THE UNIVERSITY OF CHICAGO

A STATISTICAL LEARNING THEORY FOR MODELS WITH HIGH COMPLEXITY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

LIJIA ZHOU

CHICAGO, ILLINOIS

JUNE 2023

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

When I first started my undergraduate study at UChicago, I could not have imagined that I would be writing this thesis today. It has been an incredible journey, and I am deeply grateful for the support and guidance from many people.

First and foremost, I want to thank my advisor, Nati Srebro, for his encouragement, insights, and mentorship. Even as a college senior, Nati had kindly invited me to learn about his research. I found my interest in deep learning theory in my first year of graduate school, and Nati's computational and statistical learning theory course was a great introduction. Throughout my Ph.D., I have always felt that the problems Nati gave me were exciting and meaningful. I can't thank Nati enough for pointing me to the right research questions. His insights helped me find the natural answer to research problems I didn't know I could solve. I appreciate that Nati always takes the time to explain concepts that I am unfamiliar with, and his passion, knowledge, and incredible commitment to details constantly inspire me.

I also want to thank my thesis committee members, Chao Gao and Rina Foygel Barber. I took many statistics courses from you throughout my graduate and undergraduate studies, and both of you have been great teachers and researchers that I genuinely admire. Your comments and suggestions on the draft of this thesis mean a lot to me.

This thesis would not have been possible without the help of my collaborators: Frederic Koehler, Danica Sutherland, Pragya Sur, Zhen Dai, Jamie Simon, and Gal Vardi. It has been an extraordinary and humbling experience working with you all. I especially want to thank Freddie for teaching me many things essential to my research. It was one of the most exciting moments in my Ph.D. when Freddie told me about the Gaussian minimax theorem. The tool was perfect for proving the speculative bound from my first paper (which I had been stuck on for about a year). I probably would still be stuck without his help. I also want to thank Dani for her continuous support. I am thankful for our weekly meetings during the pandemic – so many problems were solved by just talking to you. I have to thank Zhen for answering my endless questions on linear algebra.

I would come up with some random conjecture on matrices and spend hours trying to prove it, and then Zhen would come along and show me a simple counterexample to save me from my misery.

I want to express my gratitude to the graduate students, postdocs, and alumni in Nati's group: Kavya Ravichandran, Owen Melia, Anmol Kabra, Gene Li, Kumar Kshitij Patel, Omar Montasser, Donya Saless, Nirmit Joshi, Sam Buchanan, Lingxiao Wang, Akilesh Tangella, Blake Woodworth, Xiaoxia Wu, Suriya Gunasekar, Pritish Kamath, and Brian Bullins. It has been a privilege to get to know the diverse research areas in the group. Attending conferences and workshops with you all in Berkeley, New York, and New Orleans were memorable experiences for my Ph.D. I also want to thank those participating in the Machine Learning and Optimization (MLO) reading group. The reading group has been beneficial for me to learn about new research directions and broaden my scope of knowledge.

There are many more faculty members from both UChicago and TTIC that have helped me. I want to thank Professor Per Mykland, Lek-Heng Lim, and Yali Amit for allowing me to take reading and research courses with them before I settled on statistical learning theory; Professor Steve Lalley for being my first-year advisor and the comprehensive lecture notes on probability theory; my undergraduate research advisor Michael Stein for encouraging me to do a Ph.D. in statistics from the first place and introducing me to the world of spatiotemporal modeling. I have significantly benefited from the courses offered at UChicago and TTIC. I want to thank Professor Alisa Knizel, Hongyuan Mei, Yibi Huang, Yi Sun, Yuehaw Khoo, and Stephen Stigler for the interesting courses I was TA for. I also want to thank the administrative staff for making the courses and teaching run smoothly every day.

In addition, I want to thank my roommate Haochen Wang, my fellow stats and CAM Ph.D. students (especially Yi Wang), and the people who generously helped me in my industry job search. I am glad to make friends with Yanfei Zhou, Deqing Fu, Marshall Dong, and many others while they were master's students in the stats department. I had plenty of help in teaching and statistical consulting. I have met many interesting people through my high school friends in Chicago and

New York, my first year at UCLA, and my internship at Citadel. There is no way that I can adequately express my gratitude to everyone who has supported me in this short acknowledgment section.

Finally, I am grateful to Qi for the past seven years. Your love and optimism have brought so much joy into my life, and I am very excited for the next chapter of our lives together. I wish to express my deepest gratitude to my parents, You Zhou and Jieling Li. Thank you both for giving me a comfortable life and supporting me to study in the U.S. I couldn't be who I am today without your dedication to my education. Your unconditional love and support have given me the courage to pursue goals that seemed impossible to reach.

ABSTRACT

Understanding why high-dimensional estimators can generalize beyond finite training samples is a fundamental problem in statistical learning theory. The traditional intuition, as suggested by Occam's razor, is that models with low complexity tend to generalize better. We can often find simple models that explain the training data well if the high-dimensional data distribution has some hidden low-dimensional structure (for example, sparse linear regression and low-rank matrix recovery). However, contrary to our traditional intuition, complex models which interpolate noisy training labels can also enjoy good generalization in some settings. This phenomenon, which we call "interpolation learning," has significantly challenged our theoretical foundation of statistical learning. In this thesis, we present a novel Moreau envelope generalization theory to establish the concentration of measure in high dimensions. Since our result can precisely quantify the role of model complexity in generalization error, we can establish strong consistency results even though the norm of the high-dimensional interpolants that we consider diverges. In addition to proving sharp non-asymptotic bounds for interpolants in various contexts, we also recover versions of classical results from the compressed sensing and high-dimensional statistics literature. Applications of our theory include kernel ridge regression, max-margin classification, phase retrieval, matrix sensing, and some simple neural networks.

CHAPTER 1

INTRODUCTION

High-dimensional models are ubiquitous in modern statistics and machine learning applications. In atmospheric and climate sciences, measurements can be taken on a fine grid of extended spatial fields, leading to a massive collection of complex features in relatively short time horizons. In financial exchanges, millions of different instruments are traded every single day. Similar situations can be found in other areas, such as genetics and neuroscience, where the large number of predictors requires a model to have many more parameters than the number of training samples. When the number of features is small, embedding them in a higher dimensional space can allow us to learn non-linear relationships. We can understand classical machine learning algorithms such as boosting and kernel regression this way. Nowadays, machine learning models are growing increasingly high-dimensional. A state-of-the-art neural network architecture for speech recognition, natural language processing, or computer vision can contain more than billions of parameters.

When we fit a high-dimensional model on samples, the goal is usually to reveal some truth about the population distribution or to be able to make good predictions on future unseen data points. Statistical learning theory provides a rigorous mathematical framework to reason about generalization. In the statistical learning framework, a model is a function that maps an instance to some prediction, and we specify a class of models to choose from (also known as concept/hypothesis class). In most of this thesis, we will consider hypothesis classes parameterized by a high dimensional vector. Also, we have a loss function that measures the quality of the model output on a single instance. Given *independent and identically distributed* samples from some unknown distribution, we hope to find a model in the hypothesis class that is nearly optimal in terms of the expected loss with respect to the unknown population. When we have the prior knowledge that the population distribution comes from some fixed parametric family and we wish to estimate the parameters, then we can pick the loss to be the log-likelihood function, and a generalization guarantee will also be a bound on estimation error (in terms of KL divergence). More generally,

if we only care about average prediction error on a new sample, then only very mild assumptions on the data distribution are required. We just need to specify how to measure the accuracy of a prediction.

In high-dimensional settings, it is usually possible to perfectly minimize the loss function on training samples. However, if a model fits too closely to noisy training data, it can memorize noise patterns that do not exist in the population distribution. As a result, overfitting will hurt the generalization performance at test time. To avoid overfitting, classical learning theory suggests that we should constrain the complexity of the learned model. We have a rich mathematical understanding of why regularization helps. There is a long history of literature showing that the class of low-complexity models enjoys uniform convergence (also known as "concentration of measure" or "uniform law of large numbers"). Roughly speaking, uniform convergence refers to the phenomenon that *uniformly* over a class of models, the training error and population error will be close to each other. As a result, minimizing the training error in a constrained class of models is approximately the same as minimizing the test error. In addition, the complexity of a class of models can be precisely measured by quantities such as VC-subgraph dimension [52], fat-shattering dimension [1], covering numbers [52] and Rademacher complexity [3]. By balancing the approximation error from choosing a low-complexity class with the generalization error from uniform convergence, we can find the optimal amount of regularization and derive some quantitative generalization bound for regularized estimators without making any parametric assumptions on the population distribution.

On the other hand, the empirical success of deep neural networks has remarkably escaped the curse of overfitting. It is common to train a neural network without explicit regularization so that it can exactly interpolate the training data. An interpolating over-parameterized model will generally have very high model complexity. Yet, in practice, it still generalizes decently well [11, 48, 76]. One may argue that there is very little noise in applications like image classification, so interpolation is sensible. However, the thought-provoking experiment from Belkin et al. [10]

shows that interpolating classifiers can generalize well even if we randomly flip some percentages of the training labels. Though the training error stays at zero across different noise levels, the test error of interpolating predictors will miraculously adapt and remain close to the optimal error as we vary how much noise we add. The observation that complex high-dimensional models can also generalize to the population raises a fundamental question to the machine learning theory: can the concentration of measure capture learning in these modern settings? What are the proper theoretical tools to understand the generalization of high-complexity models that are becoming increasingly popular in practice?

In this thesis, we attempt to bridge the gap between our traditional understanding of learning with the mysterious success of high-dimensional interpolants. Building on the intuitions from classical learning theory, we develop a more refined uniform convergence technique to establish generalization in high-dimensional settings. In addition to showing that it is powerful enough to explain interpolation learning, we demonstrate our theoretical framework's versatility by applying it to analyze regularized estimators in various interesting settings, such as sparse linear regression, high-dimensional generalized linear model, matrix sensing, and phase retrieval.

In Chapter 2, we begin the study of interpolants by analyzing the test error of kernel ridge regression (KRR). Using the well-known omniscient risk estimate, we control the test error of kernel ridge interpolant by the error of the optimally balanced estimators and a multiplicative factor that measures the cost of interpolation. Direct analysis of the cost of interpolation reveals the necessary and sufficient condition for which the kernel ridge interpolant is asymptotically as good as the optimally balanced estimators. We then apply this result to establish benign overfitting in kernel ridge regression and recover the multiple descent phenomenon for inner-product kernels in a polynomial regime. In addition to directly analyzing the cost of interpolation, we provide another perspective to understand benign overfitting based on uniform convergence in Section 2.4. We show that the size of the RKHS norm is sufficient to explain generalization.

In Chapter 3, we consider the problem of linear regression with Gaussian features. We use the

Gaussian Minimax Theorem (GMT) to prove the same uniform convergence guarantee in Chapter 2 known as the optimistic rate for *any* linear predictors. Doing so allows us to move beyond ridge regression and analyze the minimal norm interpolant for any norm. Choosing the ℓ_1 norm gives us a novel benign overfitting result for basis pursuit. In addition to understanding interpolants, we apply our theory to recover some classical statistical guarantees for ridge and LASSO regression under random designs. We also provide a high-probability version of the precise error for Ordinary Least Squares (OLS) with isotropic features in the proportional regime. Finally, we apply our theory to analyze the minimum nuclear norm estimator in matrix sensing in Section 3.6.

In Chapter 4, we introduce the Moreau envelope generalization theory, which generalizes the result in linear regression to any generalized linear loss. If we have a globally Lipschitz loss, then our Moreau envelope theory can be translated to the classical Rademacher complexity bound (but without the loose multiplicative factor of 2). Examples of a globally Lipschitz loss include the absolute loss, the Huber loss, and the log-likelihood function for Logistic regression and Binomial GLM. More generally, even if the loss function is neither Lipschitz nor convex, we show that as long as the loss's square root is Lipschitz, we can establish a uniform convergence guarantee with an optimistic rate. As an example, we consider the problem of linear classification with the squared hinge loss, which corresponds to soft and hard margin Support Vector Machines (SVM), and show that the same argument as in Chapter 3 can be used to establish benign overfitting in linear classification. In Section 4.2, we introduce a general technique to compute the norm of the minimal norm interpolant and provide a lower bound that suggests our uniform convergence bound should always be asymptotically tight. Another application of this result is ReLU regression and phase retrieval, in which the loss is not even differentiable. Moreover, we extend our result in Chapter 4 to allow the loss function to have some non-linear component. This flexibility will enable us to establish a norm-based generalization bound for two-layer neural networks with weight-sharing in the first layer.

Finally, we discuss some empirical and theoretical evidence of universality and conjecture that

it should be possible to relax the Gaussian feature assumption in our Moreau envelope generalization theory. We numerically check the validity of our theory for discrete, asymmetric, and heavy-tailed feature distributions. We also briefly discuss some theoretical advances based on random matrix theory and the Lindeberg principle. With the broad applicability of uniform convergence in understanding high dimensional interpolants and regularized estimator, our results suggest a promising research direction for the theoretical foundation of deep learning. Moreover, our results demonstrate that the concentration of measure phenomenon remains a fundamental principle in statistical learning.

CHAPTER 2

KERNEL RIDGE REGRESSION

Kernel Ridge Regression (KRR) is a class of flexible and non-linear machine learning models. A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric and positive semi-definite function:

(A) $\forall x, x' \in \mathcal{X},\ K(x, x') = K(x', x)$

(B) $\forall n \in \mathbb{N},\ x_1, ..., x_n \in \mathcal{X},\ c_1, ..., c_n \in \mathbb{R}$, it holds that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0. \tag{2.1}$$

In the above, $\mathcal{X}$ is an abstract input space. For example, $\mathcal{X}$ can be taken to be $\mathbb{R}^d$ and commonly used kernels include: i) polynomial kernel $K(x, x') = \left(1 + \gamma \langle x, x' \rangle\right)^l$, ii) Laplacian kernel $K(x, x') = \exp\left(-\gamma \|x - x'\|_1\right)$, and iii) Gaussian kernel $K(x, x') = \exp\left(-\gamma \|x - x'\|_2^2\right)$. But $\mathcal{X}$ can also be a low-dimensional manifold, the space of probability distributions, the space of strings, or the set of nodes in a graph. Kernels can be defined on these non-Euclidean spaces too [9, 31, 34].

Given $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ sampled independently from some unknown joint distribution $\mathcal{D}$, we can define a predictor $f_\delta : \mathcal{X} \to \mathbb{R}$ for any ridge parameter $\delta \geq 0$ by

$$f_\delta(x) = K(D_n, x)^T \left(K(D_n, D_n) + \delta I_n\right)^{-1} Y \tag{2.2}$$

where $K(D_n, x) \in \mathbb{R}^n, K(D_n, D_n) \in \mathbb{R}^{n \times n}, Y \in \mathbb{R}^n$ are given by

$$[K(D_n, x)]_i = K(x_i, x), \quad [K(D_n, D_n)]_{i,j} = K(x_i, x_j), \quad [Y]_i = y_i. \tag{2.3}$$

We will assume that $\mathcal{D}$ is given by $x \sim p$ and $y = f^*(x) + \xi$ for some distribution p over $\mathcal{X}$, $f^* : \mathcal{X} \to \mathbb{R}$ and $\xi \sim \mathcal{N}(0, \sigma^2)$. It is then natural to find the expectation of the test error of f_δ defined as $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(f_\delta(x) - y)^2]$ with respect to the randomness of training data.

Though the setting for KRR is very generic, we can understand it as ordinary ridge regression. By Mercer's theorem [45], the kernel admits the decomposition

$$K(x, x') = \sum_i \lambda_i \phi_i(x) \phi(x') \tag{2.4}$$

where $\phi_i : \mathcal{X} \to \mathbb{R}$ satisfies $\mathbb{E}_{x \sim p}[\phi_i(x)\phi_j(x)] = 1$ if $i = j$ and 0 otherwise. For example, if $\mathcal{X} = \{x_1, ..., x_M\}$ has finite cardinality M, then (2.4) can be found by the spectral decomposition of the matrix $K(\mathcal{X}, \mathcal{X}) \in \mathbb{R}^{M \times M}$ given by $[K(\mathcal{X}, \mathcal{X})]_{i,j} = K(x_i, x_j)$. When p is the uniform distribution over the sphere in $\mathbb{R}^d$ or the boolean hypercube $\{-1, 1\}^d$, then $\{\phi_i\}$ can be taken to be the spherical harmonics or the Fourier-Walsh (parity) basis. In the case that K is the Gaussian kernel or polynomial kernel, the $\{\lambda_i\}$ has closed-form expression in terms of the modified Bessel function or the Gamma function [46].

Therefore, instead of viewing the feature x as an element of $\mathcal{X}$, we can consider the potentially infinite-dimensional real-valued vector $\psi(x) = (\sqrt{\lambda_1}\phi_1(x), \sqrt{\lambda_2}\phi_2(x), ...)$ and denote the design matrix $\Psi = [\psi(x_1), \psi(x_2), ...]^T$. Then we can understand (2.2) as

$$f_\delta(x) = \psi(x)^T \Psi^T (\Psi \Psi^T + \delta I_n)^{-1} Y = \langle w_\delta, \psi(x) \rangle \tag{2.5}$$

where $w_\delta = \Psi^T (\Psi \Psi^T + \delta I_n)^{-1} Y$ is simply the ridge regression estimate with respect to the data set (Ψ, Y). As the eigenfunctions $\{\phi_i\}$ forms a complete basis, we can expand

$$f^*(x) = \sum_i v_i \phi_i(x) = \sum_i \frac{v_i}{\sqrt{\lambda_i}} \psi_i(x) = \langle w^*, \psi(x) \rangle. \tag{2.6}$$

where we denote $w_i^* = v_i / \sqrt{\lambda_i}$. In general, the RKHS norm of a function f defined by $f(x) = \langle w, \psi(x) \rangle$ is given by $\|f\|_{\mathcal{K}} = \|w\|_2$.

2.1 Omniscient Risk Estimator

Many prior works (for example, see Canatar et al. [17], Hastie et al. [28], Jacot et al. [30], Loureiro et al. [39], Mel and Ganguli [42], Richards et al. [55], Simon et al. [60], Wu and Xu [75]) have shown that in high dimensional settings, it holds that

$$w_{\delta,i} \approx \frac{\lambda_i}{\lambda_i + \kappa_\delta} w_i^* \tag{2.7}$$

where κ_δ is the effective regularization defined in (2.8) below. Rigorous version of (2.7) can be proved using random matrix theory, while non-rigorous proofs based on methods in statistical physics (such as replica and cavity methods) also exist. We will adopt the eigenlearning framework [60] in this section. We note that although the proofs in this section are based on an non-rigorous result, all the theoretical proofs starting from Section 3 will be completely rigorous. As we will see, the simplicity of the eigenlearning equations can provide a good first step for an intuitive understanding of interpolation in high dimensions.

In the eigenlearning framework, the effective regularization κ_δ is defined by

$$\sum_i \frac{\lambda_i}{\lambda_i + \kappa_\delta} + \frac{\delta}{\kappa_\delta} = n. \tag{2.8}$$

Using κ_δ, we can define

$$\mathcal{L}_{i,\delta} = \frac{\lambda_i}{\lambda_i + \kappa_\delta}, \quad \mathcal{E}_\delta = \frac{n}{n - \sum_i \mathcal{L}_{i,\delta}^2}. \tag{2.9}$$

Then the population and training error of f_δ are given by

$$\mathcal{E}(f_\delta) = \mathcal{E}_\delta \left(\sum_i (1 - \mathcal{L}_{i,\delta})^2 v_i^2 + \sigma^2 \right) \quad \text{and} \quad \mathcal{E}_{\text{tr}}(f_\delta) = \frac{\delta^2}{n^2 \kappa_\delta^2} \mathcal{E}(f_\delta). \tag{2.10}$$

In the above, the v_i are defined in (2.6) and σ^2 is the variance of ξ in the population distribution $\mathcal{D}$, which is also the Bayes error in the problem. Finally, the expected RKHS norm is given by the

following formula:

$$\mathbb{E}\|f_\delta\|_{\mathcal{K}}^2 = \frac{\mathcal{E}(f_\delta)}{n}\sum_i \frac{\lambda_i}{(\lambda_i+\kappa_\delta)^2} + \sum_i \frac{\lambda_i v_i^2}{(\lambda_i+\kappa_\delta)^2}. \tag{2.11}$$

We see from above that equation (2.7) can provide a complete description of the stochastic system defined by f_δ for any $\delta \geq 0$.

2.2 Cost of Interpolation

In this section, we want to understand the test error of the kernel ridgeless regression $\mathcal{E}(f_0)$. In particular, how does it compare to the optimally balanced estimator?

Theorem 1. *Let $\delta^* = \arg\min_{\delta\geq 0}\mathcal{E}(f_\delta)$ be the optimal ridge parameter. It holds that*

$$\mathcal{E}_0\sigma^2 \leq \mathcal{E}(f_0) \leq \mathcal{E}_0\mathcal{E}(f_{\delta^*}). \tag{2.12}$$

Proof. The first inequality is straightforward from the definition. To prove the second, observe that

$$\begin{aligned}
\mathcal{E}(f_{\delta^*}) &= \inf_{\delta\geq 0}\ \mathcal{E}_\delta\left(\sum_i (1-\mathcal{L}_{i,\delta})^2 v_i^2 + \sigma^2\right) \\
&\geq \inf_{\delta\geq 0}\ \sum_i (1-\mathcal{L}_{i,\delta})^2 v_i^2 + \sigma^2 \\
&= \sum_i (1-\mathcal{L}_{i,0})^2 v_i^2 + \sigma^2
\end{aligned}$$

where we use the fact that $(1-\mathcal{L}_{i,\delta})^2$ decreases as κ_δ decreases, and κ_δ decreases as δ decreases. The proof concludes by observing $\sum_i (1-\mathcal{L}_{i,0})^2 v_i^2 + \sigma^2 = \mathcal{E}(f_0)/\mathcal{E}_0$. □

In many settings of interest, we have $\mathcal{E}(f_{\delta^*}) \to \sigma^2$ and so the factor $\mathcal{E}_0$ is actually tight and $\mathcal{E}(f_0) \approx \mathcal{E}_0\mathcal{E}(f_{\delta^*})$. Note that it is not always true that $\mathcal{E}(f_0) \to \sigma^2$ and so $\mathcal{E}_0$ measures the "cost of interpolation" relative to the optimal balance. Next, we will analyze the multiplicative factor $\mathcal{E}_0$.

For convenience, we will sort the eigenvalues $\lambda_1 \geq \lambda_2 \geq ...$ and denote the effective ranks [7] of the tail of the eigenvalues as

$$r_k = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \qquad \text{and} \qquad R_k := \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}. \tag{2.13}$$

Proposition 2. *For any k such that $k < n$ and $n < R_k$, it holds that*

$$\mathcal{E}_0 \leq \left(1 - \frac{k}{n}\right)^{-2} \left(1 - \frac{n}{R_k}\right)^{-1}. \tag{2.14}$$

All remaining proofs in this chapter can be found in Appendix C. Proposition 2 suggests if there exists a sequence $k = o(n)$ such that $R_k = \omega(n)$, then $\mathcal{E}_0 \to 1$ and so the ridgeless interpolant is asymptotically as good as the optimally ridge estimator. This is also one of the key conditions for benign overfitting in linear regression (Bartlett et al. [7]). Next, let's try to get lower bound on $\mathcal{E}_0$. Since $\mathcal{E}_0 = \left(1 - \frac{1}{n}\sum_i \mathcal{L}_{i,0}^2\right)^{-1}$, it suffices to lower bound $\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2$.

Proposition 3. *Fix any $b > 0$. If there exists $k < n$ such that $n \leq k + br_k$, then let k be the first such integer. Otherwise, pick $k = n$. It holds that*

$$\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \geq \max\left\{\frac{1}{(b+1)^2}\left(1 - \frac{k}{n}\right)^2 \frac{n}{R_k}, \left(\frac{b}{b+1}\right)^2 \frac{k}{n}\right\}. \tag{2.15}$$

Our lower bound is also heavily inspired by the lower bound of Theorem 4 in Bartlett et al. [7], though our proof technique is completely different and much simpler since we rely on the eigenlearning equations. In order for $\mathcal{E}_0$ to converge to 1, it must be the case that $\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \to 0$. Using the lower bound above, we identify the necessary and sufficient condition.

Corollary 4. *For any $n \in \mathbb{N}$, let k_n be the first integer $k < n$ such that $n \leq k + r_k$. Then $\mathcal{E}_0 \to 1$ if and only if*

$$\lim_{n\to\infty} \frac{k_n}{n} = 0 \quad \textit{and} \quad \lim_{n\to\infty} \frac{n}{R_{k_n}} = 0. \tag{2.16}$$

Next, we provide some examples of eigenvalues $\{\lambda_i\}$ where the ridgeless estimator is asymptotically as good as the optimally-tuned ridge estimator.

Example 1 (Junk features from Zhou et al. [77])**.**

$$\lambda_i = \begin{cases} 1 & \text{if } i \le d_S \\ \frac{1}{d_J} & \text{if } d_S + 1 \le i \le d_S + d_J \\ 0 & \text{if } i > d_S + d_J. \end{cases}$$

In this case, it is routine to check $R_k = d_J$ by choosing $k = d_S$ and so letting $d_S = o(n)$ and $d_J = \omega(n)$, Proposition 2 shows that $\mathcal{E}_0 \to 1$.

Example 2 (Slow eigendecay from Bartlett et al. [7])**.**

$$\lambda_i = i^{-1} \log^{-\alpha} i \quad \text{for some } \alpha > 0.$$

In this case, we can estimate

$$\sum_{i>k} \lambda_i \ge \int_{k+1}^{\infty} \frac{1}{x \log^{\alpha} x}\, dx = \frac{1}{(\alpha-1)\log^{\alpha-1}(k+1)}$$
$$\sum_{i>k} \lambda_i^2 \le \frac{1}{k+1} \int_{k}^{\infty} \frac{1}{x \log^{2\alpha} x}\, dx = \frac{1}{(k+1)(2\alpha-1)\log^{2\alpha-1}(k)}$$

and so

$$R_k \ge \frac{(k+1)(2\alpha-1)\log^{2\alpha-1}(k)}{(\alpha-1)^2 \log^{2\alpha-2}(k+1)} = \Theta\left(k \log k\right).$$

Then choosing $k = \frac{n}{\sqrt{\log n}}$, we have $k = o(n)$ and $R_k = \omega(n)$ because

$$\frac{R_k}{n} = \Theta(\log^{1/2} n).$$

In the following examples, we show that (2.16) does not hold and so overfitting is not benign.

However, we still have tempered overfitting [40] in the sense that $\mathcal{E}(f_0)$ does not diverge as $n \to \infty$.

Example 3 (Isotropic features in the proportional regime)**.**

$$\lambda_i = \begin{cases} 1 & \textit{if } i \le d \\ 0 & \textit{otherwise} \end{cases} \qquad \textit{for} \quad d = \gamma n \quad \textit{and} \quad \gamma > 1.$$

In this case, it is easy to check that $r_k = d - k$ *and so* $k + r_k = d > n$ *and* $k_n = n$*. Therefore, the first condition in* (2.16) *cannot hold because* $k_n/n = 1$*. On the other hand, we also have* $R_k = d - k$ *and plugging in* $k = 0$ *to Proposition 2, we obtain*

$$\mathcal{E}_0 \le \left(1 - \frac{n}{d}\right)^{-1} = \frac{\gamma}{\gamma - 1}.$$

The above upper bound is tight when $f^* = 0$ *because* $\mathcal{E}(f_{\delta^*}) = \sigma^2$ *can be obtained with infinite regularization and it is well-known that in the proportional regime (for example, see Hastie et al. [28] and Zhou et al. [78]), we have*

$$\lim_{n \to \infty} \mathcal{E}(f_0) = \sigma^2 \frac{\gamma}{\gamma - 1}.$$

Example 4 (Power law decay from Mallinar et al. [40])**.**

$$\lambda_i = i^{-\alpha} \quad \textit{for some } \alpha > 1.$$

In this case, we can estimate

$$\frac{1}{(\alpha-1)(k+1)^{\alpha-1}} = \int_{k+1}^{\infty} x^{-\alpha}\, dx \le \sum_{i>k} \lambda_i \le \int_{k}^{\infty} x^{-\alpha}\, dx = \frac{1}{(\alpha-1)k^{\alpha-1}}$$
$$\frac{1}{(2\alpha-1)(k+1)^{2\alpha-1}} = \int_{k+1}^{\infty} x^{-2\alpha}\, dx \le \sum_{i>k} \lambda_i^2 \le \int_{k}^{\infty} x^{-2\alpha}\, dx = \frac{1}{(2\alpha-1)k^{2\alpha-1}}$$

and so

$$\left(\frac{k+1}{k}\right)\frac{1}{\alpha-1} \leq \frac{r_k}{k} \leq \left(\frac{k+1}{k}\right)^{\alpha-1}\frac{1}{\alpha-1}$$
$$\left(\frac{k}{k+1}\right)^{2(\alpha-1)}\frac{2\alpha-1}{(\alpha-1)^2} \leq \frac{R_k}{k} \leq \left(\frac{k+1}{k}\right)^{2\alpha-1}\frac{2\alpha-1}{(\alpha-1)^2}.$$

The above implies $R_k = \Theta(k)$*. Therefore, since* $k_n \leq n$*, it must be the case that* $n/R_{k_n} = \Omega(1)$*, which violates the second condition in* (2.16)*.*

In fact, it is shown that $\mathcal{E}_0 \to \alpha$ for the power law decay in Mallinar et al. [40]. Observe that $\alpha = \lim_{k\to\infty} 1 + \frac{1}{r_k/k}$. The next proposition shows that the relationship $\mathcal{E}_0 \lesssim 1 + \frac{1}{r_k/k}$ actually holds more generally. We first state the finite-sample result and then apply it to obtain the desired asymptotic version.

Proposition 5. *Suppose there exists* $m, M > 0$ *such that* $m \leq r_k/k \leq M$ *for* $k \geq \lfloor \frac{n}{1+M} \rfloor$*. Then it holds that*

$$\mathcal{E}_0 \leq 4\frac{\left(\frac{1}{1+M} - \frac{1}{n}\right)^{-1}}{m}. \tag{2.17}$$

If $\{\lambda_i\}$ *does not change with* n *and* $\lim_{k\to\infty} r_k/k = \alpha > 0$*, then*

$$\lim_{n\to\infty} \mathcal{E}_0 \leq 4\left(1 + \frac{1}{\alpha}\right). \tag{2.18}$$

An implication of Proposition 5 is that as long as $\lim_{k\to\infty} r_k/k > 0$, then overfitting is tempered. The next proposition shows that the converse is actually true as well: if $\lim_{k\to\infty} r_k/k = 0$, then $\mathcal{E}(f_0) \to \infty$ as $n \to \infty$.

Proposition 6. *For any* $k \geq n + r_k$*, it holds that*

$$\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \geq \frac{n}{k}\left(1 - \frac{r_k}{k-n}\right)^2. \tag{2.19}$$

Therefore, if $\{\lambda_i\}$ does not change with n and $\lim_{k\to\infty} r_k/k = 0$, then, then it holds that

$$\lim_{n\to\infty} \mathcal{E}_0 = \infty.$$

Example 5 (Exponential decay)**.**

$$\lambda_i = e^{-i}.$$

In this case, we can estimate

$$\sum_{i>k} \lambda_i \leq \int_k^\infty e^{-x}\, dx = e^{-k}$$

and so $r_k \leq e$ and $r_k/k \to 0$. Proposition 6 implies that overfitting is catastrophic in this case.

Finally, we can summarize the above results into the trichotomy theorem below in terms of the effective rank r_k. We note that the more general cases where $\{\lambda_i\}$ are allowed to change with n can be analyzed using the finite-sample results in Proposition 2, 3, 5, 6 and Corollary 4.

Theorem 7. *Suppose that $\{\lambda_i\}$ does not change with n and the optimally tuned ridge regression is consistent: $\mathcal{E}(f_{\delta^*}) \to \sigma^2$, then*

(i) if $\lim_{k\to\infty} r_k/k = \infty$, then overfitting is benign:

$$\lim_{n\to\infty} \mathcal{E}(f_0) = \sigma^2 \tag{2.20}$$

(ii) if $\lim_{k\to\infty} r_k/k \in (0, \infty)$, then overfitting is tempered:

$$\sigma^2 < \lim_{n\to\infty} \mathcal{E}(f_0) < \infty \tag{2.21}$$

(ii) if $\lim_{k\to\infty} r_k/k = 0$, then overfitting is catastrophic:

$$\lim_{n\to\infty} \mathcal{E}(f_0) = \infty. \tag{2.22}$$

2.3 Optimal Ridge

In this section, we give bounds on the test error $\mathcal{E}(f_{\delta^*})$ of the optimally-tuned ridge regression. We first state the general result and then discuss the applications to special cases.

Theorem 8. *Fix any $l \in \mathbb{N} \cup \{\infty\}$ and $k < n/2$. For any $k' < n$, it holds that*

$$\mathcal{E}(f_{\delta^*}) \le \frac{1+\epsilon}{1-\epsilon-\frac{k'}{n}} \left[\sigma^2 + \sum_{i>l} v_i^2 + \frac{\left(\sum_{i>k} \lambda_i\right)\left(\sum_{i\le l} \frac{v_i^2}{\lambda_i}\right)}{n-k} \right] \tag{2.23}$$

where ϵ is defined as

$$\epsilon = \left(\frac{2\sum_{i\le l} \frac{v_i^2}{\lambda_i}}{\sigma^2}\right)^{2/3} \left(\frac{\sum_{i>k'} \lambda_i^2}{n}\right)^{1/3}. \tag{2.24}$$

If we ignore the complicated multiplicative factor in (2.23), then the bound

$$\mathcal{E}(f_{\delta^*}) \le \inf_l \left(\sigma^2 + \sum_{i>l} v_i^2 \right) + \frac{\left(\sum_{i>k} \lambda_i\right)\left(\sum_{i\le l} \frac{v_i^2}{\lambda_i}\right)}{n-k}$$

can be interpreted using uniform convergence. The term $\sigma^2 + \sum_{i>l} v_i^2$ is the level of training error that we wish to achieve and $\sum_{i\le l} \frac{v_i^2}{\lambda_i}$ is the RKHS norm sufficient to achieve this error. To see this, we can simply pick $f = \sum_{i\le l} \frac{v_i}{\sqrt{\lambda_i}}\psi_i$. The $\frac{1}{n-k}\left(\sum_{i>k} \lambda_i\right)\left(\sum_{i\le l} \frac{v_i^2}{\lambda_i}\right)$ term is the generalization error one can usually obtain with Rademacher complexity. We use k here to handle situations where there are $o(n)$ diverging eigenvalues. In the case that $\|f^*\|_{\mathcal{K}} < \infty$, we can simply take $l = \infty$. When f^* has infinite RKHS norm, we can send $l \to \infty$ as $n \to \infty$ and consistency is still possible because $\sum_{i>l} v_i^2 \to 0$.

Next, we analyze the multiplicative factor $\frac{1+\epsilon}{1-\epsilon-\frac{k'}{n}}$ by choosing k' appropriately. As we will see in the examples below, there usually exists a choice of $k' = o(n)$ such that ϵ is small. From (2.24), it is clear that the precise finite-sample rate will depend on the spectrum through the $\sqrt{\sum_{i>k'} \lambda_i^2}$

term. If the spectrum has very fast decay, then we can pick k' to be small and the finite-sample rate can approach the parametric rate of $O(n^{-1})$. In the case when both the norm of the features $\sum_i \lambda_i$ and the RKHS norm $\|f^*\|_{\mathcal{K}}$ are bounded, then the choice of $k' = \Theta(\sqrt{n})$ will imply that ϵ is at most the order of $o(n^{-1/2})$.

Agnostic rate. Suppose that $\sum_i \lambda_i < \infty$, it must be the case that $\lambda_i = o(i^{-1})$ and

$$\sum_{i>k'} \lambda_i^2 = o\left(\sum_{i>k'} i^{-2}\right) = o\left(\frac{1}{k'}\right).$$

Assuming that σ^2 is constant and $\sum_i \frac{v_i^2}{\lambda_i}$ is finite, we can take $k' = \Theta(\sqrt{n})$ and obtain from (2.24)

$$\epsilon = o\left(\left(\frac{1}{n^{\frac{1}{2}+1}}\right)^{1/3}\right) = o(n^{-1/2}).$$

Choosing $l = \infty$ and $k = 0$, we see that $\mathcal{E}(f_{\delta^*}) \to \sigma^2$ at the agnostic rate of $O(n^{-1/2})$ as $n \to \infty$

$$\mathcal{E}(f_{\delta^*}) \leq \left(1 + O\left(n^{-1/2}\right)\right)\left(\sigma^2 + \frac{\left(\sum_i \lambda_i\right)\left(\sum_i \frac{v_i^2}{\lambda_i}\right)}{n}\right). \tag{2.25}$$

Fast rate. If we assume that $\lambda_i = i^{-\alpha}$, then we can get faster finite-sample rate. In particular, we know that $\sum_{i>k} \lambda_i^2 = \Theta(k^{-(2\alpha-1)})$ then we can balance

$$\frac{k'}{n} + \epsilon = O\left(\frac{k'}{n} + \frac{1}{(k')^{2\alpha/3-1/3} n^{1/3}}\right)$$

and so we should set $k' = n^{\frac{1}{\alpha+1}}$. Plugging in, we see that both ϵ and k'/n are of order $O(n^{-\frac{\alpha}{1+\alpha}})$. As $\alpha \to \infty$, we see that the finite-sample rate approaches the parametric rate of $O(n^{-1})$.

