this mild restriction, we provide several examples where the RE condition is satisfied with high probability. For suitable choices, these same examples show that the RE condition can hold with high probability, even when the restricted isometry property (RIP) of [19] is violated with probability converging to one.

Example 1 (Toeplitz matrices)**.** Consider a covariance matrix with the Toeplitz structure $\Sigma_{ij} = a^{|i-j|}$ for some parameter $a \in [0, 1)$. This type of covariance structure arises naturally from autoregressive processes, where the parameter a allows for tuning of the memory in the process. The minimum eigenvalue of Σ is $1 - a > 0$, independent of the dimension d, so that the population matrix Σ clearly satisfies the RE condition. Since $\rho^2(\Sigma) = 1$, Theorem 5 implies that the sample covariance matrix $\widehat{\Sigma} = X^T X/n$ obtained by sampling from this distribution will also satisfy the RE condition with high probability as long as $n = \Omega(s \log d)$. This provides an example of a matrix family with substantial correlation between covariates for which the RE property still holds.

However, regardless of the sample size, the submatrices of the sample covariance $\widehat{\Sigma}$ will not satisfy restricted isometry properties (RIP) if the parameter a is sufficiently large. For instance, defining $S = \{1, 2, \ldots, s\}$, consider the sub-block $\widehat{\Sigma}_{SS}$ of the sample covariance matrix. Satisfying RIP requires that that the condition number $\lambda_{\max}(\widehat{\Sigma}_{SS})/\lambda_{\min}(\widehat{\Sigma}_{SS})$ be very close to one. As long as $n = \Omega(s \log d)$, known results in random matrix theory [27] guarantee that the eigenvalues of $\widehat{\Sigma}_{SS}$ will be very close to the population versions Σ_{SS}; see also the concrete calculation in Example 2 to follow. Consequently, imposing RIP amounts to requiring that the population condition number $\lambda_{\max}(\Sigma_{SS})/\lambda_{\min}(\Sigma_{SS})$ be very close to one. This condition number grows as the parameter $a \in [0, 1)$ increases towards one [35], so RIP will be violated once $a < 1$ is sufficiently large.

We now consider a matrix family with an even higher amount of dependency among the covariates, where the RIP constants are actually unbounded as the sparsity s increases, but the RE condition is still satisfied.

Example 2 (Spiked identity model)**.** For a parameter $a \in [0, 1)$, the spiked identity model is given by the family of covariance matrices

$$\Sigma := (1 - a) I_{d \times d} + a \vec{1}\, \vec{1}^T, \tag{3.11}$$

where $\vec{1} \in \mathbb{R}^d$ is the vector of all ones. The minimum eigenvalue of Σ is $1 - a$, so that the population covariance clearly satisfies the RE condition for any fixed $a \in [0, 1)$. Since this covariance matrix has maximum variance $\rho^2(\Sigma) = 1$, Corollary 1 implies that a sample covariance matrix $\widehat{\Sigma} = X^T X/n$ will satisfy the RE property with high probability with sample size $n = \Omega(s \log d)$.

On the other hand, the spiked identity matrix Σ has very poorly conditioned sub-matrices, which implies that a sample covariance matrix $\widehat{\Sigma} = X^T X/n$ will violate the restricted isometry property (RIP) with high probability as n grows. To see this fact, for an arbitrary subset S of size s, consider the associated $s \times s$ submatrix Σ_{SS}. An easy calculation shows that $\lambda_{\min}(\Sigma_{SS}) = 1 - a > 0$ and $\lambda_{\max}(\Sigma_{SS}) = 1 + a\,(s - 1)$, so that the population condition number of this sub-matrix is

$$\frac{\lambda_{\max}(\Sigma_{SS})}{\lambda_{\min}(\Sigma_{SS})} = \frac{1 + a\,(s-1)}{1-a}.$$

For any fixed $a \in (0, 1)$, this condition number diverges as s increases. We now show that the same statement applies to the sample covariance with high probability, showing that the RIP is violated. Let $u \in \mathbb{R}^s$ and $v \in \mathbb{R}^s$ denote (respectively) unit-norm eigenvectors corresponding to the minimum and maximum eigenvalues of Σ_{SS}, and define the random variables $Z_u = \|Xu\|_2^2/n$ and $Z_v = \|Xv\|_2^2/n$. Since $\langle X_i,\, v\rangle \sim N(0, \lambda_{\max}(\Sigma_{SS}))$ by construction, we have

$$Z_v = \frac{1}{n}\sum_{i=1}^{n} \langle X_i,\, v\rangle^2 \;\overset{d}{=}\; \lambda_{\max}(\Sigma_{SS})\,\{\frac{1}{n}\sum_{i=1}^{n} y_i^2\},$$

where $y_i \sim N(0, 1)$ are i.i.d. standard Gaussians, and $\overset{d}{=}$ denotes equality in distribution. By χ^2 tail bounds, we have $\mathbb{P}[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \geq \frac{1}{2}] \leq c_1 \exp(-c_2 n)$, so that $Z_v \geq \lambda_{\max}(\Sigma_{SS})/2$ with high probability. A similar argument shows that $Z_u \leq 2\lambda_{\min}(\Sigma_{SS})$ with high probability, and putting together the pieces shows that w.h.p.

$$\frac{\lambda_{\max}(\widehat{\Sigma}_{SS})}{\lambda_{\min}(\widehat{\Sigma}_{SS})} \geq \frac{1}{4}\frac{\lambda_{\max}(\Sigma_{SS})}{\lambda_{\min}(\Sigma_{SS})} \;\geq\; \frac{1}{4}\,\frac{1 + a(s-1)}{1-a},$$

which diverges as s increases.

In both of the preceding examples, the minimum eigenvalue of Σ was bounded from below and the diagonal entries of Σ were bounded from above, which allowed us to assert immediately that the RE condition was satisfied for the population covariance matrix. As a final example, we now consider sampling from population covariance matrices that are actually rank degenerate, but for which our theory still provides guarantees.

Example 3 (Highly degenerate covariance matrices)**.** Let Σ be any matrix with bounded diagonal that satisfies the RE property of some order s. Suppose that we sample n times from a $N(0, \Sigma)$ distribution, and then form the empirical covariance matrix $\widehat{\Sigma} = X^T X/n$. If $n < d$, then $\widehat{\Sigma}$ must be rank degenerate, but Corollary 1 guarantees that $\widehat{\Sigma}$ will satisfy the RE property of order s with high probability as long as $n = \Omega(s \log d)$. Moreover, by χ^2-tail bounds, the maximal diagonal element $\rho^2(\widehat{\Sigma})$ will be bounded with high probability under this same scaling.

Now if we condition on the original design matrix X in the high probability set, we may view $\widehat{\Sigma}$ as a fixed but highly rank-degenerate matrix. Suppose that we draw a new set of n i.i.d. vectors $\widetilde{X}_i \sim N(0, \widehat{\Sigma})$ using this degenerate covariance matrix. Such a resampling procedure could be relevant for a bootstrap-type calculation for assessing errors of the Lasso. We may then form a second empirical covariance matrix $\widetilde{\Sigma} = \frac{1}{n}\widetilde{X}^T\widetilde{X}$. Conditionally on $\widehat{\Sigma}$ having the RE property of order s and a bounded diagonal, Corollary 1 shows that the resampled empirical covariance $\widetilde{\Sigma}$ will also have the RE property of order s with high probability, again for $n = \Omega(s \log d)$.

This simple example shows that in the high-dimensional setting $d \gg n$, it is possible for the RE condition to hold with high probability even when the original population covariance matrix ($\widehat{\Sigma}$ in this example) has a $d-n$-dimensional nullspace. Note moreover that this is not an isolated phenomenon—-rather, it will hold for almost every sample covariance matrix $\widehat{\Sigma}$ constructed in the way that we have described.

3.4 Proof of Theorem 5

We now turn to the proof of Theorem 5. The main ingredients are the Gordon-Slepian comparison inequalities [34] for Gaussian processes, concentration of measure for Lipschitz functions [52], and a peeling argument. The first two ingredients underlie classical proofs on the ordinary eigenvalues of Gaussian random matrices [27], whereas the latter tool is used in empirical process theory [83].

3.4.1 Proof outline

Recall that Theorem 5 states that the condition

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2}v\|_2 - 9\,\rho(\Sigma)\,\sqrt{\frac{\log d}{n}}\,\|v\|_1 \qquad \text{for all } v \in \mathbb{R}^d, \tag{3.12}$$

holds with probability at least $1 - c'\exp(-cn)$, where c, c' are universal positive constants. Hence, we are bounding the random quantity $\|Xv\|_2$ in terms of $\|\Sigma^{1/2}v\|_2$ and $\|v\|_1$ for all v with high probability. It suffices to prove Theorem 5 for $\|\Sigma^{1/2}v\|_2 = 1$. Indeed, for any vector $v \in \mathbb{R}^d$ such that $\Sigma^{1/2}v = 0$, the claim holds holds trivially. Otherwise, we may consider the rescaled vector $\breve{v} = v/\|\Sigma^{1/2}v\|_2$, and note that $\|\Sigma^{1/2}\breve{v}\|_2 = 1$ by construction. By scale invariance of the condition (3.12), if it holds for the rescaled vector $\breve{v}$, it also holds for v.

Therefore, in the remainder of the proof, our goal is to lower bound the quantity $\|Xv\|_2$ over the set of v such that $\|\Sigma^{1/2}v\|_2 = 1$ in terms of $\|v\|_1$. At a high level, there are three main steps to the proof:

(1) We begin by considering the set $V(r) := \{v \in \mathbb{R}^d \mid \|\Sigma^{1/2}v\|_2 = 1, \|v\|_1 \leq r\}$, for a fixed radius r. Although this set may be empty for certain choices of $r > 0$, our analysis only concerns those choices for which it is non-empty. Define the random variable

$$M(r, X) := 1 - \inf_{v \in V(r)} \frac{\|Xv\|_2}{\sqrt{n}} \;=\; \sup_{v \in V(r)} \left\{1 - \frac{\|Xv\|_2}{\sqrt{n}}\right\}. \tag{3.13}$$

Our first step is to upper bound the expectation $\mathbb{E}[M(r,X)]$, where the expectation is taken over the random Gaussian matrix X.

(2) Second, we establish that $M(r,X)$ is a Lipschitz function of its Gaussian arguments, and then use concentration inequalities to assert that for each fixed $r > 0$, the random variable $M(r,X)$ is sharply concentrated around its expectation with high probability.

(3) Third, we use a peeling argument to show that our analysis holds with high probability and uniformly over all possible choice of the ℓ_1-radius r, which then implies that the condition (3.12) holds with high probability as claimed.

In the remainder of this section, we provide the details of each of these steps.

3.4.2 Bounding the expectation $\mathbb{E}[M(r,X)]$

This subsection is devoted to a proof of the following lemma:

Lemma 9. *For any radius $r > 0$ such that $V(r)$ is non-empty, we have*

$$\mathbb{E}[M(r,X)] \leq \frac{1}{4} + 3\rho(\Sigma)\sqrt{\frac{\log d}{n}}\, r. \tag{3.14}$$

Proof. : Let $S^{n-1} = \{u \in \mathbb{R}^n \mid \|u\|_2 = 1\}$ be the Euclidean sphere of radius 1, and recall the previously defined set $V(r) := \{v \in \mathbb{R}^d \mid \|\Sigma^{1/2} v\|_2 = 1, \|v\|_1 \leq r\}$. For each pair $(u,v) \in S^{n-1} \times V(r)$, we may define an associated zero-mean Gaussian random variable $Y_{u,v} := u^T X\, v$. This representation is useful, because it allows us to write the quantity of interest as a min-max problem in terms of this Gaussian process. In particular, we have

$$-\inf_{v \in V(r)} \|Xv\|_2 = -\inf_{v \in V(r)} \sup_{u \in S^{n-1}} u^T X v \;=\; \sup_{v \in V(r)} \inf_{u \in S^{n-1}} u^T X v. \tag{3.15}$$

We may now upper bound the expected value of the above quantity via a Gaussian comparison inequality; here we state a form of Gordon's inequality used in past work on Gaussian random matrices [27]. Suppose that $\{Y_{u,v}, (u,v) \in U \times V\}$ and $\{Z_{u,v}, (u,v) \in U \times V\}$ are two zero-mean Gaussian processes on $U \times V$. Using $\sigma(\cdot)$ to denote the standard deviation of its argument, suppose that these two processes satisfy the inequality

$$\sigma(Y_{u,v} - Y_{u',v'}) \leq \sigma(Z_{u,v} - Z_{u',v'}) \qquad \text{for all pairs } (u,v) \text{ and } (u',v') \text{ in } U \times V, \tag{3.16}$$

and this inequality holds with equality when $v = v'$. Then we are guaranteed that

$$\mathbb{E}[\sup_{v \in V} \inf_{u \in U} Y_{u,v}] \;\leq\; \mathbb{E}[\sup_{v \in V} \inf_{u \in U} Z_{u,v}]. \tag{3.17}$$

We use Gordon's inequality to show that

$$\mathbb{E}[M(r,X)] = 1 + \mathbb{E}[\sup_{v \in V(r)} \inf_{u \in S^{n-1}} Y_{u,v}] \leq 1 + \mathbb{E}[\sup_{v \in V(r)} \inf_{u \in S^{n-1}} Z_{u,v}],$$

where we recall that $Y_{u,v} = u^T X v$ and $Z_{u,v}$ is a different Gaussian process to be defined shortly.

We begin by computing $\sigma^2(Y_{u,v} - Y_{u',v'})$. To simplify notation, we note that the $X \in \mathbb{R}^{n \times d}$ can be written as $W\Sigma^{1/2}$, where $W \in \mathbb{R}^{n \times d}$ is a matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and $\Sigma^{1/2}$ is the symmetric matrix square root. In terms of W, we can write

$$Y_{u,v} = u^T W \Sigma^{1/2} v \;=\; u^T W \tilde{v},$$

where $\tilde{v} = \Sigma^{1/2} v$. It follows that

$$\sigma^2(Y_{u,v} - Y_{u',v'}) := \mathbb{E}\big(\sum_{i=1}^{n} \sum_{j=1}^{d} W_{i,j}(u_i \tilde{v}_j - u'_i \tilde{v}'_j)\big)^2 \;=\; |\!|\!| u\tilde{v}^T - (u')(\tilde{v}')^T |\!|\!|_F^2,$$

where $|\!|\!| \cdot |\!|\!|_F$ is the Frobenius norm (ℓ_2-norm applied elementwise to the matrix). This equality follows immediately since the W_{ij} variables are i.i.d $\mathcal{N}(0,1)$.