If we assume that $\lambda_i = e^{-i}$, then we can get even faster finite-sample rate. We know $\sum_{i>k} \lambda_i^2 =$

$\Theta(e^{-2k})$ then we can balance

$$\frac{k'}{n} + \epsilon = O\left(\frac{k'}{n} + \frac{e^{-2k'/3}}{n^{1/3}}\right).$$

Then choosing $k' = \Theta(\log n)$, we see that both ϵ and k'/n are of order at most $O\left(\frac{\log n}{n}\right)$.

2.4 Uniform Convergence

In Section 2.2 and 2.3, we analyze the test error of the ridgeless interpolant $\mathcal{E}(f_0)$ by first bounding the cost of interpolation $\frac{\mathcal{E}(f_0)}{\mathcal{E}(f_{\delta^*})}$ in terms of the spectrum $\{\lambda_i\}$, and then analyzing the test error of the optimal ridge $\mathcal{E}(f_{\delta^*})$. In this section, we consider a more traditional approach. We first consider the difference[1] between the test error and the training error $\sqrt{\mathcal{E}(f_\delta)} - \sqrt{\mathcal{E}_{\text{tr}}(f_\delta)}$ uniformly over all ridge parameter $\delta \geq 0$. Then, we will control the generalization gap using the complexity of the model as measured by RKHS norm. Finally, we will analyze the RKHS norm of the ridgeless interpolant $\mathbb{E}\|f_0\|_{\mathcal{K}}^2$ and use it to recover the same consistency results.

Using equations (2.10) and (2.11), we can show the following uniform convergence result.

Theorem 9. *For any $\delta \geq 0$ and $k \in \mathbb{N}$ such that $(k/n)^2 + 2(k/n) < 1$. Let $\epsilon = \sqrt{(k^2 + 2kn)/n^2}$, then it holds that*

$$(1-\epsilon)^2 \mathcal{E}(f_\delta) \leq \left(\sqrt{\mathcal{E}_{tr}(f_\delta)} + \sqrt{\frac{(\sum_{i>k} \lambda_i)\mathbb{E}\|f_\delta\|_{\mathcal{K}}^2}{n}}\right)^2. \tag{2.26}$$

As mentioned earlier, the term $\frac{1}{n}(\sum_{i>k} \lambda_i)\mathbb{E}\|f_\delta\|_{\mathcal{K}}^2$ corresponds to the generalization error and can be viewed as a bound on the Rademacher complexity of norm-constrained linear predictors. The choice of k allows us to only depend on $\sum_{i>k} \lambda_i$, which can be significantly smaller than $\sum_i \lambda_i$ when there are a small number of very large eigenvalues. In addition, equation (2.26)

1. The reason why we consider the difference of the square roots will be clear in Section 4.

precisely quantifies the bias-variance tradeoff in tuning the ridge parameter. Smaller δ leads to a smaller training error $\sqrt{\mathcal{E}_{\text{tr}}(f_\delta)}$ but requires a higher RKHS norm $\mathbb{E}\|f_\delta\|_{\mathcal{K}}^2$. In the extreme case $\delta = 0$, the training error is exactly 0 and the RKHS norm needs to be very high in order to interpolate all the noisy training labels. Such a small bias and high variance situation will typically lead to sub-optimal asymptotic test error. However, under the benign overfitting condition such as (2.16), it can be shown the minimal norm required to achieve zero training error is just large enough so that the generalization error in (2.26) can exactly match the Bayes error σ^2 as n goes to infinity. In particular, we have the following norm bound on the ridgeless interpolant.

Proposition 10. *For any $l \in \mathbb{N} \cup \{\infty\}$ and $k \in \mathbb{N}$ such that $R_k > n$, it holds that*

$$\mathbb{E}\|f_0\|_{\mathcal{K}}^2 \leq \sum_{i \leq l} \frac{v_i^2}{\lambda_i} + \left(1 - \frac{n}{R_k}\right)^{-1} \frac{n\left(\sigma^2 + \sum_{i>l} v_i^2\right)}{\sum_{i>k} \lambda_i}. \tag{2.27}$$

As n grows, the first term is usually a lower order term compared to the second term (which scales linearly with n). If $n/R_k \to 0$ and l is chosen so that $\sum_{i>l} v_i^2 \to 0$, then indeed

$$\mathbb{E}\|f_0\|_{\mathcal{K}}^2 \approx \frac{n\sigma^2}{\sum_{i>k} \lambda_i} \qquad \text{and} \qquad \frac{(\sum_{i>k} \lambda_i)\mathbb{E}\|f_0\|_{\mathcal{K}}^2}{n} \approx \sigma^2.$$

Plugging the norm bound in Proposition 10 to Theorem 9, we immediately have the following bound on the test error of the interpolating KRR solution.

Corollary 11. *For any $l \in \mathbb{N} \cup \{\infty\}$ and $k \in \mathbb{N}$ such that $(k/n)^2 + 2(k/n) < 1$ and $R_k > n$. Let $\epsilon = \sqrt{(k^2 + 2kn)/n^2}$, then it holds that*

$$(1-\epsilon)^2 \mathcal{E}(f_0) \leq \frac{\left(\sum_{i>k} \lambda_i\right)\left(\sum_{i \leq l} \frac{v_i^2}{\lambda_i}\right)}{n} + \left(1 - \frac{n}{R_k}\right)^{-1} \left(\sigma^2 + \sum_{i>l} v_i^2\right). \tag{2.28}$$

We note that (2.28) is quite similar to (2.23) with the main difference in the multiplicative factors. This is expected due to Theorem 1 and we can recover a version of Corollary 11 using

Proposition 2 and Theorem 8 as well. We see that $\mathcal{E}(f_0) \to \sigma^2$ if the following conditions hold:

(A) $\sum_i v_i^2 < \infty$

(B) there exists a sequence k_n such that

$$\lim_{n\to\infty} \frac{\sum_{i>k_n} \lambda_i}{n} = 0, \quad \lim_{n\to\infty} \frac{k_n}{n} = 0, \quad \lim_{n\to\infty} \frac{n}{R_{k_n}} = 0. \tag{2.29}$$

The first condition is very mild because it is weaker than requiring the label to have bounded second moment: $\sigma^2 + \sum_i v_i^2 = \mathbb{E}[y^2] < \infty$. The condition $\frac{1}{n}\sum_{i>k_n} \lambda_i \to 0$ is also very mild because usually $\sum_i \lambda_i = \mathbb{E}[K(x,x)] = 1 < \infty$. From Corollary 4, we also know that the last two conditions in (2.29) are necessary.

2.5 Example: Inner-Product Kernels in Polynomial Regime

In this section, we consider kernel ridge regression in the setting studied in Ghorbani et al. [24], Mei et al. [41] and Misiakiewicz [47]. Let's take p to be the uniformly distribution over the sphere in $\mathbb{R}^d$ or the boolean hypercube. Denote $\mathcal{V}_{\le l-1}$ to be the subspace of all polynomials of degree $\le l-1$ and $B(d,l) = \Theta_d(d^l)$ to be the dimension of the subspace $\mathcal{V}_l$ of degree-l polynomials orthogonal to $\mathcal{V}_{\le l-1}$. Let $\{Y_{ks}\}_{k\ge 0, s\in[B(d,k)]}$ be the polynomial basis with respect to $\mathcal{D}$ (e.g. spherical harmonics or parity functions). Expand the target function in this basis:

$$f^*(x) = \sum_{k=0}^{\infty} \sum_{s\in[B(d,k)]} v_{ks} Y_{ks}(x).$$

Inner-Product Kernel Decomposition. Consider kernels of the form $K(x,x') = h_d(\langle x, x'\rangle/d)$, then it admits the eigendecompositon in the polynomial basis:

$$K(x,x') = \sum_{k=0}^{\infty} \sum_{s\in[B(d,k)]} \frac{\mu_{d,k}(h)}{B(d,k)} Y_{ks}(x) Y_{ks}(x')$$

Interestingly, the eigenvalues of K with respect to $\mathcal{D}$ has a block diagonal structure. The block diagonal structure is a consequence of the rotation-invariance of the distribution p.

Covariance Splitting. Consider the regime $n \asymp d^l$ where l is not an integer. Choose k in Corollary 11 to include the first $\lfloor l \rfloor$ blocks. Then

$$k = \sum_{k=0}^{\lfloor l \rfloor} B(d,k) = \Theta\left(\sum_{k=0}^{\lfloor l \rfloor} d^k\right) = \Theta\left(d^{\lfloor l \rfloor}\right) = o(n). \tag{2.30}$$

Moreover, denote $P_{\leq \lfloor l \rfloor}$ to be the projection onto $\mathcal{V}_{\leq \lfloor l \rfloor}$ and $P_{> \lfloor l \rfloor}$ to be the projection onto its complement. Then we have

$$\sum_{k=0}^{\lfloor l \rfloor} \sum_{s \in [B(d,k)]} \frac{v_{ks}^2}{\mu_{d,k}/B(d,k)} = \|P_{\leq \lfloor l \rfloor} f^*\|_{\mathcal{K}}^2$$
$$\sum_{k > \lfloor l \rfloor} \sum_{s \in [B(d,k)]} v_i^2 = \|P_{> \lfloor l \rfloor} f^*\|^2.$$

We can compute

$$\sum_{k > \lfloor l \rfloor} \sum_{s \in [B(d,k)]} \frac{\mu_{d,k}(h)}{B(d,k)} = \sum_{k > \lfloor l \rfloor} \mu_{d,k}(h)$$

and so

$$\begin{aligned} R_k &= \frac{\left(\sum_{k > \lfloor l \rfloor} \sum_{s \in [B(d,k)]} \frac{\mu_{d,k}(h)}{B(d,k)}\right)^2}{\sum_{k > \lfloor l \rfloor} \sum_{s \in [B(d,k)]} \left(\frac{\mu_{d,k}(h)}{B(d,k)}\right)^2} = \frac{\left(\sum_{k > \lfloor l \rfloor} \mu_{d,k}(h)\right)^2}{\sum_{k > \lfloor l \rfloor} \frac{\mu_{d,k}(h)^2}{B(d,k)}} \\ &\geq \frac{\left(\sum_{k > \lfloor l \rfloor} \mu_{d,k}(h)\right)^2}{\sum_{k > \lfloor l \rfloor} \mu_{d,k}(h)^2} \cdot B(d, \lceil l \rceil) = \Omega(d^{\lceil l \rceil}) = \omega(n). \end{aligned} \tag{2.31}$$

As a result, if we have

$$\frac{\|P_{\leq \lfloor l \rfloor} f^*\|_{\mathcal{K}}^2 \cdot \left(\sum_{k > \lfloor l \rfloor} \mu_{d,k}(h)\right)}{n} \to 0, \tag{2.32}$$

then applying Corollary 11, we have shown that

$$\limsup_{n\to\infty} \mathcal{E}(f_0) - \sigma^2 \leq \|P_{>\lfloor l \rfloor} f^*\|^2. \tag{2.33}$$

In Ghorbani et al. [24] and Mei et al. [41], it is shown that the above is not just an upper bound. In fact, it holds that $\lim_{n\to\infty} \mathcal{E}(f_0) - \sigma^2 = \|P_{>\lfloor l \rfloor} f^*\|^2$ and so the application of our Collary 4 is in fact tight in this case.

CHAPTER 3

LINEAR REGRESSION IN GAUSSIAN SPACE

Linear regression is a fundamental model in statistics and machine learning. This chapter studies the generalization theory for linear regression with Gaussian features. As we will see in Section 3.1, the Gaussian feature assumption will be helpful for us to establish uniform convergence not just for predictors on the ridge path (Theorem 9), but for *any* linear predictor. The generality of our result can allow us to study predictors that do not necessarily have a closed-form expression and therefore cannot be analyzed directly like in Chapter 2 (for example, ℓ_1 or nuclear norm regularization). In addition, since the uniform convergence approach does not require the data distribution to be well-specified, we can extend the results in the previous section to more general settings. From now on, we will let the data distribution be a Gaussian multi-index model. We assume the data distribution $\mathcal{D}$ is given by:

(A) d-dimensional Gaussian features with arbitrary mean and covariance: $x \sim \mathcal{N}(\mu, \Sigma)$

(B) a generic multi-index model: there exists a low-dimensional projection $W = [w_1^*, ..., w_k^*] \in \mathbb{R}^{d\times k}$, a random variable $\xi \sim \mathcal{D}_\xi$ independent of x (not necessarily Gaussian), and an unknown link function $g : \mathbb{R}^{k+1} \to \mathbb{R}$ such that

$$\eta_i = \langle w_i^*, x\rangle, \quad y = g(\eta_1, ..., \eta_k, \xi). \tag{3.1}$$

Since g is not even required to be continuous, our assumption on y is quite general. The multi-index model (3.1) includes well-specified linear regression by setting $\mathcal{D}_\xi$ to be any distribution with zero mean and bounded second moment, $k = 1$ and $g(\eta, \xi) = \eta + \xi$. It also allows nonlinear trends and heteroskedasticity by changing the definition of g. For example, it is possible that $y = g(\langle w^*, x\rangle)$ is a deterministic function of x but with an unknown and nonlinear function g. Even though a linear model is misspecified in this case and cannot achieve zero population error,

it is still interesting to ask whether we can achieve the best population error for a linear predictor asymptotically. More generally, we can even allow g to be a neural network as long as the number of hidden units k in the first layer is small relative to the sample size n and $\mathcal{D}$ satisfies the hypercontractivity assumption (C) below. It can be challenging to find consistent estimates for the parameters $w_1^*, ..., w_k^*$, but we are only concerned with finding a linear model with good prediction error $L(w, b)$. Note that the feature vector x can also have arbitrary mean and covariance, so the only real restriction on $\mathcal{D}$ is the Gaussianity of x. In Chapter 5, we will see that sometimes the Gaussian feature assumption is unnecessary, so our theory holds more generally. This is a special case of a general phenomenon known as universality[1].

Given a training set $\{(x_i, y_i)\}_{i=1}^n$ sampled independently from $\mathcal{D}$, we consider the square loss and define the training and test error to be

$$\hat{L}(w, b) = \frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle + b - y_i)^2, \quad L(w, b) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\langle w, x\rangle + b - y)^2]. \tag{3.2}$$

We will consider loss functions other than the square loss in the next chapter.

3.1 Optimistic Rate

Our result in this section is a special case of the Moreau envelope theory introduced in Chapter 4, which requires a mild technical assumption known as hypercontractivity:

(C) There exists a universal constant $\tau > 0$ such that uniformly over all $(w, b) \in \mathbb{R}^d \times \mathbb{R}$, it holds that

$$\frac{\mathbb{E}[(\langle w, x\rangle + b - y)^8]^{1/8}}{\mathbb{E}[(\langle w, x\rangle + b - y)^2]^{1/2}} \leq \tau. \tag{3.3}$$

The assumption (3.3) is verified with an explicit choice of τ in Appendix B.1. We note that the power of 8 is not crucial here and can be replaced by any power greater than 2. As in Section 2.4,

1. However, we also prove that universality can fail in some misspecified settings (see the discussion in section 5.2 for more details)

we control the generalization gap defined by the difference of the square roots. Theorem 12 below is a special case of case (ii) in Theorem 36 in Chapter 4 and the proof can be found in Appendix E.

Theorem 12. *Assume (A), (B), and (C) holds and denote* $Q = I - W(W^T\Sigma W)^{-1}W^T\Sigma$. *For any* $\delta \in (0,1)$, *let* $C_\delta : \mathbb{R}^d \to [0,\infty]$ *be a continuous function such that with probability at least* $1-\delta/4$ *over* $x \sim \mathcal{N}(0,\Sigma)$, *uniformly over all* $w \in \mathbb{R}^d$,

$$\langle Qw, x\rangle \le C_\delta(w). \tag{3.4}$$

Then with probability at least $1-\delta$, *it holds uniformly over all* $(w,b) \in \mathbb{R}^d \times \mathbb{R}$ *that*

$$(1-\epsilon)\, L(w,b) \le \left(\sqrt{\hat{L}(w,b)} + \sqrt{\frac{C_\delta(w)^2}{n}}\right)^2 \tag{3.5}$$

where $\epsilon = O\left(\tau\sqrt{\frac{k\log(n/k)+\log(1/\delta)}{n}}\right)$.

Compared to Theorem 9, Theorem 12 requires the Gaussianity of x in assumption (A), but it can be proven rigorously and applied to misspecified models of the form in assumption (B). Recall that we assume $y = f^*(x) + \xi$ in Chapter 2 and we expand $f^* = \sum v_i\phi_i$, and this assumption is crucial for obtaining a closed-form expression such as equation (2.10). More importantly, our $\sqrt{C_\delta^2/n}$ term plays the role of Rademacher complexity in classical uniform convergence guarantees [3]. In our context, the average Rademacher complexity is given by the following:

Definition 1. *Given a positive semi-definite matrix* Σ *and sample size* $n \in \mathbb{N}$, *the* Rademacher complexity *of a hypothesis class* $\mathcal{H}$ *is given by*

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\substack{x_1,\dots,x_n\sim\mathcal{N}(0,\Sigma)\\ s\sim \mathit{Unif}(\{\pm1\}^n)}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n s_i h(x_i)\right|\right]. \tag{3.6}$$

Rademacher complexity measures the ability of $\mathcal{H}$ *to fit random Rademacher noise* (±1) *on an average training set sampled from the ground truth distribution. For more background, see for ex-*

ample the work of Bartlett and Mendelson [3], Bartlett et al. [5], Srebro et al. [63] and Wainwright [73].

A closely related geometric complexity measure is the Gaussian width [see, e.g., 3, 71]. The following definitions match the notation of Koehler et al. [33].

Definition 2. *The* Gaussian width *and the* radius *of a set $S \subset \mathbb{R}^d$ are*

$$W(S) := \mathbb{E}_{H\sim\mathcal{N}(0,I_d)} \sup_{s\in S} |\langle s, H\rangle| \quad \textit{and} \quad \operatorname{rad}(S) := \sup_{s\in S} \|s\|_2.$$

We also define the notation

$$W_\Sigma(S) := W(\Sigma^{1/2}S)$$

to represent the Gaussian width with respect to covariance matrix Σ.

As it turns out, when the hypothesis class $\mathcal{H}$ is linear, the Rademacher complexity is actually equivalent to Gaussian width (up to a scaling of $1/\sqrt{n}$).

Proposition 13. *Let $\mathcal{K}$ be an arbitrary subset of $\mathbb{R}^d$ and consider $\mathcal{H} = \{x \mapsto \langle w, x\rangle : w \in \mathcal{K}\}$. Then, for any positive semi-definite matrix Σ, it holds that*

$$\mathcal{R}_n(\mathcal{H}) = \frac{W_\Sigma(\mathcal{K})}{\sqrt{n}}. \tag{3.7}$$

Proof. Observe that for $x_1, ..., x_n \sim \mathcal{N}(0, \Sigma)$ independent of $s \sim \text{Unif}(\{\pm 1\}^n)$, we have that $\frac{1}{n}\sum_{i=1}^n s_i x_i \sim \mathcal{N}\left(0, \frac{1}{n}\Sigma\right)$. The rest just follows from definitions:

$$\begin{aligned}
\mathcal{R}_n(\mathcal{H}) &= \mathbb{E}_{\substack{x_1,\dots,x_n\sim\mathcal{N}(0,\Sigma)\\ s\sim\text{Unif}(\{\pm1\}^n)}} \left[\sup_{w\in\mathcal{K}} \left|\frac{1}{n}\sum_{i=1}^n s_i\langle w, x_i\rangle\right|\right] \\
&= \mathbb{E}_{\substack{x_1,\dots,x_n\sim\mathcal{N}(0,\Sigma)\\ s\sim\text{Unif}(\{\pm1\}^n)}} \left[\sup_{w\in\mathcal{K}} \left|\left\langle w, \frac{1}{n}\sum_{i=1}^n s_i x_i\right\rangle\right|\right] \\
&= \mathbb{E}_{H\sim\mathcal{N}(0,I_d)} \left[\sup_{w\in\mathcal{K}} \left|\left\langle w, \tfrac{1}{\sqrt{n}}\Sigma^{\frac{1}{2}}H\right\rangle\right|\right] = n^{-1/2} W_\Sigma(\mathcal{K}).
\end{aligned}$$

□

If we let $\Sigma^{\perp} = Q^T \Sigma Q$, then it is clear from definition that C_δ is just a high probability version of the Gaussian width $W_{\Sigma^{\perp}}$ (which is equivalent to Rademacher complexity after rescaling). If we consider Ridge regression and the minimal ℓ_2 norm interpolator, then we can approximately choose C_δ to be

$$C_\delta(w) \approx \|w\|_2 \cdot \mathbb{E}\|Q^T x\|_2 \tag{3.8}$$

by applying Cauchy-Schwarz inequality and standard concentration argument for the norm of $\|Q^T x\|_2$. We can view Q as a (potentially oblique) projection onto the subspace orthogonal to $\Sigma^{1/2} w_1^*, ..., \Sigma^{1/2} w_k^*$. Indeed, if we denote the corresponding orthogonal projection matrix as P, then it is clear that $Q = \Sigma^{-1/2} P \Sigma^{1/2}$. In other words, Q first applies the linear map $\Sigma^{1/2}$ to a predictor w, orthogonalizes $\Sigma^{1/2} w$ with respect to $\Sigma^{1/2} w_1^*, ..., \Sigma^{1/2} w_k^*$ and finally applies the inverse linear map $\Sigma^{-1/2}$. Since the trace of a covariance matrix is equal to the sum of its eigenvalues, $\mathbb{E}\|Q^T x\|_2^2 = \mathrm{Tr}(Q^T \Sigma Q)$ and the projection Q effectively set some eigenvalues to 0, we can view $\mathbb{E}\|Q^T x\|_2$ essentially as the term $\sqrt{\sum_{i>k} \lambda_i}$ in equation (2.26) of Theorem 9 and so our equation (3.5) can recover a similar uniform convergence guarantee. In fact, equation (3.5) holds not just for predictors on the ridge path. It holds for *any* predictors in the entire $\mathbb{R}^d$. As we will see in section 3.3, 3.5 and 3.6, other choices of C_δ might be more appropriate depending on the application. For an arbitrary norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$, we can pick

$$C_\delta(w) \approx \|w\| \cdot \mathbb{E}\|Q^T x\|_*. \tag{3.9}$$

In the application to LASSO regression in section 3.3, we will pick $\|\cdot\|$ to be the ℓ_1 norm. In the application to matrix sensing in section 3.6, we will pick $\|\cdot\|$ to be the nuclear norm. The corresponding dual norms are the ℓ_∞ norm and the spectral norm, respectively. In Section 4.2, we will show that we can always obtain sharp asymptotic risk bound for empirical risk minimizer (ERM) in a convex set by choose C_δ based on local Gaussian width.

3.2 Minimal Norm Interpolation

Now that we have proved the uniform convergence guarantee (Theorem 12), it remains to analyze the complexity of the minimal norm interpolant. In this section, we use the Gaussian comparison inequality to prove an analog of Proposition 10 in Chapter 2. We first introduce the notion of effective ranks, which is the same as the definition in equation (2.13) of the previous chapter.

Definition 3 (Bartlett et al. [7])**.** *The* effective ranks *of a covariance matrix Σ are*

$$r(\Sigma) = \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|_{op}} \quad \textit{and} \quad R(\Sigma) = \frac{\mathrm{Tr}(\Sigma)^2}{\mathrm{Tr}(\Sigma^2)}.$$

The effective ranks are related to the concentration of the ℓ_2 norm of a Gaussian vector with covariance Σ. In fact, both definitions of effective ranks can be derived by applying Bernstein's inequality to $\|x\|^2/\mathbb{E}\|x\|^2$ (for example, see Lemma 64 in the Appendix). For more on these notions of effective rank, see Bartlett et al. [7] and Koehler et al. [33]. We are now ready to state the norm bound for the minimal ℓ_2 norm interpolant.

Theorem 14. *Assume that (A) and (B) holds. Let Q be the same as in Theorem 12 and $\Sigma^{\perp} = Q^T \Sigma Q$. Fix any $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$ such that $Qw^\sharp = 0$ and suppose for some $\rho \in (0,1)$, it holds that*

$$\hat{L}(w^\sharp, b^\sharp) \le (1+\rho) L(w^\sharp, b^\sharp). \tag{3.10}$$

Then with probability at least $1-\delta$, for some $\epsilon \lesssim \rho + \log\left(\frac{1}{\delta}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma^\perp)}} + \frac{k}{n} + \frac{n}{R(\Sigma^\perp)}\right)$, it holds that

$$\min_{(w,b)\in\mathbb{R}^{d+1}:\hat{L}_f(w,b)=0} \|w\|_2 \le \|w^\sharp\|_2 + (1+\epsilon)\sqrt{\frac{nL(w^\sharp, b^\sharp)}{\mathrm{Tr}(\Sigma^\perp)}}. \tag{3.11}$$

The proof of Theorem 14 is the same as Theorem 38 in Chapter 4 and can be found in Appendix E.3. In the following paragraph, we give an intuitive explanation of the norm bound (3.11).

As mentioned earlier, Q is a projection matrix and so $\Sigma^{\perp} = Q^T \Sigma Q$ only retains the components of Σ that correspond to the tail of its eigenvalues. In Theorem 14, it is implicitly assumed that $R(\Sigma^{\perp})$ is large with respect to the sample size n and this condition effectively ensures that $Q^T x_1, ..., Q^T x_n$ are approximately orthogonal. Therefore, we have

$$\langle x_i, QQ^T x_j \rangle = \langle Q^T x_i, Q^T x_j \rangle \approx \begin{cases} \|Q^T x_i\|^2 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

To prove an upper bound on the minimal norm required to interpolate, we can fix any $(w^\sharp, b^\sharp)$ and consider

$$w \approx w^\sharp + \sum_{i=1}^{n} (y_i - \langle w^\sharp, x_i \rangle - b^\sharp) \frac{QQ^T x_i}{\|Q^T x_i\|^2}$$

because $\langle w, x_i \rangle + b^\sharp \approx y_i$. By the near orthogonality again, we have

$$\|w\|_2^2 \approx \|w^\sharp\|_2^2 + \sum_{i=1}^{n} \frac{(y_i - \langle w^\sharp, x_i \rangle - b^\sharp)^2}{\|Q^T x_i\|^2} \approx \|w^\sharp\|_2^2 + \frac{nL(w^\sharp, b^\sharp)}{\mathrm{Tr}(\Sigma^{\perp})}$$

which is essentially our bound (3.11) with $\epsilon \approx 0$.

If we assume a well-specified model: $y = \langle w^*, x \rangle + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$, then we can choose $w^\sharp = w^*$ and then $L(w^\sharp, b^\sharp) = \sigma^2$. More generally, we can choose $w^\sharp, b^\sharp = \arg\min_{w,b} L(w, b)$. The condition $Qw^\sharp = 0$ can always be satisfied because $L((I - Q)w, b) \leq L(w, b)$ for all $(w, b) \in \mathbb{R}^{d+1}$ by Jensen's inequality. Condition (3.10) can also be easily checked because we just need concentration of the empirical loss at a single non-random parameter. If $\|w^\sharp\|_2$ is too large for our bound (3.11), then it might be beneficial to approximately minimize $L(w, b)$ with a smaller norm (as in the case of Proposition 10). As we see in Chapter 2, the $\sqrt{n/\,\mathrm{Tr}(\Sigma^{\perp})}$ factor in the norm bound (3.11) conveniently cancels with the complexity term in equation (3.5). Since Theorem 14 applies to misspecified data distributions as well, we can combine it with Theorem 12 to establish benign overfitting for any Gaussian multi-index setting.

Corollary 15. *Assume that (A), (B), and (C) holds. Let* $Q = I - W(W^T\Sigma W)^{-1}W^T\Sigma$ *and* $\Sigma^\perp = Q^T\Sigma Q$. *Fix any* $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$ *such that* $Qw^\sharp = 0$. *Consider the minimal norm interpolator* $\hat{w}, \hat{b} = \arg\min_{(w,b):\hat{L}(w,b)=0} \|w\|_2$. *Then with probability at least* $1-\delta$, *for some*

$$\rho \lesssim \tau\sqrt{\frac{k\log(n/k)+\log(1/\delta)}{n}} + \log(1/\delta)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma^\perp)}} + \frac{k}{n} + \frac{n}{R(\Sigma^\perp)}\right),$$

it holds that

$$L(\hat{w},\hat{b}) \le (1+\rho)\left(\sqrt{L(w^\sharp,b^\sharp)} + \|w^\sharp\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma^\perp)}{n}}\right)^2. \tag{3.12}$$

The proof can be found in Appendix D.1. Using Corollary 15, we can establish the consistency of the minimal ℓ_2 norm interpolant in the following sense. Suppose there exists a sequence of distributions[2] $\mathcal{D}_n$ satisfying assumptions (A), (B) and (C) and the Bayes error $\sigma_n^2 := \inf_{w,b} L_{\mathcal{D}_n}(w,b)$ converges to some limit $\sigma^2 \in \mathbb{R}_+$. Define $\Sigma_n^\perp$ the same way as in Corollary 15 and assume there exists a sequence $(w_n^\sharp, b_n^\sharp)$ such that $Q_n w_n^\sharp = 0$ and

$$L_{\mathcal{D}_n}(w_n^\sharp, b_n^\sharp) \to \sigma^2 \quad \text{and} \quad \|w_n^\sharp\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma_n^\perp)}{n}} \to 0. \tag{3.13}$$

In addition, assume that

$$\frac{k_n}{n} \to 0 \quad \text{and} \quad \frac{n}{R(\Sigma_n^\perp)} \to 0, \tag{3.14}$$

then Corollary 15 suggests that $L_{\mathcal{D}_n}(\hat{w}_n, \hat{b}_n) \to \sigma^2$ in probability.

Finally, we note that our analysis can show something more than the consistency of the minimal ℓ_2 norm interpolant. In fact, any predictor on the ridge path with norm larger than $\|w^\sharp\|_2$ (or with regularization parameter sufficiently close to 0) can enjoy the same generalization guarantee as the minimal norm interpolant, which explains the flatness of the generalization curve as opposed to the

2. We need the feature dimension $d \ge n$ in order to interpolate so the distribution $\mathcal{D}$ might change as n increases.

classical U-shaped curve (for example, see Figure 5.3 in Chapter 5). In other words, once we fit all of the signals, it does not matter how much noise is fitted, and all near-interpolators can achieve consistency at the same time.

Corollary 16. *In the same setting as Corollary 15, it also holds with probability at least* $1-\delta$ *that for any* R *between* $\|\hat{w}\|_2$ *and* $\|w^\sharp\|_2$*, the constrained risk minimizer*

$$\hat{w}_R, \hat{b}_R := \underset{(w,b):\|w\|_2 \le R}{\arg\min} \hat{L}(w,b) \tag{3.15}$$

has the generalization bound

$$L(\hat{w}_R, \hat{b}_R) \le (1+\rho)\left(\sqrt{L(w^\sharp, b^\sharp)} + \|w^\sharp\|_2 \sqrt{\frac{\mathrm{Tr}(\Sigma^\perp)}{n}}\right)^2. \tag{3.16}$$

Corollary 16 is a consequence of Lemma 65 in the Appendix, which holds more generally. Using the same techniques in this section, we can also establish benign overfitting result for the minimal ℓ_1 norm interpolant (for example, see Koehler et al. [33]), but the ℓ_1 norm bound there is not always tight. Tighter analysis for the minimal ℓ_1 norm interpolant and more generally the minimal ℓ_p norm interpolant can be found in Wang et al. [74] and Donhauser et al. [21], and their proof techniques still depend on a result similar to our Theorem 12.

3.3 Optimally Regularized Estimators

In the last section, we show that the minimal norm interpolant is consistent when conditions (3.13) and (3.14) hold. However, as we see in the Examples 3, 4 and 5 in Chapter 2, the condition $n/R(\Sigma^\perp) \to 0$ does not always hold and is not necessary for the consistency of the optimally-tuned ridge regression. We only need this condition if we want the minimal norm interpolant to be as good as the optimal ridge. The following result, which is a corollary of Theorem 12, extends the analysis of optimal ridge to using any norm as a regularizer.

Corollary 17. *Assume that (A), (B), and (C) holds and let* $Q = I - W(W^T\Sigma W)^{-1}W^T\Sigma$. *Fix an arbitrary norm* $\|\cdot\|$ *and any* $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$. *There exists* $\lambda^* > 0$ *such that if we consider the regularized estimator*

$$\hat{w}_{\lambda^*}, \hat{b}_{\lambda^*} = \arg\min_{(w,b)} \sqrt{\hat{L}(w,b)} + \lambda^*\|w\|, \tag{3.17}$$

it holds with probability at least $1-\delta$ *that*

$$(1-\epsilon)L(\hat{w}_{\lambda^*}, \hat{b}_{\lambda^*}) \le \left(\sqrt{L(w^\sharp, b^\sharp)} + \left(\mathbb{E}\|Q^T x\|_* + \sup_{\|v\|\le 1}\|Qv\|_\Sigma\sqrt{2\log(16/\delta)}\right)\frac{\|w^\sharp\|}{\sqrt{n}}\right)^2$$

where ϵ *is the same as in Theorem 12.*

As before, suppose there exists a sequence of distributions $\mathcal{D}_n$ satisfying assumptions (A), (B) and (C) and the Bayes error $\sigma_n^2 := \inf_{w,b} L_{\mathcal{D}_n}(w,b)$ converges to some limit $\sigma^2 \in \mathbb{R}_+$. In order for the regularized estimator $(\hat{w}_{\lambda^*}, \hat{b}_{\lambda^*})$ to be consistent in the sense that $L_{\mathcal{D}_n}(\hat{w}_{\lambda^*}, \hat{b}_{\lambda^*}) \to \sigma^2$, we see that we only need a sequence $(w_n^\sharp, b_n^\sharp)$ such that

$$L_{\mathcal{D}_n}(w_n^\sharp, b_n^\sharp) \to \sigma^2, \quad \frac{\|w_n^\sharp\|\mathbb{E}\|Q_n^T x\|_*}{\sqrt{n}} \to 0, \quad \frac{\|w_n^\sharp\|\sup_{\|v\|\le 1}\|Q_n v\|_{\Sigma_n}}{\sqrt{n}} \to 0 \tag{3.18}$$

and $k_n/n \to 0$.

Note that in the definition of $(\hat{w}_{\lambda^*}, \hat{b}_{\lambda^*})$, we add the regularization term to the square root of the training loss. In the context of ℓ_1 regularization, this is known as the square-root LASSO in the literature. In our applications, the quantity $\sup_{\|v\|\le 1}\|Qv\|_\Sigma$ is usually dominated by the expected norm $\mathbb{E}\|Q^T x\|_*$. The conditions in (3.18) are relatively mild: as long as the dimension-free quantity $\|w^\sharp\|\mathbb{E}\|Q^T x\|_*$ is small relative to $\sqrt{n}$, then we can achieve small population error. This is property of the data distribution $\mathcal{D}$ that makes it suitable to use the norm $\|\cdot\|$ as a regularizer and we typically do not expect consistency in a high dimensional problem unless there is some latent low-dimensional structure. Moreover, the proof of Corollary 17 suggests that we can choose λ^* to be any high probability bound of $\|Q^T x\|_*$. In practice, even though the matrix Q is unknown,

we can use a potentially more conservative estimate $\frac{1}{n}\sum_{i=1}^{n}\|x_n\|_*$ as a surrogate for λ^*.

Ridge Regression. Specializing Corollary 17 to the Euclidean norm, observe that

$$\mathbb{E}\|Q^T x\|_2 \le \sqrt{\mathbb{E}\|Q^T x\|_2^2} = \sqrt{\mathrm{Tr}(\Sigma^\perp)}$$
$$\sup_{\|v\|_2\le 1}\|Qv\|_\Sigma = \|\Sigma^\perp\|_{op}^{1/2} \le \sqrt{\mathrm{Tr}(\Sigma^\perp)}$$

and so in the context of ridge regression, we can simplify the generalization bound as

$$(1-\epsilon)L(\hat{w}_{\lambda^*},\hat{b}_{\lambda^*}) \le \left(\sqrt{L(w^\sharp,b^\sharp)} + \left(1+\sqrt{2\log(16/\delta)}\right)\|w^\sharp\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma^\perp)}{n}}\right)^2.$$

Therefore, the conditions

$$L_{\mathcal{D}_n}(w_n^\sharp,b_n^\sharp)\to\sigma^2, \quad \|w_n^\sharp\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma_n^\perp)}{n}}\to 0, \quad \frac{k_n}{n}\to 0 \tag{3.19}$$

are indeed sufficient for the consistency of optimally-tuned ridge regression. Compared with conditions (3.13) and (3.14), we no longer need $n/R(\Sigma^\perp)\to 0$. However, from Corollary 16, having that condition means we no longer need to tune the ridge parameter λ: any sufficiently small λ will lead to consistency.

Slow Rate under Bounded ℓ_1 Norm. Specializing Corollary 17 to the ℓ_1 norm, observe that

$$\mathbb{E}\|Q^T x\|_\infty \le \sqrt{2\log(2d)\max_i \Sigma_{ii}^\perp}$$
$$\sup_{\|v\|_1\le 1}\|Qv\|_\Sigma = \sqrt{\max_i \Sigma_{ii}^\perp}$$

by standard argument (for example, Lemma 66 in the appendix) and so in the context of LASSO regression, then we can simplify the generalization bound as

$$(1-\epsilon)L(\hat{w}_{\lambda^*}, \hat{b}_{\lambda^*}) \le \left(\sqrt{L(w^\sharp, b^\sharp)} + \sqrt{8 \max_i \Sigma^\perp_{ii}} \cdot \|w^\sharp\|_1 \sqrt{\frac{\log(2d) + \log(16/\delta)}{n}} \right)^2 .$$

If we denote $\sigma^2 = \inf_{w,b} L(w,b)$ and let $w^\sharp, b^\sharp$ be the minimizer of population loss, then the above result implies the convergence rate of $\sigma\|w^\sharp\|_1\sqrt{\frac{\log(d)}{n}} + \|w^\sharp\|_1^2 \cdot \frac{\log(d)}{n}$ to σ^2, which is also known as the "slow" rate of LASSO. Moreover, if $w^\sharp$ is k-sparse, then we can bound

$$\|w^\sharp\|_1 \le k\|w^\sharp\|_\infty$$

and so under these assumptions, the LASSO slow rate guarantee becomes $\sigma\|w^\sharp\|_\infty\sqrt{\frac{k^2\log(d)}{n}} + \|w^\sharp\|_\infty^2 \cdot \frac{k^2\log(d)}{n}$. This analysis works for all predictors w^* of bounded ℓ_1-norm, and it is minimax optimal over this class, but when we assume that w^* is k-sparse it is generally suboptimal and in particular does not give exact recovery when $\sigma = 0$. We now explain how our theory recovers the correct behavior in the sparse and well-conditioned setting commonly studied in the sparse linear regression literature.

Performance under Sparsity. We show how to recover well-known results from compressed sensing and high-dimensional statistics about sparse linear regression with Gaussian designs. In particular, we prove a performance guarantee for the LASSO when the covariance matrix is well-conditioned, as previously analyzed by Raskutti et al. [53], or more generally satisfies a version of the *compatability condition* [67]. We start with the following well-known lemma commonly used in the analysis of the LASSO (see, e.g. Vershynin [71])

Lemma 18. *Suppose $w^\sharp$ is k-sparse, i.e. supported on coordinate set $S \subset [d]$ with $|S| \leq k$. Every w with $\|w\|_1 \leq \|w^\sharp\|_1$ satisfies*

$$\|(w - w^\sharp)_{S^C}\|_1 \leq \|(w - w^\sharp)_S\|_1. \tag{3.20}$$

The above lemma shows that the vector $w - w^\sharp$ lies in the covex cone

$$\mathcal{C}(S) := \{u : \|u_{S^C}\|_1 \leq \|u_S\|_1\},$$

where S is the support of $w^\sharp$. Now we can state the version of the *compatibility condition* [67] we use; the compatibility condition is a weakening of the *restricted eigenvalue condition* [12, 53], and the compatibility condition is known to be a sufficient and almost necessary condition for the LASSO to perform exact recovery from $O(k \log d)$ samples in the Gaussian random design setting [32].

Definition 4 (Compatibility Condition; see Van De Geer and Bühlmann [67])**.** *For a positive semidefinite matrix $\Sigma \in \mathbb{R}^{d \times d}$ and set $S \subset [d]$, we say Σ has S-*restricted ℓ_1-eigenvalue

$$\phi^2(\Sigma, S) = \min_{u \in \mathcal{C}(S)} \frac{|S| \cdot \langle u, \Sigma u \rangle}{\|u_S\|_1^2}.$$

We say the S-compatibility condition *holds if the S-restricted ℓ_1-eigenvalue is nonzero.*

Combine Lemma 18 and the compatibility condition, we obtain the following:

Theorem 19. *Under assumptions (A), (B) and (C), let ϵ, Q be the same as in Theorem 12 and denote $\Sigma^\perp = Q^T \Sigma Q$. Let $\sigma^2 = \min_{w,b} L(w, b)$ and $w^\sharp, b^\sharp = \arg\min_{w,b} L(w, b)$. Suppose that*

1. *$w^\sharp$ is a k-sparse vector with support $S \subset [d]$*
2. *the covariance matrix Σ satisfies the S-compatibility condition*

3. the number of samples n satisfies

$$n \geq \frac{32 \max_i \Sigma^{\perp}_{ii}}{\phi(\Sigma, S)^2} \cdot k \log(16d/\delta) \quad \textit{and} \quad \epsilon \leq 1/2$$

Then for any $\epsilon' < 1$, it holds with probability at least $1 - \delta$ that for all (w, b) satisfying $\|w\|_1 \leq \|w^\sharp\|_1$ and $\hat{L}(w, b) \leq (1 + \epsilon')\sigma^2$, we have

$$L(w, b) - \sigma^2 \leq 104\sigma^2 \left(\epsilon + \epsilon' + \frac{\max_i \Sigma^{\perp}_{ii}}{\phi(\Sigma, S)^2} \frac{8k \log(16d/\delta)}{n} \right). \tag{3.21}$$

In particular, when $\sigma = 0$ we have that $\|w - w^\sharp\|_\Sigma = 0$, and so if Σ is positive definite then we have $w = w^\sharp$ (exact recovery).

To interpret the above bound, observe that when we consider the regularized ERM $\hat{w}, \hat{b} = \arg\min_{w,b:\|w\|_1 \leq \|w^\sharp\|_1} \hat{L}(w, b)$, we know that $\hat{L}(\hat{w}, \hat{b}) \leq \hat{L}(w^\sharp, b^\sharp)$ and so we can choose $\epsilon' = O(1/\sqrt{n})$ based on concentration of $\hat{L}(w^\sharp, b^\sharp)$. As a result, we have $\epsilon + \epsilon' = \tilde{O}(\tau\sqrt{k/n} + \sigma^2/\sqrt{n})$ and the last term, assuming Σ is well-conditioned, is $O(\sigma^2 k \log(d/\delta)/n)$, which is the well-known minimax rate for sparse linear regression (e.g., Rigollet and Hütter [56]). The above analysis is not very careful in terms of constant factors; later we show how to get sharp constants in the isotropic and well-specified setting. Also, we show how to get rid of the $\epsilon + \epsilon'$ term on the right hand side under stronger assumptions.

From now on, we will assume that the distribution $\mathcal{D}$ is given by

$$x \sim \mathcal{N}(0, \Sigma), \quad \xi \sim \mathcal{N}(0, \sigma^2), \quad y = \langle w^*, x \rangle + \xi \tag{3.22}$$

and we ignore the bias term b for simplicity. We denote $X \in \mathbb{R}^{n \times d}$ as the design matrix and $Y \in \mathbb{R}^n$ as the vector of responses. We define the empirical and population loss $\hat{L}(w)$ and $L(w)$ the same way as in (3.2) with $b = 0$. In this case, the Bayes error is σ^2 and the minimizer of the population loss is w^*.

The model (3.22) has been studied in many prior works because it is the starting point for the analysis of linear regression and it is often possible to use the structure $y = \langle w^*, x\rangle + \xi$ to directly compute most quantities of interest. Note that the model (3.22) clearly satisfies the Gaussian multi-index assumptions (A) and (B). Theorem 59 in the appendix shows that it also satisfies the hypercontractivity assumption (C) and so all of our results in this chapter so far can be applied. In the following sections, we will use our uniform convergence framework to recover results proven in the setting of (3.22).

First, we state a version of Theorem 12 that uses a slightly different definition of C_δ.

Theorem 20. *Under the model assumption in* (3.22)*, let* $F : \mathbb{R}^d \to [0, \infty]$ *be a continuous function such that for* $x \sim N(0, \Sigma)$*, with probability at least* $1 - \delta'$*, it holds uniformly over all* $w \in \mathbb{R}^d$ *that*

$$\langle w - w^*,\, x\rangle \le F(w). \tag{3.23}$$

For any $\delta > 0$*, assume* $n \ge 196 \log(12/\delta)$*. Then there exists* $\beta_1 \le 14\sqrt{\frac{\log(12/\delta)}{n}}$ *such that with probability at least* $1 - 2(\delta' + \delta)$*, it holds uniformly over all* $w \in \mathbb{R}^d$ *that*

$$L(w) \le (1 + \beta_1)\left(\sqrt{\hat{L}(w)} + \frac{F(w)}{\sqrt{n}}\right)^2. \tag{3.24}$$

We omit the proof here because it is very similar to the proof of Theorem 12 and can be found in Zhou et al. [78]. Due to some small difference[3] in the proof strategy, we no longer need to introduce the projection matrix Q and the hypercontractivity constant τ. As a result, Theorem 20 is easier to use in our applications later.

3. The difference between Theorem 12 and Theorem 20 is that the proof of Theorem 12 first conditions on both $Xw_1^*, ..., Xw_k^*$ and ξ and then uses VC theory to establish low-dimensional concentration, whereas Theorem 20 only needs to condition on ξ because it can exploit the linear structure of (3.22) in y.

3.4 Improved Finite-Sample Rate

When we consider the empirical risk minimizer in a set $\mathcal{K}$ and the set $\mathcal{K}$ has low complexity, the optimal rate for the population loss goes at a "parametric rate" of $1/n$, faster than a $1/\sqrt{n}$ rate, as in the case of ordinary least squares when d is fairly small compared to n. At first glance, it may appear impossible to get faster than a $1/\sqrt{n}$ rate from the main optimistic rates bound because of the presence of the $\beta_1 = O(\sqrt{\log(1/\delta)/n})$ term. As we will show, one can actually get fast/optimal rates from this theorem, but there is a different sense in which the $1/\sqrt{n}$ is unavoidable: this rate is actually the best we can hope for if we are only allowed to use certain summary statistics of the predictor. Nevertheless, it is still possible to obtain fast/optimal rates for the empirical risk minimizer by a black-box application of Theorem 20. The strategy we use is to bound the error $\|w - w^*\|_{\hat{\Sigma}}$ in the empirical metric by using a direct and very simple argument based on the KKT condition, and then apply Theorem 20 to bound the error in the population metric. The general idea of analyzing the population loss by going through the empirical metric is very common in statistics and learning theory (e.g. Mendelson [43], Bartlett and Mendelson [4], Lecué and Mendelson [35]).

Theorem 21. *Let $\mathcal{K}$ be a closed convex set in $\mathbb{R}^d$ containing w^* and suppose $\delta' \geq 0, p \geq 0$ are such that with probability at least $1 - \delta'$ over the randomness of $x \sim N(0, \Sigma)$, uniformly over all $w \in \mathcal{K}$ we have*

$$\langle w - w^*, x \rangle \leq \|w - w^*\|_{\Sigma} \sqrt{p}. \tag{3.25}$$

Suppose that $\hat{w} = \arg\min_{w \in \mathcal{K}} \hat{L}(w)$ and $p/n \leq 0.999$, then for all $n \geq C \log(2/\delta)$ for some absolute constant $C > 0$, it holds with probability at least $1 - (\delta + \delta')$ that

$$L(\hat{w}) - \sigma^2 \leq (1 + \tau)\sigma^2 \cdot \frac{p}{n}. \tag{3.26}$$

where $\tau = \tau(p, n, \delta)$ is upper bounded by an absolute constant and satisfies $\tau(p, n, \delta) \to 1$ in any joint limit $[p + \log(2/\delta)]/n \to 0, n \to \infty$.

The details of the proof can be found in the appendix, where it is obtained as a special case of a more general result (Theorem 68). To illustrate the application of this result, we show how it is used in the analysis of Ordinary Least Squares estimator $\hat{w}_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$.

Corollary 22. *Under the model assumptions* (4.2) *with $d < n$ and assuming a sufficiently large n, it holds with probability at least $1 - \delta$ that*

$$L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 \lesssim \sigma^2 \left(\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(36/\delta)}{n}} \right)^2 \tag{3.27}$$

Theorem 21 can be applied in a very similar way to analyze other models in the low complexity regime, for example the LASSO when the sparsity level is small, which we illustrate below. Provided the ℓ_1-eigenvalue φ and maximum diagonal entry of Σ are constants, we recover the sharp $\Theta(\sigma^2 k \log(d)/n)$ minimax rate for sparse linear regression (which is sharp provided $k \ll d$; see, e.g., [56]). This recovers the guarantee for the LASSO in the Gaussian random design setting given by combining the result of Raskutti et al. [53] with the appropriate analysis of LASSO in the fixed design setting [e.g. 12, 67].

Corollary 23. *Applying Theorem 21 with $\mathcal{K} = \{\|w\|_1 \leq \|w^*\|_1\}$ the rescaled ℓ_1-ball and under the sparsity and compatability condition assumptions of* (19)*, we have with probability at least $1 - \delta$ that the LASSO solution*

$$\hat{w}_{LASSO} = \operatorname*{arg\,min}_{w : \|w\|_1 \leq \|w^*\|_1} \hat{L}(w)$$

satisfies

$$L(\hat{w}_{LASSO}) - \sigma^2 \lesssim \frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{\sigma^2 k \log(16d/\delta)}{n} \tag{3.28}$$

provided n is sufficiently large that

$$\sqrt{\frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{8k \log(16d/\delta)}{n}} \leq 0.999.$$

Comparing Corollary 23 to Theorem 19, we see that we indeed get rid of the $\epsilon + \epsilon'$ term in the upper bound of equation (3.21).

3.5 Precise Asymptotics with Isotropic Features

In this section, we consider the isotropic setting with $\Sigma = I_d$ in (3.22) and we study the proportional limit regime where $d/n \to \gamma \in (0, \infty)$. This is a well-studied regime because the asymptotic results in random matrix theory such as the Marchenko-Pastur law are directly applicable. In contrast to the consistency results that we prove in the previous sections, even optimally regularized estimators tend to be inconsistent in these settings. However, we show that we can still use uniform convergence to recover the precise asymptotics of various estimators such as OLS, minimal norm interpolant and LASSO. Importantly, the precise asymptotics are derived from novel non-asymptotic results implied by Theorem 20. Finally, we discuss the limitation of uniform convergence to obtain sharp finite-sample rate in section 3.5.3.

3.5.1 Empirical Risk Minimizer

First, we consider a high-dimensional setting when d is smaller than n. For example, when $d = n/2$, the ordinary least squares estimator $\hat{w}_{\mathrm{OLS}}$ is the unique minimizer of the training error, but it does not interpolate the training data and so the uniform convergence of interpolators analysis cannot be applied. As it turns out, our Theorem 20 is enough to tightly characterize the excess risk of $\hat{w}_{\mathrm{OLS}}$. If we write $x = \Sigma^{1/2} H$ with $H \sim \mathcal{N}(0, I_d)$, then by the Cauchy-Schwarz inequality, it holds that

$$\langle w^* - w,\, x \rangle \le \|H\|_2 \, \|w^* - w\|_\Sigma \,.$$

Using standard concentration inequalities and $L(w) - \sigma^2 = \|w - w^*\|_\Sigma^2$, we can choose

$$F(w) = \left(\sqrt{d} + 2\sqrt{\log(4/\delta')} \right) \sqrt{L(w) - \sigma^2}. \tag{3.29}$$

Theorem 24. *Under the model assumptions in* (3.22)*, let* $\gamma = d/n < 1$*. There exists some* $\epsilon \lesssim \left(\frac{\log(36/\delta)}{n}\right)^{1/2}$ *such that for all sufficiently large* n*, with probability* $1-\delta$ *it holds uniformly for all* $w \in \mathbb{R}^d$ *that*

$$\left| \sqrt{L(w) - \sigma^2} - \sqrt{\frac{\gamma \hat{L}(w)}{(1-\gamma)^2}} \right| \leq \epsilon \sqrt{\hat{L}(w)} + \sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma} - \sigma^2\right) + \epsilon \hat{L}(w)}. \tag{3.30}$$

For the empirical risk minimizer $\hat{w}_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$*, the right hand side of* (3.30) *is approximately zero because we also have*

$$\hat{L}(\hat{w}_{\mathrm{OLS}}) \leq \sigma^2(1-\gamma) + \sigma^2 \epsilon \sqrt{1-\gamma}. \tag{3.31}$$

Therefore, we obtain the following generalization bound:

$$L(\hat{w}_{\mathrm{OLS}}) - \frac{\sigma^2}{1-\gamma} \lesssim \sigma^2 \left(\frac{\log(36/\delta)}{n}\right)^{1/4}. \tag{3.32}$$

We have a relatively complicated expression in (3.30) because our choice of F according to (3.29) depends on the excess risk $L(w) - \sigma^2$, and so after applying (3.24) we need to solve a quadratic equation. All quantities in (3.30) are well-defined because $\hat{L} \geq 0$ and the $\epsilon \hat{L}(w)$ term inside the last square root ensures that with high probability it is positive. If we think of ϵ as zero for simplicity, then our uniform convergence guarantee (3.30) predicts that the excess risk $L(w) - \sigma^2$ of a predictor with training error $\hat{L}(w)$ cannot be larger than

$$\frac{1}{1-\gamma}\left(\sqrt{\frac{\gamma \hat{L}(w)}{1-\gamma}} + \sqrt{\frac{\hat{L}(w)}{1-\gamma} - \sigma^2}\right)^2.$$

The minimal error is approximately $\sigma^2(1-\gamma)$ and so all near empirical risk minimizer should enjoy an excess risk of $\sigma^2 \frac{\gamma}{1-\gamma}$, which agrees with the exact expectation formula in Hastie et al. [28]; see their discussion for additional references. Since our approach also gives us a lower bound for free

(by solving the quadratic equation), Theorem 24 is enough to show that $L(\hat{w}_{\mathrm{OLS}})$ converges to $\sigma^2 \frac{1}{1-\gamma}$ in probability. We see that even though the empirical risk minimizer is not consistent, our uniform convergence approach can still provide an accurate understanding of the excess risk, and our bound for OLS is tight at least for the leading term.

Moreover, the $O(n^{-1/4})$ rate of (3.32) comes from the fact that we need to take the square root of ϵ in the last term of (3.30); it is not too difficult to see that this is sub-optimal for OLS. In fact, in Theorem 30, we explicitly calculate the variance of $L(\hat{w}_{\mathrm{OLS}})$ and show that in the proportional scaling regime (e.g., $\gamma = 0.5$), the right amount of deviation is of order $O(n^{-1/2})$. In the fixed-d regime, the convergence rate can be accelerated to the more familiar rate of $O(n^{-1})$. In Theorem 31, we show how to use a more direct approach to obtain high probability bounds that match these variance calculations. Surprisingly, we can also show that the $O(n^{-1/4})$ rate is generally unavoidable for any uniform convergence analysis that only considers the size of $\hat{L}(w)$. Our analysis is tight in the sense that there are estimators whose training error is indistinguishable from $\hat{w}_{\mathrm{OLS}}$, but whose convergence rate is provably slower than $\Omega(n^{-1/4})$. For readers interested in the tightest rate of convergence, more details can be found in Section 3.5.3.

For the case $d > n$, the OLS estimator is no longer defined, and instead we study the performance of the minimum-norm interpolator of the data. Together with the previous result, we show that the optimistic-rate bound can capture the behavior of the pseudoinverse estimator $\hat{w} = X^+ Y$ on both sides of the double descent curve.

Theorem 25. *Under the model assumptions in* (4.2) *with* $\gamma = d/n > 1$ *and* $\Sigma = I_d$*, there exists* $\epsilon \lesssim \left(\frac{\log(18/\delta)}{n}\right)^{1/2}$ *such that with probability at least* $1 - \delta$*, the following holds uniformly over all* w *such that* $\hat{L}(w) = 0$*:*

$$\begin{aligned} &\left| L(w) - \left[\sigma^2 + \|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w^*\|_2^2\right]\right| \\ \leq &2\|w^*\|_2 \sqrt{\left(1 - \frac{1}{\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \frac{\sigma^2}{\gamma} + 3\epsilon\|w\|_2^2}. \end{aligned} \tag{3.33}$$

We prove that Theorem 25 captures the asymptotic behavior of the minimum-norm interpolator in Theorem 26 below by combining the generalization bound with a norm calculation, recovering the asymptotic formula for this setting computed by Hastie et al. [28] using random matrix theory techniques.

Theorem 26. *Under the model assumptions in* (4.2) *with* $\gamma = d/n > 1$ *and* $\Sigma = I_d$*, there exists* $\epsilon \lesssim \left(\frac{\log(40/\delta)}{n}\right)^{1/2}$ *such that with probability at least* $1 - \delta$*, it holds that*

$$\min_{w: Xw = Y} \|w\|_2^2 \leq (1+\epsilon)\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma - 1}\right). \tag{3.34}$$

Thus, by Theorem 25, we have

$$\begin{aligned} &L(\hat{w}) - \left[\left(1 - \frac{1}{\gamma}\right)\|w^*\|_2^2 + \sigma^2 \frac{\gamma}{\gamma - 1}\right] \\ &\leq \epsilon\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma - 1}\right) + \|w^*\|_2 \sqrt{\epsilon\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma - 1}\right)} \end{aligned} \tag{3.35}$$

where $\hat{w}$ *is the minimal-*ℓ_2 *norm interpolator. If we fix* σ^2, γ *and* $\|w^*\|_2$*, then as* $n \to \infty$

$$L(\hat{w}) \to \left(1 - \frac{1}{\gamma}\right)\|w^*\|_2^2 + \sigma^2 \frac{\gamma}{\gamma - 1} \quad \textit{in probability.} \tag{3.36}$$

Similar to the application in the last section, we also have a lower order $O(n^{-1/4})$ term. It is suboptimal, and we suspect that this is unavoidable for any uniform convergence analysis that only considers the typical size of $\|\hat{w}\|$. Nonetheless, this bound recovers the leading term, and the lower-order term is negligible if we only care about the difference with σ^2.

3.5.2 LASSO

A well-known application of the Gaussian Minmax Theorem is to the sharp analysis of the LASSO in the setting where the covariates are isotropic and Gaussian (see, e.g., Stojnic [64], Amelunxen

et al. [2]). Our optimistic rates bound Theorem 20 recovers a corresponding generalization bound for all predictors w with $\|w\|_1 \leq \|w^*\|_1$, which when specialized to the constrained ERM (i.e. the LASSO solution) recovers these results.

Theorem 27. *Using the notation of Theorem 24, we have with probability at least $1-\delta$ that for all w with $\|w\|_1 \leq \|w^*\|_1$,*

$$\left| \sqrt{L(w) - \sigma^2} - \sqrt{\frac{\gamma \hat{L}(w)}{(1-\gamma)^2}} \right| \leq \epsilon \sqrt{\hat{L}(w)} + \sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma} - \sigma^2\right) + \epsilon \hat{L}(w)} \tag{3.37}$$

provided $\gamma + 2\epsilon/\sqrt{n} < 1$, where

$$\mathcal{K}' := \{u : \|w^* + u\|_1 \leq \|w^*\|_1\} \quad \textit{and} \quad \gamma := \frac{1}{n} \cdot W(\mathcal{K}' \cap S^{n-1})^2.$$

Observe that if $\sigma = 0$ and $\hat{L}(w) = 0$ then we get exact recovery provided $\gamma + 2\epsilon/\sqrt{n} < 1$ which is sharp up to the constant in the confidence term (see, e.g., Amelunxen et al. [2], Chandrasekaran et al. [19]). Informally, exact recovery occurs when $n > \omega^2$, i.e. the number of observations exceeds the statistical dimension. Moreover, we can consider the asymptotic setting where $\sigma = o(1)$ and the proportional scaling limit where γ converges to constant. In this case, it is known (equation 40(a) of Thrampoulidis et al. [65]) that we have $\hat{L}(\hat{w}_{LASSO})/\sigma^2 \to 1 - \gamma$, so the right hand side converges to zero and we have

$$\frac{1}{\sigma^2} L(\hat{w}_{LASSO}) - 1 \to \frac{\gamma}{1-\gamma}.$$

Thus we recover the characterization of the performance of LASSO in this regime [65, 64]. It is possible, as in the OLS setting, to also derive non-asymptotic bounds on $\hat{L}(\hat{w}_{LASSO})$ and therefore obtain non-asymptotic bounds on the performance of the LASSO; we omit the details.

Remark 28. *The Gaussian width of the tangent cone $\mathcal{K}'$ has been sharply characterized in previous work [e.g. 19, 2]. In particular, from the work of Amelunxen et al. [2] we know that if w^* is k-*

sparse,

$$\omega = W(\mathcal{K}' \cap S^{n-1}) \leq W(\textit{cone}(\mathcal{K}') \cap S^{n-1}) \leq \sqrt{d\psi(s/d)}$$

where

$$\psi(\rho) := \inf_{\tau \geq 0} \left\{ \rho(1+\tau^2) + (1-\rho)\sqrt{2/\pi} \int_{\tau}^{\infty} (u-\tau)^2 e^{-u^2/2} du \right\},$$

as well as a corresponding lower bound which characterizes ω*.*

3.5.3 Sharp Rate for OLS

We now zero in on the question of sharp rates for Ordinary Least Squares. Unlike all of the previous sections, in this section we will use tools beyond uniform convergence in order to precisely compute second order terms in the generalization gap. Surprisingly, even though we can match the high probability bound with an exact calculation up to first order term, the existence of certain near-ERM can prevent us from recovering the correct variance term:

Theorem 29. *Under the model assumptions, fix* $\gamma = d/n$ *to be some value in* $(0, 1)$ *and pick any* $c > 0$*. Then there exists another absolute constant* $c' > 0$ *such that for all sufficiently large* n*, with probability at least* $1 - \delta$*, there exists a* $w \in \mathbb{R}^d$ *such that*

$$\hat{L}(w) - \hat{L}(\hat{w}_{\mathrm{OLS}}) \leq c \cdot \frac{\sigma^2}{n^{1/2}}, \tag{3.38}$$

but the population error satisfies

$$L(w) - L(\hat{w}_{\mathrm{OLS}}) \geq c' \cdot \frac{\sigma^2}{n^{1/4}}. \tag{3.39}$$

If we know that $\hat{L}(w) = \hat{L}(\hat{w}_{\mathrm{OLS}})$, then it is necessarily the case that $w = \hat{w}_{\mathrm{OLS}}$ and as we will see, we can get the tightest possible convergence rates. On the other hand, it is not difficult to see that $n\hat{L}(\hat{w}_{\mathrm{OLS}})/\sigma^2$ follows a chi-squared distribution with $n - d$ degrees of freedom, and by

the variance formula of chi-squared distributions, we have

$$\mathrm{Var}(\hat{L}(\hat{w}_{\mathrm{OLS}})) = \frac{2\sigma^4(1-\gamma)}{n}.$$

Consequently, $\hat{L}(\hat{w}_{\mathrm{OLS}})$ can in fact deviate from $\mathbb{E}\hat{L}(\hat{w}_{\mathrm{OLS}}) = \sigma^2(1-\gamma)$ by the order of $\sigma^2/\sqrt{n}$. If we only know that $\hat{L}(w)$ is within the normal range of $\hat{L}(\hat{w}_{\mathrm{OLS}})$, then the above theorem says that the sub-optimal rate of $O(n^{-1/4})$ that we show is actually tight and unavoidable. We can show a similar negative result for the fixed d regime that the convergence cannot be faster than $O(n^{-1/2})$, but as we can see from the last section, using $\|w - w^*\|^2_{\hat{\Sigma}} \approx \sigma^2\gamma$ as the empirical metric instead is enough to recover the parameteric rate $O(1/n)$. This argument fails for the proportional limit regime because the smallest eigenvalue of $\hat{\Sigma}$ is $(1-\sqrt{\gamma})^2$ and so we can only get the larger quantity $\sigma^2\frac{\gamma}{(1-\sqrt{\gamma})^2}$ which fails to capture the first order behavior of $\sigma^2\frac{\gamma}{1-\gamma}$.

Finally, we show how to prove the tight finite sample rate using more direct methods. In fact, we can use the higher order moments of the inverse Wishart distribution [72] to obtain the exact closed-form expressions for both the mean and variance of $L(\hat{w}_{\mathrm{OLS}})$ with any finite value of n and d.

Theorem 30. *Under the model assumptions in* (4.2) *with $d \leq n$, consider the ordinary least square estimator $\hat{w}_{\mathrm{OLS}} = (X^TX)^{-1}X^TY$. It holds that*

$$\begin{aligned} \mathbb{E}L(\hat{w}_{\mathrm{OLS}}) &= \sigma^2\frac{n-1}{n-d-1} \\ \mathrm{Var}(L(\hat{w}_{\mathrm{OLS}})) &= 2\sigma^4\frac{d(n-1)}{(n-d-1)^2(n-d-3)} \end{aligned} \tag{3.40}$$

Hence as $d/n \to \gamma$, it holds that

$$\mathbb{E}L(\hat{w}_{\mathrm{OLS}}) \to \frac{\sigma^2}{1-\gamma} \quad \textit{and} \quad \frac{n}{\sigma^4}\mathrm{Var}(L(\hat{w}_{\mathrm{OLS}})) \to \frac{2\gamma}{(1-\gamma)^3}. \tag{3.41}$$

If d is held constant, as $n \to \infty$, we have

$$n\mathbb{E}[L(\hat{w}_{\mathrm{OLS}}) - \sigma^2] \to \sigma^2 d \quad \textit{and} \quad \frac{n^2}{\sigma^4}\operatorname{Var}(L(\hat{w}_{\mathrm{OLS}})) \to 2d. \tag{3.42}$$

We can also show a matching high probability version based on the Gaussian minimax theorem:

Theorem 31. *Under the model assumptions in* (4.2) *with $d \le n$, consider the ordinary least square estimator $\hat{w}_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$ and denote $\gamma = d/n$. Assume that $\gamma \le 0.999$, then with probability at least $1 - \delta$, it holds that*

$$L(\hat{w}_{\mathrm{OLS}}) - \frac{\sigma^2}{1-\gamma} \lesssim \sigma^2 \sqrt{\frac{\gamma \log(36/\delta)}{n}}.$$

The full proof can be found in the appendix. As we can see from above, the variance of $L(\hat{w}_{\mathrm{OLS}})$ is of order $O(1/\sqrt{n})$ when d is proportional to n, and of order $O(1/n)$ when d is fixed. In both cases, the expectation is close to $\sigma^2/(1-\gamma)$. Theorem 31 shows exactly this and interpolates the two regimes: when γ is of constant order, then we recover the $O(1/\sqrt{n})$ rate, but when d is fixed, $\gamma = O(1/n)$ and so we can accelerate the convergence rate to $O(1/n)$.

3.6 Matrix Sensing

We now consider the problem of matrix sensing: given random matrices $A_1, ..., A_n$ (with i.i.d. standard Gaussian entries) and independent linear measurements $y_1, ..., y_n$ given by $y_i = \langle A_i, X^* \rangle + \xi_i$ where ξ_i is independent of A_i, and $\mathbb{E}\xi = 0$ and $\mathbb{E}\xi^2 = \sigma^2$, we hope to reconstruct the matrix $X^* \in \mathbb{R}^{d_1 \times d_2}$ with sample size $n \ll d_1 d_2$. We will assume X^* has low rank r. In this setting, since the measurement matrices have i.i.d. standard Gaussian entries, the test error is the same as the estimation error:

$$\begin{aligned} L(X) &= \mathbb{E}(\langle A, X \rangle - y)^2 = \mathbb{E}(\langle A, X - X^* \rangle)^2 + \sigma^2 \\ &= \mathbb{E}\|\hat{X} - X^*\|_F^2 + \sigma^2. \end{aligned}$$

Classical approach to this problem is finding the minimum nuclear norm solution:

$$\hat{X} = \underset{X \in \mathbb{R}^{d_1 \times d_2} : \langle A_i, X \rangle = y_i}{\arg\min} \|X\|_*. \tag{3.43}$$

Gunasekar et al. [26] also shows that gradient descent converges to the minimal nuclear norm solution in matrix factorization problems. It is well known that having low nuclear norm can ensure generalization [23, 62] and minimizing the nuclear norm ensures reconstruction [18, 54]. However, if the noise level σ is high, then even the minimal nuclear norm solution $\hat{X}$ can have large nuclear norm. Since our Theorem 12 and 20 can be adapted to different norms as regularizer, our uniform convergence guarantee can be directly applied. It remains to analyze the minimal nuclear norm required to interpolate the measurements y. For simplicity, we assume ξ to be Gaussian in the proof of Theorem 32 below, but we can easily relax it to be sub-Gaussian or sub-exponential.

Theorem 32. *Suppose that $d_1 d_2 > n$, then there exists some $\epsilon \lesssim \sqrt{\frac{\log(32/\delta)}{n}} + \frac{n}{d_1 d_2}$ such that with probability at least $1 - \delta$, it holds that*

$$\min_{\forall i \in [n], \langle A_i, X \rangle = y_i} \|X\|_* \leq \|X^*\|_* + (1 + \epsilon) \sqrt{\frac{n\sigma^2}{d_1 \vee d_2}}. \tag{3.44}$$

The proof can be found in Appendix D.5. Without loss of generality, we will assume $d_1 \leq d_2$ from now on because otherwise we can take the transpose of A and X. Similar to assuming $n/R(\Sigma^{\perp}) \to 0$ in linear regression, we implicitly assume that $n/d_1 d_2 \to 0$ in matrix sensing. Such scaling is necessary for benign overfitting because of the lower bound on the test error for *any* interpolant (for example, Proposition 4.3 of Zhou et al. [77]). We want to point out that the norm calculation above does not take advantage of the fact that X^* has low rank, and it is possible that we can further improve this estimate. The rank of X^* is only used in Theorem 33 below, in which we plug in our norm estimate in Theorem 32 to Theorem 20 and then uses the fact that $\|X^*\|_* \leq \sqrt{r}\|X^*\|_F$.

Theorem 33. *Fix any $\delta \in (0,1)$. There exist constants $c_1, c_2, c_3 > 0$ such that if $d_1 d_2 > c_1 n$, $d_2 > c_2 d_1$, $n > c_3 r(d_1+d_2)$, then with probability at least $1-\delta$ that*

$$\frac{\|\hat{X} - X^*\|_F^2}{\|X^*\|_F^2} \lesssim \frac{r(d_1+d_2)}{n} + \sqrt{\frac{r(d_1+d_2)}{n}} \frac{\sigma}{\|X^*\|_F} + \left(\sqrt{\frac{d_1}{d_2}} + \frac{n}{d_1 d_2}\right) \frac{\sigma^2}{\|X^*\|_F^2}. \tag{3.45}$$

From Theorem 33, we see that when the signal to noise ratio $\frac{\|X^*\|_F^2}{\sigma^2}$ is bounded away from zero, then we obtain consistency $\frac{\|\hat{X}-X^*\|_F^2}{\|X^*\|_F^2} \to 0$ if the following conditions hold:

(A) $r(d_1+d_2) = o(n)$

(B) $d_1 d_2 = \omega(n)$

(C) $d_1/d_2 \to \{0, \infty\}$

This can happen for example when $r = \Theta(1), d_1 = \Theta(n^{1/2}), d_2 = \Theta(n^{2/3})$. As discussed earlier, the second condition is necessary for benign overfitting, and the first consistency condition should be necessary even for regularized estimators. The third condition requires the ground truth matrix X^* to be very tall or very short. It is sufficient for consistency but may not be necessary.

CHAPTER 4

MOREAU ENVELOPE GENERALIZATION THEORY

In this chapter, we consider a more general setting than linear regression that allows the loss function to be a general linear objective. Given any continuous loss function $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ and i.i.d. sample pairs (x_i, y_i) from some data distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathcal{Y}$, we can learn a linear model $(\hat{w}, \hat{b})$ by minimizing the empirical loss $\hat{L}_f$ with the goal of achieving small population loss L_f:

$$\hat{L}_f(w, b) = \frac{1}{n} \sum_{i=1}^{n} f(\langle w, x_i \rangle + b, y_i), \quad L_f(w, b) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[f(\langle w, x \rangle + b, y)]. \tag{4.1}$$

In the above, $\mathcal{Y}$ is an abstract space for the labels. We will let $\mathcal{Y} = \mathbb{R}$ for regression problems and $\mathcal{Y} = \{-1, 1\}$ for binary classification. The setting that we consider in Chapter 3 corresponds to choosing $f(\hat{y}, y) = (\hat{y} - y)^2$. We make the same assumptions on $\mathcal{D}$ as in Chapter 3:

(A) d-dimensional Gaussian features with arbitrary mean and covariance: $x \sim \mathcal{N}(\mu, \Sigma)$

(B) a generic multi-index model: there exist a low-dimensional projection $W = [w_1^*, ..., w_k^*] \in \mathbb{R}^{d\times k}$, a random variable $\xi \sim \mathcal{D}_\xi$ independent of x (not necessarily Gaussian), and an unknown link function $g : \mathbb{R}^{k+1} \to \mathcal{Y}$ such that

$$\eta_i = \langle w_i^*, x \rangle, \quad y = g(\eta_1, ..., \eta_k, \xi). \tag{4.2}$$

Since the link function g is not required to be continuous, the label y can be binary. We can define the projection matrix Q and C_δ the same way as in Theorem 12. Our generalization theory crucially depends on the Moreau envelope, defined as follows.

Definition 5. *The Moreau envelope of a function $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ associated with smoothing parameter $\lambda \in \mathbb{R}_+$ is defined as*

$$f_\lambda(\hat{y}, y) = \inf_u \ f(u, y) + \lambda(u - \hat{y})^2. \tag{4.3}$$

The Moreau envelope is usually viewed as a smooth approximation to the original function f; its minimizer is known as the proximal operator. It plays an important role in convex analysis (see e.g. [8, 15, 57]), but is also useful and well-defined when f is nonconvex.

Stated informally, the core of our Moreau envelope theory is the following inequality: with high probability, it holds uniformly over all (w, b) and $\lambda \geq 0$ that

$$L_{f_\lambda}(w, b) \leq \hat{L}_f(w, b) + \lambda \frac{C_\delta(w)^2}{n}. \tag{4.4}$$

Traditionally, the technique of uniform convergence controls the difference $L_f - \hat{L}_f$, whereas (4.4) controls the difference between $\hat{L}_f$ and the smaller quality L_{f_λ}. However, as we will see in Section 4.1, one can usually obtain some approximation error bound between L_f and L_{f_λ} for any choice of λ depending on the smoothness properties of f. As we can see in the definition (4.3), a larger choice of λ leads to a smaller approximation error while the upper bound in (4.4) becomes larger. Therefore, equation (4.4) can automatically imply a relationship between L_f and $\hat{L}_f$ for different choice of f by choosing λ appropriately to balance the approximation error and generalization error.