Now consider a second zero-mean Gaussian process $Z_{u,v}$ indexed by $S^{n-1} \times V(r)$, and given by

$$Z_{u,v} \;=\; \vec{g}^T u + \vec{h}^T \Sigma^{1/2} v, \tag{3.18}$$

where $\vec{g} \sim N(0, I_{n \times n})$ and $\vec{h} \sim N(0, I_{d \times d})$ are standard Gaussian random vectors. With $\tilde{v} = \Sigma^{1/2} v$, we see immediately that

$$\sigma^2(Z_{u,v} - Z_{u',v'}) = \|u - u'\|_2^2 + \|\tilde{v} - \tilde{v}'\|_2^2. \tag{3.19}$$

Consequently, in order to apply the Gaussian comparison principle to $\{Y_{u,v}\}$ and $\{Z_{u,v}\}$, we need to show that

$$|\!|\!| u\tilde{v}^T - (u')(\tilde{v}')^T |\!|\!|_F^2 \leq \|u - u'\|_2^2 + \|\tilde{v} - \tilde{v}'\|_2^2 \tag{3.20}$$

for all pairs $(u, \tilde{v})$ and $(u', \tilde{v}')$ in the set of interest. Since the Frobenius norm $|\!|\!| \cdot |\!|\!|_F$ is simply the ℓ_2-norm on the vectorized form of a matrix, we can compute

$$\begin{aligned}
|\!|\!| u\, \tilde{v}^T - u'(\tilde{v}')^T |\!|\!|_F^2 &= |\!|\!| (u-u')\tilde{v}^T + u'(\tilde{v}-\tilde{v}')^T |\!|\!|_F^2 \\
&= \sum_{i=1}^n \sum_{j=1}^d [(u_i - u_i')\tilde{v}_j + u_i'(\tilde{v}_j - \tilde{v}_j')]^2 \\
&= \|\tilde{v}\|_2^2\, \|u-u'\|_2^2 + \|u'\|_2^2 \|\tilde{v} - \tilde{v}'\|_2^2 + 2(u^T u' - \|u'\|_2^2)(\|\tilde{v}\|_2^2 - \tilde{v}^T \tilde{v}') \\
&= \|u-u'\|_2^2 + \|\tilde{v}-\tilde{v}'\|_2^2 - 2(\|u'\|_2^2 - u^T u')(\|\tilde{v}\|_2^2 - \tilde{v}^T\tilde{v}'),
\end{aligned}$$

where we have used equalities $\|u\|_2 = \|u'\|_2 = 1$ and $\|\tilde{v}\|_2 = \|\tilde{v}'\|_2 = 1$. By the Cauchy-Schwarz inequality, we have $\|u\|_2^2 - u^T u' \geq 0$, and $\|\tilde{v}\|_2^2 - \tilde{v}^T \tilde{v}' \geq 0$, from which the claimed inequality (3.20) follows. When $v = v'$, we also have $\tilde{v} = \Sigma^{1/2} v = \Sigma^{1/2} v' = \tilde{v}'$, so that equality holds in the condition (3.20) when $\tilde{v} = \tilde{v}'$.

Consequently, we may apply Gordon's inequality to conclude that

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

We now observe that by definition of $V(r)$, we have

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

Each element $(\Sigma^{1/2}\vec{h})_j$ is zero-mean Gaussian with variance Σ_{jj}. Consequently, known results on Gaussian maxima (c.f. [53], equation (3.13)) imply that $\mathbb{E}[\|\Sigma^{1/2} h\|_\infty] \leq 3\sqrt{\rho^2(\Sigma)\,\log d}$, where $\rho^2(\Sigma) = \max_j \Sigma_{jj}$. Noting[2] that $\mathbb{E}[\|\vec{g}\|_2] \geq \frac{3}{4}\sqrt{n}$ for all $n \geq 10$ by standard χ^2 tail bounds and putting together the pieces, we obtain the bound

$$\mathbb{E}[-\inf_{v \in V(r)} \|Xv\|_2] \;\leq\; -\frac{3}{4}\sqrt{n} + 3\big[\rho^2(\Sigma)\,\log d\big]^{1/2}\, r.$$

Dividing by $\sqrt{n}$ and adding 1 to both sides yields

$$\mathbb{E}[M(r, X)] = \mathbb{E}[1 - \inf_{v \in V(r)} \|Xv\|_2/\sqrt{n}] \leq 1/4 + 3\,\rho(\Sigma)\,\sqrt{\frac{\log d}{n}}\; r,$$

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

as claimed.

□

3.4.3 Concentration around the mean for $M(r, X)$

Having controlled the expectation, the next step is to establish concentration of $M(r, X)$ around its mean. Note that Lemma 9 shows that $\mathbb{E}[M(r, X)] \leq t_\ell(r)$, where

$$t_\ell(r) := \frac{1}{4} + 3\,r\,\rho(\Sigma)\,\sqrt{\frac{\log d}{n}}. \tag{3.21}$$

Now we prove the following claim:

Lemma 10. *For any r such that $V(r)$ is non-empty, we have*

$$\mathbb{P}\Big[M(r, X) \geq \frac{3t_\ell(r)}{2}\Big] \leq 2\exp(-nt_\ell^2(r)/8).$$

Proof. In order to prove this lemma, it suffices to show that

$$\mathbb{P}\big[|M(r, X) - \mathbb{E}[M(r, X)]| \geq t_\ell(r)/2\big] \leq 2\exp(-nt_\ell^2(r)/8),$$

and use the upper bound on $\mathbb{E}[M(r, X)]$ derived in Lemma 9.

By concentration of measure for Lipschitz functions of Gaussians (see Appendix B of our journal paper [67]), this tail bound will follow if we show that the Lipschitz constant of $M(r, X)$ as a function of the Gaussian random matrix is less than $1/\sqrt{n}$. To make this functional dependence explicit, let us write $M(r, X)$ as the function $h(W) = \sup_{v \in V(r)} \big(1 - \|W\Sigma^{1/2}v\|_2/\sqrt{n}\big)$. We find that

$$\sqrt{n}\big[h(W) - h(W')\big] = \sup_{v \in V(r)} -\|W\Sigma^{1/2}v\|_2 - \sup_{v \in V(r)} -\|W'\Sigma^{1/2}v\|_2.$$

Since $V(r)$ is closed and bounded and the objective function is continuous, there exists $\hat{v} \in V(r)$ such that $\hat{v} = \arg\max_{v \in V(r)} -\|W\Sigma^{1/2}v\|_2$. Therefore

$$\begin{aligned}
\sup_{v \in V(r)} \big(-\|W\Sigma^{1/2}v\|_2\big) - \sup_{v \in V(r)} \big(-\|W'\Sigma^{1/2}v\|_2\big) &= -\|W\Sigma^{1/2}\hat{v}\|_2 - \sup_{v \in V(r)} \big(-\|W'\Sigma^{1/2}v\|_2\big) \\
&\leq \|W'\Sigma^{1/2}\hat{v}\|_2 - \|W\Sigma^{1/2}\hat{v}\|_2 \\
&\leq \sup_{v \in V(r)} \big(\|(W' - W)\Sigma^{1/2}v\|_2\big).
\end{aligned}$$

For a matrix A, we define its spectral norm $|||A|||_2 = \sup_{\|u\|_2=1} \|Au\|_2$. With this notation, we can bound the Lipschitz constant of h as

$$\begin{aligned} \sqrt{n}\big[h(W) - h(W')\big] &\leq \sup_{v \in V(r)} \big(\|(W-W')\Sigma^{1/2}v\|_2\big) \\ &\overset{(a)}{\leq} \Big\{\sup_{v \in V(r)} \big(\|\Sigma^{1/2}v\|_2\big)\Big\} |||(W-W')|||_2 \\ &\overset{(b)}{\leq} \Big\{\sup_{v \in V(r)} \big(\|\Sigma^{1/2}v\|_2\big)\Big\} \; |||(W-W')|||_F \\ &\overset{(c)}{=} |||W-W'|||_F. \end{aligned}$$

In this argument, inequality (a) follows by definition of the matrix spectral norm $|||\cdot|||_2$; inequality (b) follows from the bound $|||(W-W')|||_2 \leq |||(W-W')|||_F$ between the spectral and Frobenius matrix norms [45]; and equality (c) follows since $\|\Sigma^{1/2}v\|_2 = 1$ for all $v \in V(r)$. Thus, we have shown that h has Lipschitz constant $L \leq 1/\sqrt{n}$ with respect to the Euclidean norm on W (viewed as a vector with nd entries). Finally we use a standard result on the concentration for Lipschitz functions of Gaussian random variables [52, 57]—see Appendix B of Raskutti et al. [67] for one statement. Applying the concentration result Eq. (9) in from Appendix B of Raskutti et al. [67] with $m = np$, $\tilde{g} = W$, and $t = t(r)/2$ completes the proof. □

3.4.4 Extension to all vectors via peeling

Thus far, we have shown that

$$M(r, X) = 1 - \inf_{v \in V(r)} \frac{\|Xv\|_2}{\sqrt{n}} \; = \; \sup_{v \in V(r)} \Big\{1 - \frac{\|Xv\|_2}{\sqrt{n}}\Big\} \geq 3t_\ell(r)/2, \tag{3.22}$$

with probability no larger than $2\exp(-nt_\ell^2(r)/8)$ where $t_\ell(r) = \frac{1}{4} + 3\, r\, \rho(\Sigma)\, \sqrt{\frac{\log d}{n}}$. The set $V(r)$ requires that $\|v\|_1 \leq r$ for some *fixed* radius r, whereas the claim of Theorem 5 applies to all vectors v. Consequently, we need to extend the bound (3.22) to an arbitrary ℓ_1 radius.

In order do so, we define the event

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

Note that there is no r in the definition of $\mathcal{T}$, because we are setting $\|v\|_1$ to be the argument of the function t_ℓ. We claim that there are constants positive constants c, c' such that $\mathbb{P}[\mathcal{T}] \leq c\exp(-c'n)$, from which Theorem 5 will follow. We establish this claim by using a device known as peeling [4, 83]; for the version used here, see Lemma 3 proved in the

Appendix in Raskutti et al. [67]. In particular, we apply Lemma 3 with the functions

$$f(v, X) = 1 - \|Xv\|_2/\sqrt{n}, \qquad h(v) = \|v\|_1, \quad \text{and} \quad g(r) = 3t_\ell(r)/2,$$

the sequence $a_n = n$, and the set $A = \{v \in \mathbb{R}^d \mid \|\Sigma^{1/2}v\|_2 = 1\}$. Recall that the quantity t_ℓ, as previously defined (3.21), satisfies $t_\ell(r) \geq 1/4$ for all $r > 0$ and is strictly increasing. Therefore, the function $g(r) = 3t_\ell(r)/2$ is non-negative and strictly increasing as a function of r, and moreover satisfies $g(r) \geq 3/8$, so that Lemma 3 is applicable with $\mu = 3/8$. We can thus conclude that $\mathbb{P}[\mathcal{T}^c] \geq 1 - c\exp(-c'n)$ for some numerical constants c and c'.

Finally, conditioned on the event $\mathcal{T}^c$, for all $v \in \mathbb{R}^d$ with $\|\Sigma^{1/2}v\|_2 = 1$, we have

$$1 - \|Xv\|_2/\sqrt{n} \;\; \leq \;\; 3t_\ell(\|v\|_1) \;\; = \;\; \frac{3}{4} + 9\,\|v\|_1\,\rho(\Sigma)\sqrt{\frac{\log d}{n}},$$

which implies that

$$\|Xv\|_2/\sqrt{n} \geq \frac{1}{4} - 9\,\|v\|_1\,\rho(\Sigma)\,\sqrt{\frac{\log d}{n}}.$$

As noted in the proof outline, this suffices to establish the general claim.

3.5 Conclusion

Methods based on ℓ_1-relaxations are very popular, and the weakest possible conditions on the design matrix X required to provide performance guarantees—namely, the restricted nullspace and eigenvalue conditions—are well-understood. In this chapter, we have proved that these conditions hold with high probability for a broad class of Gaussian design matrices allowing for quite general dependency among the columns, as captured by a covariance matrix Σ representing the dependence among the different covariates. As a corollary, our result guarantees that known performance guarantees for ℓ_1-relaxations such as basis pursuit and Lasso hold with high probability for such problems, provided the population matrix Σ satisfies the RE condition. Interestingly, our theory shows that ℓ_1-methods can perform well when the covariates are sampled from a Gaussian distribution with a degenerate covariance matrix. Some follow-up work [98] has extended these results to random matrices with sub-Gaussian rows. In addition, there are a number of other ways in which this work could be extended. One is to incorporate additional dependence across the rows of the design matrix, as would arise in modeling time series data for example. It would also be interesting to relate the allowable degeneracy structures of Σ to applications involving real data. Finally, although this chapter provides various conditions under which the RE condition holds with high probabability, it does not address the issue of how to determine whether a given sample covariance matrix matrix $\widehat{\Sigma} = X^T X/n$ satisfies the RE condition. It would be interesting to study if there are computationally efficient methods for verifying the RE condition.

Chapter 4

Minimax-Optimal Rates For Sparse Additive Models Over Kernel Classes Via Convex Programming

4.1 Introduction

While a large body of work has focused on sparse linear models, many applications call for the additional flexibility provided by non-parametric models. In the general setting, a non-parametric regression model takes the form $y = f^*(x_1, \ldots, x_d) + w$, where $f^* : \mathbb{R}^d \to \mathbb{R}$ is the unknown regression function, and w is scalar observation noise. Unfortunately, this general non-parametric model is known to suffer severely from the so-called "curse of dimensionality", in that for most natural function classes (e.g., twice differentiable functions), the sample size n required to achieve any given error grows exponentially in the dimension d. Given this curse of dimensionality, it is essential to further constrain the complexity of possible functions f^*. One attractive candidate is the class of *additive non-parametric models* [43], in which the function f^* has an additive decomposition of the form

$$f^*(x_1, x_2, \ldots, x_d) = \sum_{j=1}^{d} f_j^*(x_j), \tag{4.1}$$

where each component function f_j^* is univariate. Given this additive form, this function class no longer suffers from the exponential explosion in sample size of the general non-parametric model. Nonetheless, one still requires a sample size $n \gg d$ for consistent estimation; note that this is true even for the linear model, which is a special case of Equation (4.1).

A natural extension of sparse linear models is the class of *sparse additive models*, in which the unknown regression function is assumed to have a decomposition of the form

$$f^*(x_1, x_2 \ldots, x_d) = \sum_{j \in S} f_j^*(x_j), \tag{4.2}$$

where $S \subseteq \{1, 2, \ldots, d\}$ is some unknown subset of cardinality $|S| = s$. Of primary interest is the case when the decomposition is genuinely sparse, so that $s \ll d$. To the best of our knowledge, this model class was first introduced by [54], and has since been studied by various researchers [50, 58, 71, 94]. Note that the sparse additive model (4.2) is a natural generalization of the sparse linear model, to which it reduces when each univariate function is constrained to be linear.