As a first example, we consider the square loss $f(\hat{y}, y) = (\hat{y} - y)^2$. It can be easily checked that the Moreau envelope is exactly proportional to itself:

$$f_\lambda(\hat{y}, y) = \inf_u (u - y)^2 + \lambda(u - \hat{y})^2 = \frac{\lambda}{1 + \lambda} f(\hat{y}, y) \tag{4.5}$$

and so $L_{f_\lambda} = \frac{\lambda}{1+\lambda} L_f$. Multiplying $\frac{1+\lambda}{\lambda}$ on both hand sides of (4.4) and choosing $\lambda = \sqrt{\frac{n\hat{L}_f(w,b)}{C_\delta(w)^2}}$ to minimizes the upper bound yields a generalization of our Theorem 12:

$$\begin{aligned} L_f(w, b) &\leq \frac{1 + \lambda}{\lambda} \hat{L}_f(w, b) + (1 + \lambda) \frac{C_\delta(w)^2}{n} \\ &= \left(\sqrt{\hat{L}_f(w, b)} + \sqrt{\frac{C_\delta(w)^2}{n}} \right)^2. \end{aligned}$$

4.1 Lower Bounds on Moreau Envelope

In this section, we study the relationship between L_f and L_{f_λ} for general loss function f. We first introduce the standard definition of Lipschitz and smooth functions.

Definition 6. *A function $f : \mathbb{R} \to \mathbb{R}$ is M-Lipschitz if for all x, y in $\mathbb{R}$,*

$$|f(x) - f(y)| \leq M|x - y|.$$

Definition 7. *A twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ is H-smooth if for all x in $\mathbb{R}$*

$$|f''(x)| \leq H.$$

It is well-known that non-negative and smooth functions are square-root Lipschitz (for example, Lemma 2.1 in Srebro et al. [63]).

Proposition 34. *For a H-smooth and non-negative function $f : \mathbb{R} \to \mathbb{R}$, it holds that*

$$|f'(t)| \leq \sqrt{2Hf(t)} \quad \textit{and} \quad \left|\frac{d}{dt}\sqrt{f(t)}\right| \leq \sqrt{\frac{H}{2}}. \tag{4.6}$$

Therefore, $\sqrt{f}$ is $\sqrt{H/2}$-Lipschitz.

Proof. Since f is H-smooth and non-negative, by Taylor's theorem, for any $s, t \in \mathbb{R}$, we have

$$\begin{aligned} 0 &\leq f(s) \\ &= f(t) + f'(t)(s-t) + \frac{f''(a)}{2}(s-t)^2 \\ &\leq f(t) + f'(t)(s-t) + \frac{H}{2}(s-t)^2 \end{aligned}$$

where $a \in [\min(s,t), \max(s,t)]$. Setting $s = t - \frac{f'(t)}{H}$ yields the desired bound. The second inequality follows by the chain rule. □

If the loss function f is Lipschitz or smooth (or square-root Lipschitz more generally), we can have the following lower bound on the Moreau envelope.

Proposition 35. *If for each $y \in \mathcal{Y}$, $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is M-Lipschitz with respect to the first argument, then for any $\lambda \geq 0$, it holds that*

$$f(\hat{y}, y) \leq f_\lambda(\hat{y}, y) + \frac{M^2}{4\lambda}. \tag{4.7}$$

If for each $y \in \mathcal{Y}$, $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is non-negative and $\sqrt{f}$ is $\sqrt{H}$-Lipschitz with respect to the first argument, then for any $\lambda \geq 0$, it holds that

$$f(\hat{y}, y) \leq \frac{\lambda + H}{\lambda} f_\lambda(\hat{y}, y). \tag{4.8}$$

Proof. If f is Lipschitz, then

$$\begin{aligned}
f_\lambda(\hat{y}, y) &= \inf_u \ f(u, y) + \lambda(u - \hat{y})^2 \\
&\geq \inf_u \ f(\hat{y}, y) - M|u - \hat{y}| + \lambda(u - \hat{y})^2 \\
&= f(\hat{y}, y) + \inf_u \ Mu + \lambda u^2 \\
&= f(\hat{y}, y) - \frac{M^2}{4\lambda}.
\end{aligned}$$

If $\sqrt{f}$ is Lipschitz, then by Lemma 75 in Appendix E

$$\begin{aligned}
f_\lambda(\hat{y}, y) &= \inf_u \ f(u, y) + \lambda(u - \hat{y})^2 \\
&\geq \inf_u \ \left(\sqrt{f(\hat{y}, y)} - \sqrt{H}|u - \hat{y}|\right)_+^2 + \lambda(u - \hat{y})^2 \\
&= \inf_u \sup_{\lambda' \geq 0} \ -\lambda'|u - \hat{y}|^2 + \frac{\lambda'}{H + \lambda'} f(\hat{y}, y) + \lambda(u - \hat{y})^2 \\
&\geq \frac{\lambda}{\lambda + H} f(\hat{y}, y).
\end{aligned}$$

□

Plugging in the bound (4.7) into equation (4.4) and choosing $\lambda = \frac{1}{2}\sqrt{\frac{nM^2}{C_\delta(w)^2}}$ to minimize the upper bound, we obtain

$$\begin{aligned} L_f(w,b) &\le \hat{L}_f(w,b) + \lambda \frac{C_\delta(w)^2}{n} + \frac{M^2}{4\lambda} \\ &= \hat{L}_f(w,b) + M\sqrt{\frac{C_\delta(w)^2}{n}}. \end{aligned} \tag{4.9}$$

Plugging in the bound (4.8) into equation (4.4) and choosing $\lambda = \sqrt{\frac{nH\hat{L}_f(w,b)}{C_\delta(w)^2}}$ to minimize the upper bound, we obtain

$$\begin{aligned} L_f(w,b) &\le \frac{\lambda + H}{\lambda}\hat{L}_f(w,b) + (\lambda + H)\frac{C_\delta(w)^2}{n} \\ &= \left(\sqrt{\hat{L}_f(w,b)} + \sqrt{\frac{HC_\delta(w)^2}{n}}\right)^2. \end{aligned} \tag{4.10}$$

Note that the term $\sqrt{\frac{C_\delta(w)^2}{n}}$ can be interpreted as Rademacher complexity. In comparison to the standard symmetrization and contraction argument [3], our equation (4.9) remove the factor of two and can be shown to be tight [79]. In comparison to the optimistic rate in Srebro et al. [63], our more refined bound (4.10) avoids the hidden constant and logarithmic factor.

We are now ready to state our formal result. As in Theorem 12, we will make two additional mild technical assumptions:

(C) hypercontractivity: there exists a universal constant $\tau > 0$ such that uniformly over all $(w,b) \in \mathbb{R}^d \times \mathbb{R}$, it holds that

$$\frac{\mathbb{E}[f(\langle w,x\rangle + b, y)^4]^{1/4}}{\mathbb{E}[f(\langle w,x\rangle + b, y)]} \le \tau. \tag{4.11}$$

(D) the class of functions on $\mathbb{R}^k \times \mathcal{Y}$ defined below has VC dimension at most h:

$$\{(x,y) \to \mathbb{1}\{f(\langle w,x\rangle + b, y) > t\} : (w,b,t) \in \mathbb{R}^k \times \mathbb{R} \times \mathbb{R}\}. \tag{4.12}$$

We note that the class of functions in assumption (D) is defined on $\mathbb{R}^k \times \mathcal{Y}$ instead of $\mathbb{R}^d \times \mathcal{Y}$. Therefore, h is completely independent of the dimension d and can typically be chosen to be $O(k)$. We state the formal result below.

Theorem 36. *Assume that (A), (B), (C), and (D) holds, and let* $Q = I - W(W^T \Sigma W)^{-1} W^T \Sigma$. *For any* $\delta \in (0,1)$, *let* $C_\delta : \mathbb{R}^d \to [0, \infty]$ *be a continuous function such that with probability at least* $1 - \delta/4$ *over* $x \sim \mathcal{N}(0, \Sigma)$, *uniformly over all* $w \in \mathbb{R}^d$,

$$\langle Qw, x \rangle \leq C_\delta(w). \tag{4.13}$$

Then it holds that

(i) if for each $y \in \mathcal{Y}$, f *is* M*-Lipschitz with respect to the first argument, then with probability at least* $1 - \delta$, *it holds that uniformly over all* $(w, b) \in \mathbb{R}^d \times \mathbb{R}$, *we have*

$$(1 - \epsilon)\, L_f(w, b) \leq \hat{L}_f(w, b) + M \sqrt{\frac{C_\delta(w)^2}{n}} \tag{4.14}$$

(ii) if for each $y \in \mathcal{Y}$, f *is non-negative and* $\sqrt{f}$ *is* $\sqrt{H}$*-Lipschitz with respect to the first argument, then with probability at least* $1 - \delta$, *it holds that uniformly over all* $(w, b) \in \mathbb{R}^d \times \mathbb{R}$, *we have*

$$(1 - \epsilon)\, L_f(w, b) \leq \left(\sqrt{\hat{L}_f(w, b)} + \sqrt{\frac{H\, C_\delta(w)^2}{n}} \right)^2 \tag{4.15}$$

where $\epsilon = O\left(\tau \sqrt{\frac{h \log(n/h) + \log(1/\delta)}{n}} \right)$.

4.2 Upper Bounds on Training Error

We note that we only requires f to be Lipschitz or square-root Lipschitz in Section 4.1. In particular, f does not need to be convex or differentiable. This observation will be useful for our applications in Section 4.4 and 4.5. On the other hand, we will show in this section that if f is

assumed to be convex, then our moreau envelope theory is always asymptotically tight with an appropriate choice of C_δ. This is because if we interpret equation (4.4) as a lower bound on the training error, then a matching upper bound also holds. Finally, we show how to use this upper bound on the training error to derive bounds on the norm of the minimal norm interpolant.

Given a data distribution $\mathcal{D}$ satisfying assumptions (A) and (B), we can define a distribution $\tilde{\mathcal{D}}$ over $(\tilde{x}, \tilde{y}) \in \mathbb{R}^{k+1} \times \mathcal{Y}$ by

$$\tilde{x} = \begin{pmatrix} \tilde{x}_{1:k} \\ \tilde{x}_{k+1} \end{pmatrix} \sim \mathcal{N}(0, I_{k+1}), \quad \xi \sim \mathcal{D}_\xi, \quad \text{and} \quad \tilde{y} = g(W^T\mu + (W^T\Sigma W)^{1/2}\tilde{x}_{1:k}, \xi) \tag{4.16}$$

and we define a mapping $\phi : \mathbb{R}^d \to \mathbb{R}^{k+1}$ by

$$\phi(w) := \begin{pmatrix} (W^T\Sigma W)^{-1/2} W^T \Sigma w \\ \|\Sigma^{1/2} Q w\|_2 \end{pmatrix}. \tag{4.17}$$

We will show that for any $w \in \mathbb{R}^d$, the distribution of $f(\langle w, x\rangle + b, y)$ with $(x, y) \sim \mathcal{D}$ is the same as $f(\langle \phi(w), \tilde{x}\rangle + w^T\mu + b, \tilde{y})$ with $(\tilde{x}, \tilde{y}) \sim \tilde{\mathcal{D}}$. Therefore, we can understand $\tilde{\mathcal{D}}$ as an equivalent low-dimensional distribution for which the concentration of measure can be easily established using assumption (C) and (D).

We note that y only depends on x through $W^T x$ and for any $w \in \mathbb{R}^d$, we have

$$\begin{pmatrix} w^T x \\ W^T x \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} w^T\mu \\ W^T\mu \end{pmatrix}, \begin{pmatrix} w^T\Sigma w & w^T\Sigma W \\ W^T\Sigma w & W^T\Sigma W \end{pmatrix} \right)$$

and so by the conditional distribution of multivariate Gaussian, it is straightforward to check that

$$w^T x \,|\, W^T x = \eta \sim \mathcal{N}\left(w^T\mu + w^T\Sigma W (W^T\Sigma W)^{-1}(\eta - W^T\mu), \|\Sigma^{1/2} Q w\|_2^2 \right).$$

As a result, we can write $W^T x = W^T \mu + (W^T \Sigma W)^{1/2} \tilde{x}_{1:k}$ and

$$\begin{aligned} w^T x &= w^T \mu + w^T \Sigma W (W^T \Sigma W)^{-1/2} \tilde{x}_{1:k} + \|\Sigma^{1/2} Q w\|_2 \tilde{x}_{k+1} \\ &= w^T \mu + \langle \phi(w), \tilde{x} \rangle. \end{aligned}$$

Indeed, the joint distribution of $(w^T \mu + \langle \phi(w), \tilde{x} \rangle, \tilde{y})$ is the same as $(\langle w, x \rangle, y)$. We are now ready to state the upper bound for training error.

Theorem 37. *Under assumptions (A) and (B), let $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be convex with respect to the first argument for any $y \in \mathcal{Y}$. Fix $\mathcal{K} \subset \mathbb{R}^d, \mathcal{B} \subset \mathbb{R}$ to be any bounded convex sets. Suppose that τ is such that with probability at least $1 - \delta/2$ over $x \sim \mathcal{N}(0, \Sigma)$ and $(\tilde{x}_i, \tilde{y}_i)_{i=1}^n$ sampled i.i.d. from $\tilde{\mathcal{D}}$ defined in* (4.16)*, it holds that*

$$\min_{(w,b) \in \mathcal{K} \times \mathcal{B}} \sup_{\lambda \geq 0} \frac{1}{n} \sum_{i=1}^n f_\lambda(\langle \phi(w), \tilde{x}_i \rangle + w^T \mu + b, \tilde{y}_i) - \frac{\lambda}{n} \langle Qw, x \rangle^2 \tag{4.18}$$

Then with probability at least $1 - \delta$, it holds that

$$\min_{(w,b) \in \mathcal{K} \times \mathcal{B}} \hat{L}_f(w, b) \leq \tau. \tag{4.19}$$

We explain why Theorem 37 suggests the tightness of our Moreau envelope theory (4.4) below. Note that the first term in equation (4.18) only depends[1] on w through $\phi(w)$ and we can write it as

$$\min_{(\tilde{w},b) \in \phi(\mathcal{K}) \times \mathcal{B}} \max_{\lambda \geq 0} \frac{1}{n} \sum_{i=1}^n f_\lambda(\langle \tilde{w}, \tilde{x}_i \rangle + w^T \mu + b, \tilde{y}_i) - \frac{\lambda}{n} \left[\max_{w \in \mathcal{K} : \phi(w) = \tilde{w}} \langle Qw, x \rangle^2 \right]. \tag{4.20}$$

It is obvious that for $w \in \mathcal{K}$

$$\langle Qw, x \rangle \leq \max_{w' \in \mathcal{K} : \phi(w') = \phi(w)} \langle Qw', x \rangle \approx \mathbb{E} \left[\max_{w' \in \mathcal{K} : \phi(w') = \phi(w)} \langle Qw', x \rangle \right] := C(\phi(w)) \tag{4.21}$$

1. Using $\phi(w)$, we can recover $w^T \Sigma W$. We can assume μ lies in the span of ΣW by increasing k by 1 and letting $w^*_{k+1} = \Sigma^{-1} \mu$.

by the concentration of Lipschitz functions (we ignore the confidence term δ in C_δ for now). Then by (4.4), it should hold uniformly over all $(w, b) \in \mathcal{K} \times \mathcal{B}$ that

$$\hat{L}_f(w, b) \geq \max_{\lambda \geq 0} L_{f_\lambda}(w, b) - \frac{\lambda C(\phi(w))^2}{n}.$$

On the other hand, equation (4.20) only involves minimization over a low-dimensional space and it implicitly suggests a concentration assumption: if we consider the constrained ERM

$$\hat{w}, \hat{b} = \operatorname*{arg\,min}_{(w,b) \in \mathcal{K} \times \mathcal{B}} \hat{L}_f(w, b),$$

then Theorem 37 implies

$$\begin{aligned}
\hat{L}_f(\hat{w}, \hat{b}) &= \min_{(w,b) \in \mathcal{K} \times \mathcal{B}} \hat{L}_f(w, b) \leq \tau \\
&\approx \min_{(\tilde{w},b) \in \phi(\mathcal{K}) \times \mathcal{B}} \max_{\lambda \geq 0} \mathbb{E} f_\lambda(\langle \tilde{w}, \tilde{x} \rangle + w^T \mu + b, \tilde{y}) - \frac{\lambda C(\tilde{w})^2}{n} \\
&\leq \max_{\lambda \geq 0} \mathbb{E} f_\lambda(\langle \phi(\hat{w}), \tilde{x} \rangle + w^T \mu + \hat{b}, \tilde{y}) - \frac{\lambda C(\phi(\hat{w}))^2}{n} \\
&= \max_{\lambda \geq 0} L_{f_\lambda}(\hat{w}, \hat{b}) - \frac{\lambda C(\phi(\hat{w}))^2}{n}.
\end{aligned}$$

Hence, we have shown (informally) that

$$\hat{L}_f(\hat{w}, \hat{b}) \approx \max_{\lambda \geq 0} L_{f_\lambda}(\hat{w}, \hat{b}) - \frac{\lambda C(\phi(\hat{w}))^2}{n} \tag{4.22}$$

which explains the tightness of our moreau envelope theory (4.4). In the following settings:

(i) $f(\hat{y}, y) = (\hat{y} - y)^2$ with $y \in \mathbb{R}$,

(ii) $f(\hat{y}, y) = (1 - \hat{y}y)_+^2$ with $y \in \{-1, 1\}$,

it can be shown that $f_\lambda = \frac{\lambda}{\lambda+1}$ and so the maximization in (4.22) can be solved exactly. In

particular, by Lemma 75, we have

$$\hat{L}_f(\hat{w}, \hat{b}) \approx \left(\sqrt{L_f(\hat{w}, \hat{b})} - \frac{C(\phi(\hat{w}))}{\sqrt{n}} \right)_+^2 \tag{4.23}$$

and so if $(\hat{w}, \hat{b})$ is an interpolant: $\hat{L}_f(\hat{w}, \hat{b}) \approx 0$, then we should have asymptotically

$$L_f(\hat{w}, \hat{b}) \approx \frac{C(\phi(\hat{w}))^2}{n}. \tag{4.24}$$

The choice of C in (4.21) is known as *Local Gaussian Width*. In fact, all of the results in section 3.5 can be understood by this calculation (e.g., see Example 4 of Zhou et al. [78]). Finally, we note that Theorem 37 is also useful for computing the minimal norm required to achieve zero training error. In particular, we can let $\mathcal{K} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$ and we can find the minimal B such that we can choose $\tau = 0$ in (4.18). This argument is used in the proof of Theorem 14 in Chapter 3 and Theorem 38 in the next section.

4.3 Linear Classification with SVM

In this section, we consider binary classification with the squared hinge loss. As mentioned in the previous section, we can use Theorem 37 to obtain a norm bound for the max-margin classifier.

Theorem 38. *Under assumptions (A) and (B), let $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be the squared hinge loss $f(\hat{y}, y) = (1 - \hat{y}y)_+^2$ with $\mathcal{Y} = \{-1, 1\}$. Let Q be the same as in Theorem 36 and $\Sigma^\perp = Q^T \Sigma Q$. Fix any $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$ such that $Qw^\sharp = 0$ and for some $\rho \in (0, 1)$, it holds that*

$$\hat{L}_f(w^\sharp, b^\sharp) \leq (1 + \rho) L_f(w^\sharp, b^\sharp). \tag{4.25}$$

Then with probability at least $1 - \delta$, for some $\epsilon \lesssim \rho + \log\left(\frac{1}{\delta}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma^\perp)}} + \frac{k}{n} + \frac{n}{R(\Sigma^\perp)}\right)$, it holds that

$$\min_{(w,b)\in\mathbb{R}^{d+1}:\hat{L}_f(w,b)=0} \|w\|_2 \leq \|w^\sharp\|_2 + (1+\epsilon)\sqrt{\frac{nL_f(w^\sharp,b^\sharp)}{\mathrm{Tr}(\Sigma^\perp)}}. \tag{4.26}$$

Combining the above norm bound with Theorem 36, we can follow the same lines of proof of Corollary 15 and 16 to show the following:

Corollary 39. *Under assumptions (A), (B) and (C) with $\mathcal{Y} = \{-1, 1\}$ and $f(\hat{y}, y) = (1 - \hat{y}y)_+^2$, denote $W \in \mathbb{R}^{d\times k}$ by $W = [w_1^*, ..., w_k^*]$ and let $Q = I - W(W^T\Sigma W)^{-1}W^T\Sigma$. Let $\Sigma^\perp = Q^T\Sigma Q$ and fix any $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$ such that $Qw^\sharp = 0$. Then with probability at least $1 - \delta$, for some*

$$\rho \lesssim \tau\sqrt{\frac{k\log(n/k) + \log(1/\delta)}{n}} + \log(1/\delta)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma^\perp)}} + \frac{k}{n} + \frac{n}{R(\Sigma^\perp)}\right),$$

it holds that for any SVM solution of the form $\hat{w}, \hat{b} = \arg\min_{w,b} \hat{L}_f(w, b) + \lambda\|w\|_2^2$ such that $\|\hat{w}\|_2 \geq \|w^\sharp\|_2$, we have

$$L(\hat{w}, \hat{b}) \leq (1+\rho)\left(\sqrt{L(w^\sharp, b^\sharp)} + \|w^\sharp\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma^\perp)}{n}}\right)^2. \tag{4.27}$$

Experimental results on SVM can be found in Figure 5.8 of Chapter 5.1. As theory predicts, the generalization curve looks very similar to linear regression with the square loss.

4.4 Applications to Non-Convex Problems

Moreover, we show that our Theorem 36 can be applied to not only linear regression and max-margin classification. Since we only need $\sqrt{f}$ to be Lipschitz, for any 1-Lipschitz function $\sigma : \mathbb{R} \to \mathbb{R}$, we can consider

(i) $f(\hat{y}, y) = (\sigma(\hat{y}) - y)^2$ when $\mathcal{Y} = \mathbb{R}$

(ii) $f(\hat{y}, y) = (1 - \sigma(\hat{y})y)_+^2$ when $\mathcal{Y} = \{-1, 1\}$.

We can interpret the above loss function f as learning a neural network with a single hidden unit. Common choices of 1-Lipschitz and non-linear σ include

(i) sigmoid activation: $\sigma(\hat{y}) = \frac{1}{1+e^{-\hat{y}}}$ or hyperbolic tangent activation: $\sigma(\hat{y}) = \tanh \hat{y}$

(ii) any Leaky ReLU activation with $|\alpha| < 1$:

$$\sigma(\hat{y}) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0. \end{cases}$$

Indeed, Theorem 36 can be straightforwardly applied to these situations and can be useful if we consider optimally regularized estimators. However, we typically do not expect benign overfitting under these loss functions f for the following simple reason (which is also pointed out in [59]): interpolating with $f(\hat{y}, y) = (\sigma(\hat{y}) - y)^2$ is the same as interpolating with $f(\hat{y}, y) = (\hat{y} - \sigma^{-1}(y))^2$ when σ is invertible (as in the case of sigmoid, hyperbolic tangent and Leaky ReLU activations with $\alpha > 0$), and so we can apply our square loss theory to the distribution over $(x, \sigma^{-1}(y))$. In general, the predictor that minimizes the population loss $\mathbb{E}(\sigma(\langle w, x \rangle + b) - y)^2$ is going to be different to the minimizer of $\mathbb{E}(\langle w, x \rangle + b - \sigma^{-1}(y))^2$.

In contrast, the situation with $\sigma(\hat{y}) = \max\{\hat{y}, 0\}$ is still interesting because σ is not invertible, and this setting is known as ReLU regression. In order to be able to interpolate, we must have $y \geq 0$. If $y > 0$ with probability 1, then $\sigma(\hat{y}) = y$ is the same as $\hat{y} = y$ and we are back to the square loss case. However, if there is some probability mass at $y = 0$, then the minimal norm interpolant can be very different. Interestingly, even though the loss function $f(\hat{y}, y) = (\sigma(\hat{y}) - y)^2$ is 1 square-root Lipschitz, the minimal norm interpolant will typically be inconsistent under this choice of f. As it turns out, the more appropriate loss is

$$f(\hat{y}, y) = \begin{cases} (\hat{y} - y)^2 & \text{if} \quad y > 0 \\ \sigma(\hat{y})^2 & \text{if} \quad y = 0. \end{cases} \tag{4.28}$$

Even though the above choice of f is non-convex and so Theorem 37 cannot be directly applied, it is still possible to compute the minimal norm required to interpolate the data. The loss (4.28) shows up naturally in our norm calculation (e.g., Lemma 82 in the Appendix) by applying CGMT to the quantity $\min_{w:\forall i\in[n],\sigma(\langle w,x_i\rangle)=y_i} \|w\|_2$. Another reason why the loss (4.28) is natural is because it also satisfies the relationship $f_\lambda = \frac{\lambda}{1+\lambda} f$ while ensuring $f(\hat{y}, y) = 0$ is equivalent to $\sigma(\hat{y}) = y$ for any $y \geq 0$. We state the norm bound below, which is completely analogous to Theorem 38. Moreover, since the loss defined in (4.28) is also 1 square-root Lipschitz, it can be combined with Theorem 36 to establish benign overfitting for ReLU regression just like Corollary 39.

Theorem 40. *Under assumptions (A) and (B), let $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be the loss defined in* (4.28) *with $\mathcal{Y} = \mathbb{R}_{\geq 0}$. Let Q be the same as in Theorem 36 and $\Sigma^\perp = Q^T \Sigma Q$. Fix any $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$ such that $Qw^\sharp = 0$ and for some $\rho \in (0, 1)$, it holds that*

$$\hat{L}_f(w^\sharp, b^\sharp) \leq (1+\rho) L_f(w^\sharp, b^\sharp). \tag{4.29}$$

Then with probability at least $1 - \delta$, for some $\epsilon \lesssim \rho + \log\left(\frac{1}{\delta}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma^\perp)}} + \frac{k}{n} + \frac{n}{R(\Sigma^\perp)}\right)$, it holds that

$$\min_{(w,b)\in\mathbb{R}^{d+1}:\hat{L}_f(w,b)=0} \|w\|_2 \leq \|w^\sharp\|_2 + (1+\epsilon)\sqrt{\frac{nL_f(w^\sharp, b^\sharp)}{\mathrm{Tr}(\Sigma^\perp)}}. \tag{4.30}$$

Another example of non-invertible activation is $\sigma(\hat{y}) = |\hat{y}|$. This is related to the problem of phase retrieval. For the problem of phase retrieval, the correct loss should be $f(\hat{y}, y) = (|\hat{y}| - y)^2$ because we can check that it satisfies the relationship $f_\lambda = \frac{\lambda}{1+\lambda} f$. Even though the norm calculation is not as simple as the case of ReLU regression, it should be possible to show benign overfitting results for the phase retrieval problem as well.

4.5 Single Index Model

Finally, we show that our results can be extended to a even more general setting than (4.1), and the more general result can allow us to establish uniform convergence bound for two-layer neural networks with weight sharing. Suppose that we have a parameter space $\Theta \subseteq \mathbb{R}^p$ and a continuous mapping w from $\theta \in \Theta$ to a linear predictor $w(\theta) \in \mathbb{R}^d$. Given a function $f : \mathbb{R} \times \mathcal{Y} \times \Theta \to \mathbb{R}$ that is continuous with respect to the first and third argument and i.i.d. sample pairs (x_i, y_i) drawn from distribution $\mathcal{D}$, we can define the training and population error to be

$$\hat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(\langle w(\theta), x_i\rangle, y_i, \theta) \quad \text{and} \quad L(\theta) = \mathbb{E}[f(\langle w(\theta), x\rangle, y, \theta)]. \tag{4.31}$$

We make the same assumptions (A) and (B) on $\mathcal{D}$. Assumptions (C) and (D) can be naturally extended to the following:

(E) there exists a universal constant $\tau > 0$ such that uniformly over all $\theta \in \Theta$, it holds that

$$\frac{\mathbb{E}[f(\langle w(\theta), x\rangle, y, \theta)^4]^{1/4}}{\mathbb{E}[f(\langle w(\theta), x\rangle, y, \theta)]} \leq \tau. \tag{4.32}$$

(F) bounded VC dimensions: the class of functions on $\mathbb{R}^{k+1} \times \mathcal{Y}$ defined by

$$\{(x, y) \to \mathbb{1}\{f(\langle w, x\rangle + b, y, \theta) > t\} : (w, b, t, \theta) \in \mathbb{R}^{k+1} \times \mathbb{R} \times \mathbb{R} \times \Theta\} \tag{4.33}$$

has VC dimension at most h.

Now we can state the extension of Theorem 36.

Theorem 41. *Suppose that assumptions (A), (B), (E) and (F) hold. Denote $W \in \mathbb{R}^{d\times k}$ by $W = [w_1^*, ..., w_k^*]$ and let $Q = I - W(W^T\Sigma W)^{-1}W^T\Sigma$. For any $\delta \in (0, 1)$, let $C_\delta : \mathbb{R}^d \to [0, \infty]$ be a continuous function such that with probability at least $1 - \delta/4$ over $x \sim \mathcal{N}(0, \Sigma)$, uniformly over*

all $\theta \in \Theta$,

$$\langle Qw(\theta), x\rangle \leq C_\delta(w(\theta)). \tag{4.34}$$

Then it holds that

(i) if for each $\theta \in \Theta$ and $y \in \mathcal{Y}$, f is M_θ-Lipschitz with respect to the first argument and M_θ is continuous in θ, then with probability at least $1-\delta$, it holds that uniformly over all $\theta \in \Theta$, we have

$$(1-\epsilon)\, L(\theta) \leq \hat{L}(\theta) + M_\theta \sqrt{\frac{C_\delta(w(\theta))^2}{n}} \tag{4.35}$$

(ii) if for each $\theta \in \Theta$ and $y \in \mathcal{Y}$, f is non-negative and $\sqrt{f}$ is $\sqrt{H_\theta}$-Lipschitz with respect to the first argument, and H_θ is continuous in θ, then with probability at least $1-\delta$, it holds that uniformly over all $\theta \in \Theta$, we have

$$(1-\epsilon)\, L(\theta) \leq \left(\sqrt{\hat{L}(\theta)} + \sqrt{\frac{H_\theta\, C_\delta(w(\theta))^2}{n}}\right)^2 \tag{4.36}$$

where $\epsilon = O\left(\tau\sqrt{\frac{h\log(n/h)+\log(1/\delta)}{n}}\right)$.

Single-Index Neural Network. In this example, we consider two-layer neural networks with weight sharing in the first layer. More precisely, we have $\theta = (w, a, b) \in \mathbb{R}^{d+2N}$ where N is the number of hidden units. Let $\sigma(x) = \max(x, 0)$ be the ReLU activation function. The hypothesis given by θ is

$$h_\theta(x) := \sum_{i=1}^{N} a_i \sigma(\langle w, x\rangle - b_i)$$

In the context of regression, we can consider the square loss and f is given by

$$f(\langle w, x\rangle, y, \theta) := (h_\theta(x) - y)^2 \tag{4.37}$$

and in the context of classification, we can consider the squared hinge loss and f is given by

$$f(\langle w, x\rangle, y, \theta) := (1 - h_\theta(x)y)_+^2. \tag{4.38}$$

The above f is well-defined because h_θ depends on x only through $\langle w, x\rangle$. If we write $\hat{y} = \langle w, x\rangle$, then we have

$$f(\hat{y}, y, \theta) = \left(\sum_{i=1}^N a_i\sigma(\hat{y} - b_i) - y\right)^2 \qquad \text{or} \qquad f(\hat{y}, y, \theta) = \left(1 - \sum_{i=1}^N a_i\sigma(\hat{y} - b_i)y\right)_+^2.$$

Without loss of generality, we can assume that $b_1 \leq ... \leq b_N$, then it is easy to see that $\sqrt{f}$ is $\max_{j\in[N]}\left|\sum_{i=1}^j a_i\right|$ Lipschitz. Applying Theorem 41, we obtain the following corollary.

Corollary 42. *Fix an arbitrary norm $\|\cdot\|$ and consider f as defined in* (4.37) *or* (4.38)*. Assume that the data distribution $\mathcal{D}$ satisfy (A), (B), and (E). Without loss of generality, we further assume that b_i are sorted, then with probability at least $1 - \delta$, it holds that uniformly over all $\theta = (w, a, b) \in \mathbb{R}^{d+2N}$, we have*

$$(1 - \epsilon)\, L(\theta) \leq \left(\sqrt{\hat{L}(\theta)} + \frac{\max_{j\in[N]}\left|\sum_{i=1}^j a_i\right| \, \|Qw\| \left(\mathbb{E}\|x\|_* + \epsilon'\right)}{\sqrt{n}}\right)^2 \tag{4.39}$$

where ϵ is the same as in Theorem 41 with $h = O((k+N)\log(k+N))$ and the norm concentration term is $\epsilon' = O\left(\sup_{\|u\|\leq 1}\|u\|_\Sigma\sqrt{\log(1/\delta)}\right)$.

The above theorem says given a network $h_\theta(x) := \sum_{i=1}^N a_i\sigma(\langle w, x\rangle - b_i)$, after sorting the b_i's, a good complexity measure to look at is $\max_{j\in[N]}\left|\sum_{i=1}^j a_i\right| \cdot \|Qw\|$ and equation (4.39) precisely quantify how the complexity of a network controls generalization.

CHAPTER 5

UNIVERSALITY

In this chapter, we discuss the extent to which the Gaussian feature assumption made in Chapter 3 and 4 can be relaxed. As we see in Chapter 2, the proof of Proposition 9, which is an important special case of our optimistic rate theory Theorem 12 and Theorem 36 in Chapter 3 and 4, only depends on the omniscient risk estimator introduced in Section 2.1. Rigorous version of the omniscient risk estimator can be proved using random matrix theory while assuming $x = \Sigma^{1/2}z$ and z has i.i.d. coordinates with zero mean, unit variance, and bounded 12th absolute central moment (for example, see Wu and Xu [75] and Hastie et al. [28]). In similar settings, it is possible to directly establish the universality of Gaussian Minimax Theorem with Lindeberg's method [27]. In addition, even though the eigenfunctions for kernel ridge regression are nearly always not independent, many works have suggested that its asymptotic test error should only depend on the spectrum and we can usually define an equivalent Gaussian model. This is known as the Gaussian equivalence conjecture and has been rigorously proven in the context of inner product kernels and random features model (for example, Hu and Lu [29]).

In Section 5.1, we provide empirical evidence that our generalization theory Theorem 36 should continue to hold when $x = \Sigma^{1/2}z$ and z has i.i.d. coordinates even though the coordinate distribution is not necessarily Gaussian. However, in section 5.2, we show that there is a relatively simple setting where the coordinates of x are dependent (but uncorrelated), and universality fails in a meaningful way.

5.1 Numerical Experiments

Feature Distribution. In this section, the marginal distribution of x is always given by $x = \Sigma^{1/2}z$, where z is a random vector with i.i.d. coordinates that have mean 0 and variance 1. The coordinate distributions of z that we consider in the simulations include:

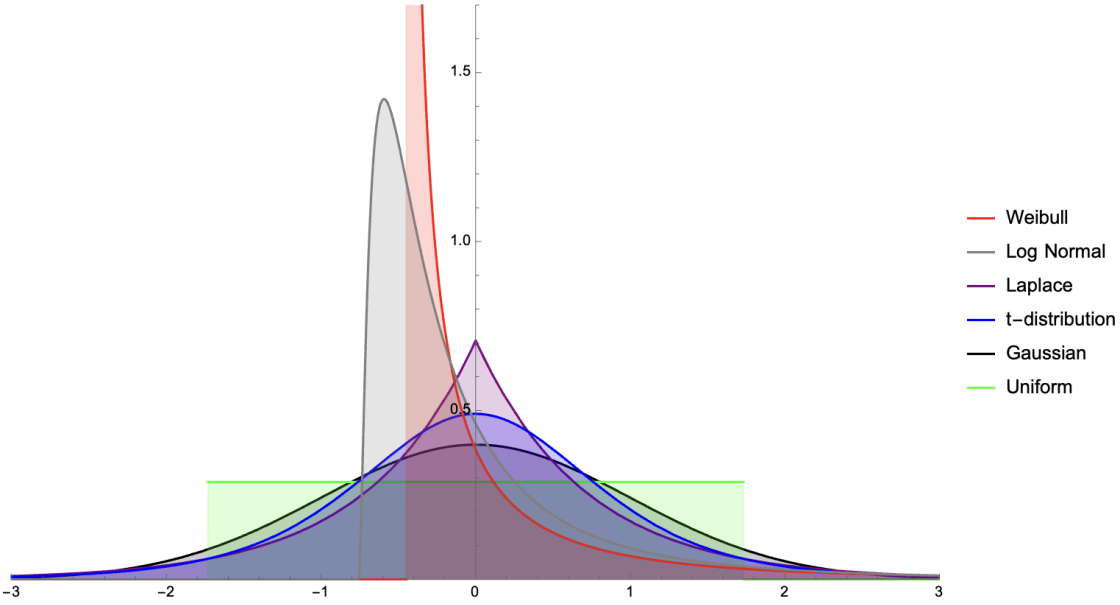

Figure 5.1: Probability density plot for coordinate distributions of z.

- standard Gaussian distribution with density $p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$,
- uniform distribution over $[-\sqrt{3}, \sqrt{3}]$,
- Laplace distribution with density $p(z) = \frac{1}{2b} e^{-\frac{|z|}{b}}$ and scale parameter $b = \frac{1}{\sqrt{2}}$.

We also consider discrete distributions:

- Rademacher distribution with equal chance of being -1 or 1,
- Poisson distribution with rate parameter 1 and probability mass function $\Pr(\tilde{z} = k) = \frac{e^{-1}}{k!}$ (we take $z = \tilde{z} - 1$ to center the mean),

and heavy-tailed distributions:

- Student's t-distribution
  - t-distribution with 5 degrees of freedom has density $p(\tilde{z}) = \frac{8}{3\sqrt{5}\pi\left(1+\frac{\tilde{z}^2}{5}\right)^3}$
  - It has variance $\frac{5}{3}$ and so we let $z = \sqrt{\frac{3}{5}}\tilde{z}$. It is symmetric, unbounded and has finite fourth moment. However, moments of order 5 or higher do not exist.

- Weibull distribution
  - Weibull distribution with scale parameter $\lambda = 1$ and shape parameter $k = 0.5$ has density $p(\tilde{z}) = \frac{e^{-\sqrt{\tilde{z}}}}{2\sqrt{\tilde{z}}} \mathbb{1}_{\{\tilde{z} \geq 0\}}$. It has mean 2 and variance 20 and so we take $\tilde{z} = \frac{z-2}{\sqrt{20}}$
- Log-Normal distribution
  - the distribution of e^Z, where Z follows the standard Gaussian distribution. It has mean $\sqrt{e}$ and variance $e(e-1)$, and so we can choose $z = \frac{e^Z - \sqrt{e}}{\sqrt{e(e-1)}}$

Covariance Matrix and Scaling. For simplicity, we choose Σ to be diagonal and consider

- Isotropic features $\Sigma = I_d$ in the proportional scaling $(n = 300, d = 350)$
- Junk features in the over-parameterized scaling $(n = 300, d = 3000)$

$$\Sigma_{kk} = \begin{cases} 1 & \text{if } k = 1, 2, 3 \\ 0.05^2 & \text{otherwise} \end{cases}$$

- Non-benign features in the over-parameterized scaling $(n = 300, d = 3000)$

$$\Sigma_{kk} = \begin{cases} 1 & \text{if } k = 1, 2, 3 \\ \frac{1}{k^2} & \text{otherwise} \end{cases}$$

As we see in Example 1 of Chapter 2, the junk features setting is known to satisfy the benign overfitting conditions [7, 77], by which the minimal ℓ_2-norm interpolator is consistent. In contrast, our Example 4 in Chapter 2 also shows that overfitting is not benign in the second case, but our theory from section 3.3 shows that the optimally-tuned ridge regression can be consistent.

5.1.1 Linear Regression

We fit linear models to minimize the square loss with ℓ_1 and ℓ_2 penalty. For simplicity, we ignore the intercept term in this section, but we will consider models with intercept in the context of linear classification. We can obtain many data distributions by combining the different options below:

Conditional Distribution of y. Let

$$w^* = (1.5, 0, ..., 0)$$

$$\xi \sim \mathcal{N}(0, 0.5)$$

and consider

- a well-specified linear model:

$$y = \langle w^*, x \rangle + \xi$$

- a mis-specified model:

$$y = \underbrace{\langle w^*, x \rangle}_{\text{linear signal}} + \underbrace{|x_1| \cdot \cos x_2}_{\text{non-linear term}} + \underbrace{x_3 \cdot \xi}_{\text{heteroscedasticity}}$$

The second model does not satisfy the classical assumptions for linear regression because the Bayes predictor

$$\mathbb{E}[y|x] = \langle w^*, x \rangle + |x_1| \cdot \cos x_2$$

is non-linear and the variance of the residual also depends on x_3. Even though statistical inference can be challenging for models like this, we can hope to learn a model that competes with the optimal linear predictor (which is not necessarily the same as w^*) in terms of prediction error.

Though our theory is restricted to Gaussian features, we conjecture that it can be extended to a more general class of distributions and we use numerical simulations to confirm our conjecture.

Ridge Regression

1. Isotropic features: similar to the choice of F in equation (3.29) in Chapter 3, we can choose C_δ by the simple Cauchy-Schwarz bound

$$\langle Qw, x\rangle \le \|Qw\|_2 \cdot \|x\|_2 \approx \sqrt{d}\|Qw\|_2$$

resulting in the following bound

$$L_f(w) \le (1+o(1))\left(\sqrt{\hat{L}_f(w)} + \sqrt{\frac{d}{n}}\cdot\|Qw\|_2\right)^2 \tag{5.1}$$

2. Junk and non-benign features: similar to the analysis in section 3.2, we can choose C_δ by

$$\langle Qw, x\rangle \le \|w\|_2 \cdot \|Q^T x\|_2 \approx \|w\|_2\sqrt{\mathrm{Tr}(\Sigma^\perp)}$$

resulting in the following bound

$$L_f(w) \le (1+o(1))\left(\sqrt{\hat{L}_f(w)} + \|w\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma^\perp)}{n}}\right)^2 \tag{5.2}$$

In all of the experiments, we use a constant close to 1 to replace the $1+o(1)$ factor in our generalization bounds. Note that (5.2) can be interpreted in terms of Rademacher complexity:

$$\begin{aligned}\mathbb{E}_{\substack{x_1,\ldots,x_n\sim\mathcal{D}\\ s\sim\mathrm{Unif}(\{\pm1\}^n)}}\left[\sup_{\|w\|_2\le B}\left|\frac{1}{n}\sum_{i=1}^n s_i\langle w, Q^T x_i\rangle\right|\right] &= \frac{B}{n}\cdot\mathbb{E}_{\substack{x_1,\ldots,x_n\sim\mathcal{D}\\ s\sim\mathrm{Unif}(\{\pm1\}^n)}}\left[\left\|\sum_{i=1}^n s_i Q^T x_i\right\|_2\right]\\ &\le B\cdot\sqrt{\frac{\mathrm{Tr}(\Sigma^\perp)}{n}}\end{aligned}$$

The last inequality holds generally for any distribution with $\mathbb{E}_{x\sim\mathcal{D}}[xx^T] = \Sigma$ by Cauchy-Schwarz inequality. In our examples, $x = \Sigma^{1/2}z$ and z is scaled to satisfy $\mathbb{E}[zz^T] = I_d$. Therefore,

we will use equation (5.1) and (5.2) even for non-Gaussian data.

LASSO Regression Similar to the case of ridge regression, we use the analogy to Rademacher complexity to extend our theory to the ℓ_1 case. Since we can no longer bound the ℓ_∞ norm of a sum using the Cauchy-Schwarz inequality, it is easier to directly work with the empirical Rademacher complexity (which also should be similar to the expected Rademacher complexity in the settings that we consider)

$$\frac{\|w\|_1}{n} \cdot \mathbb{E}_{s\sim \text{Unif}(\{\pm 1\}^n)} \left[\left\| \sum_{i=1}^{n} s_i Q^T x_i \right\|_\infty \right]$$

and we can estimate the expected norm by

$$\frac{1}{B} \sum_{k=1}^{B} \left\| \sum_{i=1}^{n} s_{k,i} Q^T x_i \right\|_\infty$$

for a large value of B and $s_1, ..., s_B$ sampled independently from Unif$(\{\pm 1\}^n)$. In our implementation, $s_1, ..., s_B$ are fresh samples each time the risk bound is computed. To summarize, we use the following expression for the calculation of risk bound:

1. Isotropic features:

$$\left(\sqrt{\hat{L}_f(w)} + \|Qw\|_1 \cdot \frac{1}{nB} \sum_{k=1}^{B} \left\| \sum_{i=1}^{n} s_{k,i} x_i \right\|_\infty \right)^2 \tag{5.3}$$

2. Junk and non-benign features:

$$\left(\sqrt{\hat{L}_f(w)} + \|w\|_1 \cdot \frac{1}{nB} \sum_{k=1}^{B} \left\| \sum_{i=1}^{n} s_{k,i} Q^T x_i \right\|_\infty \right)^2 \tag{5.4}$$

which are analogous to (5.1) and (5.2).

We note that it is important to use the Rademacher complexity to extend to non-Gaussian

features in the ℓ_1 case, rather than a bound similar to $\frac{\|w\|_1 \mathbb{E}\|x\|_\infty}{\sqrt{n}}$. Empirically, the latter is too small to provide a valid upper bound on the test loss. This is because $\|x\|_\infty$ is deterministic for distributions like the Rademacher distribution, while the random signs in the definition of Rademacher complexity allows a tail behavior more similar to Gaussian and so we can regain a log factor in the norm component.

For both ridge and LASSO regression, risk curves measured in the square loss are shown in three figures corresponding to the different data covariances. Within each figure, there are 16 subplots corresponding to the different combinations of one of the eight feature distributions and label generating process (well-specified vs mis-specified) as defined at the beginning of the section.

Similar to the situation in the rest of the experiments, the training error is close to 0 with sufficiently small regularization, and the confidence bands are wider with heavy-tailed distributions. Also, the null risk and the Bayes risk are different across different feature distributions when there is model misspecification (see the calculation in the appendix for more details).

The plots for isotropic, junk and non-benign features in the ridge regression setting can be found in figures 5.2, 5.3 and 5.4, respectively. Generally speaking, the experiments confirm the tightness and wide applicability of our generalization guarantees. The specific feature distribution and model misspecification do not seem to affect the shape of test error curve.

- Figure 5.2: Ridge regression with isotropic data $(n = 300, d = 350)$. As proved by theorem 26, the risk bound (5.1) follows the test error curve closely. This is true even in the non-Gaussian and mis-specified settings. Note that we do not have benign-overfitting because we are in the proportional scaling regime with d close to n, and the population risk of the minimal-ℓ_2 norm interpolator is even worse than the null-risk (more significantly so with misspecification). The optimally-tuned ridge regression has risk better than the null risk, but it is still far from the Bayes risk because the consistency result of optimally-tuned ridge regression assumes $\mathrm{Tr}(\Sigma)/n \to 0$.

- Figure 5.3: Ridge regression with junk features $(n = 300, d = 3000)$. In the junk features

setting, as predicted in Theorem 16, the test error curve is essentially flat once the regularization is small enough to fit the signal, and we get nearly optimal population risk as long as we do not over-regularize the predictor. The test error curve can be expected to be more flat with increasing d. This phenomenon is also consistent across different feature distributions and label generating processes. Our bound (5.2) closely tracks the performance of ridge regression along the entire regularization path.

- Figure 5.4: Ridge regression with non-benign features $(n = 300, d = 3000)$. In the non-benign features setting, as proved by Corollary 17, the optimally-tuned ridge regression achieves nearly optimal prediction risk. Our risk bound is tight up to the point up to the point where the test error starts to increase. As expected, the minimal norm interpolator fails to achieve consistency even though we are in the overparameterized regime. Once again, the distribution and model misspecification has no effect on the shape of the test error curve.

The plots for isotropic, junk, and non-benign features in the LASSO regression setting can be found in figures 5.5, 5.6 and 5.7. The risk bounds in the ℓ_1 case are not as tight as in the ℓ_2 case because they are only expected to be tight in certain parts of the entire regularization path. Similar results and experiments were obtained by Wang et al. [74], Donhauser et al. [21].

- Figure 5.5: LASSO regression with isotropic data $(n = 300, d = 350)$. Contrary to the inconsistency of optimally-tuned ridge regression in this setting, the regularized LASSO estimator can achieve nearly optimal population risk thanks to sparsity. The risk bound (5.3) appears to be valid and sufficient for the consistency of optimal LASSO in the distributions that we consider, though it is not very tight for interpolation. Recall that the minimal-ℓ_1 norm interpolator suffers from an exponentially slow convergence rate when $d = n^\alpha$ [74] and observe that the population risk of the minimal-ℓ_1 norm interpolator is again worse than the null-risk.

- Figure 5.6: LASSO regression with junk features $(n = 300, d = 3000)$. Similar to the

isotropic setting, the regularized LASSO can achieve nearly optimal prediction risk and the risk bound (D.17) is sufficient to explain this phenomenon. Once again, the data distribution and model misspecification appear to have no effect on the shape of the test error curve. It is theoretically possible to use a nearly identical risk bound to show the consistency of minimal-ℓ_1 norm interpolator when n is large and d is super-exponential in n [33], but as we can see, $n = 300$ and $d = 3000$ is not quite large enough yet. On the other hand, overfitting is more benign than what theory predicts, suggesting a better analysis may yield a weaker condition required for consistency.

- Figure 5.7: LASSO regression with non-benign features $(n = 300, d = 3000)$. Though the population risk and the associated risk bound of regularized LASSO can be quite close to the Bayes risk, overfitting with minimal-ℓ_1 norm interpolator does not appear to be benign (and there is no existing theoretical result suggesting that consistency is possible with a larger n or d). In particular, its ℓ_1 norm increases much more quickly than the junk-features case. Though the (D.17) is not tight throughout the entire regularization path, it is still a valid upper bound on the test error across different feature distributions and label generating processes.

5.1.2 Linear Classification

Similarly, we fit linear models to minimize the squared hinge loss with ℓ_2 and ℓ_1 penalty. We can consider the same feature distributions and data covariance structure as in the preceding section. For faster computation (because margin classifiers can be slower to compute than regressors), we take $k = 1$, and $n = 100, d = 120$ in the proportional scaling and $n = 100, d = 2000$ in the overparameterized scaling. The label y is generated by the following model:

$$\eta = \langle w^*, x \rangle + b^*, \quad \Pr(y = 1 \,|\, x) = 1 - \Pr(y = -1 \,|\, x) = g(\eta)$$

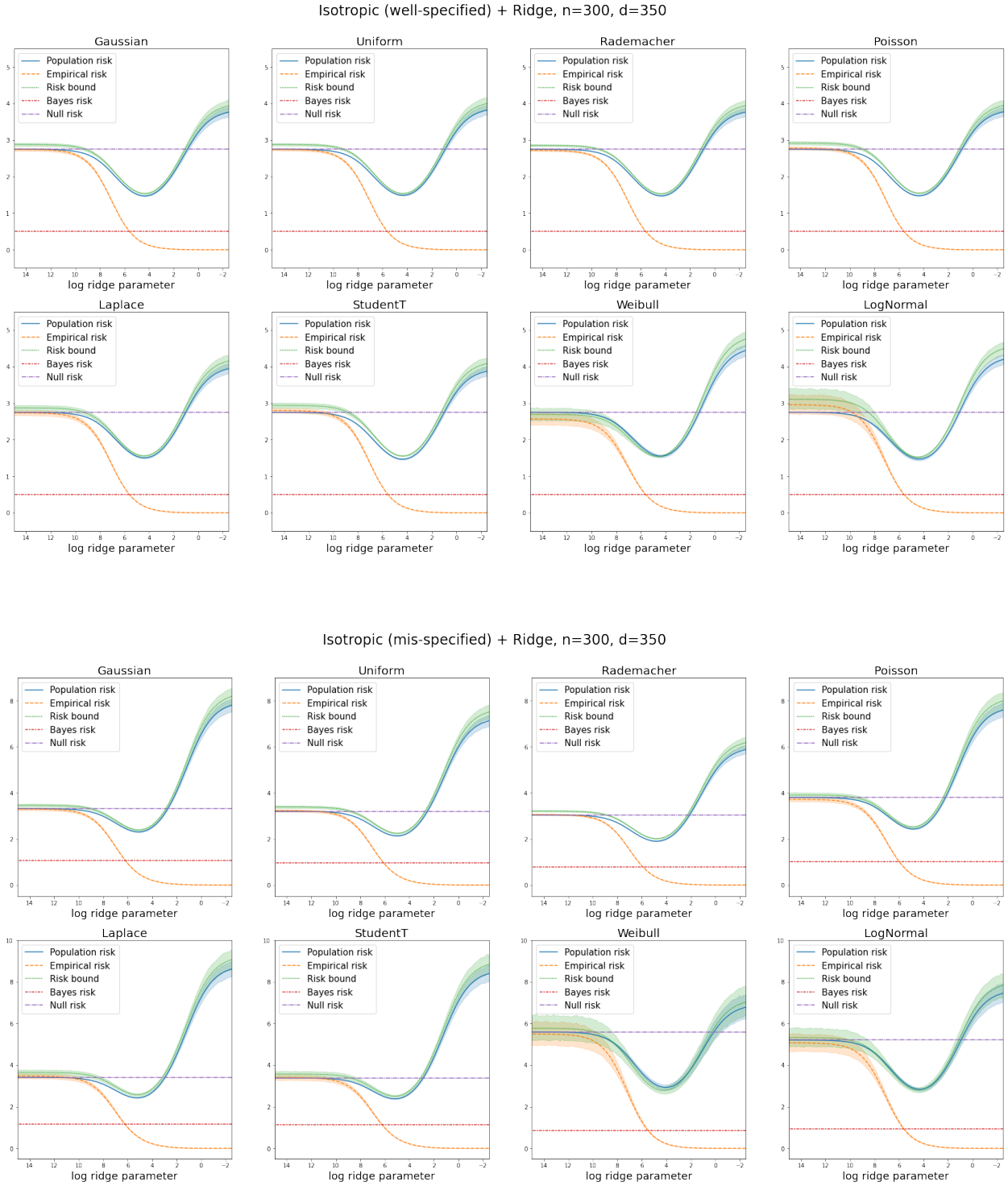

Figure 5.2: Ridge regression with isotropic data $(n = 300, d = 350)$.

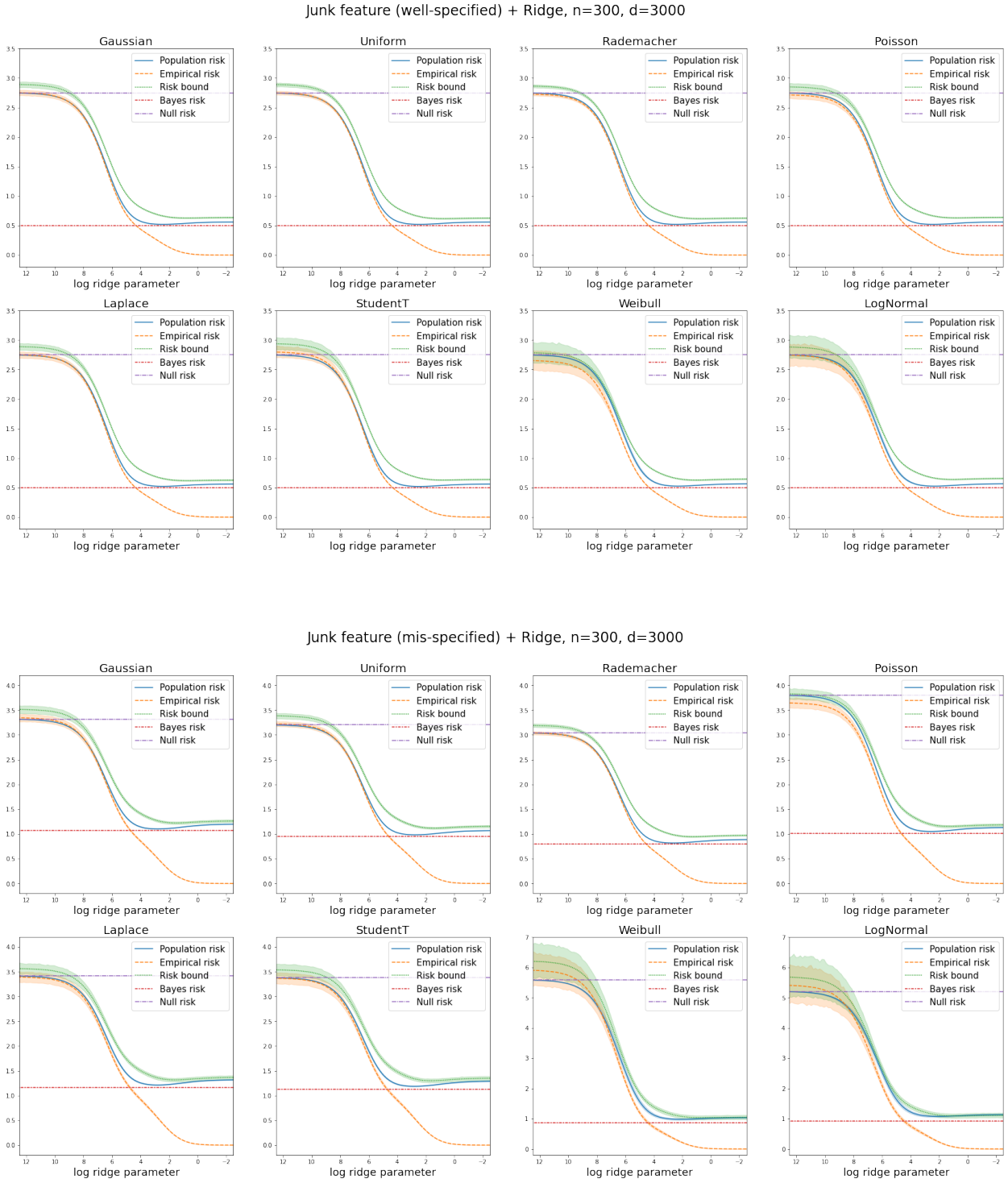

Figure 5.3: Ridge regression with junk features $(n = 300, d = 3000)$.

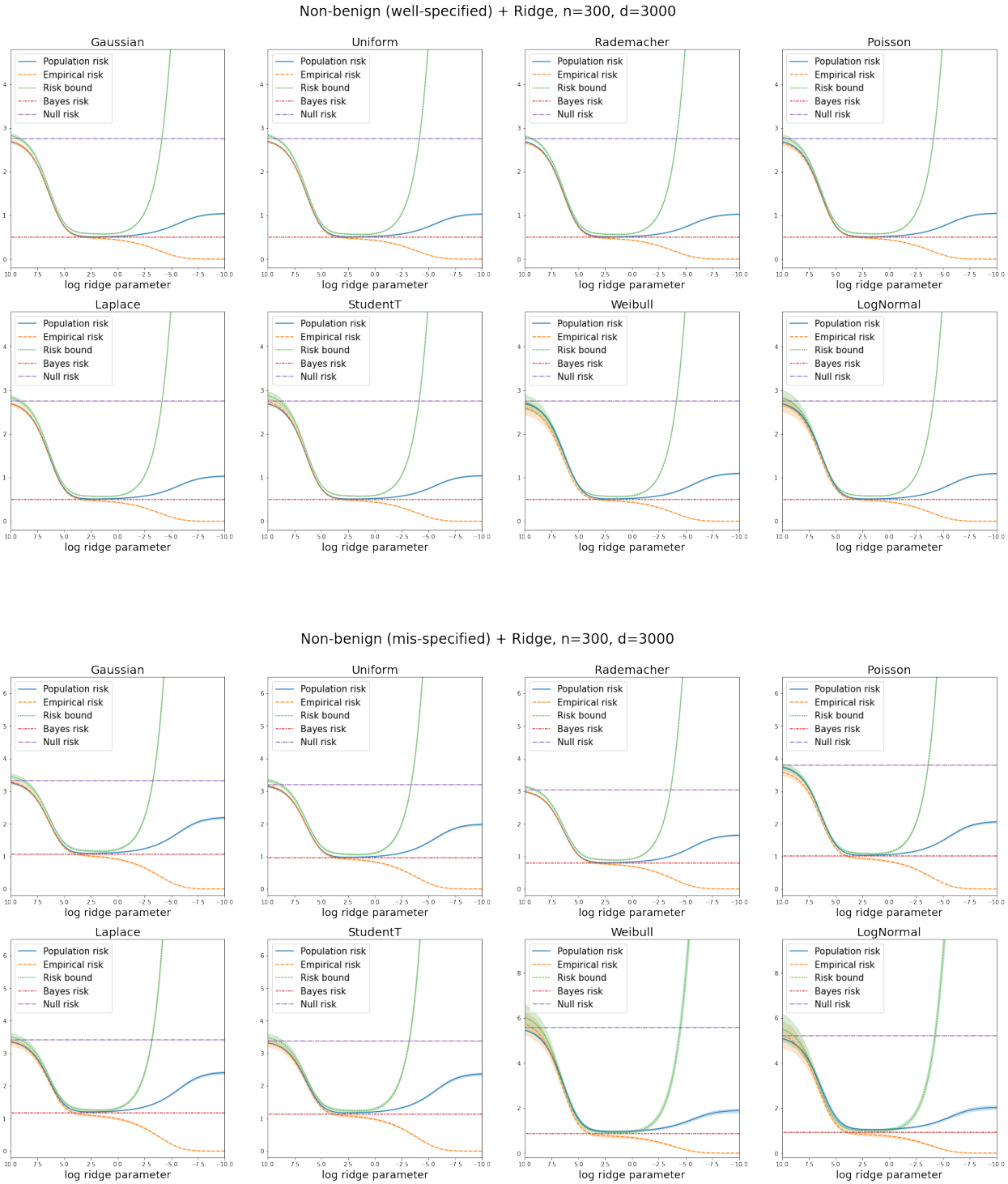

Figure 5.4: Ridge regression with non-benign features $(n = 300, d = 3000)$.

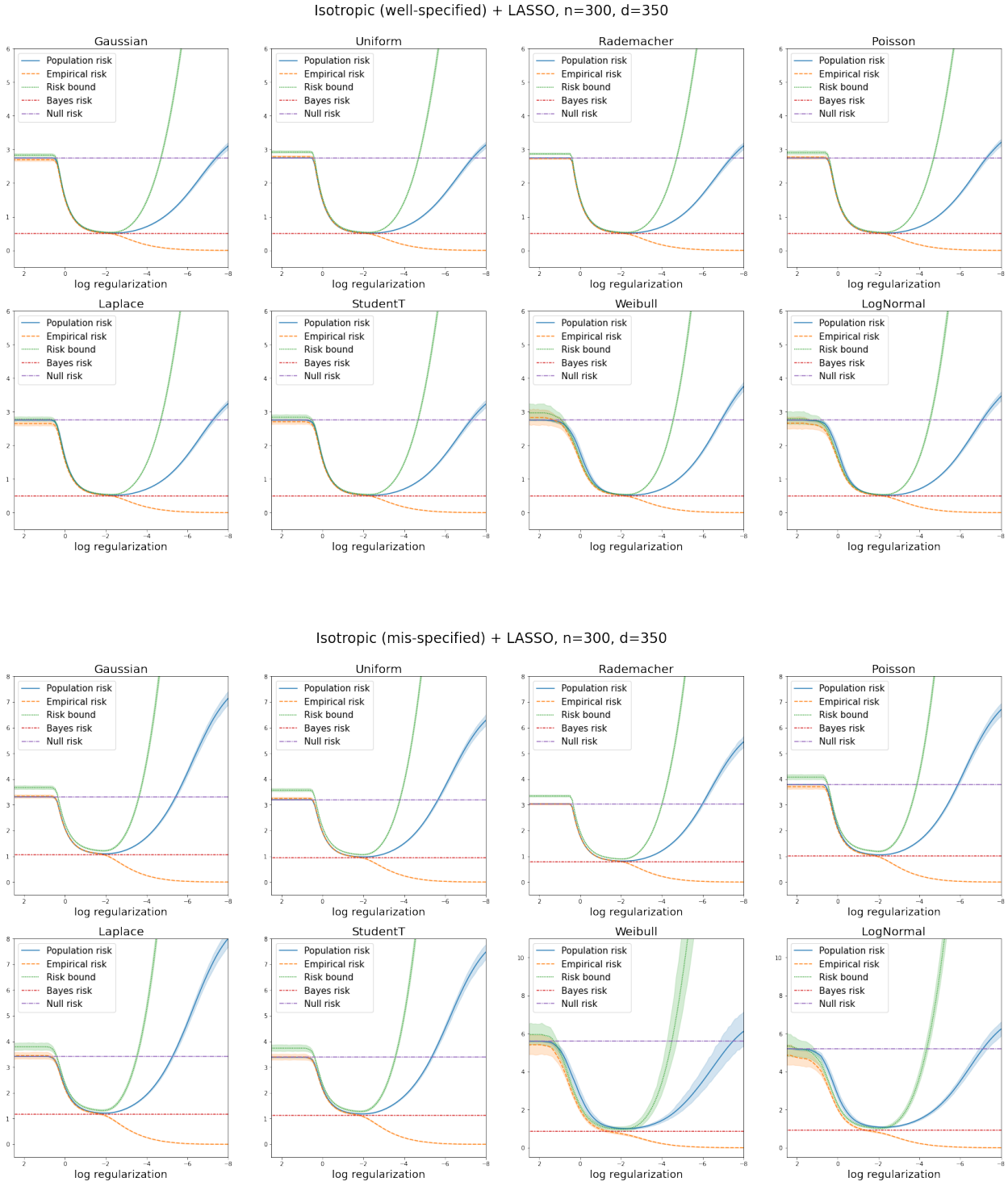

Figure 5.5: LASSO regression with isotropic data $(n = 300, d = 350)$.

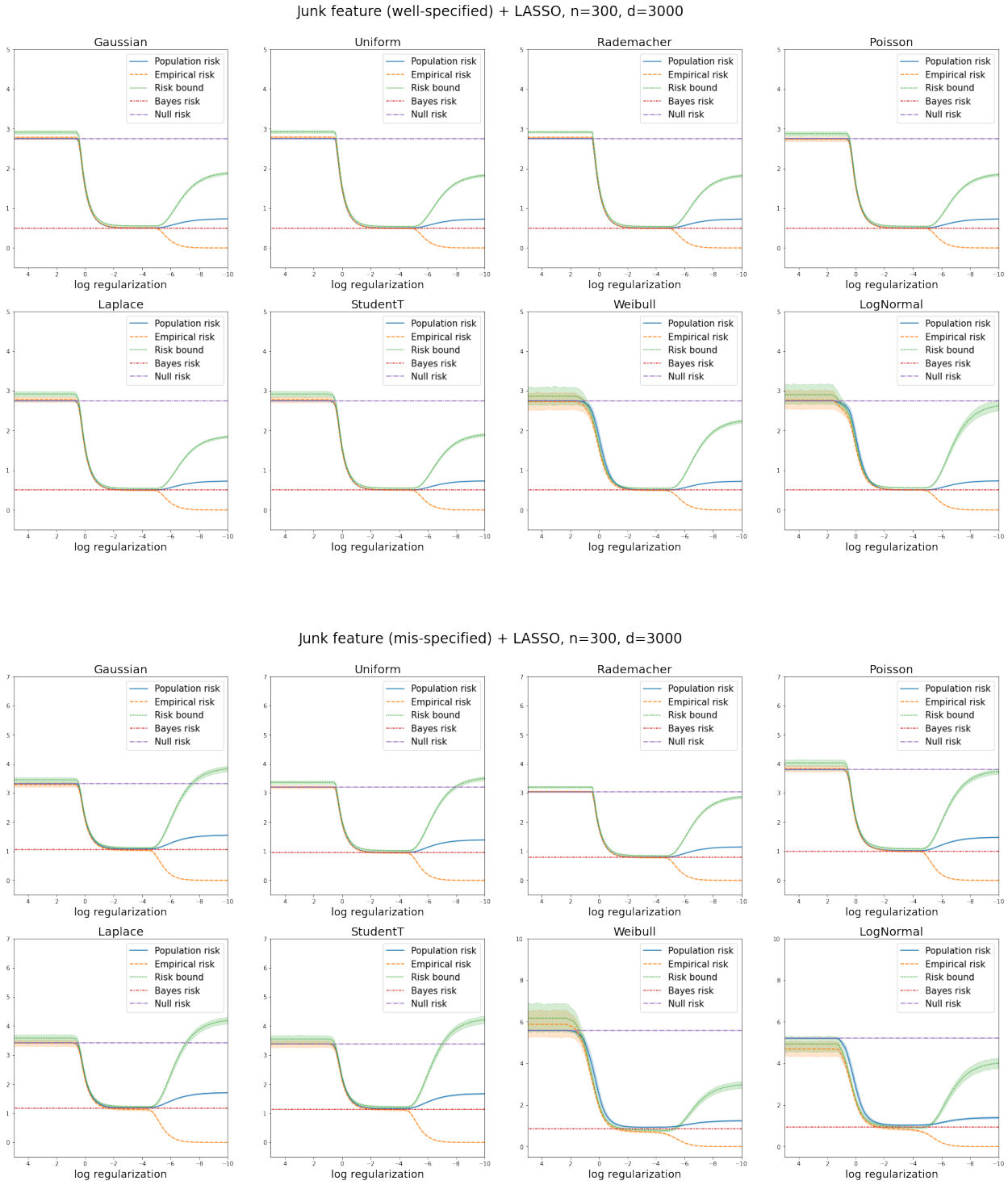

Figure 5.6: LASSO regression with junk features $(n = 300, d = 3000)$.

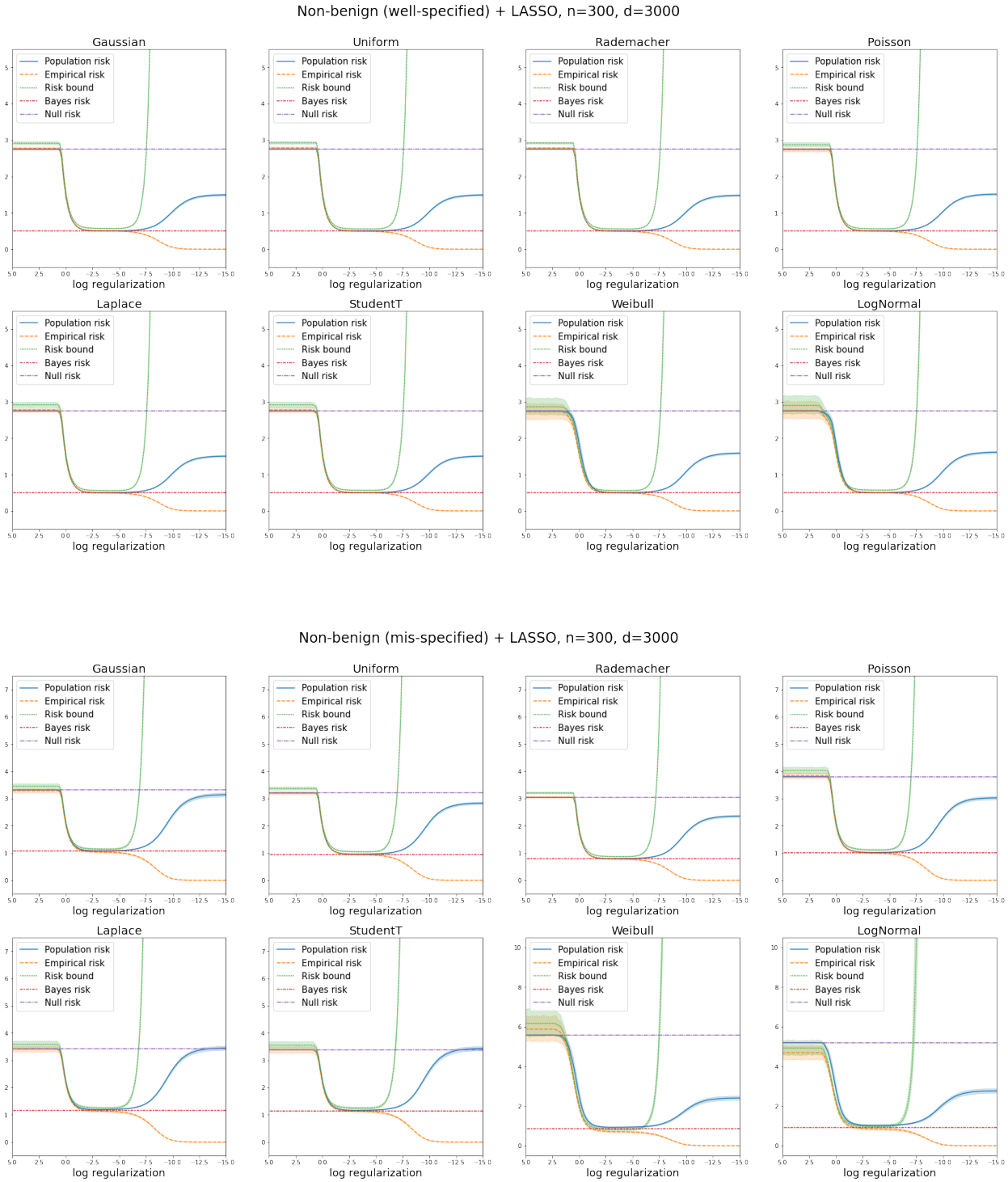

Figure 5.7: LASSO regression with non-benign features $(n = 300, d = 3000)$.

where $g : \mathbb{R} \to [0, 1]$ is the logistic link function. Since we use the squared hinge loss for learning (which is not the negative log-likelihood function), the linear model that we learn is not necessarily well-calibrated and so this can also be considered as a misspecified setting. Therefore, we will only consider one label generating process in the classification context. Finally, by our Moreau envelope theory, we can use completely the same risk bounds for ℓ_2 and ℓ_1 margin classifiers.

The plots for ℓ_2 and ℓ_1 margin classifiers can be found in Figures 5.8 and 5.9. Each figure contain three subplots, and each subplot corresponds to one of the data covariance and contains the risk curves measured in squared hinge loss for the eight feature distributions.

ℓ_2-Margin Classifiers. As in the regression case, overfitting is not benign when the features are isotropic and the population risk of ℓ_2 max-margin classifier can be worse than the null risk. The risk bounds tightly control the test errors across different feature distributions. The difference between risk bound and the actual test error is larger when the feature distribution is heavy-tailed, but the confidence interval is also wider due to the relatively small sample size.

In the junk feature setting, the under-regularized part of the regularization path is essentially flat for all feature distributions. Overall, the experimental result is very similar to 5.3, as predicted by our theory in Section 4.3. The non-benign case is also similar to 5.4 except that the U-shape curve is quite narrower near the optimal amount of regularization.

ℓ_1-Margin Classifiers. In each of the subplots, the risk bound is tight only up to a certain point before the ℓ_1 norm starts to increase quite a lot, leading to loose bound near interpolation. However, the risk bound is tight enough to establish consistency of optimally-tuned predictor in the junk and non-benign features setting. Again, the population risk of ℓ_1 max-margin classifier can be worse than the null risk even in the junk features setting. Observe that different distributions do not seem to change the shape of generalization curve, and there is an interesting multiple descent phenomenon in the non-benign feature case, which has already been discovered in previous literature [20, 37, 38].