In past work, several groups have proposed computationally efficient methods for estimating sparse additive models (4.2). Just as ℓ_1-based relaxations such as the Lasso have desirable properties for sparse parametric models, more general ℓ_1-based approaches have proven to be successful in this setting. [54] proposed the COSSO method, which extends the Lasso to cases where the component functions f_j^* lie in a reproducing kernel Hilbert space (RKHS); see also [94] for a similar extension of the non-negative garrote [16]. [7] analyzes a closely related method for the RKHS setting, in which least-squares loss is penalized by an ℓ_1-sum of Hilbert norms, and establishes consistency results in the classical (fixed d) setting. Other related ℓ_1-based methods have been proposed in independent work by [49], [71] and [58], and analyzed under high-dimensional scaling ($d \gg n$). As we describe in more detail in Section 4.3.4, each of the above chapters establish consistency and convergence rates for the prediction error under certain conditions on the covariates as well as the sparsity s and dimension d. However, it is not clear whether the rates obtained in these chapters are sharp for the given methods, nor whether the rates are minimax-optimal. Past work by [50] establishes rates for sparse additive models with an additional global boundedness condition, but as will be discussed at more length in the sequel, these rates are not minimax optimal in general.

This chapter makes three main contributions to this line of research. Our first contribution is to analyze a simple polynomial-time method for estimating sparse additive models and provide upper bounds on the error in the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norms. The estimator[1] that we analyze is based on a combination of least-squares loss with two ℓ_1-based sparsity penalty terms, one corresponding to an $\ell_1/L^2(\mathbb{P}_n)$ norm and the other an $\ell_1/\|\cdot\|_{\mathcal{H}}$ norm. Our first main result (Theorem 6) shows that with high probability, if we assume the univariate functions are bounded and independent, the error of our procedure in the squared $L^2(\mathbb{P}_n)$ and $L^2(\mathbb{P})$ norms is bounded by $\mathcal{O}\big(\frac{s \log d}{n} + s\nu_n^2\big)$, where the quantity ν_n^2 corresponds to the optimal rate for estimating a single univariate function. Importantly, our analysis does *not* require a global boundedness condition on the class $\mathcal{F}_{d,s,\mathcal{H}}$ of all s-sparse models, an assumption that

[1] The same estimator was proposed concurrently by [50]; we discuss connections to this work in the sequel.

is often imposed in classical non-parametric analysis. Indeed, as we discuss below, when such a condition is imposed, then significantly faster rates of estimation are possible. The proof of Theorem 6 involves a combination of techniques for analyzing M-estimators with decomposable regularizers [64], combined with various techniques in empirical process theory for analyzing kernel classes [9, 61, 83]. Our second contribution is complementary in nature, in that it establishes algorithm-independent minimax lower bounds on $L^2(\mathbb{P})$ error. These minimax lower bounds, stated in Theorem 7, are specified in terms of the metric entropy of the underlying univariate function classes. For both finite-rank kernel classes and Sobolev-type classes, these lower bounds match our achievable results, as stated in Corollaries 2 and 3, up to constant factors in the regime of sub-linear sparsity ($s = o(d)$). Thus, for these function classes, we have a sharp characterization of the associated minimax rates. The lower bounds derived in this chapter initially appeared in the Proceedings of the NIPS Conference (December 2009). The proofs of Theorem 2 is based on characterizing the packing entropies of the class of sparse additive models, combined with classical information theoretic techniques involving Fano's inequality and variants [42, 92, 93].

Our third contribution is to determine upper bounds on minimax $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ error when we impose a global boundedness assumption on the class $\mathcal{F}_{d,s,\mathcal{H}}$. More precisely, a global boundedness condition means that the quantity $B(\mathcal{F}_{d,s,\mathcal{H}}) = \sup_{f \in \mathcal{F}_{d,s,\mathcal{H}}} \sup_x |\sum_{j=1}^d f_j(x_j)|$ is assumed to be bounded independently of (s, d). As mentioned earlier, our upper bound in Theorem 6 does *not* impose a global boundedness condition, whereas in contrast, the analysis of [50], or KY for short, does impose such a global boundedness condition. Under global boundedness, their work provides rates on the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norm that are of the same order as the results presented here. It is natural to wonder whether or not this difference is actually significant—that is, do the minimax rates for the class of sparse additive models depend on whether or not global boundedness is imposed? In Section 4.3.5, we answer this question in the affirmative. In particular, Theorem 8 and Corollary 4 provide upper bounds on the minimax rates, as measured in either the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$-norms, under a global boundedness condition. These rates are faster than those of Theorem 3 in the KY chapter, in particular showing that the KY rates are not optimal for problems with $s = \Omega(\sqrt{n})$. In this way, we see that the assumption of global boundedness, though relatively innocuous for classical (low-dimensional) non-parametric problems, can be quite limiting in high dimensions.

The remainder of the chapter is organized as follows. In Section 4.2, we provide background on kernel spaces and the class of sparse additive models considered in this chapter. Section 4.3 is devoted to the statement of our main results and discussion of their consequences; it includes description of our method, the upper bounds on the convergence rate that it achieves, and a matching set of minimax lower bounds. Section 4.3.5 investigates the restrictiveness of the global uniform boundedness assumption and in particular, Theorem 8 and Corollary 4 demonstrate that there are classes of globally bounded functions for which faster rates are possible. Section 4.4 is devoted to the proofs of our three main theorems, with the more technical details deferred to the Appendices in our journal version [69]. We conclude with a discussion in Section 4.5.

4.2 Background and Problem Set-up

We begin with some background on reproducing kernel Hilbert spaces, before providing a precise definition of the class of sparse additive models studied in this chapter.

4.2.1 Reproducing Kernel Hilbert Spaces

Given a subset $\mathcal{X} \subset \mathbb{R}$ and a probability measure $\mathbb{Q}$ on $\mathcal{X}$, we consider a Hilbert space $\mathcal{H} \subset L^2(\mathbb{Q})$, meaning a family of functions $g : \mathcal{X} \to \mathbb{R}$, with $\|g\|_{L^2(\mathbb{Q})} < \infty$, and an associated inner product $\langle \cdot, \, \cdot \rangle_{\mathcal{H}}$ under which $\mathcal{H}$ is complete. The space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if there exists a symmetric function $\ker : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that for each $x \in \mathcal{X}$: (a) the function $\ker(\cdot, x)$ belongs to the Hilbert space $\mathcal{H}$, and (b) we have the reproducing relation $f(x) = \langle f, \, \ker(\cdot, x)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Any such kernel function must be positive semidefinite; under suitable regularity conditions, Mercer's theorem ([63]) guarantees that the kernel has an eigen-expansion of the form

$$\ker(x, x') = \sum_{k=1}^{\infty} \mu_k \phi_k(x) \phi_\ell(x'), \tag{4.3}$$

where $\mu_1 \geq \mu_2 \geq \mu_3 \geq \ldots \geq 0$ are a non-negative sequence of eigenvalues, and $\{\phi_k\}_{k=1}^{\infty}$ are the associated eigenfunctions, taken to be orthonormal in $L^2(\mathbb{Q})$. The decay rate of these eigenvalues will play a crucial role in our analysis, since they ultimately determine the rate ν_n for the univariate RKHS's in our function classes.

Since the eigenfunctions $\{\phi_k\}_{k=1}^{\infty}$ form an orthonormal basis, any function $f \in \mathcal{H}$ has an expansion of the $f(x) = \sum_{k=1}^{\infty} a_k \phi_k(x)$, where $a_k = \langle f, \, \phi_k \rangle_{L^2(\mathbb{Q})} = \int_{\mathcal{X}} f(x) \phi_k(x) \, d\,\mathbb{Q}(x)$ are (generalized) Fourier coefficients. Associated with any two functions in $\mathcal{H}$—say $f = \sum_{k=1}^{\infty} a_k \phi_k$ and $g = \sum_{k=1}^{\infty} b_k \phi_k$—are two distinct inner products. The first is the usual inner product in $L^2(\mathbb{Q})$, $\langle f, \, g \rangle_{L^2(\mathbb{Q})} := \int_{\mathcal{X}} f(x) g(x) \, d\,\mathbb{Q}(x)$. By Parseval's theorem, it has an equivalent representation in terms of the expansion coefficients—namely

$$\langle f, \, g \rangle_{L^2(\mathbb{Q})} = \sum_{k=1}^{\infty} a_k b_k.$$

The second inner product, denoted $\langle f, g\rangle_{\mathcal{H}}$, is the one that defines the Hilbert space; it can be written in terms of the kernel eigenvalues and generalized Fourier coefficients as

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} \frac{a_k b_k}{\mu_k}.$$

Using this definition, the Hilbert ball of unit radius for a kernel with eigenvalues $\{\mu_k\}_{k=1}^{\infty}$ and eigenfunctions $\{\phi_k\}_{k=1}^{\infty}$ is given by

$$\mathbb{B}_{\mathcal{H}}(1) := \big\{ f = \sum_{k=1}^{\infty} a_k \phi_k \mid \sum_{k=1}^{\infty} \frac{a_k^2}{\mu_k} \leq 1 \big\}.$$

For more background on reproducing kernel Hilbert spaces, we refer the reader to various standard references [6, 76, 77, 88].

4.2.2 Sparse Additive Models Over RKHS

For each $j = 1, \ldots, d$, let $\mathcal{H}_j \subset L^2(\mathbb{Q})$ be a reproducing kernel Hilbert space of univariate functions on the domain $\mathcal{X} \subset \mathbb{R}$. We assume that

$$\mathbb{E}[f_j(x)] = \int_{\mathcal{X}} f_j(x) d\,\mathbb{Q}(x) \;=\; 0 \qquad \text{for all } f_j \in \mathcal{H}_j, \text{ and for each } j = 1, 2, \ldots, d.$$

As will be clarified momentarily, our observation model (4.5) allows for the possibility of a non-zero mean $\bar{f}$, so that there is no loss of generality in this assumption. For a given subset $S \subset \{1, 2, \ldots, d\}$, we define

$$\mathcal{H}(S) := \big\{ f = \sum_{j \in S} f_j \mid f_j \in \mathcal{H}_j, \text{ and } f_j \in \mathbb{B}_{\mathcal{H}_j}(1) \;\forall\, j \in S \big\},$$

corresponding to the class of functions $f : \mathcal{X}^d \to \mathbb{R}$ that decompose as sums of univariate functions on co-ordinates lying within the set S. Note that $\mathcal{H}(S)$ is also (a subset of) a reproducing kernel Hilbert space, in particular with the norm

$$\|f\|_{\mathcal{H}(S)}^2 = \sum_{j \in S} \|f_j\|_{\mathcal{H}_j}^2,$$

where $\|\cdot\|_{\mathcal{H}_j}$ denotes the norm on the univariate Hilbert space $\mathcal{H}_j$. Finally, for $s \in \{1, 2, \ldots, \lfloor d/2 \rfloor\}$, we define the function class

$$\mathcal{F}_{d,s,\mathcal{H}} := \bigcup_{\substack{S \subset \{1,2,\ldots,d\} \\ |S| = s}} \mathcal{H}(S). \tag{4.4}$$

To ease notation, we frequently adopt the shorthand $\mathcal{F} = \mathcal{F}_{d,s,\mathcal{H}}$, but the reader should recall that $\mathcal{F}$ depends on the choice of Hilbert spaces $\{\mathcal{H}_j\}_{j=1}^d$, and moreover, that we are actually studying a *sequence of function classes* indexed by (d, s).

Now let $\mathbb{P} = \mathbb{Q}^d$ denote the product measure on the space $\mathcal{X}^d \subseteq \mathbb{R}^d$. Given an arbitrary $f^* \in \mathcal{F}$, we consider the observation model

$$y_i = \bar{f} + f^*(x_i) + w_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{4.5}$$

where $\{w_i\}_{i=1}^n$ is an i.i.d. sequence of standard normal variates, and $\{x_i\}_{i=1}^n$ is a sequence of design points in $\mathbb{R}^d$, sampled in an i.i.d. manner from $\mathbb{P}$.

Given an estimate $\widehat{f}$, our goal is to bound the error $\widehat{f} - f^*$ under two norms. The first is the *usual* $L^2(\mathbb{P})$ *norm* on the space $\mathcal{F}$; given the product structure of $\mathbb{P}$ and the additive nature of any $f \in \mathcal{F}$, it has the additive decomposition $\|f\|^2_{L^2(\mathbb{P})} = \sum_{j=1}^d \|f_j\|^2_{L^2(\mathbb{Q})}$. In addition, we consider the error in the *empirical* $L^2(\mathbb{P}_n)$*-norm* defined by the sample $\{x_i\}_{i=1}^n$, defined as

$$\|f\|^2_{L^2(\mathbb{P}_n)} := \frac{1}{n}\sum_{i=1}^n f^2(x_i).$$

Unlike the $L^2(\mathbb{P})$ norm, this norm does not decouple across the dimensions, but part of our analysis will establish an approximate form of such decoupling. For shorthand, we frequently use the notation $\|f\|_2 = \|f\|_{L^2(\mathbb{P})}$ and $\|f\|_n = \|f\|_{L^2(\mathbb{P}_n)}$ for a d-variate function $f \in \mathcal{F}$. With a minor abuse of notation, for a univariate function $f_j \in \mathcal{H}_j$, we also use the shorthands $\|f_j\|_2 = \|f_j\|_{L^2(\mathbb{Q})}$ and $\|f_j\|_n = \|f_j\|_{L^2(\mathbb{Q}_n)}$.

4.3 Main Results and Their Consequences

This section is devoted to the statement of our three main results, and discussion of some of their consequences. We begin in Section 4.3.1 by describing a regularized M-estimator for sparse additive models, and we state our upper bounds for this estimator in Section 4.3.2. We illustrate our upper bounds for various concrete instances of kernel classes. In Section 4.3.3, we state minimax lower bounds on the $L^2(\mathbb{P})$ error over the class $\mathcal{F}_{d,s,\mathcal{H}}$, which establish the optimality of our procedure. In Section 4.3.4, we provide a detailed comparison between our results to past work, and in Section 4.3.5 we discuss the effect of global boundedness conditions on optimal rates.

4.3.1 A Regularized M-Estimator For Sparse Additive Models

For any function of the form $f = \sum_{j=1}^{d} f_j$, the $(L^2(\mathbb{Q}_n), 1)$ and $(\mathcal{H}, 1)$-norms are given by

$$\|f\|_{n,1} := \sum_{j=1}^{d} \|f_j\|_n, \quad \text{and} \quad \|f\|_{\mathcal{H},1} := \sum_{j=1}^{d} \|f_j\|_{\mathcal{H}},$$

respectively. Using this notation and defining the sample mean $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$, we define the cost functional

$$\mathcal{L}(f) = \frac{1}{2n}\sum_{i=1}^{n}\big(y_i - \bar{y}_n - f(x_i)\big)^2 + \lambda_n \|f\|_{n,1} + \rho_n \|f\|_{\mathcal{H},1}.$$

The cost functional $\mathcal{L}(f)$ is least-squares loss with a sparsity penalty $\|f\|_{n,1}$ and a smoothness penalty $\|f\|_{\mathcal{H},1}$. Here (λ_n, ρ_n) are a pair of positive regularization parameters whose choice will be specified by our theory. Given this cost functional, we then consider the M-estimator

$$\widehat{f} \in \arg\min_{f} \mathcal{L}(f) \quad \text{subject to } f = \textstyle\sum_{j=1}^{d} f_j \text{ and } \|f_j\|_{\mathcal{H}} \leq 1 \text{ for all } j = 1, 2, \ldots, d. \tag{4.6}$$

In this formulation (4.6), the problem is infinite-dimensional in nature, since it involves optimization over Hilbert spaces. However, an attractive feature of this M-estimator is that, as a consequence of the representer theorem [48], it can be reduced to an equivalent convex program in $\mathbb{R}^n \times \mathbb{R}^d$. In particular, for each $j = 1, 2, \ldots, d$, let $\ker^j$ denote the kernel function for co-ordinate j. Using the notation $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ for the i^{th} sample, we define the collection of empirical kernel matrices $K^j \in \mathbb{R}^{n \times n}$, with entries $K^j_{i\ell} = \ker^j(x_{ij}, x_{\ell j})$. By the representer theorem, any solution $\widehat{f}$ to the variational problem (4.6) can be written in the form

$$\widehat{f}(z_1, \ldots, z_d) = \sum_{i=1}^{n}\sum_{j=1}^{d} \widehat{\alpha}_{ij} \ker^j(z_j, x_{ij}),$$

for a collection of weights $\big\{\widehat{\alpha}_j \in \mathbb{R}^n,\ j = 1, \ldots, d\big\}$. The optimal weights $(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_d)$ are any minimizer to the following convex program:

$$\arg \min_{\substack{\alpha_j \in \mathbb{R}^n \\ \alpha_j^T K^j \alpha_j \leq 1}} \left\{ \frac{1}{2n}\|y - \bar{y}_n - \sum_{j=1}^{d} K^j \alpha_j\|_2^2 + \lambda_n \sum_{j=1}^{d} \sqrt{\frac{1}{n}\|K^j \alpha_j\|_2^2} + \rho_n \sum_{j=1}^{d} \sqrt{\alpha_j^T K^j \alpha_j} \right\}. \tag{4.7}$$

This problem is a second-order cone program (SOCP), and there are various algorithms for finding a solution to arbitrary accuracy in time polynomial in (n, d), among them interior point methods (e.g., see §11 in [15]).

Various combinations of sparsity and smoothness penalties have been used in past work on sparse additive models. For instance, the method of [71] is based on least-squares loss regularized with single sparsity constraint, and separate smoothness constraints for each univariate function. They solve the resulting optimization problem using a back-fitting procedure. [49] develop a method based on least-squares loss combined with a single penalty term $\sum_{j=1}^{d} \|f_j\|_{\mathcal{H}}$. Their method also leads to an SOCP if $\mathcal{H}$ is a reproducing kernel Hilbert space, but differs from the program (4.7) in lacking the additional sparsity penalties. [58] analyzed least-squares regularized with a penalty term of the form $\sum_{j=1}^{d} \sqrt{\lambda_1 \|f_j\|_n^2 + \lambda_2 \|f_j\|_{\mathcal{H}}^2}$, where λ_1 and λ_2 are a pair of regularization parameters. In their method, λ_1 controls the sparsity while λ_2 controls the smoothness. If $\mathcal{H}$ is an RKHS, the method in [58] reduces to an ordinary group Lasso problem on a different set of variables, which can be cast as a quadratic program. The more recent work of [50] is based on essentially the same estimator as problem (4.6), except that we allow for a non-zero mean for the function, and estimate it as well. In addition, the KY analysis involves a stronger condition of global boundedness. We provide a more in-depth comparison of our analysis and results with the past work listed above in Sections 4.3.4 and 4.3.5.

4.3.2 Upper Bound

We now state a result that provides upper bounds on the estimation error achieved by the estimator (4.6), or equivalently (4.7). To simplify presentation, we state our result in the special case that the univariate Hilbert space $\mathcal{H}_j, j = 1, \ldots, d$ are all identical, denoted by $\mathcal{H}$. However, the analysis and results extend in a straightforward manner to the general setting of distinct univariate Hilbert spaces, as we discuss following the statement of Theorem 6.

Let $\mu_1 \geq \mu_2 \geq \ldots \geq 0$ denote the non-negative eigenvalues of the kernel operator defining the univariate Hilbert space $\mathcal{H}$, as defined in Equation (4.3), and define the function

$$\mathcal{Q}_{\sigma,n}(t) := \frac{1}{\sqrt{n}} \big[\sum_{\ell=1}^{\infty} \min\{t^2, \mu_\ell\} \big]^{1/2}.$$

Let $\nu_n > 0$ be the smallest positive solution to the inequality

$$40\nu_n^2 \geq \mathcal{Q}_{\sigma,n}(\nu_n), \tag{4.8}$$

where the 40 is simply used for technical convenience. We refer to ν_n as the *critical univariate rate*, as it is the minimax-optimal rate for $L^2(\mathbb{P})$-estimation of a single univariate function in the Hilbert space $\mathcal{H}$ [61, 83]. This quantity will be referred to throughout the remainder of the chapter.

Our choices of regularization parameters are specified in terms of the quantity

$$\gamma_n := \kappa \max\big\{\nu_n, \sqrt{\frac{\log d}{n}}\big\}, \tag{4.9}$$

where κ is a fixed constant that we choose later. We assume that each function within the unit ball of the univariate Hilbert space is uniformly bounded by a constant multiple of its Hilbert norm—that is, for each $j = 1, \ldots, d$ and each $f_j \in \mathcal{H}$,

$$\|f_j\|_\infty := \sup_{x_j} |f_j(x_j)| \leq c\, \|f_j\|_{\mathcal{H}}. \tag{4.10}$$

This condition is satisfied for many kernel classes including Sobolev spaces, and any univariate RKHS in which the kernel function[2] bounded uniformly by c. Such a condition is routinely imposed for proving upper bounds on rates of convergence for non-parametric least squares in the univariate case $d = 1$ [80, 83]. Note that this univariate boundedness does not imply that the multivariate functions $f = \sum_{j \in S} f_j$ in $\mathcal{F}$ are uniformly bounded independently of (d, s); rather, since such functions are the sum of s terms, they can take on values of the order $\sqrt{s}$.

The following result applies to any class $\mathcal{F}_{d,s,\mathcal{H}}$ of sparse additive models based on a univariate Hilbert space satisfying condition (4.10), and to the estimator (4.6) based on n i.i.d. samples $(x_i, y_i)_{i=1}^n$ from the observation model (4.5).

Theorem 6. *Let $\widehat{f}$ be any minimizer of the convex program* (4.6) *with regularization parameters $\lambda_n \geq 16\gamma_n$ and $\rho_n \geq 16\gamma_n^2$. Then provided that $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$, there are universal constants (C, c_1, c_2) such that*

$$\mathbb{P}\Bigg[\max\{\|\widehat{f} - f^*\|_2^2,\ \|\widehat{f} - f^*\|_n^2\} \geq C\big\{s\lambda_n^2 + s\rho_n\big\}\Bigg] \leq c_1 \exp(-c_2 n\gamma_n^2).$$

We provide the proof of Theorem 6 in Section 4.4.1.

Remarks

First, the technical condition $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$ is quite mild, and satisfied in most cases of interest, among them the kernels considered below in Corollaries 2 and 3.

Second, note that setting $\lambda_n = c\gamma_n$ and $\rho_n = c\gamma_n^2$ for some constant $c \in [16, \infty)$ yields the rate $\Theta(s\gamma_n^2 + s\rho_n) = \Theta(\frac{s \log d}{n} + s\nu_n^2)$. This rate may be interpreted as the sum of a

[2] Indeed, we have

$$\sup_{x_j} |f_j(x_j)| = \sup_{x_j} |\langle f_j(.), \ker(., x_j)\rangle_{\mathcal{H}}| \ \leq\ \sup_{x_j} \sqrt{\ker(x_j, x_j)}\|f_j\|_{\mathcal{H}}.$$

subset selection term ($\frac{s \log d}{n}$) and an s-dimensional estimation term ($s\nu_n^2$). Note that the subset selection term ($\frac{s \log d}{n}$) is independent of the choice of Hilbert space $\mathcal{H}$, whereas the s-dimensional estimation term is independent of the ambient dimension d. Depending on the scaling of the triple (n, d, s) and the smoothness of the univariate RKHS $\mathcal{H}$, either the subset selection term or function estimation term may dominate. In general, if $\frac{\log d}{n} = o(\nu_n^2)$, the s-dimensional estimation term dominates, and vice versa otherwise. At the boundary, the scalings of the two terms are equivalent.

Finally, for clarity, we have stated our result in the case where the univariate Hilbert space $\mathcal{H}$ is identical across all co-ordinates. However, our proof extends with only notational changes to the general setting, in which each co-ordinate j is endowed with a (possibly distinct) Hilbert space $\mathcal{H}_j$. In this case, the M-estimator returns a function $\widehat{f}$ such that (with high probability)

$$\max\big\{\|\widehat{f} - f^*\|_n^2,\ \|\widehat{f} - f^*\|_2^2\big\} \ \leq\ C\Big\{\frac{s \log d}{n} + \sum_{j \in S} \nu_{n,j}^2\Big\},$$

where $\nu_{n,j}$ is the critical univariate rate associated with the Hilbert space $\mathcal{H}_j$, and S is the subset on which f^* is supported.

Theorem 6 has a number of corollaries, obtained by specifying particular choices of kernels. First, we discuss m-rank operators, meaning that the kernel function ker can be expanded in terms of m eigenfunctions. This class includes linear functions, polynomial functions, as well as any function class based on finite dictionary expansions. First we present a corollary for finite-rank kernel classes.

Corollary 2. *Under the same conditions as Theorem 6, consider an univariate kernel with finite rank m. Then any solution $\widehat{f}$ to the problem (4.6) with $\lambda_n = c\gamma_n$ and $\rho_n = c\gamma_n^2$ with $16 \leq c < \infty$ satisfies*

$$\mathbb{P}\Big[\max\big\{\|\widehat{f} - f^*\|_n^2, \|\widehat{f} - f^*\|_2^2\big\} \geq C\big\{\frac{s \log d}{n} + s\frac{m}{n}\big\}\Big] \leq c_1 \exp\big(-c_2(m + \log d)\big). \quad (4.11)$$

Proof. : It suffices to show that the critical univariate rate (4.8) satisfies the scaling $\nu_n^2 = \mathcal{O}(m/n)$. For a finite-rank kernel and any $t > 0$, we have

$$\mathcal{Q}_{\sigma,n}(t) = \frac{1}{\sqrt{n}}\sqrt{\sum_{j=1}^{m} \min\{t^2, \mu_j\}} \ \leq\ t\sqrt{\frac{m}{n}},$$

from which the claim follows by the definition (4.8). □

Next, we present a result for the RKHS's with infinitely many eigenvalues, but whose eigenvalues decay at a rate $\mu_k \simeq (1/k)^{2\alpha}$ for some parameter $\alpha > 1/2$. Among other examples, this type of scaling covers the case of Sobolev spaces, say consisting of functions with α derivatives [14, 37].

Corollary 3. *Under the same conditions as Theorem 6, consider an univariate kernel with eigenvalue decay $\mu_k \simeq (1/k)^{2\alpha}$ for some $\alpha > 1/2$. Then the kernel estimator defined in* (4.6) *with $\lambda_n = c\gamma_n$ and $\rho_n = c\gamma_n^2$ with $16 \leq c < \infty$ satisfies*

$$\mathbb{P}\Big[\max\big\{\|\widehat{f} - f^*\|_n^2, \|\widehat{f} - f^*\|_2^2\big\} \geq C\big\{\frac{s\log d}{n} + s\big(\frac{1}{n}\big)^{\frac{2\alpha}{2\alpha+1}}\big\}\Big] \leq c_1 \exp\big(-c_2(n^{\frac{1}{2\alpha+1}} + \log d)\big). \tag{4.12}$$

Proof. : As in the previous corollary, we need to compute the critical univariate rate ν_n. Given the assumption of polynomial eigenvalue decay, a truncation argument shows that $\mathcal{Q}_{\sigma,n}(t) = \mathcal{O}\big(\frac{t^{1-\frac{1}{2\alpha}}}{\sqrt{n}}\big)$. Consequently, the critical univariate rate (4.8) satisfies the scaling $\nu_n^2 \asymp \nu_n^{1-\frac{1}{2\alpha}}/\sqrt{n}$, or equivalently, $\nu_n^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}}$. □

4.3.3 Minimax Lower Bounds

In this section, we derive lower bounds on the minimax error in the $L^2(\mathbb{P})$-norm that complement the achievability results derived in Theorem 6. Given the function class $\mathcal{F}$, we define the minimax $L^2(\mathbb{P})$-error $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}})$ to be the largest quantity such that

$$\inf_{\widehat{f}_n} \sup_{f^* \in \mathcal{F}} \mathbb{P}_{f^*}[\|\widehat{f}_n - f^*\|_2^2 \geq \mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}})] \geq 1/2, \tag{4.13}$$

where the infimum is taken over all measurable functions of the n samples $\{(x_i, y_i)\}_{i=1}^n$, and $\mathbb{P}_{f^*}$ denotes the data distribution when the unknown function is f^*. Given this definition, note that Markov's inequality implies that

$$\inf_{\widehat{f}_n} \sup_{f^* \in \mathcal{F}} \mathbb{E}\|\widehat{f}_n - f^*\|_2^2 \geq \frac{\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}})}{2}.$$

Central to our proof of the lower bounds is the metric entropy structure of the univariate reproducing kernel Hilbert spaces. More precisely, our lower bounds depend on the *packing entropy,* defined as follows. Let $(\mathcal{S}, \rho)$ be a totally bounded metric space, consisting of a set $\mathcal{S}$ and a metric $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$. An ϵ-packing of $\mathcal{S}$ is a collection $\{f^1, \ldots, f^M\} \subset \mathcal{S}$ such that $\rho(f^i, f^j) \geq \epsilon$ for all $i \neq j$. The ϵ-packing number $M(\epsilon; \mathcal{S}, \rho)$ is the cardinality of the largest ϵ-packing. The packing entropy is the simply the logarithm of the packing number, namely the quantity $\log M(\epsilon; \mathcal{S}, \rho)$, to which we also refer as the metric entropy. In this chapter, we derive explicit minimax lower bounds for two different scalings of the univariate metric entropy.