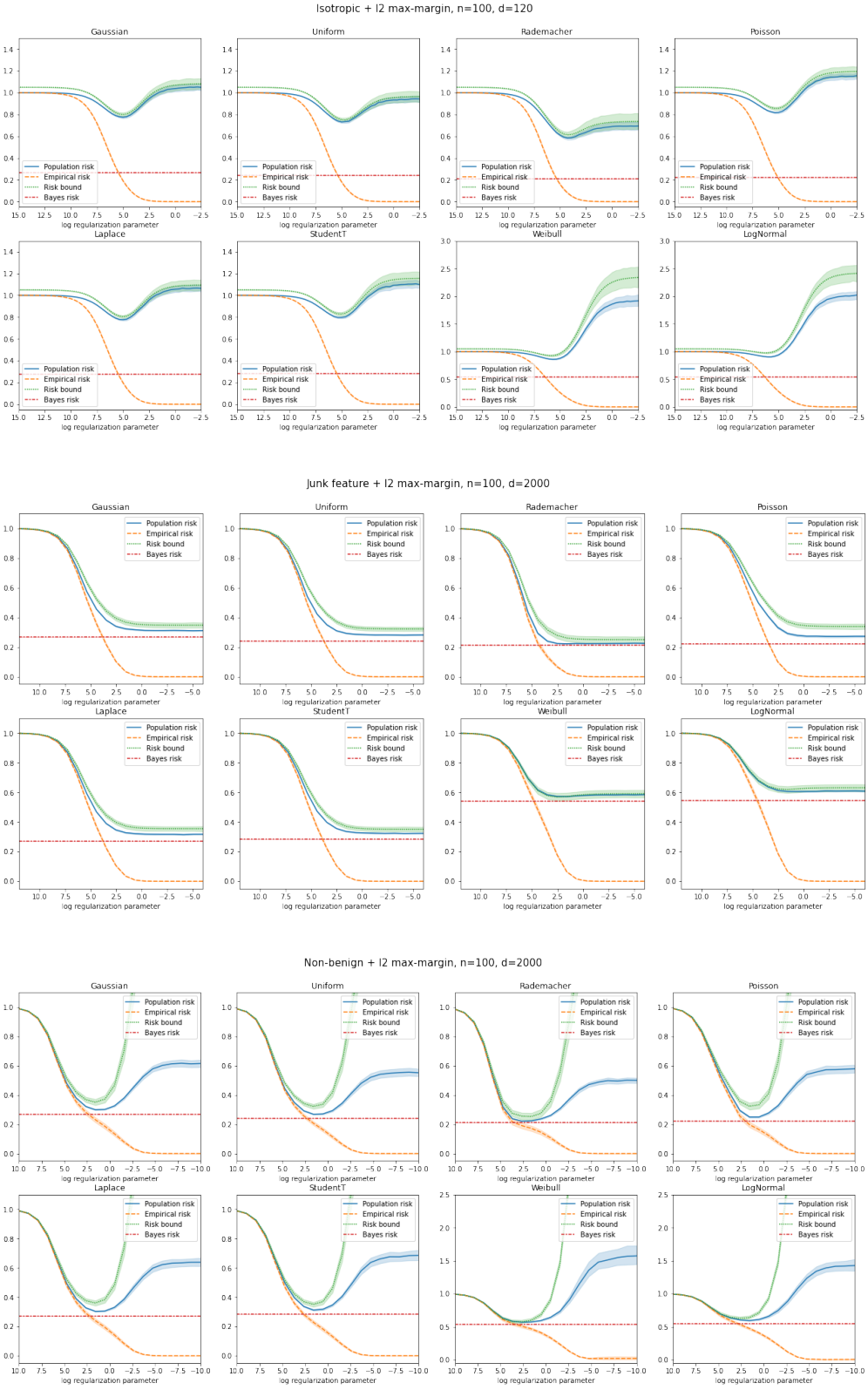

Figure 5.8: ℓ_2 margin classification: isotropic, junk and non-benign features.

Figure 5.9: ℓ_1 margin classification: isotropic, junk and non-benign features.

5.2 Provable Failure of Universality

In this section, we discuss an example where universality provably fails. The counterexample in this section is motivated by example 2 of Shamir [59]. For non-Gaussian features, it is possible to introduce strong dependence between the tail and the leading component of x while ensuring that they are uncorrelated. The dependence will prevent the norm of the tail from concentrating around its mean. In contrast, for Gaussian features with matching covariance, the two components must be independent. So the norm of the tail will concentrate — such discrepancy results in an over-optimistic bound for non-Gaussian data. In particular, we will assume the data distribution $\mathcal{D}$ over (x, y) is given by:

(A) $x = (x_{|k}, x_{|d-k})$ where $x_{|k} \sim \mathcal{N}(0, \Sigma_{|k})$ and there exists a function $h : \mathbb{R}^k \to \mathbb{R}$ such that

$$x_{|d-k} = h(x_{|k}) \cdot z \quad \text{with} \quad z \sim \mathcal{N}\left(0, \Sigma_{|d-k}\right) \quad \text{independent of } x_{|k} \tag{5.5}$$

(B) there exists a function $g : \mathbb{R}^{k+1} \to \mathbb{R}$ such that

$$y = g(x_{|k}, \xi) \tag{5.6}$$

where $\xi \sim \mathcal{D}_\xi$ is independent of x (but not necessarily Gaussian).

By the independence between z and $x_{|k}$, we can easily see that $x_{|k}$ and $x_{|d-k}$ are uncorrelated. Moreover, the covariance matrix of $x_{|d-k}$ is

$$\mathbb{E}[x_{|d-k} x_{|d-k}^T] = \mathbb{E}[h(x_{|k})^2] \cdot \Sigma_{|d-k} \tag{5.7}$$

which is just a re-scaling of $\Sigma_{|d-k}$ and therefore has the same effective rank as $R(\Sigma_{|d-k})$.

Optimistic rate. Even though the feature x in $\mathcal{D}$ is non-Gaussian, its tail $x_{|d-k}$ is Gaussian conditioned on the low-dimensional component $x_{|k}$. Therefore, our proof technique is still applicable

after a conditioning step, and we can handle the low-dimensional part with VC theory as before. However, it turns out that the uniform convergence bound for 1 square-root Lipschitz loss is no longer valid. Instead, we have to re-weight the loss by $\frac{1}{h(x_{|k})^2}$. More precisely, the application of our Gaussian comparison techniques shows the following:

Theorem 43. *Consider dataset (X, Y) drawn i.i.d. from the data distribution $\mathcal{D}$ according to (A) and (B), and fix any $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ such that $\sqrt{f}$ is 1-Lipschitz for any $y \in \mathcal{Y}$. Fix any $\delta > 0$ and suppose there exists $\epsilon_\delta < 1$ and $C_\delta : \mathbb{R}^{d-k} \to [0, \infty]$ such that*

(i) with probability at least $1 - \delta/2$ over (X, Y) and $G \sim \mathcal{N}(0, I_n)$, it holds uniformly over all $w_{|k} \in \mathbb{R}^k$ and $\|w_{|d-k}\|_{\Sigma_{|d-k}} \in \mathbb{R}_{\geq 0}$ that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_{i|k})^2} f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})\|w_{|d-k}\|_{\Sigma_{|d-k}} G_i, y_i) \geq (1-\epsilon_\delta)\mathbb{E}_{\mathcal{D}}\left[\frac{1}{h(x_{|k})^2} f(\langle w, x\rangle, y)\right]$$

(ii) with probability at least $1 - \delta/2$ over $z_{|d-k} \sim \mathcal{N}(0, \Sigma_{|d-k})$, it holds uniformly over all $w_{|d-k} \in \mathbb{R}^{d-k}$ that

$$\langle w_{|d-k}, z_{|d-k}\rangle \leq C_\delta(w_{|d-k}) \tag{5.8}$$

then with probability at least $1 - \delta$, it holds uniformly over all $w \in \mathbb{R}^d$ that

$$(1 - \epsilon_\delta)\mathbb{E}\left[\frac{1}{h(x_{|k})^2} f(\langle w, x\rangle, y)\right] \leq \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_{i|k})^2} f(\langle w, x_i\rangle, y_i) + \frac{C_\delta(w_{|d-k})}{\sqrt{n}}\right)^2. \tag{5.9}$$

The first assumption is mild because we only require uniform convergence over a low-dimensional concept class. Low-dimensional concentration can typically be verified using hypercontractivity and VC argument like in Chapter 4. The choice of C_δ is defined similarly. The only difference is that we need to upper bound $\langle w_{|d-k}, z_{|d-k}\rangle$ instead of $\langle w_{|d-k}, x_{|d-k}\rangle$, which is helpful because the norm of $z_{|d-k}$ will usually concentrate while the norm of $x_{|d-k}$ does not. Though the above generalization bound applies to any 1 square-root Lipschitz function, we will focus on the square

loss because this section's primary goal is constructing a counterexample to universality.

Computing the minimum norm. Next, we will compute the minimal norm to achieve zero training error for linear regression.

Theorem 44. *Under assumptions (A) and (B), fix any* $w^*_{|k} \in \mathbb{R}^k$ *and suppose for some* $\rho \in (0,1)$*, it holds with probability at least* $1-\delta/8$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \langle w^*_{|k}, x_{i|k}\rangle}{h(x_{i|k})}\right)^2 \le (1+\rho)\cdot \mathbb{E}\left[\left(\frac{y - \langle w^*_{|k}, x_{|k}\rangle}{h(x_{|k})}\right)^2\right]. \tag{5.10}$$

Then with probability at least $1-\delta$*, for some* $\epsilon \lesssim \rho + \log\left(\frac{1}{\delta}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma_{|d-k})}} + \frac{n}{R(\Sigma_{|d-k})}\right)$*, it holds that*

$$\min_{w\in\mathbb{R}^d:\forall i,\langle w,x_i\rangle = y_i} \|w\|_2^2 \le \|w^*_{|k}\|_2^2 + (1+\epsilon)\frac{n\mathbb{E}\left[\left(\frac{y-\langle w^*_{|k},x_{|k}\rangle}{h(x_{|k})}\right)^2\right]}{\mathrm{Tr}(\Sigma_{|d-k})} \tag{5.11}$$

It is easy to see that the w^* that minimizes the population weighted square loss satisfy $w^*_{|d-k} = 0$, and so we can let $w^*_{|k}$ to be the minimizer. Again, the first assumption is mild because we only require concentration at one single hypothesis.

Benign overfitting. Consider the minimal norm interpolant $\hat{w} = \arg\min_{w\in\mathbb{R}^d: Xw=Y} \|w\|_2$ and define

$$L^* = \inf_w \mathbb{E}\left[\left(\frac{y - \langle w, x\rangle}{h(x_{|k})}\right)^2\right].$$

Similar to Corollary 39 in Chapter 4, plugging in Theorem 44 to the uniform convergence guarantee Theorem 43, we have (ignoring lower order terms)

$$\begin{aligned}\mathbb{E}\left[\frac{1}{h(x_{|k})^2}(\langle \hat{w}, x\rangle - y)^2\right] &\leq \frac{\mathrm{Tr}(\Sigma_{|d-k}) \cdot \|\hat{w}\|_2^2}{n} \\ &\leq \frac{\mathrm{Tr}(\Sigma_{|d-k}) \cdot \|w^*\|_2^2}{n} + (1+\epsilon)\, L^*.\end{aligned}$$

Therefore, if we choose $k = 1$ and $\Sigma_{|d-1} = \frac{1}{d-1} I_{d-1}$ with $n \ll d$, then it holds that

$$\mathbb{E}\left[\frac{1}{h(x_{|k})^2}(\langle \hat{w}, x\rangle - y)^2\right] \to L^* \tag{5.12}$$

which recovers the result of Shamir [59].

Breaking Uniform Convergence. Finally, we explain why this is a counterexample to universality. If we pretend that the features are Gaussian, then since we choose $\Sigma_{|d-k}$ to be benign, our theory in Chapter 4 predicts that

$$\begin{aligned}\inf_w \mathbb{E}[(\langle w, x\rangle - y)^2] &\leq \mathbb{E}[(\langle \hat{w}, x\rangle - y)^2] \\ &\leq (1+o(1))\frac{\|\hat{w}\|_2^2 \cdot \mathbb{E}\|x_{|d-k}\|_2^2}{n} \\ &= (1+o(1))\frac{\|\hat{w}\|_2^2 \cdot \mathbb{E}[h(x_{|k})^2]\,\mathrm{Tr}(\Sigma_{|d-k})}{n}.\end{aligned}$$

Therefore, for sufficiently large n, it must be the case that

$$\inf_w \mathbb{E}[(\langle w, x\rangle - y)^2] \leq \inf_w\ \mathbb{E}[h(x_{|k})^2] \cdot \mathbb{E}\left[\left(\frac{y - \langle w, x\rangle}{h(x_{|k})}\right)^2\right]. \tag{5.13}$$

However, this cannot be always true when y and $h(x_{|k})$ are dependent. For example, let's consider the case where $k = 1$, $x_1 \sim \mathcal{N}(0,1)$, and h only depends on x_1 through $|x_1|$. We consider the case

where $y = h(|x_1|)^2$. Then

$$\mathbb{E}[x_1 y] = 0 \quad \text{and} \quad \mathbb{E}[x_{|d-k} y] = \mathbb{E}[h(|x_1|)^3 z] = 0$$

and so

$$\inf_w \mathbb{E}[(\langle w, x\rangle - y)^2] = \mathbb{E}[y^2] = \mathbb{E}[h(|x_1|)^4]. \tag{5.14}$$

On the other hand, we can also check

$$\mathbb{E}\left[\frac{x_1 y}{h(|x_1|)^2}\right] = \mathbb{E}[x_1] = 0 \quad \text{and} \quad \mathbb{E}\left[\frac{x_{|d-k} y}{h(|x_1|)^2}\right] = \mathbb{E}[x_{|d-k}] = 0$$

and so

$$\inf_w \mathbb{E}\left[\left(\frac{y - \langle w, x\rangle}{h(x_{|k})}\right)^2\right] = \mathbb{E}\left[\left(\frac{y}{h(|x_1|)}\right)^2\right] = \mathbb{E}[h(|x_1|)^2]. \tag{5.15}$$

Then equation (5.13) predicts that $\mathbb{E}[h(|x_1|)^4] \leq \mathbb{E}[h(|x_1|)^2]^2$, but this is impossible because

$$\mathbb{E}[h(|x_1|)^4] - \mathbb{E}[h(|x_1|)^2]^2 = \mathrm{Var}(h(|x_1|)^2) > 0$$

as long as $h(|x_1|)$ is not degenerate. For example, we can take $h(|x|) = 1 + |x|$. In general, $h(|x|)^2$ can follow any non-negative distribution by choosing $\sqrt{h}$ to the inverse CDF and so we can make (5.13) arbitrarily loose.

CHAPTER 6

CONCLUSION

In this thesis, we study the statistical learning theory for models with high complexity, with particular emphasis on understanding the generalization error of high-dimensional interpolants in non-realizable settings. A theoretical understanding of machine learning is instrumental — it can provide a tool to predict the generalization performance and guide the design of better algorithms and model selection procedures in practice; it also allows us to understand, on a fundamental level, what properties of the data distribution are essential to the success of machine learning.

As we can see in our applications in Chapter 2, 3 and 4, even though the norm of the interpolants that we consider can diverge as the sample size goes to infinity, we are still able to establish strong consistency results through a uniform convergence argument. We accomplish this by proving uniform convergence bounds with the optimal dependence on model complexity. Our theory provides a unifying perspective on benign overfitting and optimal regularization in statistical learning, and it has many benefits from a technical point of view — it is non-asymptotic and does not require any particular scaling between the sample size and feature dimension; it requires very mild assumptions and can handle any Gaussian multi-index setting; it does not depend on any specific training algorithm and can be easily adapted to different loss functions (for example, max-margin classification, phase retrieval, and ReLU regression) and complexity measures (for example, ℓ_1, ℓ_2 and nuclear norm). Nevertheless, this generalization theory crucially depends on the assumption that the features are Gaussian. In Chapter 5, we see that we should be able to significantly relax this assumption, at least in some settings of kernel ridge regression and random feature regression, as well as the setting where the feature vector is a linear transformation of independent variables. However, as we see in section 5.2, universality can fail unexpectedly. It is an important open question to understand when we can safely assume the features are Gaussian.

There are many potential exciting extensions to our results in Chapter 4. For example, we consider the applications of our Moreau envelope generalization theory to Lipschitz or square-root

Lipschitz loss functions. However, if we use the log link in many generalized linear models, the loss function is neither Lipschitz nor square-root Lipschitz globally. In these situations, some other properties of the loss function (such as convexity) might allow us to prove sharp generalization bound for regularized estimators. In addition, we have ignored the optimization concerns in most of this thesis. When we consider convex losses such as the square loss or the squared hinge loss, many first-order methods (such as gradient descent with or without momentum and stochasticity) will converge to the minimal norm interpolant. However, the optimization landscape for non-convex losses is more complicated. In these situations, even though the uniform convergence bounds can still be applied, we need to take the training dynamics into account as well. Finally, we can only prove generalization bound for single-index neural networks in section 4.5. It would be interesting to establish a similar result for arbitrary neural networks. Finding the right complexity measure and the optimal dependence in uniform convergence bounds will be a significant breakthrough in our mathematical understanding of deep learning.

REFERENCES

[1] Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[2] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[3] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[4] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability theory and related fields*, 135(3):311–334, 2006.

[5] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[6] Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

[7] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[8] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[9] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, 2001.

[10] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, 2018.

[11] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[12] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

[13] Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)*, 50(4):506–519, 2003.

[14] Anselm Blumer, Andrzej Ehrenfeucht, David Henry Haussler, and Manfred Klaus Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

[15] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[16] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in optimization*, 8(3-4):231–358, 2015.

[17] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):1–12, 2021.

[18] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[19] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6): 805–849, 2012.

[20] Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve. In *Advances in Neural Information Processing Systems*, 2021.

[21] Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effects of inductive bias. In *International Conference on Machine Learning*, 2022.

[22] Rick Durrett. *Probability: theory and examples*. Cambridge University Press, 2019.

[23] Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *COLT*, pages 315–340, 2011.

[24] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

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

[26] Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, 2017.

[27] Qiyang Han and Yandi Shen. Universality of regularized regression estimators in high dimensions. 2022.

[28] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 2019.

[29] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023.

[30] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. In *Advances in Neural Information Processing Systems*, 2020.

[31] Tony Jebara and Risi Kondor. Bhattacharyya and expected likelihood kernels. In *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), 2777*, 2003.

[32] Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression, 2021.

[33] Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. In *Advances in Neural Information Processing Systems*, 2021.

[34] Risi Kondor and John Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *International Conference on Machine Learning*, 2002.

[35] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds, 2013.

[36] Michel Ledoux. A heat semigroup approach to concentration on the sphere and on a compact riemannian manifold. *Geometric & Functional Analysis GAFA*, 2(2):221–224, 1992.

[37] Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. 2021.

[38] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, 2020.

[39] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34: 18137–18151, 2021.

[40] Neil Rohit Mallinar, James B Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. In *Advances in Neural Information Processing Systems*, 2022.

[41] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

[42] Gabriel C. Mel and Surya Ganguli. A theory of high dimensional regression with arbitrary correlations between input features and target functions: Sample complexity, multiple descent curves and a hierarchy of phase transitions. In *International Conference on Machine Learning*, volume 139, page 7578–7587, 2021.

[43] Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39. PMLR, 2014.

[44] Shahar Mendelson. Extending the scope of the small-ball method. 2017.

[45] James Mercer. Functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London*, 209:4–415, 1909.

[46] Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, 2006.

[47] Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv:2204.10425*, 2022.

[48] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations – Workshop*, 2015.

[49] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[50] Dmitry Panchenko. Some extensions of an inequality of vapnik and chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002.

[51] Dmitry Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *The Annals of Probability*, 31(4):2068–2081, 2003.

[52] David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[53] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[54] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[55] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge(less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, volume 130, page 3889–3897, 2021.

[56] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813:814, 2015.

[57] Ralph Tyrrell Rockafellar. *Convex analysis*. Princeton university press, 1970.

[58] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[59] Ohad Shamir. The implicit bias of benign overfitting. 2022.

[60] James B Simon, Madeline Dickens, Dhruva Karkada, and Michael R. DeWeese. The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks. *arXiv:2110.03922*, 2021.

[61] Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

[62] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560, 2005.

[63] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss, 2010.

[64] Mihailo Stojnic. A framework to characterize performance of LASSO algorithms, 2013.

[65] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity, 2014.

[66] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR, 2015.

[67] Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[68] Ramon van Handel. Probability in high dimension. Lecture notes, Princeton University, 2014. URL `https://web.math.princeton.edu/~rvan/APC550.pdf`.

[69] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 1982.

[70] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2010.

[71] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

[72] Dietrich von Rosen. Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics*, 15(2):97–109, 1988.

[73] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[74] Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum l1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[75] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, 2020.

[76] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[77] Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. On uniform convergence and low-norm interpolation learning. In *Advances in Neural Information Processing Systems*, 2020.

[78] Lijia Zhou, Frederic Koehler, Danica J. Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. In *ACM / IMS Journal of Data Science*, 2021.

[79] Lijia Zhou, Frederic Koehler, Pragya Sur, Danica J. Sutherland, and Nathan Srebro. A non-asymptotic moreau envelope theory for high-dimensional generalized linear models. In *Advances in Neural Information Processing Systems*, 2022.

APPENDIX A

GAUSSIAN MINIMAX THEOREM

In this appendix, we introduce some key technical tools and useful lemmas.

Gaussian Minmax Theorem. The following result is Theorem 3 of Thrampoulidis et al. [66], known as the Convex Gaussian Minmax Theorem or CGMT (see also Theorem 1 in the same reference). As explained there, it is a consequence of the main result of Gordon [25], known as Gordon's Theorem or the Gaussian Minmax Theorem. Despite the name, convexity is only required for one of the theorem's conclusions.

Theorem 45 (Convex Gaussian Minmax Theorem; [66, 25])**.** *Let $Z : n \times d$ be a matrix with i.i.d. $\mathcal{N}(0,1)$ entries and suppose $G \sim \mathcal{N}(0, I_n)$ and $H \sim \mathcal{N}(0, I_d)$ are independent of Z and each other. Let S_w, S_u be compact sets and $\psi : S_w \times S_u \to \mathbb{R}$ be an arbitrary continuous function. Define the Primary Optimization (PO) problem*

$$\Phi(Z) := \min_{w \in S_w} \max_{u \in S_u} \langle u, Zw \rangle + \psi(w, u) \tag{A.1}$$

and the Auxiliary Optimization (AO) problem

$$\phi(G, H) := \min_{w \in S_w} \max_{u \in S_u} \|w\|_2 \langle G, u \rangle + \|u\|_2 \langle H, w \rangle + \psi(w, u). \tag{A.2}$$

Under these assumptions, $\Pr(\Phi(Z) < c) \leq 2 \Pr(\phi(G, H) \leq c)$ *for any* $c \in \mathbb{R}$.

Furthermore, if we suppose that S_w, S_u are convex sets and $\psi(w, u)$ is convex in w and concave in u, then $\Pr(\Phi(Z) > c) \leq 2 \Pr(\phi(G, H) \geq c)$.

In other words, the first conclusion says that high probability lower bounds on the auxiliary optimization $\phi(G, H)$ imply high probability lower bounds on the primary optimization $\Phi(Z)$. Importantly, this direction holds without any convexity assumptions. Under the additional con-

vexity assumptions, the second conclusion gives a similar comparison of high probability upper bounds.

In our analysis, we need a slightly more general statement of the Gaussian Minmax Theorem than 45: we need the minmax formulation to include additional variables which only affect the deterministic term in the minmax problem. It's straightforward to prove this result by repeating the argument in Thrampoulidis et al. [66]; below we give an alternative proof which reduces to 45, by introducing extremely small extra dimensions to contain the extra variables. Intuitively, this works because the statement of the GMT allows for arbitrary continuous functions ψ, with no dependence on their quantitative smoothness.

Theorem 46 (Variant of GMT)**.** *Let $Z : n \times d$ be a matrix with i.i.d. $\mathcal{N}(0,1)$ entries and suppose $G \sim \mathcal{N}(0, I_n)$ and $H \sim \mathcal{N}(0, I_d)$ are independent of Z and each other. Let S_W, S_U be compact sets in $\mathbb{R}^d \times \mathbb{R}^{d'}$ and $\mathbb{R}^n \times \mathbb{R}^{n'}$ respectively, and let $\psi : S_W \times S_U \to \mathbb{R}$ be an arbitrary continuous function. Define the* Primary Optimization (PO) *problem*

$$\Phi(Z) := \min_{(w,w') \in S_W} \max_{(u,u') \in S_U} \langle u, Zw \rangle + \psi((w,w'),(u,u')) \tag{A.3}$$

and the Auxiliary Optimization (AO) *problem*

$$\phi(G,H) := \min_{(w,w') \in S_W} \max_{(u,u') \in S_U} \|w\|_2 \langle G, u \rangle + \|u\|_2 \langle H, w \rangle + \psi((w,w'),(u,u')). \tag{A.4}$$

Under these assumptions, $\Pr(\Phi(Z) < c) \leq 2\Pr(\phi(G,H) \leq c)$ *for any* $c \in \mathbb{R}$.

Proof. Let $\epsilon \in (0,1)$ be arbitrary and

$$S_{W,\epsilon} := \{(w, \epsilon w') : (w,w') \in S_W\}, \quad S_{U,\epsilon} := \{(u, \epsilon u') : (u,u') \in S_U\}.$$

Define $\psi_\epsilon((w,w'),(u,u')) := \psi((w, \frac{1}{\epsilon}w'),(u, \frac{1}{\epsilon}u'))$ so that if $W = (w, \epsilon w')$ and $U = (u, \epsilon u')$, then $\psi_\epsilon(W,U) = \psi((w,w'),(u,u'))$. We also define $S_w = \{w \in \mathbb{R}^d : \exists w' \, s.t. \, (w,w') \in S_W\}$.

The other sets $S_{w'}, S_u$ and $S_{u'}$ are defined similarly. It is clear that $S_w, S_{w'}, S_u, S_{u'}, S_{W,\epsilon}$ and $S_{U,\epsilon}$ are all still compact in their respective topology, and ψ_ϵ is continuous for every $\epsilon > 0$.

Let $Z' : (n + n') \times (d + d')$ be a matrix with i.i.d. $N(0, 1)$ entries such that the top left $n \times d$ matrix is Z. Similarly, we define G' to be a $(n+n')$-dimensional Gaussian vector with independent coordinates such that the first n coordinates are G, and H' to be a $(d + d')$-dimensional Gaussian vector with independent coordinates such that the first d coordinates are H. Next, consider the augmented PO and AO:

$$\begin{aligned} \Phi_\epsilon(Z') &:= \min_{W \in S_{W,\epsilon}} \max_{U \in S_{U,\epsilon}} \langle U, Z'W \rangle + \psi_\epsilon(W, U) \\ \phi_\epsilon(G', H') &:= \min_{W \in S_{W,\epsilon}} \max_{U \in S_{U,\epsilon}} \|W\|_2 \langle G', U \rangle + \|U\|_2 \langle H', W \rangle + \psi_\epsilon(W, U) \end{aligned} \tag{A.5}$$

It is clear that for a small value of ϵ, the augmented problem will be close to the original problem. More precisely, for every $(w, w') \in S_W$ and $(u, u') \in S_U$

$$\begin{aligned} &|\langle (w, \epsilon w'), Z'(u, \epsilon u') \rangle - \langle w, Zu \rangle| \\ &\quad = |\epsilon \langle (0, w'), Z'(u, 0) \rangle + \epsilon \langle (w, 0), Z'(0, u') \rangle + \epsilon^2 \langle (0, w'), Z'(0, u') \rangle| \\ &\quad \le \epsilon (R(S_w) + R(S_{w'}))(R(S_u) + R(S_{u'})) \|Z'\|_{op} = \epsilon A \|Z'\|_{op} \end{aligned} \tag{A.6}$$

where $A := (R(S_w) + R(S_{w'}))(R(S_u) + R(S_{u'}))$ is deterministic and does not depend on ϵ. Similarly, it is routine to check

$$\begin{aligned} \|w\|_2 \langle G, u \rangle &= \|w\|_2 (\langle G', (u, \epsilon u') \rangle - \epsilon \langle G', (0, u') \rangle) \\ \|u\|_2 \langle H, w \rangle &= \|u\|_2 (\langle H', (w, \epsilon w') \rangle - \epsilon \langle H', (0, w') \rangle) \end{aligned}$$

so by the triangle inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\big| \|(w, \epsilon w')\|_2 \langle G', (u, \epsilon u') \rangle - \|w\|_2 \langle G, u \rangle \big| \\ &\quad \le \epsilon R(S_{w'}) \|G'\|_2 (R(S_u) + \epsilon R(S_{u'})) + \epsilon R(S_w) \|G'\|_2 R(S_{u'}) \le \epsilon A \|G'\|_2 \end{aligned} \tag{A.7}$$

and

$$\begin{aligned}&\left|\|(u,\epsilon u')\|_2\langle H',(w,\epsilon w')\rangle - \|u\|_2\langle H,w\rangle\right| \\ &\quad\le \epsilon R(S_{u'})\|H'\|_2(R(S_w)+\epsilon R(S_{w'})) + \epsilon R(S_u)\|H'\|_2 R(S_{w'}) \le \epsilon A\|H'\|_2\end{aligned} \tag{A.8}$$

From (A.6), it follows that

$$\left|\Phi_\epsilon(Z') - \Phi(Z)\right| \le \epsilon A\|Z'\|_{op}. \tag{A.9}$$

Similarly, from (A.7) and (A.8), it follows that

$$|\phi_\epsilon(G',H') - \phi(G,H)| \le \epsilon A(\|G'\|_2 + \|H'\|_2). \tag{A.10}$$

Approximating the original PO and AO by (A.5) allows us to directly apply the Gaussian Minmax Theorem. For any $c \in \mathbb{R}$, we have

$$\begin{aligned}\Pr(\Phi(Z) < c) &\le \Pr(\Phi_\epsilon(Z') < c+\sqrt{\epsilon}) + \Pr(\epsilon A\|Z'\|_{op} > \sqrt{\epsilon}) \\ &\le 2\Pr(\phi_\epsilon(G',H') \le c+\sqrt{\epsilon}) + \Pr(\epsilon A\|Z'\|_{op} > \sqrt{\epsilon}) \\ &\le 2\Pr(\phi(G',H') \le c+2\sqrt{\epsilon}) + 2\Pr\left(\epsilon A(\|G'\|_2 + \|H'\|_2) > \sqrt{\epsilon}\right) \\ &\quad + \Pr(\epsilon A\|Z'\|_{op} > \sqrt{\epsilon}) \\ &\le 2\Pr(\phi(G',H') \le c+2\sqrt{\epsilon}) + 2\Pr\left(\|G'\|_2 > \frac{1}{2A\sqrt{\epsilon}}\right) \\ &\quad + 2\Pr\left(\|H'\|_2 > \frac{1}{2A\sqrt{\epsilon}}\right) + \Pr\left(\|Z'\|_{op} > \frac{1}{A\sqrt{\epsilon}}\right)\end{aligned}$$

where we used (A.9) in the first inequality, 45 in the second inequality, and (A.10) in the last inequality. This holds for arbitrary $\epsilon > 0$ and taking the limit $\epsilon \to 0$ shows the result, because the CDF is right continuous [22] and the remaining terms go to zero by standard concentration inequalities (see Lemma 52 and Theorem 53 below). □

Truncation Lemmas. Theorem 46 requires S_W and S_U to be compact. However, we can usually get around the compactness requirement by a truncation argument in our applications.

Lemma 47. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary function and $\mathcal{S}_r^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$, then for any set $\mathcal{K}$, it holds that*

$$\lim_{r\to\infty} \sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) = \sup_{w\in\mathcal{K}} f(w). \tag{A.11}$$

If f is a random function, then for any $t \in \mathbb{R}$

$$\Pr\left(\sup_{w\in\mathcal{K}} f(w) > t\right) = \lim_{r\to\infty} \Pr\left(\sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) > t\right). \tag{A.12}$$

Proof. We consider two cases:

1. Suppose that $\sup_{w\in\mathcal{K}} f(w) = \infty$. Then for any $M > 0$, there exists $x_M \in \mathcal{K}$ such that $f(x_M) > M$. Hence for any $r > \|x_M\|_2$, it holds that

$$\sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) > M \implies \liminf_{r\to\infty} \sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) \geq M$$

   As the choice of M is arbitrary, we have $\lim_{r\to\infty} \sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) = \infty$ as desired.

2. Suppose that $\sup_{w\in\mathcal{K}} f(w) = M < \infty$. Then for any $\epsilon > 0$, there exists $x_\epsilon \in \mathcal{K}$ such that $f(x_\epsilon) > M - \epsilon$. Hence for any $r > \|x_\epsilon\|_2$, it holds that

$$\sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) > M - \epsilon \implies \liminf_{r\to\infty} \sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) \geq M - \epsilon$$

   As the choice of ϵ is arbitrary, we have $\liminf_{r\to\infty} \sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) \geq M$. On the other hand, it must be the case (by definition of supremum) that

$$\sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) \leq M \implies \limsup_{r\to\infty} \sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) \leq M$$

Consequently, the limit of $\sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w)$ exists and equals M.

Finally, by the fact that the supremum is increasing in r and the continuity of probability measure, we have

$$\begin{aligned}
\Pr\left(\sup_{w\in\mathcal{K}} f(w) > t\right) &= \Pr\left(\lim_{r\to\infty} \sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) > t\right) \\
&= \Pr\left(\bigcup_{r\in\mathbb{N}} \bigcap_{R\geq r} \sup_{w\in\mathcal{K}\bigcap\mathcal{S}_R^d} f(w) > t\right) \\
&= \lim_{r\to\infty} \Pr\left(\bigcap_{R\geq r} \sup_{w\in\mathcal{K}\bigcap\mathcal{S}_R^d} f(w) > t\right) \\
&= \lim_{r\to\infty} \Pr\left(\sup_{w\in\mathcal{K}\cap\mathcal{S}_r^d} f(w) > t\right). \qquad \square
\end{aligned}$$

Lemma 48. *Let $\mathcal{K}$ be a compact set and f, g be continuous real-valued functions on $\mathbb{R}^d$. Then it holds that*

$$\lim_{r\to\infty} \sup_{w\in\mathcal{K}} \inf_{0\leq\lambda\leq r} \lambda f(w) + g(w) = \sup_{w\in\mathcal{K}: f(w)\geq 0} g(w). \tag{A.13}$$

If f and g are random functions, then for any $t \in \mathbb{R}$

$$\Pr\left(\sup_{w\in\mathcal{K}: f(w)\geq 0} g(w) \geq t\right) = \lim_{r\to\infty} \Pr\left(\sup_{w\in\mathcal{K}} \inf_{0\leq\lambda\leq r} \lambda f(w) + g(w) \geq t\right). \tag{A.14}$$

Proof. We consider two cases:

1. The limiting problem is infeasible: $\forall w \in \mathcal{K}, f(w) < 0$. Then by compactness and the continuity of f, there exists $\mu < 0$ such that for all $w \in \mathcal{K}$

$$f(w) < \mu \implies \sup_{w\in\mathcal{K}} \inf_{0\leq\lambda\leq r} \lambda f(w) + g(w) \leq r\mu + \sup_{w\in\mathcal{K}} g(w).$$

By compactness and the continuity of g again, we have $\sup_{w\in\mathcal{K}} g(w) < \infty$ and so

$$\lim_{r\to\infty} \sup_{w\in\mathcal{K}} \inf_{0\leq\lambda\leq r} \lambda f(w) + g(w) = -\infty$$

as desired.

2. The limiting problem is feasible: $\exists w_0 \in \mathcal{K}, f(w_0) \geq 0$. In this case, let

$$\begin{aligned} w_r &= \arg\max_{w\in\mathcal{K}} \inf_{0\leq\lambda\leq r} \lambda f(w) + g(w) \\ &= \arg\max_{w\in\mathcal{K}} r \cdot f(w)\mathbb{1}_{\{f(w)\leq 0\}} + g(w) \end{aligned}$$

be an arbitrary maximizer for each r. Note that a maximizer necessarily exists in $\mathcal{K}$ by compactness of $\mathcal{K}$ and the continuity of f and g. By compactness of $\mathcal{K}$ again, the sequence $\{w_r\}$ at positive integer values of r has a subsequential limit: $\exists r_n \to \infty$ and $w_\infty \in \mathcal{K}$ such that $w_{r_n} \to w_\infty$.

For the sake of contradiction, assume that $f(w_\infty) < 0$, then by continuity, there exists $\mu < 0$ such that for all sufficiently large n

$$f(w_{r_n}) < \mu \implies \sup_{w\in\mathcal{K}} \inf_{0\leq\lambda\leq r_n} \lambda f(w) + g(w) = r_n \cdot f(w_{r_n}) + g(w_{r_n}) \leq r_n\mu + \sup_{w\in\mathcal{K}} g(w)$$

which is unbounded from below as $n \to \infty$. On the other hand, we have

$$\sup_{w\in\mathcal{K}} \inf_{0\leq\lambda\leq r_n} \lambda f(w) + g(w) \geq g(w_0)$$

and so we have reached a contradiction; thus $f(w_\infty) \geq 0$. Observe that

$$\sup_{w\in\mathcal{K}} \inf_{0\leq\lambda\leq r_n} \lambda f(w) + g(w) = r_n \cdot f(w_{r_n})\mathbb{1}_{\{f(w_{r_n})\leq 0\}} + g(w_{r_n}) \leq g(w_{r_n})$$

and so by continuity of g

$$\limsup_{n\to\infty} \sup_{w\in\mathcal{K}} \inf_{0\le\lambda\le r_n} \lambda f(w) + g(w) \le g(w_\infty) \le \sup_{w\in\mathcal{K}: f(w)\ge 0} g(w).$$

The $\liminf$ direction follows immediately from the definition, and so the limit exists and equals $\sup_{w\in\mathcal{K}:f(w)\ge 0} g(w)$. We can conclude that

$$\lim_{r\to\infty} \sup_{w\in\mathcal{K}} \inf_{0\le\lambda\le r} \lambda f(w) + g(w) = \sup_{w\in\mathcal{K}: f(w)\ge 0} g(w)$$

because it is a monotonic sequence.