Logarithmic Metric Entropy

There exists some $m > 0$ such that

$$\log M(\epsilon; \mathbb{B}_{\mathcal{H}}(1), L^2(\mathbb{P})) \simeq m \,\log(1/\epsilon) \qquad \text{for all } \epsilon \in (0,1). \tag{4.14}$$

Function classes with metric entropy of this type include linear functions (for which $m = k$), univariate polynomials of degree k (for which $m = k+1$), and more generally, any function space with finite VC-dimension [86]. This type of scaling also holds for any RKHS based on a kernel with rank m [22], and these finite-rank kernels include both linear and polynomial functions as special cases.

Polynomial Metric Entropy

There exists some $\alpha > 0$ such that

$$\log M(\epsilon; \mathbb{B}_{\mathcal{H}}(1), L^2(\mathbb{P})) \simeq (1/\epsilon)^{1/\alpha} \qquad \text{for all } \epsilon \in (0,1). \tag{4.15}$$

Various types of Sobolev/Besov classes exhibit this type of metric entropy decay [14, 37]. In fact, any RKHS in which the kernel eigenvalues decay at a rate $k^{-2\alpha}$ have a metric entropy with this scaling [21, 22].

We are now equipped to state our lower bounds on the minimax risk (4.13):

Theorem 7. *Given n i.i.d. samples from the sparse additive model* (4.5) *with sparsity $s \leq d/4$, there is an universal constant $C > 0$ such that:*

(a) For a univariate class $\mathcal{H}$ with logarithmic metric entropy (4.14) *indexed by parameter m, we have*

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}}) \;\geq\; C\Big\{\frac{s\log(d/s)}{n} + s\,\frac{m}{n}\Big\}.$$

(b) For a univariate class $\mathcal{H}$ with polynomial metric entropy (4.15) *indexed by α, we have*

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}}) \;\geq\; C\left\{\frac{s\log(d/s)}{n} + s\,\big(\frac{1}{n}\big)^{\frac{2\alpha}{2\alpha+1}}\right\}. \tag{4.16}$$

The proof of Theorem 7 is provided in Section 4.4.2. The most important consequence of Theorem 7 is in establishing the minimax-optimality of the results given in Corollary 2 and 3; in particular, in the regime of sub-linear sparsity (i.e., for which $\log d = \mathcal{O}(\log(d/s))$), the combination of Theorem 7 with these corollaries identifies the minimax rates up to constant factors.

4.3.4 Comparison With Other Estimators

It is interesting to compare these convergence rates in $L^2(\mathbb{P}_n)$ error with those established in the past work. [71] show that any solution to their back-fitting method is consistent in terms of mean-squared error risk (see Theorem 3 in their chapter), but their analysis does not allow $s \to \infty$. The method of [49] is based regularizing the least-squares loss with the $(\mathcal{H}, 1)$-norm penalty—that is, the regularizer $\sum_{j=1}^{d} \|f_j\|_{\mathcal{H}}$; Theorem 2 in their chapter provides a rate that holds for the triple (n, d, s) tending to infinity. In quantitative terms, however, their rates are looser than those given here; in particular, their bound includes a term of the order $\frac{s^3 \log d}{n}$, which is larger than the bound in Theorem 6. [58] analyze a different M-estimator to the one we analyze in this chapter. Rather than adding two separate $(\mathcal{H}, 1)$-norm and an $(\|.\|_n, 1)$-norm penalties, they combine the two terms into a single sparsity and smoothness penalty. For their estimator, [58] establish a convergence rate of the form $\mathcal{O}(s(\frac{\log d}{n})^{\frac{2\alpha}{2\alpha+1}})$ in the case of α-smooth Sobolev spaces (see Theorem 1 in their chapter). Note that relative to optimal rates given here in Theorem 7(b), this scaling is sub-optimal: more precisely, we either have $\frac{\log d}{n} < (\frac{\log d}{n})^{\frac{2\alpha}{2\alpha+1}}$, when the subset selection term dominates, or $(\frac{1}{n})^{\frac{2\alpha}{2\alpha+1}} < (\frac{\log d}{n})^{\frac{2\alpha}{2\alpha+1}}$, when the s-dimensional estimation term dominates. In all of the above-mentioned methods, it is unclear whether or not a sharper analysis would yield better rates. Finally, [50] analyze the same estimator as the M-estimator (4.6), and for the case of polynomial metric entropy, establish the same rates Theorem 6, albeit under a global boundedness condition. In the following section, we study the implications of this assumption.

4.3.5 Upper Bounds Under A Global Boundedness Assumption

As discussed previously in the introduction, the chapter of [50], referred to as KY for short, is based on the M-estimator (4.6). In terms of rates obtained, they establish a convergence rate based on two terms as in Theorem 6, but with a pre-factor that depends on the global quantity

$$B = \sup_{f \in \mathcal{F}_{d,s,\mathcal{H}}} \|f\|_\infty = \sup_{f \in \mathcal{F}_{d,s,\mathcal{H}}} \sup_{x} |f(x)|,$$

assumed to be bounded independently of dimension and sparsity. Such types of global boundedness conditions are fairly standard in classical non-parametric estimation, where they have no effect on minimax rates. In sharp contrast, the analysis of this section shows that for sparse additive models in the regime $s = \Omega(\sqrt{n})$, such global boundedness can *substantially speed up* minimax rates, showing that the rates proven in KY are not minimax optimal for these classes. The underlying insight is as follows: when the sparsity grows, imposing global boundedness over s-variate functions substantially reduces the effective dimension from its original size s to a lower dimensional quantity, which we denote by $sK_B(s, n)$, and moreover, the quantity $K_B(s, n) \to 0$ when $s = \Omega(\sqrt{n})$ as described below.

Recall the definition (4.4) of the function class $\mathcal{F}_{d,s,\mathcal{H}}$. The model considered in the KY chapter is the smaller function class

$$\mathcal{F}^*_{d,s,\mathcal{H}}(B) := \bigcup_{\substack{S\subset\{1,2,\dots,d\}\\ |S|=s}} \mathcal{H}(S,B),$$

where $\mathcal{H}(S,B) := \big\{f = \sum_{j\in S} f_j \mid f_j \in \mathcal{H}, \text{ and } f_j \in \mathbb{B}_{\mathcal{H}}(1)\ \forall\, j \in S \text{ and } \|f\|_\infty \leq B\big\}$.

The following theorem provides sharper rates for the Sobolev case, in which each univariate Hilbert space has eigenvalues decaying as $\mu_k \simeq k^{-2\alpha}$ for some smoothness parameter $\alpha > 1/2$. Our probabilistic bounds involve the quantity

$$\delta_n := \max\big(\sqrt{\frac{s\log(d/s)}{n}}, B(\frac{s^{\frac{1}{\alpha}}\log s}{n})^{1/4}\big), \tag{4.17}$$

and our rates are stated in terms of the function

$$K_B(s,n) := B\sqrt{\log s}(s^{-1/2\alpha}n^{1/(4\alpha+2)})^{2\alpha-1},$$

where it should be noted that $K_B(s,n) \to 0$ if $s = \Omega(\sqrt{n})$.

With this notation, we have the following *upper bound* on the minimax risk over the function class $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$.

Theorem 8. *Consider any RKHS $\mathcal{H}$ with eigenvalue decay $k^{-2\alpha}$, and uniformly bounded eigenfunctions (i.e., $\|\phi_k\|_\infty \leq C < \infty$ for all k). Then there are universal constants (c_1, c_2, κ) such that with probability greater than $1 - 2\exp\big(-c_1 n\delta_n^2\big)$, we have*

$$\min_{\hat{f}} \max_{f^*\in\mathcal{F}^*_{d,s,\mathcal{H}}(B)} \|\widehat{f} - f^*\|_2^2 \leq \underbrace{\kappa^2(1+B)Csn^{-\frac{2\alpha}{2\alpha+1}}\bigg(K_B(s,n) + n^{-1/(2\alpha+1)}\log(d/s)\bigg)}_{\mathfrak{M}_{\mathbb{P}}(\mathcal{F}^*_{d,s,\mathcal{H}}(B))}, \tag{4.18}$$

as long as $n\delta_n^2 = \Omega(\log(1/\delta_n))$.

We provide the proof of Theorem 8 in Section 4.4.3; it is based on analyzing directly the least-squares estimator over $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$. The assumption that $\|\phi_k\|_\infty \leq C < \infty$ for all k includes the usual Sobolev spaces in which ϕ_k are (rescaled) Fourier basis functions. An immediate consequence of Theorem 8 is that the minimax rates over the function class $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$ can be strictly faster than minimax rates for the class $\mathcal{F}_{d,s,\mathcal{H}}$, which does not impose global boundedness. Recall that the minimax lower bound from Theorem 7 (b) is based on the quantity

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}}) \ := \ C_1\big\{s\big(\frac{1}{n}\big)^{\frac{2\alpha}{2\alpha+1}} + \frac{s\log(d/s)}{n}\big\} \ = \ C_1 sn^{-\frac{2\alpha}{2\alpha+1}}\bigg(1 + n^{-1/(2\alpha+1)}\log(d/s)\bigg),$$

for a universal constant C_1. Note that up to constant factors, the achievable rate (4.18) from Theorem 8 is the same except that the term 1 is replaced by the function $K_B(s,n)$. Consequently, for scalings of (s,n) such that $K_B(s,n) \to 0$, global boundedness conditions lead to strictly faster rates.

Corollary 4. *Under the conditions of Theorem 8, we have*

$$\frac{\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}})}{\mathfrak{M}_{\mathbb{P}}(\mathcal{F}^*_{d,s,\mathcal{H}}(B))} \geq \frac{C_1(1 + n^{-1/(2\alpha+1)} \log(d/s))}{C\,\kappa^2(1+B)\,(K_B(s,n) + n^{-1/(2\alpha+1)} \log(d/s))} \to +\infty$$

whenever $B = \mathcal{O}(1)$ and $K_B(s,n) \to 0$.

Remarks

The quantity $K_B(s,n)$ is guaranteed to decay to zero as long as the sparsity index s grows in a non-trivial way with the sample size. For instance, if we have $s = \Omega(\sqrt{n})$ for a problem of dimension $d = \mathcal{O}(n^\beta)$ for any $\beta \geq 1/2$, then it can be verified that $K_B(s,n) = o(1)$. As an alternative view of the differences, it can be noted that there are scalings of (n,s,d) for which the minimax rate $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}})$ over $\mathcal{F}_{d,s,\mathcal{H}}$ is constant—that is, does not vanish as $n \to +\infty$—while the minimax rate $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}^*_{d,s,\mathcal{H}}(B))$ does vanish. As an example, consider the Sobolev class with smoothness $\alpha = 2$, corresponding to twice-differentiable functions. For a sparsity index $s = \Theta(n^{4/5})$, then Theorem 7(b) implies that $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}}) = \Omega(1)$, so that the minimax rate over $\mathcal{F}_{d,s,\mathcal{H}}$ is strictly bounded away from zero for all sample sizes. In contrast, under a global boundedness condition, Theorem 8 shows that the minimax rate is upper bounded as $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}^*_{d,s,\mathcal{H}}(B)) = \mathcal{O}\big(n^{-1/5}\sqrt{\log n}\big)$, which tends to zero.

In summary, Theorem 8 and Theorem 7(b) together show that the minimax rates over $\mathcal{F}_{d,s,\mathcal{H}}$ and $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$ can be drastically different. Thus, global boundedness is a stringent condition in the high-dimensional setting; in particular, the rates given in Theorem 3 of [50] are not minimax optimal when $s = \Omega(\sqrt{n})$.

4.4 Proofs

In this section, we provide the proofs of our three main theorems. For clarity in presentation, we split the proofs up into a series of lemmas, with the bulk of the more technical arguments deferred to the appendices. This splitting allows our presentation in Section 4.4 to be relatively streamlined.

4.4.1 Proof of Theorem 6

At a high-level, Theorem 6 is based on an appropriate adaptation to the non-parametric setting of various techniques that have been developed for sparse linear regression [11, 64].

In contrast to the parametric setting where classical tail bounds are sufficient, controlling the error terms in the non-parametric case requires more advanced techniques from empirical process theory. In particular, we make use of various concentration theorems for Gaussian and empirical processes [52, 56, 65, 83], as well as results on the Rademacher complexity of kernel classes [9, 61].

At the core of the proof are three technical lemmas. First, Lemma 11 provides an upper bound on the Gaussian complexity of any function of the form $f = \sum_{j=1}^{d} f_j$ in terms of the norms $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{n,1}$ previously defined. Lemma 12 exploits the notion of decomposability [64], as applied to these norms, in order to show that the error function belongs to a particular cone-shaped set. Finally, Lemma 13 establishes an upper bound on the $L^2(\mathbb{P})$ error of our estimator in terms of the $L^2(\mathbb{P}_n)$ error. The latter lemma can be interpreted as proving that our problem satisfies non-parametric analog of a restricted eigenvalue condition [11], or more generally, of a restricted strong convexity condition [64]. The proof of Lemma 13 involves a new approach that combines the Sudakov minoration [65] with a one-sided tail bound for non-negative random variables [24, 31].

Throughout the proof, we use C and c_i, $i = 1, 2, 3, 4$ to denote universal constants, independent of (n, d, s). Note that the precise numerical values of these constants may change from line to line. The reader should recall the definitions of ν_n and γ_n from Equations (4.8) and (4.9) respectively. For a subset $A \subseteq \{1, 2, \ldots, d\}$ and a function of the form $f = \sum_{j=1}^{d} f_j$, we adopt the convenient notation

$$\|f_A\|_{n,1} \; := \sum_{j \in A} \|f_j\|_n, \quad \text{and} \quad \|f_A\|_{\mathcal{H},1} \; := \sum_{j \in A} \|f_j\|_{\mathcal{H}}. \tag{4.19}$$

We begin by establishing an inequality on the error function $\widehat{\Delta} := \widehat{f} - f^*$. Since $\widehat{f}$ and f^* are, respectively, optimal and feasible for the problem (4.6), we are guaranteed that $\mathcal{L}(\widehat{f}) \leq \mathcal{L}(f^*)$, and hence that the error function $\widehat{\Delta}$ satisfies the bound

$$\frac{1}{2n}\sum_{i=1}^{n}(w_i + \bar{f} - \bar{y}_n - \widehat{\Delta}(x_i))^2 + \lambda_n\|\widehat{f}\|_{n,1} + \rho_n\|\widehat{f}\|_{\mathcal{H},1} \leq \frac{1}{2n}\sum_{i=1}^{n}(w_i + \bar{f} - \bar{y}_n)^2 + \lambda_n\|f^*\|_{n,1} + \rho_n\|f^*\|_{\mathcal{H},1}.$$

Some simple algebra yields the bound

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \Big|\frac{1}{n}\sum_{i=1}^{n} w_i\widehat{\Delta}(x_i)\Big| + |\bar{y}_n - \bar{f}|\Big|\frac{1}{n}\sum_{i=1}^{n}\widehat{\Delta}(x_i)\Big| + \lambda_n\|\widehat{\Delta}\|_{n,1} + \rho_n\|\widehat{\Delta}\|_{\mathcal{H},1}. \tag{4.20}$$

Following the terminology of [83], we refer to this bound as our *basic inequality.*

Controlling Deviation From The Mean

Our next step is to control the error due to estimating the mean $|\bar{y}_n - \bar{f}|$. We begin by observing that this error term can be written as $\bar{y}_n - \bar{f} = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{f})$. Next we observe that $y_i - \bar{f} = \sum_{j \in S} f_j^*(x_{ij}) + w_i$ is the sum of the s independent random variables $f_j^*(x_{ij})$, each bounded in absolute value by one, along with the independent sub-Gaussian noise term w_i; consequently, the variable $y_i - \bar{f}$ is sub-Gaussian with parameter at most $\sqrt{s+1}$. (See, for instance, Lemma 1.4 in [17]). By applying standard sub-Gaussian tail bounds, we have $\mathbb{P}(|\bar{y}_n - \bar{f}| > t) \leq 2\exp(-\frac{nt^2}{2(s+1)})$, and hence, if we define the event $\mathcal{C}(\gamma_n) = \{|\bar{y}_n - \bar{f}| \leq \sqrt{s}\gamma_n\}$, we are guaranteed

$$\mathbb{P}[\mathcal{C}(\gamma_n)] \geq 1 - 2\exp(-\frac{n\gamma_n^2}{4}).$$

For the remainder of the proof, we condition on the event $\mathcal{C}(\gamma_n)$. Under this conditioning, the bound (4.20) simplifies to:

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \big|\frac{1}{n}\sum_{i=1}^n w_i \widehat{\Delta}(x_i)\big| + \sqrt{s}\gamma_n \|\widehat{\Delta}\|_n + \lambda_n \|\widehat{\Delta}\|_{n,1} + \rho_n \|\widehat{\Delta}\|_{\mathcal{H},1},$$

where we have applied the Cauchy-Schwarz inequality to write $\big|\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}(x_i)\big| \leq \|\widehat{\Delta}\|_n$.

Controlling the Gaussian Complexity Term

The following lemma provides control the Gaussian complexity term on the right-hand side of inequality (4.20) by bounding the Gaussian complexity for the univariate functions $\widehat{\Delta}_j$, $j = 1, 2, \ldots, d$ in terms of their $\|\cdot\|_n$ and $\|\cdot\|_{\mathcal{H}}$ norms. In particular, recalling that $\gamma_n = \kappa \max\{\sqrt{\frac{\log d}{n}},\ \nu_n\}$, we have the following lemma.

Lemma 11. *Define the event*

$$\mathcal{T}(\gamma_n) := \Big\{\forall\ \ j = 1, 2, \ldots, d,\ \big|\frac{1}{n}\sum_{i=1}^n w_i \widehat{\Delta}_j(x_{ij})\big| \leq 8\gamma_n^2\ \|\widehat{\Delta}_j\|_{\mathcal{H}} + 8\gamma_n\ \|\widehat{\Delta}_j\|_n\Big\}. \tag{4.21}$$

Then under the condition $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$, we have

$$\mathbb{P}(\mathcal{T}(\gamma_n)) \geq 1 - c_1 \exp(-c_2 n\gamma_n^2).$$

The proof of this lemma, provided in Appendix B of Raskutti et al. [69], uses concentration of measure for Lipschitz functions of Gaussian random variables [52], combined with peeling and weighting arguments from empirical process theory [5, 83]. In particular, the subset selection term $(\frac{s\log d}{n})$ in Theorem 6 arises from taking the maximum over all d components.

The remainder of our analysis involves conditioning on the event $\mathcal{T}(\gamma_n) \cap \mathcal{C}(\gamma_n)$. Using Lemma 11, when conditioned on the event $\mathcal{T}(\gamma_n) \cap \mathcal{C}(\gamma_n)$ we have:

$$\|\widehat{\Delta}\|_n^2 \;\leq\; 2\sqrt{s}\gamma_n\|\widehat{\Delta}\|_n + (16\gamma_n + 2\lambda_n)\|\widehat{\Delta}\|_{n,1} + (16\gamma_n^2 + 2\rho_n)\|\widehat{\Delta}\|_{\mathcal{H},1}. \tag{4.22}$$

Exploiting Decomposability

Recall that S denotes the true support of the unknown function f^*. By the definition (4.19), we can write $\|\widehat{\Delta}\|_{n,1} = \|\widehat{\Delta}_S\|_{n,1} + \|\widehat{\Delta}_{S^c}\|_{n,1}$, where $\widehat{\Delta}_S := \sum_{j \in S} \widehat{\Delta}_j$ and $\widehat{\Delta}_{S^c} := \sum_{j \in S^c} \widehat{\Delta}_j$. Similarly, we have an analogous representation for $\|\widehat{\Delta}\|_{\mathcal{H},1}$. The next lemma shows that conditioned on the event $\mathcal{T}(\gamma_n)$, the quantities $\|\widehat{\Delta}\|_{\mathcal{H},1}$ and $\|\widehat{\Delta}\|_{n,1}$ are not significantly larger than the corresponding norms as applied to the function $\widehat{\Delta}_S$.

Lemma 12. *Conditioned on the events $\mathcal{T}(\gamma_n)$ and $\mathcal{C}(\gamma_n)$, and with the choices $\lambda_n \geq 16\gamma_n$ and $\rho_n \geq 16\gamma_n^2$, we have*

$$\lambda_n\|\widehat{\Delta}\|_{n,1} + \rho_n\|\widehat{\Delta}\|_{\mathcal{H},1} \leq 4\lambda_n\|\widehat{\Delta}_S\|_{n,1} + 4\rho_n\|\widehat{\Delta}_S\|_{\mathcal{H},1} + \frac{1}{2}s\gamma_n^2. \tag{4.23}$$

The proof of this lemma, provided in Appendix C, is based on the decomposability (see [64]) of the $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{n,1}$ norms. This lemma allows us to exploit the sparsity assumption, since in conjunction with Lemma 11, we have now bounded the right-hand side of the inequality (4.22) by terms involving only $\widehat{\Delta}_S$.

For the remainder of the proof of Theorem 6, we assume $\lambda_n \geq 16\gamma_n$ and $\rho_n \geq 16\gamma_n^2$. In particular, still conditioning on $\mathcal{C}(\gamma_n) \cap \mathcal{T}(\gamma_n)$ and applying Lemma 12 to inequality (4.22), we obtain

$$\begin{aligned}
\|\widehat{\Delta}\|_n^2 &\leq 2\sqrt{s}\gamma_n\|\widehat{\Delta}\|_n + 3\lambda_n\|\widehat{\Delta}\|_{n,1} + 3\rho_n\|\widehat{\Delta}\|_{\mathcal{H},1} \\
&\leq 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 12\lambda_n\|\widehat{\Delta}_S\|_{n,1} + 12\rho_n\|\widehat{\Delta}_S\|_{\mathcal{H},1} + \frac{3}{32}s\rho_n,
\end{aligned}$$

Finally, since both $\widehat{f}_j$ and f_j^* belong to $\mathbb{B}_{\mathcal{H}}(1)$, we have $\|\widehat{\Delta}_j\|_{\mathcal{H}} \leq \|\widehat{f}_j\|_{\mathcal{H}} + \|f_j^*\|_{\mathcal{H}} \leq 2$, which implies that $\|\widehat{\Delta}_S\|_{\mathcal{H},1} \leq 2s$, and hence

$$\|\widehat{\Delta}\|_n^2 \leq 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 12\lambda_n\|\widehat{\Delta}_S\|_{n,1} + 25s\rho_n. \tag{4.24}$$

Upper Bounding $\|\widehat{\Delta}_S\|_{n,1}$

The next step is to control the term $\|\widehat{\Delta}_S\|_{n,1} = \sum_{j \in S}\|\widehat{\Delta}_j\|_n$ that appears in the upper bound (4.24). Ideally, we would like to upper bound it by a quantity of the order $\sqrt{s}\|\widehat{\Delta}_S\|_2 = \sqrt{s}\sqrt{\sum_{j \in S}\|\widehat{\Delta}_j\|_2^2}$. Such an upper bound would follow immediately if it were phrased in terms of the population $\|\cdot\|_2$-norm rather than the empirical-$\|\cdot\|_n$ norm, but there are additional cross-terms with the empirical norm. Accordingly, a somewhat more delicate argument is required, which we provide here. First define the events

$$\mathcal{A}_j(\lambda_n) := \{\|\widehat{\Delta}_j\|_n \leq 2\|\widehat{\Delta}_j\|_2 + \lambda_n\},$$

and $\mathcal{A}(\lambda_n) = \cap_{j=1}^d \mathcal{A}_j(\lambda_n)$. By applying Lemma 7 from Appendix A with $t = \lambda_n \geq 16\gamma_n$ and $b = 2$, we conclude that $\|\widehat{\Delta}_j\|_n \leq 2\|\widehat{\Delta}_j\|_2 + \lambda_n$ with probability greater than $1 - c_1 \exp(-c_2 n\lambda_n^2)$. Consequently, if we define the event $\mathcal{A}(\lambda_n) = \cap_{j\in S}\mathcal{A}_j(\lambda_n)$, then this tail bound together with the union bound implies that

$$\mathbb{P}[\mathcal{A}^c(\lambda_n)] \leq s\, c_1 \exp(-c_2 n\lambda_n^2) \;\leq\; c_1 \exp(-c_2' n\lambda_n^2), \tag{4.25}$$

where we have used the fact that $\lambda_n = \Omega(\sqrt{\frac{\log s}{n}})$. Now, conditioned on the event $\mathcal{A}(\lambda_n)$, we have

$$\begin{aligned} \|\widehat{\Delta}_S\|_{n,1} = \sum_{j\in S}\|\widehat{\Delta}_j\|_n \;&\leq\; 2\sum_{j\in S}\|\widehat{\Delta}_j\|_2 + s\lambda_n \\ &\leq\; 2\sqrt{s}\|\widehat{\Delta}_S\|_2 + s\lambda_n \leq 2\sqrt{s}\|\widehat{\Delta}\|_2 + s\lambda_n. \end{aligned} \tag{4.26}$$

Substituting this upper bound (4.26) on $\|\widehat{\Delta}_S\|_{n,1}$ into our earlier inequality (4.24) yields

$$\|\widehat{\Delta}\|_n^2 \;\leq\; 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 24\sqrt{s}\lambda_n\|\widehat{\Delta}\|_2 + 12s\lambda_n^2 + 25s\rho_n. \tag{4.27}$$

At this point, we encounter a challenge due to the unbounded nature of our function class. In particular, if $\|\widehat{\Delta}\|_2$ were upper bounded by $C\max(\|\widehat{\Delta}\|_n, \sqrt{s}\lambda_n, \sqrt{s\rho_n})$, then the upper bound (4.27) would immediately imply the claim of Theorem 6. If one were to assume global boundedness of the multivariate functions $\widehat{f}$ and f^*, as done in past work of [50], then an upper bound on $\|\widehat{\Delta}\|_2$ of this form would directly follow from known results (e.g., Theorem 2.1 in [9].) However, since we do not impose global boundedness, we need to develop a novel approach to this final hurdle.

Controlling $\|\widehat{\Delta}\|_2$ For Unbounded Classes

For the remainder of the proof, we condition on the event $\mathcal{A}(\lambda_n) \cap \mathcal{T}(\gamma_n) \cap \mathcal{C}(\gamma_n)$. We split our analysis into three cases. Throughout the proof, we make use of the quantity

$$\tilde{\delta}_n := B\max(\sqrt{s}\lambda_n, \sqrt{s\rho_n}), \tag{4.28}$$

where $B \in (1, \infty)$ is a constant to be chosen later in the argument.

Case 1: If $\|\widehat{\Delta}\|_2 < \|\widehat{\Delta}\|_n$, then combined with inequality (4.27), we conclude that

$$\|\widehat{\Delta}\|_n^2 \le 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 24\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 12s\lambda_n^2 + 25s\rho_n.$$

This is a quadratic inequality in terms of the quantity $\|\widehat{\Delta}\|_n$, and some algebra shows that it implies the bound $\|\widehat{\Delta}\|_n \le 15\max(\sqrt{s}\lambda_n, \sqrt{s\rho_n})$. By assumption, we then have $\|\widehat{\Delta}\|_2 \le 15\max(\sqrt{s}\lambda_n, \sqrt{s\rho_n})$ as well, thereby completing the proof of Theorem 6.

Case 2: If $\|\widehat{\Delta}\|_2 < \tilde{\delta}_n$, then together with the bound (4.27), we conclude that

$$\|\widehat{\Delta}\|_n^2 \le 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 24\sqrt{s}\lambda_n\tilde{\delta}_n + 12s\lambda_n^2 + 25s\rho_n.$$

This inequality is again a quadratic in $\|\widehat{\Delta}\|_n$; moreover, note that by definition (4.28) of $\tilde{\delta}_n$, we have $s\lambda_n^2 + s\rho_n = \mathcal{O}(\tilde{\delta}_n^2)$. Consequently, this inequality implies that $\|\widehat{\Delta}\|_n \le C\tilde{\delta}_n$ for some constant C. Our starting assumption implies that $\|\widehat{\Delta}\|_2 \le \tilde{\delta}_n$, so that the claim of Theorem 6 follows in this case.