Finally, by the fact that the supremum is decreasing in r and the continuity of probability measure, we have

$$\begin{aligned}
\Pr\left(\sup_{w\in\mathcal{K}:f(w)\ge 0} g(w) \ge t\right) &= \Pr\left(\lim_{r\to\infty} \sup_{w\in\mathcal{K}} \inf_{0\le\lambda\le r} \lambda f(w) + g(w) \ge t\right) \\
&= \Pr\left(\cap_r \sup_{w\in\mathcal{K}} \inf_{0\le\lambda\le r} \lambda f(w) + g(w) \ge t\right) \\
&= \lim_{r\to\infty} \Pr\left(\sup_{w\in\mathcal{K}} \inf_{0\le\lambda\le r} \lambda f(w) + g(w) \ge t\right).
\end{aligned}$$

□

Concentration of Lipschitz functions. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the norm $\|\cdot\|$ if it holds for all $x, y \in \mathbb{R}^n$ that $|f(x) - f(y)| \le L\|x - y\|$. We use the concentration of Lipschitz functions of a Gaussian.

Theorem 49 (van Handel [68], Theorem 3.25)**.** *If f is L-Lipschitz with respect to the Euclidean norm and $Z \sim \mathcal{N}(0, I_n)$, then*

$$\Pr(|f(Z) - \mathbb{E}f(Z)| \ge t) \le 2e^{-t^2/2L^2}. \tag{A.15}$$

We also use a similar result for functions of a uniformly spherical vector (see Theorem 5.1.4 and Exercise 5.1.12 of Vershynin [71]); we cite a result with sharp constant factor from Ledoux [36].

Theorem 50 (Spherical concentration; Ledoux [36])**.** *If f is L-Lipschitz with respect to the Euclidean norm and $Z \sim$ Uni(S^{n-1}) where $S^{n-1} = \{u \in \mathbb{R}^n : \|u\| = 1\}$ is the unit sphere, Uni(S^{n-1}) is the uniform measure on the sphere, and $n \geq 3$, then*

$$\Pr(|f(Z) - \mathbb{E}f(Z)| \geq t) \leq 2e^{-(n-2)t^2/2L^2}. \tag{A.16}$$

The following lemma says that a $o(n)$-dimensional subspace cannot align with a random spherically symmetric vector.

Lemma 51. *Suppose that S is a fixed subspace of dimension d in $\mathbb{R}^n$ with $n \geq 4$, P_S is the orthogonal projection onto S, and V is a spherically symmetric random vector (i.e. $V/\|V\|_2$ is uniform on the sphere). Then*

$$\frac{\|P_S V\|_2}{\|V\|_2} \leq \sqrt{d/n} + 2\sqrt{\log(2/\delta)/n}. \tag{A.17}$$

with probability at least $1 - \delta$. Conditional on this inequality holding, we therefore have uniformly for all $s \in S$ that

$$|\langle s, V\rangle| = |\langle s, P_S V\rangle| \leq \|s\|_2 \|P_S V\|_2 \leq \|s\|_2 \|V\|_2 \left(\sqrt{d/n} + 2\sqrt{\log(2/\delta)/n}\right). \tag{A.18}$$

Proof. This is trivial if $d \geq n$, since the left-hand side is at most 1. Thus assume without loss of generality that $d < n$. By symmetry, it suffices to fix S to be the span of basis vectors $e_1, \ldots, e_d$ and to bound $\|P_S V\|_2$ for V a uniformly random chosen vector from the unit sphere in $\mathbb{R}^n$. Recall that for any coordinate i, we have $\mathbb{E}V_i^2 = 1/n$ by symmetry among the coordinates and the fact that $\|V\|_2^2 = 1$ almost surely. The function $v \mapsto \|P_S v\|_2$ is a 1-Lipschitz function and $\mathbb{E}\|P_S V\|_2 \leq$

$\sqrt{\mathbb{E}\|P_S V\|_2^2} = \sqrt{d/n}$, so by Theorem 50 above

$$\|P_S V\|_2 \le \sqrt{d/n} + \sqrt{2\log(2/\delta)/(n-2)})$$

with probability at least $1-\delta$. Using $n \ge 4$ gives the result. □

The concentration of the Euclidean norm of a Gaussian vector follows from Theorem 49; we state it explicitly below.

Lemma 52. *Suppose that $Z \sim N(0, I_n)$. Then*

$$\Pr(\left|\|Z\|_2 - \sqrt{n}\right| \ge t) \le 4e^{-t^2/4}. \tag{A.19}$$

Proof. First we recall the standard fact (see e.g. Chandrasekaran et al. [19]) that

$$\sqrt{n} - 1 \le \frac{n}{\sqrt{n+1}} \le \mathbb{E}\|Z\|_2 \le \sqrt{n}.$$

Because the norm is 1-Lipschitz, it follows from Theorem 49 that

$$\Pr(|\|Z\|_2 - \mathbb{E}\|Z\|_2| \ge t) \le 2e^{-t^2/2}$$

so

$$\Pr(\left|\|Z\|_2 - \sqrt{n}\right| \ge t+1) \le 2e^{-t^2/2}.$$

Now using that $(t-1)^2 \ge t^2/2 - 1$ shows

$$\Pr(\left|\|Z\|_2 - \sqrt{n}\right| \ge t) \le 2e^{-(t^2/2-1)/2} \le 4e^{-t^2/4}.$$ □

Wishart Concentration. Let $\sigma_{\min}(A)$ denote the minimum singular value of an arbitrary matrix A, and $\sigma_{\max}$ the maximum singular value. We use $\|A\|_{op} = \sigma_{\max}(A)$ to denote the operator norm of matrix A.

Theorem 53 (Vershynin [70], Corollary 5.35)**.** *Let $n, N \in \mathbb{N}$. Let $A \in \mathbb{R}^{N \times n}$ be a random matrix with entries i.i.d. $\mathcal{N}(0,1)$. Then for any $t > 0$, it holds with probability at least $1 - 2\exp(-t^2/2)$ that*

$$\sqrt{N} - \sqrt{n} - t \leq \sigma_{min}(A) \leq \sigma_{max}(A) \leq \sqrt{N} + \sqrt{n} + t. \tag{A.20}$$

APPENDIX B

VC THEORY AND HYPERCONTRACTIVITY

Recall the following definition of VC-dimension from Shalev-Shwartz and Ben-David [58].

Definition 8. *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ and let $C = \{c_1, ..., c_m\} \subset \mathcal{X}$. The restriction of $\mathcal{H}$ to C is*

$$\mathcal{H}_C = \{(h(c_1), ..., h(c_m)) : h \in \mathcal{H}\}.$$

A hypothesis class $\mathcal{H}$ shatters *a finite set $C \subset \mathcal{X}$ if $|\mathcal{H}_C| = 2^{|C|}$. The VC-dimension of $\mathcal{H}$ is the maximal size of a set that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrary large size, we say $\mathcal{H}$ has infinite VC-dimension.*

Also, we have the following well-known result for the class of nonhomogenous halfspaces in $\mathbb{R}^d$ (Theorem 9.3 of Shalev-Shwartz and Ben-David [58]), and the result on VC-dimension of the union of two hypothesis classes (Lemma 3.2.3 of Blumer et al. [14]):

Theorem 54. *The class $\{x \mapsto sign(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$ has VC-dimension $d + 1$.*

Theorem 55. *Let $\mathcal{H}$ a hypothesis classes of finite VC-dimension $d \geq 1$. Let $\mathcal{H}_2 := \{\max(h_1, h_2) : h_1, h_2 \in \mathcal{H}\}$ and $\mathcal{H}_3 := \{\min(h_1, h_2) : h_1, h_2 \in \mathcal{H}\}$. Then, both the VC-dimension of $\mathcal{H}_2$ and the VC-dimension of $\mathcal{H}_3$ are $O(d)$.*

The VC dimension of neural networks is also well understood (equation 2 of Bartlett et al. [6]):

Theorem 56. *The VC-dimension of a neural network with piecewise linear activation function, W parameters, and L layers has VC-dimension $O(WL \log W)$.*

Using Theorem 54, 55, and 56, we can verify the VC dimension assumptions made in Chapter 3 and 4. For example, the VC dimension for the square loss and the squared hinge loss is analyzed in Appendix E.2 and E.3 of Zhou et al. [79]. We can easily establish low-dimensional concentration due to the following result:

Theorem 57 (Vapnik [69], Special case of Assertion 4 in Chapter 7.8; see also Theorem 7.6)**.** *Suppose that the loss function $l : \mathcal{Z} \times \Theta \to \mathbb{R}_{\geq 0}$ satisfies*

(i) for every $\theta \in \Theta$, the function $l(\cdot, \theta)$ is measurable with respect to the first argument

(ii) the class of functions $\{z \mapsto \mathbb{1}\{l(z, \theta) > t\} : (\theta, t) \in \Theta \times \mathbb{R}\}$ has VC-dimension at most h

and the distribution $\mathcal{D}$ over $\mathcal{Z}$ satisfies for every $\theta \in \Theta$

$$\frac{\mathbb{E}_{z\sim\mathcal{D}}[l(z,\theta)^4]^{1/4}}{\mathbb{E}_{z\sim\mathcal{D}}[l(z,\theta)]} \leq \tau, \tag{B.1}$$

then for any $n > h$, with probability at least $1 - \delta$ over the choice of $(z_1, \ldots, z_n) \sim \mathcal{D}^n$, it holds uniformly over all $\theta \in \Theta$ that

$$\frac{1}{n}\sum_{i=1}^{n} l(z_i, \theta) \geq \left(1 - 8\tau\sqrt{\frac{h(\log(2n/h) + 1) + \log(12/\delta)}{n}}\right) \mathbb{E}_{z\sim\mathcal{D}}[l(z, \theta)]. \tag{B.2}$$

The assumption (B.1) is standard (indeed, this is the setting primarily focused on in [69]) and is sometimes referred to as *hypercontractivity* or *norm equivalence* in the literature; a variant of the result holds with 4 replaced by $1+\epsilon$. In many settings of interest, this can be directly checked using the fact that x is Gaussian (for instance, see section B.1 and B.2 below). Of course, our general result can be applied without this assumption, by using low-dimensional concentration under an alternative assumption: Vapnik [69], Panchenko [50, 51], Mendelson [44] have further discussion and alternative results; in particular, Assertion 3 of Vapnik [69, Chapter 7.8] gives a bound based on a fourth-moment assumption, and Panchenko [51, Theorem 3] gives one based on a version of Rademacher complexity.

B.1 Square Loss

The following theorem is the Gaussian space analogue of Theorem 9.21 in O'Donnell [49] and can be proved using the same argument by Theorem 11.23 and replacing the Fourier basis on $\{-1,1\}^n$ with the Hermite polynomials on $\mathbb{R}^n$.

Theorem 58 (O'Donnell [49])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a polynomial of degree at most k. Then for any $q \geq 2$, it holds that*

$$\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[|f(z)|^q]^{1/q} \leq (q-1)^{k/2}\mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[|f(z)|^2]^{1/2}. \tag{B.3}$$

We will now check the hypercontractivity assumption (C) when both the mean and standard deviation of y is a polynomial of $\eta_1, ..., \eta_k$.

Theorem 59. *Suppose that in* (4.2)*, we have*

$$y = m(\eta_1, ..., \eta_k) + s(\eta_1, ..., \eta_k) \cdot \xi$$

where m, s are both polynomials of degree at most l and ξ has finite eighth moment, then

$$\frac{\mathbb{E}[(\langle w, x\rangle + b - y)^8]^{1/8}}{\mathbb{E}[(\langle w, x\rangle + b - y)^2]^{1/2}} \leq \sqrt{2}\cdot\sqrt{7}^l\left(\frac{E[\xi^8]^{1/8}}{E[\xi^2]^{1/2}}\right). \tag{B.4}$$

Proof. By triangular inequality in the ℓ_p space and independence between x and ξ

$$\begin{aligned}\mathbb{E}[(\langle w, x\rangle + b - y)^8]^{1/8} &\leq \mathbb{E}[(\langle w, x\rangle + b - m(\eta_1, ..., \eta_k))^8]^{1/8} + \mathbb{E}[(s(\eta_1, ..., \eta_k)\cdot\xi)^8]^{1/8} \\ &= \mathbb{E}[(\langle w, x\rangle + b - m(\eta_1, ..., \eta_k))^8]^{1/8} + \mathbb{E}[s(\eta_1, ..., \eta_k)^8]^{1/8}\cdot\mathbb{E}[\xi^8]^{1/8}\end{aligned}$$

Since $\langle w, x\rangle, \eta_1, ..., \eta_k$ are jointly Gaussian, we can apply Theorem 58 and upper bound the above

by

$$\begin{aligned}
&\sqrt{7}^l \left(\mathbb{E}[(\langle w, x \rangle + b - m(\eta_1, ..., \eta_k))^2]^{1/2} + \mathbb{E}[s(\eta_1, ..., \eta_k)^2]^{1/2} \cdot \mathbb{E}[\xi^8]^{1/8} \right) \\
&\leq \sqrt{7}^l \left(\frac{E[\xi^8]^{1/8}}{E[\xi^2]^{1/2}} \right) \left(\mathbb{E}[(\langle w, x \rangle + b - m(\eta_1, ..., \eta_k))^2]^{1/2} + \mathbb{E}[s(\eta_1, ..., \eta_k)^2]^{1/2} \cdot \mathbb{E}[\xi^2]^{1/2} \right) \\
&\leq \sqrt{7}^l \left(\frac{E[\xi^8]^{1/8}}{E[\xi^2]^{1/2}} \right) \sqrt{2} \cdot \sqrt{\mathbb{E}[(\langle w, x \rangle + b - m(\eta_1, ..., \eta_k))^2] + \mathbb{E}[s(\eta_1, ..., \eta_k)^2] \cdot \mathbb{E}[\xi^2]}
\end{aligned}$$

where we use $E[\xi^8]^{1/8} \geq \mathbb{E}[\xi^2]^{1/2}$ in the second inequality and $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ in the last inequality. The desired conclusion follows by observing

$$\mathbb{E}[(\langle w, x \rangle + b - y)^2] = \mathbb{E}[(\langle w, x \rangle + b - m(\eta_1, ..., \eta_k))^2] + \mathbb{E}[s(\eta_1, ..., \eta_k)^2] \cdot \mathbb{E}[\xi^2]$$

because x and ξ are independent. □

Remark 60. *The assumption that ξ has finite eighth moment can be significantly relaxed because there is a version of Theorem 57 in Vapnik [69] that replaces the exponent of 4 by $1 + \epsilon$. However, allowing heavier tails of ξ comes at the cost of a larger constant in front of τ or a slower convergence rate with respect to n in the low-dimensional concentration term.*

B.2 Squared Hinge Loss

For illustration, we show how to check hypercontractivity (C) for the squared hinge loss under some example generative assumptions on y. In the first and simpler example, suppose that there is an arbitrary constant $\eta > 0$ such that

$$\min\{\Pr(y = 1 \mid x), \Pr(y = -1 \mid x)\} \geq \eta$$

almost surely. This assumption is satisfied, for example, if the data is generated by an arbitrary function of $\eta_1, \ldots, \eta_k$ combined with Random Classification Noise (see e.g. Blum et al. [13]), i.e.

the label is flipped with some probability. Then if $\hat{y} = \langle w, x \rangle + b$ is the prediction, we have

$$\mathbb{E} \max(0, 1 - y\hat{y})^2 \geq \eta \mathbb{E}(1 + |\hat{y}|)^2 \geq \eta(1 + \mathbb{E}[\hat{y}^2]),$$

and on the other hand we always have

$$\mathbb{E} \max(0, 1 - y\hat{y})^8 \leq \mathbb{E}(1 + |\hat{y}|)^8 \leq 2^8(1 + \mathbb{E}[\hat{y}^8]) \leq 2^{16}(1 + \mathbb{E}[\hat{y}^2]^4) \leq 2^{16}(1 + \mathbb{E}[\hat{y}^2])^4$$

where the second-to-last inequality follows from the fact that $\hat{y}$ is marginally Gaussian and using standard formula for the moments of a Gaussian. It follows that

$$\frac{\mathbb{E}[\max(0, 1 - y\hat{y})^8]^{1/8}}{\mathbb{E}[\max(0, 1 - y\hat{y})^2]^{1/2}} \leq \frac{4}{\sqrt{\eta}}$$

which verifies (4.11) in this setting.

We now consider a more general situation and show that if there is a *non-negligible* portion of x's such that that y is noisy, hypercontractivity is still guaranteed to hold. Let A_η be the event that $\min\{\Pr(y = 1 \mid x), \Pr(y = -1 \mid x)\} \geq \eta$. Then

$$\begin{aligned}\mathbb{E} \max(0, 1 - y\hat{y})^2 \geq \mathbb{E}[\mathbb{1}(A_\eta) \max(0, 1 - y\hat{y})^2] &\geq \eta \mathbb{E}[\mathbb{1}(A_\eta)(1 + |\hat{y}|)^2] \\ &\geq \eta Q(\Pr(A_\eta)) \mathbb{E}[(1 + |\hat{y}|)^2]\end{aligned}$$

where Q is defined below. In the last step, we considered the worst case event A_η for given $\Pr(A_\eta)$, which corresponds to chopping the tails off of $\hat{y}$; considering this example, we see the inequality holds where where $Q : (0, 1] \to (0, 1]$ is an explicit function

$$Q(p) := \min \left\{ \frac{\int_{-z_p}^{z_p} |x| e^{-x^2/2} dx}{2}, \frac{\int_{-z_p}^{z_p} x^2 e^{-x^2/2} dx}{\sqrt{2\pi}} \right\} \tag{B.5}$$

and z_p is defined such that $\Pr_{g \sim N(0,1)}[|g| > z_p] = p$. Repeating the argument above yields the

following result:

Theorem 61. *Suppose that under* (4.2)*, there exists* $\eta > 0$ *such that* $p_\eta := \Pr(\min\{\Pr(y = 1 \mid x), \Pr(y = -1 \mid x)\} \geq \eta) > 0$*. Then for any* w, b *we have that for* $\hat{y} = \langle w, x \rangle + b$*,*

$$\frac{\mathbb{E}[\max(0, 1 - y\hat{y})^8]^{1/8}}{\mathbb{E}[\max(0, 1 - y\hat{y})^2]^{1/2}} \leq \frac{4}{\sqrt{\eta Q(p_\eta)}}$$

For another example, if y follows a logistic regression model $\mathbb{E}[y \mid x] = \tanh(\beta w_1^* \cdot x)$ with normalization $\langle w_1^*, \Sigma w_1^* \rangle = 1$, then by Theorem 61 with e.g. $\eta = 1/2$, we verify (4.11) with τ a constant depending only on β. The result also holds for more general models like $\mathbb{E}[y \mid x] = \tanh(f(\eta_1, \ldots, \eta_k))$ as long as f is not always very large.

APPENDIX C

PROOFS FOR SECTION 2

C.1 Proofs for Section 2.2

Lemma 62. *For any $k \in \mathbb{N}$, it holds that*

$$\kappa_0 \geq \left(1 - \frac{n}{R_k}\right) \frac{\sum_{i>k} \lambda_i}{n} \quad \textit{and} \quad \kappa_0 \geq \lambda_{k+1} \left(\frac{k + r_k}{n} - 1\right). \tag{C.1}$$

Moreover, for any $k < n$, it holds that

$$\kappa_0 \leq \left(1 - \frac{k}{n}\right)^{-1} \frac{\sum_{i>k} \lambda_i}{n} \tag{C.2}$$

Proof. Observe the following Cauchy-Schwarz inequality:

$$\begin{aligned}
\left(\sum_{i>k} \lambda_i\right)^2 &= \left(\sum_{i>k} \sqrt{\frac{\lambda_i}{\lambda_i + \kappa_0}} \sqrt{\lambda_i(\lambda_i + \kappa_0)}\right)^2 \\
&\leq \left(\sum_{i>k} \frac{\lambda_i}{\lambda_i + \kappa_0}\right) \left(\sum_{i>k} \lambda_i(\lambda_i + \kappa_0)\right) \\
&\leq \left(\sum_{i} \frac{\lambda_i}{\lambda_i + \kappa_0}\right) \left(\sum_{i>k} \lambda_i(\lambda_i + \kappa_0)\right) \\
&= n \left(\sum_{i>k} \lambda_i^2 + \kappa_0 \sum_{i>k} \lambda_i\right).
\end{aligned}$$

Rearranging in terms of κ_0 proves the first inequality. Moreover, it holds that

$$\begin{aligned}
n &= \sum_{i \leq k} \frac{\lambda_i}{\lambda_i + \kappa_0} + \sum_{i>k} \frac{\lambda_i}{\lambda_i + \kappa_0} \\
&\geq \frac{k\lambda_{k+1}}{\lambda_{k+1} + \kappa_0} + \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1} + \kappa_0}.
\end{aligned}$$

which can be rearranged to the second lower bound. Finally, observe that

$$n = \sum_i \frac{\lambda_i}{\lambda_i + \kappa_0} \leq k + \frac{\sum_{i>k} \lambda_i}{\kappa_0}$$

and rearranging concludes the proof of the last inequality. □

Lemma 63. *For any $\delta \geq 0$ and any $k \in \mathbb{N}$ such that $k < n$, it holds that*

$$\frac{\delta}{n} \leq \kappa_\delta \leq \frac{\sum_{i>k} \lambda_i + \delta}{n-k} \tag{C.3}$$

Proof. Observe that

$$\frac{\delta}{\kappa_\delta} \leq \sum_i \frac{\lambda_i}{\lambda_i + \kappa_\delta} + \frac{\delta}{\kappa_\delta} \leq k + \sum_{i>k} \frac{\lambda_i}{\kappa_\delta} + \frac{\delta}{\kappa_\delta}$$

and the proof concludes by plugging in (2.8) and some rearrangement. □

Proposition 2. *For any k such that $k < n$ and $n < R_k$, it holds that*

$$\mathcal{E}_0 \leq \left(1 - \frac{k}{n}\right)^{-2} \left(1 - \frac{n}{R_k}\right)^{-1}. \tag{2.14}$$

Proof. By definition, we have

$$\begin{aligned} n - \frac{\delta}{\kappa_\delta} &= \sum_i \frac{\lambda_i}{\lambda_i + \kappa_\delta} \\ &\leq \sum_{i \leq k} \frac{\lambda_i}{\lambda_i + \kappa_\delta} + \sum_{i>k} \frac{\sqrt{\lambda_i}}{\lambda_i + \kappa_\delta} \sqrt{\lambda_i} \\ &\leq k + \sqrt{\sum_{i>k} \frac{\lambda_i}{(\lambda_i + \kappa_\delta)^2} \sum_{i>k} \lambda_i}. \end{aligned}$$

Rearranging, we get

$$\frac{\left(n - k - \frac{\delta}{\kappa_\delta}\right)^2}{\sum_{i>k} \lambda_i} \leq \sum_{i>k} \frac{\lambda_i}{(\lambda_i + \kappa_\delta)^2}.$$

At the same time, observe that

$$\begin{aligned} 1 - \frac{1}{n}\sum_i \mathcal{L}_{i,\delta}^2 &= \frac{1}{n}\left(\sum_i \frac{\lambda_i}{\lambda_i + \kappa_\delta} + \frac{\delta}{\kappa_\delta} - \left(\frac{\lambda_i}{\lambda_i + \kappa_\delta}\right)^2\right) \\ &= \frac{\kappa_\delta}{n}\sum_i \frac{\lambda_i}{(\lambda_i + \kappa_\delta)^2} + \frac{\delta}{n\kappa_\delta} \\ &\geq \frac{\kappa_\delta}{n}\frac{\left(n - k - \frac{\delta}{\kappa_\delta}\right)^2}{\sum_{i>k}\lambda_i} + \frac{\delta}{n\kappa_\delta}. \end{aligned} \tag{C.4}$$

Plugging in $\delta = 0$ and Lemma 62, we have

$$\mathcal{E}_0 = \left(1 - \frac{1}{n}\sum_i \mathcal{L}_{i,0}^2\right)^{-1} \leq \left(\frac{\kappa_0}{n}\frac{(n-k)^2}{\sum_{i>k}\lambda_i}\right)^{-1} = \left(1 - \frac{k}{n}\right)^{-2}\left(1 - \frac{n}{R_k}\right)^{-1}.$$

□

Proposition 3. *Fix any $b > 0$. If there exists $k < n$ such that $n \leq k + br_k$, then let k be the first such integer. Otherwise, pick $k = n$. It holds that*

$$\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \geq \max\left\{\frac{1}{(b+1)^2}\left(1 - \frac{k}{n}\right)^2 \frac{n}{R_k}, \left(\frac{b}{b+1}\right)^2 \frac{k}{n}\right\}. \tag{2.15}$$

Proof. Observe that for any $k < n$

$$\begin{aligned} \sum_i \mathcal{L}_{i,0}^2 &\geq \sum_{i:\lambda_i \leq \lambda}\left(\frac{\lambda_i}{\lambda_i + \kappa_0}\right)^2 \\ &\geq \frac{\sum_{i:\lambda_i \leq \lambda}\lambda_i^2}{(\lambda + \frac{\sum_{i>k}\lambda_i}{n-k})^2} \end{aligned}$$

Without loss of generality, sort the eigenvalues in descending order. Then pick the first $k' < n$

such that $\lambda_{k'+1} \leq b\frac{\sum_{i>k'}\lambda_i}{n-k'}$ and set $\lambda = \lambda_{k'+1}$. By the choice of λ, we have

$$\sum_i \mathcal{L}_{i,0}^2 \geq \frac{1}{(b+1)^2}\frac{\sum_{i>k'}\lambda_i^2}{\left(\frac{\sum_{i>k'}\lambda_i}{n-k'}\right)^2} \tag{C.5}$$

which can be rearranged to the first part of the desired inequality. Moreover, we know that for any $k < k' < n$, we have $\lambda_{k+1} > b\frac{\sum_{i>k}\lambda_i}{n-k}$ and so by the second part of Lemma 62, we have

$$\kappa_0 \leq \frac{\sum_{i>k}\lambda_i}{n-k} \leq \frac{1}{b}\lambda_{k+1}.$$

Therefore, we have

$$\begin{aligned}\sum_i \mathcal{L}_{i,0}^2 &\geq \sum_{k\leq k'}\left(\frac{\lambda_k}{\lambda_k+\kappa_0}\right)^2 \\ &\geq k'\left(\frac{b}{b+1}\right)^2.\end{aligned} \tag{C.6}$$

and we are done with the case when there exists such k'. Finally, if there is no such k', then by the same reasoning, we have

$$\begin{aligned}\sum_i \mathcal{L}_{i,0}^2 &\geq \sum_{k\leq n}\left(\frac{\lambda_k}{\lambda_k+\kappa_0}\right)^2 \\ &\geq n\left(\frac{b}{b+1}\right)^2\end{aligned}$$

and we are done. □

Proposition 5. *Suppose there exists $m, M > 0$ such that $m \leq r_k/k \leq M$ for $k \geq \lfloor\frac{n}{1+M}\rfloor$. Then it holds that*

$$\mathcal{E}_0 \leq 4\frac{\left(\frac{1}{1+M}-\frac{1}{n}\right)^{-1}}{m}. \tag{2.17}$$

If $\{\lambda_i\}$ *does not change with* n *and* $\lim_{k\to\infty} r_k/k = \alpha > 0$*, then*

$$\lim_{n\to\infty} \mathcal{E}_0 \leq 4\left(1 + \frac{1}{\alpha}\right). \tag{2.18}$$

Proof. Let k be the largest integer such that $\lambda_k \geq \kappa_0$. Then it holds that

$$\begin{aligned} 1 - \frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 &= \frac{1}{n}\sum_i \mathcal{L}_{i,0} - \mathcal{L}_{i,0}^2 \\ &= \frac{\kappa_0}{n}\sum_i \frac{\lambda_i}{(\lambda_i + \kappa_0)^2} \\ &\geq \frac{\kappa_0}{n}\sum_{i>k} \frac{\lambda_i}{(\lambda_i + \kappa_0)^2} \\ &\geq \frac{\kappa_0}{n}\frac{\lambda_k}{4\kappa_0^2}\frac{\sum_{i>k}\lambda_i}{\lambda_k} \geq \frac{1}{4}\frac{mk}{n}. \end{aligned} \tag{C.7}$$

It remains to lower bound k and verify that $k \geq \lfloor \frac{n}{1+M} \rfloor$. By construction, it holds that $\lambda_{k+1} < \kappa_0$. Also, by Lemma 62 we know that

$$\kappa_0 \leq \frac{1}{n-k'}\sum_{i>k'} \lambda_i = \frac{r_{k'}}{n-k'}\lambda'_k \leq \frac{Mk'}{n-k'}\lambda'_k.$$

Choosing $k' = \lfloor \frac{n}{1+M} \rfloor$, then the above become

$$\lambda_{k+1} < \kappa_0 \leq M\frac{k'}{n-k'}\lambda'_k \leq \lambda'_k.$$

Since the λ_i are sorted, it must be the case that

$$k \geq k' \implies k \geq \frac{n}{1+M} - 1 \tag{C.8}$$

and so we have

$$1 - \frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \geq \frac{1}{4}m\left(\frac{1}{1+M} - \frac{1}{n}\right).$$

Taking the inverse concludes the proof of the first part. Finally, suppose that $\lim_{k\to\infty} r_k/k = \alpha$, then fix any $\epsilon \in (0, \alpha)$, there exists N such that $\forall k \geq N$, $\alpha - \epsilon \leq r_k/k \leq \alpha + \epsilon$. Then for $n \geq (1 + 2\alpha)(N + 1)$, we can choose $m = \alpha - \epsilon$ and $M = \alpha + \epsilon$ because

$$\lfloor \frac{n}{1+M} \rfloor \geq \frac{n}{1+M} - 1 \geq \frac{n}{1+2\alpha} - 1 \geq N.$$

Applying the non-asymptotic result and sending $n \to \infty$, we obtain

$$\lim_{n\to\infty} \mathcal{E}_0 \leq 4\frac{1+\alpha+\epsilon}{\alpha-\epsilon}.$$

Finally, we conclude the proof by sending $\epsilon \to 0$. □

Proposition 6. *For any $k \geq n + r_k$, it holds that*

$$\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \geq \frac{n}{k}\left(1 - \frac{r_k}{k-n}\right)^2. \tag{2.19}$$

Therefore, if $\{\lambda_i\}$ does not change with n and $\lim_{k\to\infty} r_k/k = 0$, then, then it holds that

$$\lim_{n\to\infty} \mathcal{E}_0 = \infty.$$

Proof. By Cauchy-Schwarz, we have

$$\begin{aligned} n &= \sum_i \frac{\lambda_i}{\lambda_i + \kappa_0} \\ &= \sum_{i>k} \frac{\lambda_i}{\lambda_i + \kappa_0} + \sum_{i\leq k} \frac{\lambda_i}{\lambda_i + \kappa_0} \\ &\leq \frac{\sum_{i>k}\lambda_i}{\kappa_0} + \sqrt{k}\sqrt{\sum_{i\leq k}\left(\frac{\lambda_i}{\lambda_i+\kappa_0}\right)^2} \end{aligned}$$

By Lemma 62, we have $\kappa_0 \geq \lambda_{k+1}\left(\frac{k}{n}-1\right)$. Combine with above, we obtain

$$n \leq r_k \frac{n}{k-n} + \sqrt{k}\sqrt{\sum_{i\leq k}\left(\frac{\lambda_i}{\lambda_i+\kappa_0}\right)^2}$$

Rearranging gives us

$$\frac{n\left(1-\frac{r_k}{k-n}\right)}{\sqrt{k}} \leq \sqrt{\sum_{i\leq k}\left(\frac{\lambda_i}{\lambda_i+\kappa_0}\right)^2}$$

$$\implies \frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \geq \frac{1}{n}\sum_{i\leq k}\left(\frac{\lambda_i}{\lambda_i+\kappa_0}\right)^2 \geq \frac{n}{k}\left(1-\frac{r_k}{k-n}\right)^2$$

For any $\epsilon > 0$, choose $k = (1+\epsilon)n$, we get

$$\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \geq \frac{1}{1+\epsilon}\left(1-\frac{r_k}{k}\frac{1+\epsilon}{\epsilon}\right)^2$$

Therefore, if $r_k = o(k)$, then

$$\lim_{n\to\infty}\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \geq \frac{1}{1+\epsilon}$$

However, since the choice of ϵ is arbitrary, then we can send $\epsilon \to 0$. The desired conclusion follows by $\mathcal{E}_0 = \left(1-\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2\right)^{-1}$. □

Theorem 7. *Suppose that $\{\lambda_i\}$ does not change with n and the optimally tuned ridge regression is consistent: $\mathcal{E}(f_{\delta^*}) \to \sigma^2$, then*

(i) if $\lim_{k\to\infty} r_k/k = \infty$, then overfitting is benign:

$$\lim_{n\to\infty}\mathcal{E}(f_0) = \sigma^2 \tag{2.20}$$

(ii) if $\lim_{k\to\infty} r_k/k \in (0,\infty)$, then overfitting is tempered:

$$\sigma^2 < \lim_{n\to\infty}\mathcal{E}(f_0) < \infty \tag{2.21}$$

(ii) if $\lim_{k\to\infty} r_k/k = 0$*, then overfitting is catastrophic:*

$$\lim_{n\to\infty} \mathcal{E}(f_0) = \infty. \tag{2.22}$$

Proof. First, assume that $r_k/k \to \infty$. For any $\epsilon > 0$, we can pick $k = \epsilon n$ in Proposition 2 and obtain the following:

$$\mathcal{E}_0 \leq \frac{1}{(1-\epsilon)^2}\left(1 - \frac{1}{\epsilon}\frac{k}{R_k}\right)^{-1}. \tag{C.9}$$

Since we have

$$\sum_{i>k} \lambda_i^2 \leq \lambda_{k+1} \sum_{i>k} \lambda_i \implies R_k \geq r_k,$$

we can send $n \to \infty$ and $k/R_k \leq k/r_k \to 0$. Therefore, it holds that

$$\lim_{n\to\infty} \mathcal{E}_0 \leq \frac{1}{(1-\epsilon)^2}.$$

Since the choice of $\epsilon > 0$ can be made arbitrarily small, we have the desired conclusion. On the other hand, suppose that r_k/k is bounded, then there exists $M > 0$ such that $r_k < kM$ for all k. If we let $b = 1/(3M)$, then for all $k \leq n/2$, it holds that

$$k + br_k < k(1 + bM) \leq \frac{1+bM}{2} n \leq \frac{2n}{3} < n.$$

Then we can apply Theorem 3 to show that

$$\frac{1}{n}\sum_i \mathcal{L}_{i,0}^2 \geq \frac{1}{2(1+3M)^2}$$

and so $\lim_{n\to\infty} \mathcal{E}_0 > 1$. Finally, (ii) and (iii) follows from Theorem 1, Proposition 5 and 6. □

C.2 Proofs for Section 2.3 and 2.4

Theorem 8. *Fix any $l \in \mathbb{N} \cup \{\infty\}$ and $k < n/2$. For any $k' < n$, it holds that*

$$\mathcal{E}(f_{\delta^*}) \leq \frac{1+\epsilon}{1-\epsilon-\frac{k'}{n}} \left[\sigma^2 + \sum_{i>l} v_i^2 + \frac{\left(\sum_{i>k} \lambda_i\right)\left(\sum_{i\leq l} \frac{v_i^2}{\lambda_i}\right)}{n-k} \right] \tag{2.23}$$

where ϵ is defined as

$$\epsilon = \left(\frac{2\sum_{i\leq l} \frac{v_i^2}{\lambda_i}}{\sigma^2}\right)^{2/3} \left(\frac{\sum_{i>k'} \lambda_i^2}{n}\right)^{1/3}. \tag{2.24}$$

Proof. For any $k < n$, it holds that

$$\begin{aligned}
\sum_i (1-\mathcal{L}_{i,\delta})^2 v_i^2 &= \kappa_\delta \sum_{i\leq l} \frac{\kappa_\delta \lambda_i}{(\lambda_i+\kappa_\delta)^2} \frac{v_i^2}{\lambda_i} + \sum_{i>l} v_i^2 \\
&\leq \kappa_\delta \sum_{i\leq l} \frac{v_i^2}{\lambda_i} + \sum_{i>l} v_i^2 \\
&\leq \frac{\delta + \sum_{i>k} \lambda_i}{n-k} \sum_{i\leq l} \frac{v_i^2}{\lambda_i} + \sum_{i>l} v_i^2 \\
&= \frac{\sum_{i\leq l} \frac{v_i^2}{\lambda_i}}{n-k}\delta + \frac{\left(\sum_{i>k} \lambda_i\right)\left(\sum_{i\leq l} \frac{v_i^2}{\lambda_i}\right)}{n-k} + \sum_{i>l} v_i^2
\end{aligned}$$

and for any $k' < n$

$$\frac{1}{n}\sum_i \left(\frac{\lambda_i}{\lambda_i+\kappa_\delta}\right)^2 \leq \frac{k'}{n} + \frac{\sum_{i>k'} \lambda_i^2}{n\kappa_\delta^2} \leq \frac{k'}{n} + \frac{n\sum_{i>k'} \lambda_i^2}{\delta^2}.$$