Case 3: Otherwise, we may assume that $\|\widehat{\Delta}\|_2 \ge \tilde{\delta}_n$ and $\|\widehat{\Delta}\|_2 \ge \|\widehat{\Delta}\|_n$. In this case, the inequality (4.27) together with the bound $\|\widehat{\Delta}\|_2 \ge \|\widehat{\Delta}\|_n$ implies that

$$\|\widehat{\Delta}\|_n^2 \le 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_2 + 24\sqrt{s}\lambda_n\|\widehat{\Delta}\|_2 + 12s\lambda_n^2 + 25s\rho_n. \tag{4.29}$$

Our goal is to establish a lower bound on the left-hand-side—namely, the quantity $\|\widehat{\Delta}\|_n^2$—in terms of $\|\widehat{\Delta}\|_2^2$. In order to do so, we consider the function class $\mathcal{G}(\lambda_n, \rho_n)$ defined by functions of the form $g = \sum_{j=1}^d g_j$, and such that

$$\begin{aligned} \lambda_n\|g\|_{n,1} + \rho_n\|g\|_{\mathcal{H},1} \le&\ 4\lambda_n\|g_S\|_{n,1} + 4\rho_n\|g_S\|_{\mathcal{H},1} + \frac{1}{32}s\rho_n, && (4.30)\\ \|g_S\|_{1,n} \le&\ 2\sqrt{s}\|g_S\|_2 + s\lambda_n \qquad \text{and} && (4.31)\\ \|g\|_n \le&\ \|g\|_2. && (4.32) \end{aligned}$$

Conditioned on the events $\mathcal{A}(\gamma_n)$, $\mathcal{T}(\gamma_n)$ and $\mathcal{C}(\gamma_n)$, and with our choices of regularization parameter, we are guaranteed that the error function $\widehat{\Delta}$ satisfies all three of these constraints, and hence that $\widehat{\Delta} \in \mathcal{G}(\lambda_n, \rho_n)$. Consequently, it suffices to establish a lower bound on $\|g\|_n$ that holds uniformly over the class $\mathcal{G}(\lambda_n, \rho_n)$. In particular, define the event

$$\mathcal{B}(\lambda_n, \rho_n) := \Big\{ \|g\|_n^2 \ge \|g\|_2^2/2 \quad \text{for all } g \in \mathcal{G}(\lambda_n, \rho_n) \quad \text{such that} \quad \|g\|_2 \ge \tilde{\delta}_n \Big\}. \tag{4.33}$$

The following lemma shows that this event holds with high probability.

Lemma 13. *Under the conditions of Theorem 6, there are universal constants c_i such that*

$$\mathbb{P}[\mathcal{B}(\lambda_n, \rho_n)] \geq 1 - c_1 \exp(-c_2 n \gamma_n^2). \tag{4.34}$$

We note that this lemma can be interpreted as guaranteeing a version of restricted strong convexity [64] for the least-squares loss function, suitably adapted to the non-parametric setting. Since we do not assume global boundedness, the proof of this lemma requires a novel technical argument, one which combines a one-sided tail bound for non-negative random variables [24, 31] with the Sudakov minoration [65] for the Gaussian complexity. We refer the reader to Appendix D for the details of the proof.

Using Lemma 13 and conditioning on the event $\mathcal{B}(\lambda_n, \rho_n)$, we are guaranteed that $\|\widehat{\Delta}\|_n^2 \geq \|\widehat{\Delta}\|_2^2/2$, and hence, combined with our earlier bound (4.29), we conclude that

$$\|\widehat{\Delta}\|_2^2 \leq 4\sqrt{s}\lambda_n\|\widehat{\Delta}\|_2 + 48\sqrt{s}\lambda_n\|\widehat{\Delta}\|_2 + 24s\lambda_n^2 + 50s\rho_n.$$

Hence $\|\widehat{\Delta}\|_n \leq \|\widehat{\Delta}\|_2 \leq C \max(\sqrt{s}\lambda_n, \sqrt{s\rho_n})$, completing the proof of the claim in the third case.

In summary, the entire proof is based on conditioning on the three events $\mathcal{T}(\gamma_n)$, $\mathcal{A}(\lambda_n)$ and $\mathcal{B}(\lambda_n, \rho_n)$. From the bound (4.25) as well as Lemmas 11 and 13, we have

$$\mathbb{P}\big[\mathcal{T}(\gamma_n) \cap \mathcal{A}(\lambda_n) \cap \mathcal{B}(\lambda_n, \rho_n) \cap \mathcal{C}(\gamma_n)\big] \geq 1 - c_1 \exp\big(-c_2 n \gamma_n^2\big),$$

thereby showing that $\max\{\|\widehat{f} - f^*\|_n^2, \|\widehat{f} - f^*\|_2^2\} \leq C \max(s\lambda_n^2, s\rho_n)$ with the claimed probability. This completes the proof of Theorem 6.

4.4.2 Proof of Theorem 7

We now turn to the proof of the minimax lower bounds stated in Theorem 7. For both parts (a) and (b), the first step is to follow a standard reduction to testing (see e.g. [42, 92, 93]) so as to obtain a lower bound on the minimax error $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}})$ in terms of the probability of error in a multi-way hypothesis testing. We then apply different forms of the Fano inequality [92, 93] in order to lower bound the probability of error in this testing problem. Obtaining useful bounds requires a precise characterization of the metric entropy structure of $\mathcal{F}_{d,s,\mathcal{H}}$, as stated in Lemma 14.

Reduction to Testing

We begin with the reduction to a testing problem. Let $\{f^1, \ldots, f^M\}$ be a δ_n-packing of $\mathcal{F}$ in the $\|\cdot\|_2$-norm, and let Θ be a random variable uniformly distributed over the index set $[M] := \{1, 2, \ldots, M\}$. Note that we are using M as a shorthand for the packing number

$M(\delta_n; \mathcal{F}, \|\cdot\|_2)$. A standard argument [42, 92, 93] then yields the lower bound

$$\inf_{\widehat{f}} \sup_{f^* \in \mathcal{F}} \mathbb{P}\big[\|\widehat{f} - f^*\|_2^2 \geq \delta_n^2/2\big] \geq \inf_{\widehat{\Theta}} \mathbb{P}[\widehat{\Theta} \neq \Theta],$$

where the infimum on the right-hand side is taken over all estimators $\widehat{\Theta}$ that are measurable functions of the data, and take values in the index set $[M]$.

Note that $\mathbb{P}[\widehat{\Theta} \neq \Theta]$ corresponds to the error probability in a multi-way hypothesis test, where the probability is taken over the random choice of Θ, the randomness of the design points $X_1^n := \{x_i\}_{i=1}^n$, and the randomness of the observations $Y_1^n := \{y_i\}_{i=1}^n$. Our initial analysis is performed conditionally on the design points, so that the only remaining randomness in the observations Y_1^n comes from the observation noise $\{w_i\}_{i=1}^n$. From Fano's inequality [26], for any estimator $\widehat{\Theta}$, we have $\mathbb{P}\big[\widehat{\Theta} \neq \Theta \mid X_1^n\big] \geq 1 - \frac{I_{X_1^n}(\Theta; Y_1^n) + \log 2}{\log M}$, where $I_{X_1^n}(\Theta; Y_1^n)$ denotes the mutual information between Θ and Y_1^n with X_1^n fixed. Taking expectations over X_1^n, we obtain the lower bound

$$\mathbb{P}\big[\widehat{\Theta} \neq \Theta\big] \geq 1 - \frac{\mathbb{E}_{X_1^n}\big[I_{X_1^n}(\Theta; Y_1^n)\big] + \log 2}{\log M}. \tag{4.35}$$

The remainder of the proof consists of constructing appropriate packing sets of $\mathcal{F}$, and obtaining good upper bounds on the mutual information term in the lower bound (4.35).

Constructing Appropriate Packings

We begin with results on packing numbers. Recall that $\log M(\delta; \mathcal{F}, \|\cdot\|_2)$ denotes the δ-packing entropy of $\mathcal{F}$ in the $\|\cdot\|_2$ norm.

Lemma 14. *(a) For all $\delta \in (0,1)$ and $s \leq d/4$, we have*

$$\log M(\delta; \mathcal{F}, \|\cdot\|_2) = \mathcal{O}\big(s \ \log M(\frac{\delta}{\sqrt{s}}; \mathbb{B}_{\mathcal{H}}(1), \|\cdot\|_2) + s \log \frac{d}{s}\big).$$

(b) For a Hilbert class with logarithmic metric entropy (4.14) *and such that $\|f\|_2 \leq \|f\|_{\mathcal{H}}$, there exists set $\{f^1, \ldots, f^M\}$ with $\log M \ \geq \ C\big\{s \log(d/s) + sm\big\}$, and*

$$\delta \ \leq \ \|f^k - f^\ell\|_2 \ \leq \ 8\delta \qquad \text{for all } k \neq \ell \in \{1, 2, \ldots, M\}.$$

The proof, provided in Appendix E, is combinatorial in nature. We now turn to the proofs of parts (a) and (b) of Theorem 7.

Proof of Theorem 7(a)

In order to prove this claim, it remains to exploit Lemma 14 in an appropriate way, and to upper bound the resulting mutual information. For the latter step, we make use of the generalized Fano approach [93].

From Lemma 14, we can find a set $\{f^1, \ldots, f^M\}$ that is a δ-packing of $\mathcal{F}$ in ℓ_2-norm, and such that $\|f^k - f^\ell\|_2 \leq 8\delta$ for all $k, \ell \in [M]$. For $k = 1, \ldots, M$, let $\mathbb{Q}^k$ denote the conditional distribution of Y_1^n conditioned on X_1^n and the event $\{\Theta = k\}$, and let $D(\mathbb{Q}^k \,\|\, \mathbb{Q}^\ell)$ denote the Kullback-Leibler divergence. From the convexity of mutual information [26], we have the upper bound $I_{X_1^n}(\Theta; Y_1^n) \leq \frac{1}{\binom{M}{2}} \sum_{k,\ell=1}^M D(\mathbb{Q}^k \,\|\, \mathbb{Q}^\ell)$. Given our linear observation model (4.5), we have

$$D(\mathbb{Q}^k \,\|\, \mathbb{Q}^\ell) = \frac{1}{2\sigma^2} \sum_{i=1}^n \big(f^k(x_i) - f^\ell(x_i)\big)^2 \; = \; \frac{n\,\|f^k - f^\ell\|_n^2}{2},$$

and hence

$$\mathbb{E}_{X_1^n}\big[I_{X_1^n}(Y_1^n; \Theta)\big] \leq \frac{n}{2} \frac{1}{\binom{M}{2}} \sum_{k \neq \ell} \mathbb{E}_{X_1^n}[\|f^k - f^\ell\|_n^2] \; = \; \frac{n}{2} \frac{1}{\binom{M}{2}} \sum_{k \neq \ell} \|f^k - f^\ell\|_2^2.$$

Since our packing satisfies $\|f^k - f^\ell\|_2^2 \leq 64\delta^2$, we conclude that

$$\mathbb{E}_{X_1^n}\big[I_{X_1^n}(Y_1^n; \Theta)\big] \leq 32n\delta^2.$$

From the Fano bound (4.35), for any $\delta > 0$ such that $\frac{32n\delta^2 + \log 2}{\log M} < \frac{1}{4}$, then we are guaranteed that $\mathbb{P}[\widehat{\Theta} \neq \Theta] \geq \frac{3}{4}$. From Lemma 14(b), our packing set satisfies $\log M \geq C\big\{sm + s\log(d/s)\big\}$, so that so that the choice $\delta^2 = C'\left\{\frac{sm}{n} + \frac{s\log(d/s)}{n}\right\}$, for a suitably small $C' > 0$, can be used to guarantee the error bound $\mathbb{P}[\widehat{\Theta} \neq \Theta] \geq \frac{3}{4}$.

Proof of Theorem 7(b)

In this case, we use an upper bounding technique due to [92] in order to upper bound the mutual information. Although the argument is essentially the same, it does not follow verbatim from their claims—in particular, there are some slight differences due to our initial conditioning—so that we provide the details here. By definition of the mutual information, we have

$$I_{X_1^n}(\Theta; Y_1^n) = \frac{1}{M} \sum_{k=1}^M D(\mathbb{Q}^k \,\|\, \mathbb{P}_Y),$$

where $\mathbb{Q}^k$ denotes the conditional distribution of Y_1^n given $\Theta = k$ and still with X_1^n fixed, whereas $\mathbb{P}_Y$ denotes the marginal distribution of $\mathbb{P}_Y$.

Let us define the notion of a covering number, in particular for a totally bounded metric space $(\mathcal{S}, \rho)$, consisting of a set $\mathcal{S}$ and a metric $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$. An ϵ-covering set of $\mathcal{S}$ is a collection $\{f^1, \ldots, f^N\}$ of functions such that for all $f \in \mathcal{S}$ there exists $k \in \{1, 2, ..., N\}$ such that $\rho(f, f^k) \leq \epsilon$. The ϵ-covering number $N(\epsilon; \mathcal{S}, \rho)$ is the cardinality of the smallest ϵ-covering set.

Now let $\{g^1, \ldots, g^N\}$ be an ϵ-cover of $\mathcal{F}$ in the $\|\cdot\|_2$ norm, for a tolerance ϵ to be chosen. As argued in [92], we have

$$I_{X_1^n}(\Theta; Y_1^n) \;=\; \frac{1}{M}\sum_{j=1}^{M} D(\mathbb{Q}^j \,\|\, \mathbb{P}_Y) \leq D(\mathbb{Q}^k \,\|\, \frac{1}{N}\sum_{k=1}^{N} \mathbb{P}^k),$$

where $\mathbb{P}^\ell$ denotes the conditional distribution of Y_1^n given g^ℓ and X_1^n. For each ℓ, let us choose $g^{\ell^*(k)}$ as follows: $\ell^*(k) \in \arg\min_{\ell=1,\ldots,N} \|g^\ell - f^k\|_2$. We then have the upper bound

$$I_{X_1^n}(\Theta; Y_1^n) \leq \frac{1}{M}\sum_{k=1}^{M} \big\{ \log N + \frac{n}{2}\|g^{\ell^*(k)} - f^k\|_n^2 \big\}.$$

Taking expectations over X_1^n, we obtain

$$\begin{aligned}\mathbb{E}_{X_1^n}[I_{X_1^n}(\Theta; Y_1^n)] &\leq \frac{1}{M}\sum_{k=1}^{M} \big\{ \log N + \frac{n}{2}\mathbb{E}_{X_1^n}[\|g^{\ell^*(k)} - f^k\|_n^2] \big\} \\ &\leq \log N + \frac{n}{2}\,\epsilon^2,\end{aligned}$$

where the final inequality follows from the choice of our covering set.

From this point, we can follow the same steps as [92]. The polynomial scaling (4.15) of the metric entropy guarantees that their conditions are satisfied, and we conclude that the minimax error is lower bounded by any $\delta_n > 0$ such that $n\delta_n^2 \;\geq\; C \log N(\delta_n; \mathcal{F}, \|\cdot\|_2)$. From Lemma 14 and the assumed scaling (4.15), it is equivalent to solve the equation

$$n\delta_n^2 \;\geq\; C\bigg\{ s\log(d/s) + s(\sqrt{s}/\delta_n)^{1/\alpha} \bigg\},$$

from which some algebra yields $\delta_n^2 = C\big\{\frac{s\log(d/s)}{n} + s\big(\frac{1}{n}\big)^{\frac{2\alpha}{2\alpha+1}}\big\}$ as a suitable choice.

4.4.3 Proof of Theorem 8

Recall the definition of $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$ and $\mathcal{H}(S,B)$ from Section 4.3.5; note that it guarantees that $\|f^*\|_\infty \leq B$. In order to establish upper bounds on the minimax rate in $L^2(\mathbb{P})$-error over $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$, we analyze a least-squares estimator—albeit *not* the same as the original M-estimator (4.6)—constrained to $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$, namely

$$\widehat{f} \in \arg \min_{f \in \mathcal{F}^*_{d,s,\mathcal{H}}(B)} \sum_{i=1}^n (y_i - \bar{y}_n - f(x_i))^2. \tag{4.36}$$

Since our goal is to upper bound the minimax rate in $L^2(\mathbb{P})$ error, it is sufficient to upper bound the $L^2(\mathbb{P})$-norm of $\widehat{f} - f^*$ where $\widehat{f}$ is any solution to (4.36). The proof shares many steps with the proof of Theorem 6. First, the same reasoning shows that the error $\widehat{\Delta} := \widehat{f} - f^*$ satisfies the basic inequality

$$\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}^2(x_i) \leq \frac{2}{n}|\sum_{i=1}^n w_i \widehat{\Delta}(x_i)| + |\bar{y}_n - \bar{f}||\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}(x_i)|.$$

Recall the definition (4.17) of the critical rate δ_n. Once again, we first control the term error due to estimating the mean $|\bar{y}_n - \bar{f}| = |\frac{1}{n}\sum_{i=1}^n (y_i - \bar{f})|$. Since $|f^*(x_i)|$ is at most B and w_i is standard Gaussian and independent, the random variable $y_i - \bar{f} = f^*(x_i) + w_i$ is sub-Gaussian with parameter $\sqrt{B^2+1}$. The samples are all i.i.d., so that by standard sub-Gaussian tail bounds, we have

$$\mathbb{P}[|\bar{y}_n - \bar{f}| > t] \leq 2\exp(-\frac{nt^2}{2(B^2+1)}).$$

Setting $\mathcal{A}(\delta_n) = \{|\bar{y}_n - \bar{f}| \leq B\delta_n\}$, it is clear that

$$\mathbb{P}[\mathcal{A}(\delta_n)] \geq 1 - 2\exp(-\frac{n\delta_n^2}{4}).$$

For the remainder of the proof, we condition on the event $\mathcal{A}(\delta_n)$, in which case Equation (4.20) simplifies to

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \big|\frac{1}{n}\sum_{i=1}^n w_i \widehat{\Delta}(x_i)\big| + B\delta_n\|\widehat{\Delta}\|_n. \tag{4.37}$$

Here we have used the fact that $\left|\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}(x_i)\right| \leq \|\widehat{\Delta}\|_n$, by the Cauchy-Schwartz inequality.

Now we control the Gaussian complexity term $\big|\frac{1}{n}\sum_{i=1}^n w_i\widehat{\Delta}(x_i)\big|$. For any fixed subset S, define the random variable

$$\widehat{Z}_n(w,t;\mathcal{H}(S,2B)) := \sup_{\substack{\Delta\in\mathcal{H}(S,2B)\\ \|\Delta\|_n\le t}} \big|\frac{1}{n}\sum_{i=1}^n w_i\Delta(x_i)\big|. \tag{4.38}$$

We first bound this random variable for a fixed subset S of size $2s$, and then take the union bound over all $\binom{d}{2s}$ possible subsets.

Lemma 15. *Assume that the RKHS $\mathcal{H}$ has eigenvalues $(\mu_k)_{k=1}^\infty$ that satisfy $\mu_k \simeq k^{-2\alpha}$ and eigenfunctions such that $\|\phi_k\|_\infty \le C$. Then we have*

$$\mathbb{P}\big[\exists t>0 \textit{ such that } \widehat{Z}_n(w,t;\mathcal{H}(S,2B)) \ge 16B\,C\sqrt{\frac{s^{1/\alpha}\log s}{n}} + 3t\delta_n\big] \le c_1\exp(-9n\delta_n^2). \tag{4.39}$$

The proof of Lemma 15 is provided Appendix F. Returning to inequality (4.37), we note that by definition,

$$\frac{2}{n}|\sum_{i=1}^n w_i\widehat{\Delta}(x_i)| \le \max_{|S|=2s}\widehat{Z}_n(w,\|\widehat{\Delta}\|_n;\mathcal{H}(S,2B)).$$

Lemma 15 combined with the union bound implies that

$$\max_{|S|=2s}\widehat{Z}_n(w,\|\widehat{\Delta}\|_n;\mathcal{H}(S,2B)) \le 16B\,C\sqrt{\frac{s^{1/\alpha}\log s}{n}} + 3\delta_n\|\widehat{\Delta}\|_n$$

with probability at least $1-c_1\binom{d}{2s}\ \exp(-3n\delta_n^2)$. Our choice (4.17) of δ_n ensures that this probability is at least $1-c_1\exp(-c_2n\delta_n^2)$. Combined with the basic inequality (4.37), we conclude that

$$\|\widehat{\Delta}\|_n^2 \le 32B\,C\sqrt{\frac{s^{1/\alpha}\log s}{n}} + 7B\delta_n\,\|\widehat{\Delta}\|_n \tag{4.40}$$

with probability $1-c_1\exp(-c_2n\delta_n^2)$.

By definition (4.17) of δ_n, the bound (4.40) implies that $\|\widehat{\Delta}\|_n = \mathcal{O}(\delta_n)$ with high probability. In order to translate this claim into a bound on $\|\widehat{\Delta}\|_2$, we require the following result:

Lemma 16. *There exist universal constants (c, c_1, c_2) such that for all $t \ge c\delta_n$, we have*

$$\frac{\|g\|_2}{2} \ \le\ \|g\|_n \ \le\ \frac{3}{2}\|g\|_2 \qquad \textit{for all } g\in\mathcal{H}(S,2B) \textit{ with } \|g\|_2\ge t \tag{4.41}$$

with probability at least $1 - c_1 \exp(-c_2 n t^2)$.

Proof. The bound (4.41) follows by applying Lemma 7 in Appendix A with $\mathcal{G} = \mathcal{H}(S, 2B)$ and $b = 2B$. The critical radius from (35) in Raskutti et al. [69] needs to satisfy the relation $\mathcal{Q}_{w,n}(\epsilon_n; \mathcal{H}(S, 2B)) \leq \frac{\epsilon_n^2}{40}$. From Lemma 11 in Raskutti et al. [69], the choice $\epsilon_n^2 = 320B\,C\sqrt{\frac{s^{1/\alpha}\log s}{n}}$ satisfies this relation. By definition (4.17) of δ_n, we have $\delta_n \geq c\epsilon_n$ for some universal constant c, which completes the proof. □

This lemma implies that with probability at least $1 - c_1 \exp(-c_2 B n \delta_n^2)$, we have $\|\widehat{\Delta}\|_2 \leq 2\|\widehat{\Delta}\|_n + C\delta_n$. Combined with our earlier upper bound on $\|\widehat{\Delta}\|_n$, this completes the proof of Theorem 8.

4.5 Discussion

In this chapter, we have studied estimation in the class of sparse additive models in which each univariate function lies within a reproducing kernel Hilbert space. In conjunction, Theorems 6 and 7 provide a precise characterization of the minimax-optimal rates for estimating f^* in the $L^2(\mathbb{P})$-norm for various kernel classes with bounded univariate functions. These classes include finite-rank kernels (with logarithmic metric entropy), as well as kernels with polynomially decaying eigenvalues (and hence polynomial metric entropy). In order to establish achievable rates, we analyzed a simple M-estimator based on regularizing the least-squares loss with two kinds of ℓ_1-based norms, one defined by the univariate Hilbert norm and the other by the univariate empirical norm. On the other hand, we obtained our lower bounds by a combination of approximation-theoretic and information-theoretic techniques.

An important feature of our analysis is we assume only that each univariate function is bounded, but do not assume that the multivariate function class is bounded. As discussed in Section 4.3.5, imposing a global boundedness condition in the high-dimensional setting can lead to a substantially smaller function classes; for instance, for Sobolev classes and sparsity $s = \Omega(\sqrt{n})$, Theorem 8 shows that it is possible to obtain much faster rates than the optimal rates for the class of sparse additive models with univariate functions bounded. Theorem 8 in our chapter shows that the rates obtained under global boundedness conditions are not minimax optimal for Sobolev spaces in the regime $s = \Omega(\sqrt{n})$.

There are a number of ways in which this work could be extended. Our work considered only a hard sparsity model, in which at most s co-ordinate functions were non-zero, whereas it could be realistic to use a "soft" sparsity model involving ℓ_q-norms. Some recent work by [81] has studied some extensions of this type. In addition, the analysis here was based on assuming independence of the covariates x_j, $j = 1, 2, \ldots d$; it would be interesting to investigate the case when the random variables are endowed with some correlation structure. One might expect some changes in the optimal rates, particularly if many of the variables are strongly dependent. Finally, this work considered only the function class consisting of sums of co-ordinate functions, whereas a natural extension would be to consider nested non-parametric classes formed of sums over hierarchies of subsets of variables.

Bibliography

[1] R. Adamczak, A. Litvak, N. Tomczak-Jaegermann, and A. Pajor. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. Technical report, University of Alberta, 2009.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 1971.

[3] M. Akcakaya and V. Tarokh. Shannon theoretic limits on noisy compressive sampling. Technical Report arXiv:cs.IT:0711.0366, Harvard University, November 2007.

[4] 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.

[5] K. S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

[6] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[7] F. Bach. Consistency of the Group Lasso and Multiple Kernel Learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

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

[9] P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 2005. To appear.

[10] P. Bickel, Y. Ritov, and A. Tsybakov. Aggregation for gaussian regression. *Annals of Statistics*, 34, 2007.

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

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

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

[14] M. S. Birman and M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Math. USSR-Sbornik*, 2(3):295–317, 1967.

[15] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[16] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, (37):373–384, 1995.

[17] V. V. Buldygin and Y. V. Kozachenko. *Metric characterization of random variables and random processes*. American Mathematical Society, Providence, RI, 2000.

[18] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, pages 169–194, 2007.

[19] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info Theory*, 51(12):4203–4215, December 2005.

[20] 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.

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

[22] B. Carl and H. Triebel. Inequalities between eigenvalues, entropy numbers and related quantities of compact operators in banach spaces. *Annals of Mathematics*, 251:129–133, 1980.

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

[24] F. Chung and L. Lu. Concentration inequalities and martingale inequalities. *Internet Mathematics*, 3:79–127, 2006.

[25] 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.

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

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

[28] D. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, April 2006.

[29] D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Info Theory*, 47(7):2845–2862, 2001.

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

[31] U. Einmahl and D. M. Mason. Some universal results on the behavior of the increments of partial sums. *Annals of Probability*, 24:1388–1407, 1996.

[32] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Info Theory*, 48(9):2558–2567, September 2002.

[33] A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Trans. Info Theory*, 49(6):1579–1581, 2003.

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

[35] R. M. Gray. Toeplitz and Circulant Matrices: A Review. Technical report, Stanford University, Information Systems Laboratory, 1990.

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

[37] C. Gu. *Smoothing spline ANOVA models.* Springer Series in Statistics. Springer, New York, NY, 2002.

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

[39] O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Subspaces and orthogonal decompositions generated by bounded orthogonal systems. *Journal of Positivity*, 11(2):269–283, 2007.

[40] O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Majorizing measures and proportional subsets of bounded orthonormal systems. *Journal of Rev. Mat. Iberoam*, 24(3):1075–1095, 2008.

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

[42] 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.

[43] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

[44] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. Technical report, University of Wisconsin-Madison, 2010.

[45] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

[46] A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* Springer, 2009.

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

[48] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Jour. Math. Anal. Appl.*, 33:82–95, 1971.

[49] V. Koltchinskii and M. Yuan. Sparse Recovery in Large Ensembles of Kernel Machines. In *Proceedings of COLT*, 2008.

[50] V. Koltchinskii and M. Yuan. Sparsity in Multiple Kernel Learning. *Annals of Statistics*, 38:3660–3695, 2010.

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

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

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

[54] Y. Lin and H. H. Zhang. Component Selection and Smoothing in Multivariate Nonparametric Regression. *Annals of Statistics*, 34:2272–2297, 2006.

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

[56] P. Massart. About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28(2):863–884, 2000.

[57] P. Massart. *Concentration Inequalties and Model Selection*. Ecole d'Eté de Probabilités, Saint-Flour. Springer, New York, 2003.

[58] L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional Additive Modeling. *Annals of Statistics*, 37:3779–3821, 2009.

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

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

[61] S. Mendelson. Geometric parameters of kernel machines. In *Proceedings of COLT*, pages 29–43, 2002.

[62] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Journal of Constr. Approx.*, 28(3):277–289, 2008.

[63] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:415–446, 1909.

[64] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Proceedings of NIPS*, December 2009.

[65] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, UK, 1989.

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

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

[68] G. Raskutti, M. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57:3841–3863, 2011.

[69] G. Raskutti, M. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13:389–427, 2012.

[70] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. Technical report, U. C. Berkeley, October 2009. Posted as http://arxiv.org/abs/0910.2042.

[71] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: Sparse Additive Models. *Journal of the Royal Statistical Society, Series B*, 71(5):1009–1030, 2009.

[72] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Annals of Statistics*, 39:731–771, 2011.

[73] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, September 1978.

[74] Justin Romberg. Compressive sensing by random convolution. *SIAM Journal of Imaging Science*, 2(4):1098–1128, 2009.

[75] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure and Appl. Math.*, 61(8):1025–1045, 2008.

[76] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, Harlow, UK, 1988.

[77] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

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

[79] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–468, March 1978.

[80] C. J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 13(2):689–705, 1985.

[81] T. Suzuki and M. Sugiyama. Fast Learning Rate of Multiple Kernel Learning: Trade-off Between Sparsity and Smoothness. In *AISTATS Conference*, 2012.

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

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

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

[85] S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[86] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes.* Springer-Verlag, New York, NY, 1996.

[87] V.Q. Vu, P. Ravikumar, T. Naselaris, K. Kay, J. Gallant, and B. Yu. Encoding and decoding v1 fmri responses to natural images with sparse nonparametric models. *Annals of Applied Statistics*, 5(2B):1159–1182, 2011.

[88] G. Wahba. *Spline models for observational data.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PN, 1990.

[89] 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.

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

[91] W. Wang, M. J. Wainwright, and K. Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. Technical Report arXiv:0806.0604, UC Berkeley, June 2008. Presented at ISIT 2008, Toronto, Canada.

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

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

[94] M. Yuan. Nonnegative Garrote Component Selection in Functional ANOVA Models. In *Conference on Artificial Intelligence and Statistics*, pages 660–666, 2007.

[95] 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.

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

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

[98] S. Zhou. Restricted eigenvalue conditions on subgaussian random matrices. Technical report, Department of Mathematics, ETH Zürich, December 2009.