To balance the two terms, we choose δ such that

$$\sigma^2 \frac{n\sum_{i>k'} \lambda_i^2}{\delta^2} = \frac{\sum_{i\leq l} \frac{v_i^2}{\lambda_i}}{n-k}\delta \implies \delta = \sigma^{2/3} \left(\frac{n(n-k)\sum_{i>k'} \lambda_i^2}{\sum_{i\leq l} \frac{v_i^2}{\lambda_i}}\right)^{1/3}$$

and so using $k \leq n/2$, we have

$$\frac{\sum_{i \leq l} \frac{v_i^2}{\lambda_i}}{n-k}\delta = \left(\frac{\sigma\sqrt{\sum_{i>k'} \lambda_i^2}\sum_{i \leq l} \frac{v_i^2}{\lambda_i}}{(n-k)/\sqrt{n}}\right)^{2/3} \leq \sigma^2\epsilon.$$

The proof concludes by plugging into the definition of $\mathcal{E}(f_{\delta^*})$. □

Theorem 9. *For any $\delta \geq 0$ and $k \in \mathbb{N}$ such that $(k/n)^2 + 2(k/n) < 1$. Let $\epsilon = \sqrt{(k^2+2kn)/n^2}$, then it holds that*

$$(1-\epsilon)^2\mathcal{E}(f_\delta) \leq \left(\sqrt{\mathcal{E}_{tr}(f_\delta)} + \sqrt{\frac{(\sum_{i>k}\lambda_i)\mathbb{E}\|f_\delta\|_{\mathcal{K}}^2}{n}}\right)^2. \tag{2.26}$$

Proof. Applying equation (2.10) and (2.8), we can write the difference

$$\begin{aligned}\left(\sqrt{\mathcal{E}(f_\delta)} - \sqrt{\mathcal{E}_{\text{tr}}(f_\delta)}\right)^2 &= \left(1 - \frac{\delta}{n\kappa_\delta}\right)^2 \mathcal{E}(f_\delta) \\ &\leq \frac{1}{n}\left(\sum_i \frac{\lambda_i}{\lambda_i + \kappa_\delta}\right)^2 \frac{\mathcal{E}(f_\delta)}{n}.\end{aligned}$$

We can do covariance splitting by

$$\begin{aligned}\left(\sum_i \frac{\lambda_i}{\lambda_i+\kappa_\delta}\right)^2 &\leq \left(k + \sum_{i>k}\frac{\lambda_i}{\lambda_i+\kappa_\delta}\right)^2 \\ &= k^2 + 2k\left(\sum_{i>k}\frac{\lambda_i}{\lambda_i+\kappa_\delta}\right) + \left(\sum_{i>k}\frac{\sqrt{\lambda_i}}{\lambda_i+\kappa_\delta}\sqrt{\lambda_i}\right)^2 \\ &\leq k^2 + 2kn + \left(\sum_{i>k}\frac{\lambda_i}{(\lambda_i+\kappa_\delta)^2}\right)\left(\sum_{i>k}\lambda_i\right)\end{aligned}$$

where in the last inequality we use (2.8) and Cauchy-Schwarz, and so by (2.11)

$$\left(\sqrt{\mathcal{E}(f_\delta)} - \sqrt{\mathcal{E}_{\text{tr}}(f_\delta)}\right)^2 \leq \frac{k^2+2kn}{n^2}\mathcal{E}(f_\delta) + \left(\frac{\mathcal{E}(f_\delta)}{n}\sum_{i>k}\frac{\lambda_i}{(\lambda_i+\kappa_\delta)^2}\right)\left(\frac{1}{n}\sum_{i>k}\lambda_i\right)$$
$$\leq \frac{k^2+2kn}{n^2}\mathcal{E}(f_\delta) + \frac{\mathbb{E}\|f_\delta\|_{\mathcal{K}}^2(\sum_{i>k}\lambda_i)}{n}$$

then using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$

$$\sqrt{\mathcal{E}(f_\delta)} - \sqrt{\mathcal{E}_{\text{tr}}(f_\delta)} \leq \sqrt{\frac{k^2+2kn}{n^2}\mathcal{E}(f_\delta) + \frac{\mathbb{E}\|f_\delta\|_{\mathcal{K}}^2(\sum_{i>k}\lambda_i)}{n}}$$
$$\leq \sqrt{\frac{k^2+2kn}{n^2}\mathcal{E}(f_\delta)} + \sqrt{\frac{\mathbb{E}\|f_\delta\|_{\mathcal{K}}^2(\sum_{i>k}\lambda_i)}{n}}.$$

Re-arranging concludes the proof. □

Proposition 10. *For any $l \in \mathbb{N} \cup \{\infty\}$ and $k \in \mathbb{N}$ such that $R_k > n$, it holds that*

$$\mathbb{E}\|f_0\|_{\mathcal{K}}^2 \leq \sum_{i\leq l}\frac{v_i^2}{\lambda_i} + \left(1 - \frac{n}{R_k}\right)^{-1}\frac{n\left(\sigma^2 + \sum_{i>l}v_i^2\right)}{\sum_{i>k}\lambda_i}. \tag{2.27}$$

Proof. When $\delta = 0$, it holds that

$$\frac{n}{\mathcal{E}_0} = n - \sum_i \mathcal{L}_{i,0}^2 = \sum_i \frac{\lambda_i}{\lambda_i+\kappa_0} - \frac{\lambda_i^2}{(\lambda_i+\kappa_0)^2}$$
$$= \sum_i \frac{\lambda_i(\lambda_i+\kappa_0) - \lambda_i^2}{(\lambda_i+\kappa_0)^2}$$
$$= \kappa_0\left(\sum_i \frac{\lambda_i}{(\lambda_i+\kappa_0)^2}\right)$$

by applying (2.9) and (2.8). Therefore, the first term in (2.11) can be simplified as

$$\begin{aligned}\frac{\mathcal{E}(f_0)}{n}\sum_i \frac{\lambda_i}{(\lambda_i+\kappa_0)^2} &= \frac{\mathcal{E}_0\left(\sum_i (1-\mathcal{L}_{i,0})^2 v_i^2+\sigma^2\right)}{n}\sum_i \frac{\lambda_i}{(\lambda_i+\kappa_0)^2} \\ &= \sum_i \frac{(1-\mathcal{L}_{i,0})^2}{\kappa_0} v_i^2 + \frac{\sigma^2}{\kappa_0} \\ &= \sum_i \frac{\kappa_0}{(\lambda_i+\kappa_0)^2} v_i^2 + \frac{\sigma^2}{\kappa_0}\end{aligned}$$

by the definition in (2.10) and (2.9). Plugging in, we arrive at

$$\mathbb{E}\|f_0\|_{\mathcal{K}}^2 = \sum_i \frac{v_i^2}{\lambda_i+\kappa_0} + \frac{\sigma^2}{\kappa_0} \tag{C.10}$$

To handle situations where f^* is not in the RKHS, observe that for any l, we have

$$\begin{aligned}\sum_i \frac{v_i^2}{\lambda_i+\kappa_0} &= \sum_{i\le l} \frac{v_i^2}{\lambda_i+\kappa_0} + \sum_{i>l} \frac{v_i^2}{\lambda_i+\kappa_0} \\ &\le \sum_{i\le l} \frac{v_i^2}{\lambda_i} + \frac{1}{\kappa_0}\sum_{i>l} v_i^2\end{aligned}$$

and so

$$\mathbb{E}\|f_0\|_{\mathcal{K}}^2 \le \sum_{i\le l} \frac{v_i^2}{\lambda_i} + \frac{1}{\kappa_0}\left(\sigma^2 + \sum_{i>l} v_i^2\right). \tag{C.11}$$

The proof concludes by plugging in Lemma 62. □

APPENDIX D

PROOFS FOR SECTION 3

D.1 Proofs for Section 3.2

We will make use of the following standard concentration result.

Lemma 64. *For any covariance matrix Σ, it holds that with probability at least $1-\delta$,*

$$1 - \frac{\|\Sigma^{1/2}H\|_2^2}{\mathrm{Tr}(\Sigma)} \lesssim \frac{\log(4/\delta)}{\sqrt{R(\Sigma)}} \tag{D.1}$$

and

$$\|\Sigma H\|_2^2 \lesssim \log(4/\delta)\,\mathrm{Tr}(\Sigma^2). \tag{D.2}$$

Therefore, provided that $R(\Sigma) \gtrsim \log(4/\delta)^2$, it holds that

$$\left(\frac{\|\Sigma H\|_2}{\|\Sigma^{1/2}H\|_2}\right)^2 \lesssim \log(4/\delta)\frac{\mathrm{Tr}(\Sigma^2)}{\mathrm{Tr}(\Sigma)}. \tag{D.3}$$

Proof. Because we are considering ℓ_2 norm and H is standard Gaussian, without loss of generality we can assume that Σ is diagonal and we denote the diagonals of Σ as $\lambda_1, ..., \lambda_d$. By the sub-exponential Bernstein inequality [71, Corollary 2.8.3], we have with probability at least $1-\delta/2$

$$\left|\frac{\|\Sigma^{1/2}H\|_2^2}{\mathrm{Tr}(\Sigma)} - 1\right| = \left|\sum_{i=1}^p \frac{\lambda_i}{\sum_j \lambda_j}(H_i^2 - 1)\right| \lesssim \sqrt{\frac{\log(4/\delta)}{R(\Sigma)}} \vee \frac{\log(4/\delta)}{r(\Sigma)} \leq \frac{\log(4/\delta)}{\sqrt{R(\Sigma)}}$$

where the last inequality uses that $R(\Sigma) \leq r(\Sigma)^2$, shown in Lemma 5 of Bartlett et al. [7]. Using the sub-exponential Bernstein inequality again, we show with probability at least $1-\delta/2$

$$\left|\frac{\|\Sigma H\|_2^2}{\mathrm{Tr}(\Sigma^2)} - 1\right| \lesssim \sqrt{\frac{\log(4/\delta)}{R(\Sigma^2)}} \vee \frac{\log(4/\delta)}{r(\Sigma^2)}$$

From Lemma 5 of Bartlett et al. [7], we know that the effective ranks are at least 1. This implies

$$\|\Sigma H\|_2^2 \lesssim \log(4/\delta)\,\mathrm{Tr}(\Sigma^2).$$

Provided that $R(\Sigma) \gtrsim \log(4/\delta)^2$, we have

$$\|\Sigma^{1/2} H\|_2^2 \geq \frac{1}{2}\,\mathrm{Tr}(\Sigma)$$

in which case it holds that

$$\frac{\|\Sigma H\|_2^2}{\|\Sigma^{1/2} H\|_2^2} \lesssim \log(4/\delta)\frac{\mathrm{Tr}(\Sigma^2)}{\mathrm{Tr}(\Sigma)}. \qquad \square$$

Corollary 15. *Assume that (A), (B), and (C) holds. Let* $Q = I - W(W^T\Sigma W)^{-1}W^T\Sigma$ *and* $\Sigma^\perp = Q^T\Sigma Q$. *Fix any* $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$ *such that* $Qw^\sharp = 0$. *Consider the minimal norm interpolator* $\hat{w}, \hat{b} = \arg\min_{(w,b):\hat{L}(w,b)=0} \|w\|_2$. *Then with probability at least* $1-\delta$, *for some*

$$\rho \lesssim \tau\sqrt{\frac{k\log(n/k) + \log(1/\delta)}{n}} + \log(1/\delta)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma^\perp)}} + \frac{k}{n} + \frac{n}{R(\Sigma^\perp)}\right),$$

it holds that

$$L(\hat{w}, \hat{b}) \leq (1+\rho)\left(\sqrt{L(w^\sharp, b^\sharp)} + \|w^\sharp\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma^\perp)}{n}}\right)^2. \tag{3.12}$$

Proof. By Lemma 64, there exists some constant $C > 0$ such that with probability at least $1 - \delta/8$

$$\begin{aligned}\langle Qw, x\rangle &\leq \|w\|_2 \cdot \|Q^T x\|_2 \\ &\leq \|w\|_2 \cdot \left(1 + C\frac{\log(32/\delta)}{\sqrt{R(\Sigma^\perp)}}\right)^{1/2}\sqrt{\mathrm{Tr}(\Sigma^\perp)}\end{aligned}$$

and so we can choose C_δ to be the above upper bound in Theorem 12. It holds that with probability

at least $1-\delta/2$, uniformly over all $(\hat{w},\hat{b})$ such that $\hat{L}(\hat{w},\hat{b})=0$, we have

$$L(\hat{w},\hat{b}) \le (1-\epsilon)^{-1}\left(1+C\frac{\log(32/\delta)}{\sqrt{R(\Sigma^{\perp})}}\right)\frac{\|\hat{w}\|_2^2\operatorname{Tr}(\Sigma^{\perp})}{n}.$$

Moreover, we apply Theorem 14 to show with probability at least $1-\delta/2$

$$\|\hat{w}\|_2 \le \|w^{\sharp}\|_2 + (1+\epsilon')\sqrt{\frac{nL_f(w^{\sharp},b^{\sharp})}{\operatorname{Tr}(\Sigma^{\perp})}}$$

then by a union bound, we have with probability at least $1-\delta$, it holds that

$$L(\hat{w},\hat{b}) \le \frac{(1+\epsilon')^2\left(1+C\frac{\log(32/\delta)}{\sqrt{R(\Sigma^{\perp})}}\right)}{1-\epsilon}\left(\|w^{\sharp}\|_2\sqrt{\frac{\operatorname{Tr}(\Sigma^{\perp})}{n}}+\sqrt{L_f(w^{\sharp},b^{\sharp})}\right)^2.$$

□

Flatness of generalization curve. Next, we prove a general lemma that shows any constrained ERM with a sufficiently small regularization can enjoy the same generalization guarantee as the minimal norm interpolator under the benign overfitting conditions. Note that we only require $\sqrt{f}$ to be convex below and so this result can be applied to linear classification with the squared hinge loss as well as linear regression with the square loss in Chapter 3.

Lemma 65. *Suppose that $\sqrt{f}$ is convex and there exists a convex function $C:\mathbb{R}^d\to\mathbb{R}$ such that for any $(w,b)\in\mathbb{R}^{d+1}$*

$$(1-\epsilon)L_f(w,b) \le \left(\sqrt{\hat{L}_f(w,b)}+\frac{C(w)}{\sqrt{n}}\right)^2.$$

Let $(w',b')\in\mathbb{R}^{d+1}$ be any interpolator: $f(\langle w',x_i\rangle+b',y_i)=0$ for all $i\in[n]$. Then for any $(w,b)\in\mathbb{R}^{d+1}$ and any R between $C(w)$ and $C(w')$, consider the constrained empirical risk

minimizer of the form

$$w_R, b_R := \arg \min_{(w,b):C(w)\leq R} \hat{L}_f(w, b).$$

It holds that

$$(1-\epsilon)L_f(w_R, b_R) \leq \max\left\{\left(\sqrt{\hat{L}_f(w,b)} + \frac{C(w)}{\sqrt{n}}\right)^2, \frac{C(w')^2}{n}\right\} \tag{D.4}$$

Proof. For any R between $C(w)$ and $C(w')$, we can write

$$R = (1-\alpha)C(w) + \alpha C(w')$$

for some $\alpha \in [0,1]$. If we define $w_\alpha := (1-\alpha)w + \alpha w'$ and $b_\alpha := (1-\alpha)b + \alpha b'$, then by convexity, we have

$$C(w_\alpha) \leq (1-\alpha)C(w) + \alpha C(w') = R$$

and by definition, we have $\hat{L}_f(w_R, b_R) \leq \hat{L}_f(w_\alpha, b_\alpha)$. In addition, by the convexity of $\sqrt{f}$ and Jensen's inequality, we have

$$\begin{aligned}
\hat{L}_f(w_\alpha, b_\alpha) &= \frac{1}{n}\sum_{i=1}^n f(\langle w_\alpha, x_i\rangle + b_\alpha, y_i) \\
&= \frac{1}{n}\sum_{i=1}^n \sqrt{f((1-\alpha)(\langle w, x_i\rangle + b) + \alpha(\langle w', x_i\rangle + b'), y_i)}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n \left((1-\alpha)\sqrt{f(\langle w, x_i\rangle + b, y_i)} + \alpha\sqrt{f(\langle w', x_i\rangle + b', y_i)}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (1-\alpha)^2 f(\langle w, x_i\rangle + b, y_i) = (1-\alpha)^2 \hat{L}_f(w, b)
\end{aligned}$$

and so

$$\begin{aligned}
\sqrt{(1-\epsilon)L_f(w_R,b_R)} &\le \sqrt{\hat{L}_f(w_R,b_R)} + \frac{C(w_R)}{\sqrt{n}} \\
&\le \sqrt{\hat{L}_f(w_\alpha,b_\alpha)} + \frac{R}{\sqrt{n}} \\
&\le (1-\alpha)\sqrt{\hat{L}_f(w,b)} + \frac{(1-\alpha)C(w)+\alpha C(w')}{\sqrt{n}} \\
&= (1-\alpha)\left(\sqrt{\hat{L}_f(w,b)} + \frac{C(w)}{\sqrt{n}}\right) + \alpha\frac{C(w')}{\sqrt{n}} \\
&\le \max\left\{\sqrt{\hat{L}_f(w,b)} + \frac{C(w)}{\sqrt{n}}, \frac{C(w')}{\sqrt{n}}\right\}.
\end{aligned}$$

Taking the square on both hand side concludes the proof.

□

D.2 Proofs for Section 3.3

Corollary 17. *Assume that (A), (B), and (C) holds and let $Q = I - W(W^T\Sigma W)^{-1}W^T\Sigma$. Fix an arbitrary norm $\|\cdot\|$ and any $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$. There exists $\lambda^* > 0$ such that if we consider the regularized estimator*

$$\hat{w}_{\lambda^*}, \hat{b}_{\lambda^*} = \arg\min_{(w,b)} \sqrt{\hat{L}(w,b)} + \lambda^*\|w\|, \tag{3.17}$$

it holds with probability at least $1-\delta$ that

$$(1-\epsilon)L(\hat{w}_{\lambda^*}, \hat{b}_{\lambda^*}) \le \left(\sqrt{L(w^\sharp, b^\sharp)} + \left(\mathbb{E}\|Q^T x\|_* + \sup_{\|v\|\le 1}\|Qv\|_\Sigma\sqrt{2\log(16/\delta)}\right)\frac{\|w^\sharp\|}{\sqrt{n}}\right)^2$$

where ϵ is the same as in Theorem 12.

Proof. Since we can write

$$\|Q^T x\|_* = \sup_{\|v\|\le 1}\langle v, Q^T\Sigma^{1/2}H\rangle$$

with $H \sim \mathcal{N}(0, I_d)$ and we can use Cauchy-Schwarz inequality to show that

$$\begin{aligned}
&\sup_{\|v\|\leq 1} \langle v, Q^T\Sigma^{1/2}H\rangle - \sup_{\|v\|\leq 1} \langle v, Q^T\Sigma^{1/2}H'\rangle \\
\leq &\sup_{\|v\|\leq 1} \langle v, Q^T\Sigma^{1/2}H\rangle - \langle v, Q^T\Sigma^{1/2}H'\rangle = \sup_{\|v\|\leq 1} \langle \Sigma^{1/2}Qv, H - H'\rangle \\
\leq &\sup_{\|v\|\leq 1} \|Qv\|_\Sigma \cdot \|H - H'\|_2,
\end{aligned}$$

by concentration of Lipschitz function (Lemma 49), it holds with probability at least $1 - \delta/8$ that

$$\|Q^T x\|_* \leq \mathbb{E}\|Q^T x\|_* + \sup_{\|v\|\leq 1} \|Qv\|_\Sigma \sqrt{2\log(16/\delta)}.$$

Using the definition of dual norm, we have

$$\begin{aligned}
\langle Qw, x\rangle &\leq \|w\|\|Q^T x\|_* \\
&\leq \|w\| \left(\mathbb{E}\|Q^T x\|_* + \sup_{\|v\|\leq 1} \|Qv\|_\Sigma \sqrt{2\log(16/\delta)}\right)
\end{aligned}$$

and we can choose C_δ to be the above upper bound in Theorem 12. Applying Theorem 12, we show that

$$\begin{aligned}
(1-\epsilon)L(\hat{w}, \hat{b}) &\leq \left(\sqrt{\hat{L}(\hat{w}, \hat{b})} + \left(\frac{\mathbb{E}\|Q^T x\|_*}{\sqrt{n}} + \sup_{\|v\|\leq 1} \|Qv\|_\Sigma \sqrt{\frac{2\log(16/\delta)}{n}}\right) \|\hat{w}\|\right)^2 \\
&\leq \left(\sqrt{\hat{L}(w^\sharp, b^\sharp)} + \left(\frac{\mathbb{E}\|Q^T x\|_*}{\sqrt{n}} + \sup_{\|v\|\leq 1} \|Qv\|_\Sigma \sqrt{\frac{2\log(16/\delta)}{n}}\right) \|w^\sharp\|\right)^2
\end{aligned}$$

Suppose that $\hat{L}(w^\sharp, b^\sharp) \leq (1+\rho)L(w^\sharp, b^\sharp)$, then it follows that

$$L(\hat{w}, \hat{b}) \leq \frac{1+\rho}{1-\epsilon}\left(\sqrt{L(w^\sharp, b^\sharp)} + \left(\frac{\mathbb{E}\|Q^T x\|_*}{\sqrt{n}} + \sup_{\|v\|\leq 1} \|Qv\|_\Sigma \sqrt{\frac{2\log(16/\delta)}{n}}\right) \|w^\sharp\|\right)^2. \quad \square$$

Next, we prove a standard expectation bound and concentration inequality for the ℓ_∞ norm of

Gaussian vectors.

Lemma 66. *For any covariance matrix Σ, consider $z \sim \mathcal{N}(0, \Sigma)$. Then it holds that*

$$\mathbb{E}\|z\|_\infty \leq \sqrt{2\log(2d)\max_i \Sigma_{ii}} \tag{D.5}$$

and with probability at least $1-\delta$, we have

$$\|z\|_\infty \leq \sqrt{2\log(2d/\delta)\max_i \Sigma_{ii}}. \tag{D.6}$$

In addition, it holds that

$$\sup_{\|v\|_1 \leq 1} \|v\|_\Sigma = \sqrt{\max_i \Sigma_{ii}}. \tag{D.7}$$

Proof. By Jensen's inequality, for any $\lambda > 0$, it holds that

$$\begin{aligned} e^{\lambda \mathbb{E}\|z\|_\infty} &\leq \mathbb{E} e^{\lambda \|z\|_\infty} \leq \mathbb{E}\left[\sum_i e^{\lambda |z_i|}\right] \\ &\leq 2\sum_i \mathbb{E} e^{\lambda z_i} = 2\sum_i e^{\lambda^2 \Sigma_{ii}/2} \leq 2de^{\lambda^2 \max_i \Sigma_{ii}/2} \end{aligned}$$

and so rearranging gives

$$\mathbb{E}\|z\|_\infty \leq \frac{\log(2d)}{\lambda} + \lambda \frac{\max_i \Sigma_{ii}}{2}.$$

Optimizing over λ proves the first desired bound. To show the high probability bound, we can use a union bound and the standard Gaussian tail bound $\Pr(|Z| \geq t) \leq 2e^{-t^2/2}$ to show that

$$\begin{aligned} \Pr(\|z\|_\infty > t) &\leq \sum_i \Pr(|z_i| > t) \leq \sum_i 2e^{-t^2/2\Sigma_{ii}} \\ &\leq 2de^{-t^2/2\max_i \Sigma_{ii}} \end{aligned}$$

and setting $\delta = 2de^{-t^2/2\max_i \Sigma_{ii}}$, we have $t = \sqrt{2\max_i \Sigma_{ii}\log(2d/\delta)}$. Finally, for $A =$

$[a_1, ..., a_d]$, we show that

$$\begin{aligned}\max_{\|v\|_1 \le 1} \|Av\|_2 &= \max_{\|v\|_1 \le 1, \|u\|_2 \le 1} \langle u, Av \rangle \\ &= \max_{\|u\|_2 \le 1} \|A^T u\|_\infty = \max_{\|u\|_2 \le 1} \max_i |a_i^T u| \\ &= \max_i \max_{\|u\|_2 \le 1} |a_i^T u| = \max_i \|a_i\|_2 = \sqrt{\max_i (A^T A)_{ii}}. \end{aligned}$$ □

Lemma 18. *Suppose $w^\sharp$ is k-sparse, i.e. supported on coordinate set $S \subset [d]$ with $|S| \le k$. Every w with $\|w\|_1 \le \|w^\sharp\|_1$ satisfies*

$$\|(w - w^\sharp)_{S^C}\|_1 \le \|(w - w^\sharp)_S\|_1. \tag{3.20}$$

Proof. Note that over this set, we have

$$\begin{aligned}\|(w - w^\sharp)_{S^C}\|_1 = \|w_{S^C}\|_1 = \|w\|_1 - \|w_S\|_1 &\le \|w^\sharp\|_1 - \|w_S\|_1 \\ &\le \|(w^\sharp - w_S)\|_1 = \|(w - w^\sharp)_S\|_1\end{aligned}$$

where the first inequality uses $\|w\|_1 \le \|w^\sharp\|_1$ and the second inequality follows by the triangle inequality. □

Theorem 19. *Under assumptions (A), (B) and (C), let ϵ, Q be the same as in Theorem 12 and denote $\Sigma^\perp = Q^T \Sigma Q$. Let $\sigma^2 = \min_{w,b} L(w, b)$ and $w^\sharp, b^\sharp = \arg\min_{w,b} L(w, b)$. Suppose that*

1. *$w^\sharp$ is a k-sparse vector with support $S \subset [d]$*
2. *the covariance matrix Σ satisfies the S-compatibility condition*
3. *the number of samples n satisfies*

$$n \ge \frac{32 \max_i \Sigma^\perp_{ii}}{\phi(\Sigma, S)^2} \cdot k \log(16d/\delta) \quad \textit{and} \quad \epsilon \le 1/2$$

Then for any $\epsilon' < 1$, it holds with probability at least $1 - \delta$ that for all (w, b) satisfying $\|w\|_1 \leq \|w^\sharp\|_1$ and $\hat{L}(w, b) \leq (1 + \epsilon')\sigma^2$, we have

$$L(w, b) - \sigma^2 \leq 104\sigma^2 \left(\epsilon + \epsilon' + \frac{\max_i \Sigma^\perp_{ii}}{\phi(\Sigma, S)^2} \frac{8k \log(16d/\delta)}{n} \right). \tag{3.21}$$

In particular, when $\sigma = 0$ we have that $\|w - w^\sharp\|_\Sigma = 0$, and so if Σ is positive definite then we have $w = w^\sharp$ (exact recovery).

Proof. Note that $L(w, b) = \mathbb{E}[(\langle w, x \rangle + b - y)^2]$ is a quadratic function in both w and b with

$$\nabla^2 L(w, b) = 2 \begin{pmatrix} \mu\mu^T + \Sigma & \mu \\ \mu^T & 1 \end{pmatrix}.$$

Since the second-order Taylor expansion is exact for quadratic functions, if we let $\sigma^2 = \min_{w,b} L(w, b)$ and $w^\sharp, b^\sharp = \arg\min_{w,b} L(w, b)$, it must be the case that

$$\begin{aligned} L(w, b) &= \sigma^2 + (w - w^\sharp)^T(\mu\mu^T + \Sigma)(w - w^\sharp) + 2(b - b^\sharp)\langle \mu, w - w^\sharp \rangle + (b - b^\sharp)^2 \\ &= \sigma^2 + \|w - w^\sharp\|^2_\Sigma + (\langle \mu, w - w^\sharp \rangle + b - b^\sharp)^2. \end{aligned}$$

By Lemma 18, Lemma 66, the compatibility condition, and a union bound, it holds for $x \sim \mathcal{N}(0, \Sigma)$ that with probability at least $1 - \delta/8$

$$\begin{aligned} \langle Qw, x \rangle &= \langle Q(w - w^\sharp), x \rangle = \langle w - w^\sharp, Q^T x \rangle \\ &\leq \|w - w^\sharp\|_1 \|Q^T x\|_\infty \\ &\leq 2\|(w - w^\sharp)_S\|_1 \sqrt{2 \log(16d/\delta) \max_i \Sigma^\perp_{ii}} \\ &\leq \frac{k^{1/2}\|w - w^\sharp\|_\Sigma}{\phi(\Sigma, S)} \sqrt{8 \log(16d/\delta) \max_i \Sigma^\perp_{ii}} \end{aligned} \tag{D.8}$$

and so we can apply Theorem 12 with $C_\delta(w)$ equal to the right hand side of (D.8). Therefore, we

have shown that

$$(1-\epsilon)\,L(w,b) \le \left(\sqrt{\hat{L}(w,b)} + \frac{k^{1/2}\|w-w^\sharp\|_\Sigma}{\phi(\Sigma,S)}\sqrt{8\frac{\log(16d/\delta)}{n}\max_i \Sigma^\perp_{ii}}\right)^2$$
$$\le \left(\sigma\sqrt{1+\epsilon'} + \sqrt{L(w,b)-\sigma^2}\sqrt{\frac{8k\log(16d/\delta)}{n}\frac{\max_i \Sigma^\perp_{ii}}{\phi(\Sigma,S)^2}}\right)^2.$$

For the simplicity of notation, we denote

$$x = \sqrt{L(w,b)-\sigma^2}, \quad a = \sqrt{\frac{1+\epsilon'}{1-\epsilon}}, \quad b = \sqrt{\frac{8k\log(16d/\delta)}{(1-\epsilon)n}\frac{\max_i \Sigma^\perp_{ii}}{\phi(\Sigma,S)^2}}$$

then we have shown

$$\sigma^2 + x^2 \le (a\sigma + bx)^2.$$

Solving this quadratic equation, the above becomes

$$x \le \frac{\sigma}{1-b^2}\left(ab + \sqrt{b^2+a^2-1}\right)$$

and so using the inequality $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and the AM-GM inequality, we can show

$$\begin{aligned}
x^2 &\le \frac{\sigma^2}{(1-b^2)^2}\left(ab + \sqrt{b^2+a^2-1}\right)^2 \\
&= \frac{\sigma^2}{(1-b^2)^2}\left(a^2b^2 + b^2 + a^2 - 1 + 2ab\sqrt{b^2+a^2-1}\right) \\
&\le \frac{\sigma^2}{(1-b^2)^2}\left(a^2b^2 + b^2 + a^2 - 1 + 2ab^2 + 2ab\sqrt{a^2-1}\right) \\
&\le \frac{\sigma^2}{(1-b^2)^2}\left(a^2b^2 + b^2 + a^2 - 1 + 2ab^2 + a^2b^2 + a^2 - 1\right) \\
&= \frac{\sigma^2}{(1-b^2)^2}\left((1+2a+2a^2)b^2 + 2(a^2-1)\right).
\end{aligned}$$

Therefore, if $a \le 2$ and $b^2 \le \frac{1}{2}$ (which is guaranteed by the assumption), then $x^2 \le 52\sigma^2(b^2 +$

$a^2 - 1)$ and we are done. $\square$

Remark 67 (Generalization Bound for Larger Cones)**.** *For simplicity, in the above analysis we gave a generalization bound for predictors w satisfying $\|w\|_1 \leq \|w^*\|_1$, or more generally $\|(w - w^*)_{S^C}\|_1 \leq \|(w - w^*)_S\|_1$, which covers the case of the LASSO with oracle regularization commonly considered in the literature (see, e.g., Vershynin [71]). In situations where adaptivity to the unknown value of $\|w^*\|_1$ is important, the relevant predictor w may only be guaranteed to satisfy the weaker bound $\|(w - w^*)_{S^C}\|_1 \leq C\|(w - w^*)_S\|_1$ for some $C > 1$ and the analogous version of the compatibility condition/restricted eigenvalue condition over this cone is assumed (see, e.g. Bickel et al. [12], Van De Geer and Bühlmann [67], Rigollet and Hütter [56], Wainwright [73]); adopting the analysis to predictors in this larger cone is straightforward and we omit the details.*

D.3 Proofs for Section 3.4

Lemma 68. *Under the assumptions of Theorem 20 and with the definition of β_1 there, with probability at least $1 - 4(\delta + \delta')$*

$$L(\hat{w}) \leq \sigma^2 + (1 + 2\beta_1)\left(\sqrt{\sigma F(\hat{w})/\sqrt{n}} + F(\hat{w})/\sqrt{n}\right)^2$$

where $\hat{w}$ is any empirical risk minimizer over a closed convex set $\mathcal{K}$ containing w^*, i.e. $\hat{L}(\hat{w}) = \min_{w \in \mathcal{K}} \hat{L}(w)$.

Proof. Write $X = Z\Sigma^{1/2}$ with Z a matrix of i.i.d. Gaussians, and observe

$$\frac{1}{n}\langle Z^T\xi, \Sigma^{1/2}(w - w^*)\rangle = \frac{1}{n}\langle \xi, Z\Sigma^{1/2}(w - w^*)\rangle = \frac{1}{n}\langle \xi, X(w - w^*)\rangle$$

Note that conditional on ξ, $Z^T\xi$ is just a standard Gaussian $N(0, \|\xi\|_2^2 I_d)$. So with probability at

least $1-\delta'$ (recalling the defining property of the complexity functional F) we have

$$\frac{1}{n}\langle Z^T\xi, \Sigma^{1/2}(w-w^*)\rangle \leq \frac{\|\xi\|_2}{n}F(w). \tag{D.9}$$

Observe that

$$\nabla_w \hat{L}(w) = \frac{1}{n}\nabla_w\|Y-Xw\|_2^2 = -\frac{2}{n}X^T(Y-Xw) = -\frac{2}{n}(X^T\xi + X^TX(w^*-w))$$

so from the KKT condition $\langle w^*-\hat{w}, \nabla_w\hat{L}(\hat{w})\rangle \geq 0$ we have

$$\langle w^*-\hat{w}, X^T\xi\rangle + \langle w^*-w, X^TX(w^*-w)\rangle \leq 0$$

so rearranging gives the first inequality, and using (D.9) gives the second inequality in

$$\|w^*-\hat{w}\|_{\hat{\Sigma}} \leq \sqrt{\frac{1}{n}\langle \xi, X(\hat{w}-w^*)\rangle} \leq \sqrt{\frac{\|\xi\|_2}{n}F(w)}.$$

By Theorem 20 (defining $F(w)=\infty$ outside of $\mathcal{K}$), for all $w \in \mathcal{K}$

$$\|w^*-w\|_\Sigma \leq (1+\beta_1)\left[\|w^*-w\|_{\hat{\Sigma}} + F(w)/\sqrt{n}\right]$$

and so for $\hat{w}$ we have

$$\begin{aligned}\|w^*-\hat{w}\|_\Sigma &\leq (1+\beta_1)\left[\|w^*-\hat{w}\|_{\hat{\Sigma}} + F(\hat{w})/\sqrt{n}\right] \\ &\leq (1+\beta_1)\left[\sqrt{\frac{\|\xi\|_2}{n}F(\hat{w})} + F(\hat{w})/\sqrt{n}\right]\end{aligned}$$

and using the fact that the norm $\|\xi\|_2$ concentrates about $\sigma\sqrt{n}$ by Lemma 52 and recalling the

definition of β_1, we have

$$\|w^* - \hat{w}\|_\Sigma^2 \le (1 + 2\beta_1) \left(\sqrt{\sigma F(\hat{w})/\sqrt{n}} + F(\hat{w})/\sqrt{n} \right)^2.$$

Finally, recalling that $L(\hat{w}) = \sigma^2 + \|w - \hat{w}\|_\Sigma^2$ gives the bound as claimed. □

Theorem 21. *Let $\mathcal{K}$ be a closed convex set in $\mathbb{R}^d$ containing w^* and suppose $\delta' \ge 0, p \ge 0$ are such that with probability at least $1 - \delta'$ over the randomness of $x \sim N(0, \Sigma)$, uniformly over all $w \in \mathcal{K}$ we have*

$$\langle w - w^*, x \rangle \le \|w - w^*\|_\Sigma \sqrt{p}. \tag{3.25}$$

Suppose that $\hat{w} = \arg\min_{w \in \mathcal{K}} \hat{L}(w)$ and $p/n \le 0.999$, then for all $n \ge C \log(2/\delta)$ for some absolute constant $C > 0$, it holds with probability at least $1 - (\delta + \delta')$ that

$$L(\hat{w}) - \sigma^2 \le (1 + \tau)\sigma^2 \cdot \frac{p}{n}. \tag{3.26}$$

where $\tau = \tau(p, n, \delta)$ is upper bounded by an absolute constant and satisfies $\tau(p, n, \delta) \to 1$ in any joint limit $[p + \log(2/\delta)]/n \to 0, n \to \infty$.

Proof. Defining $\rho := \sqrt{p/n}$ and Theorem 68 gives

$$\begin{aligned} \|w^* - \hat{w}\|_\Sigma &\le (1 + 2\beta_1)^{1/2} \left(\sqrt{\sigma F(\hat{w})/\sqrt{n}} + F(\hat{w})/\sqrt{n} \right) \\ &= (1 + 2\beta_1)^{1/2} \left(\sqrt{\sigma\rho \|w - w^*\|_\Sigma} + \rho \|w - w^*\|_\Sigma \right) \end{aligned}$$

hence

$$(1 - (1 + 2\beta_1)^{1/2}\rho)\|w - w^*\|_\Sigma \le (1 + 2\beta_1)^{1/2} \sqrt{\sigma\rho \|w - w^*\|_\Sigma}$$

which is equivalent to

$$\|w - w^*\|_\Sigma \le \frac{(1 + 2\beta_1)\sigma\rho}{(1 - (1 + 2\beta_1)^{1/2}\rho)^2}$$

and this in turn is equivalent to the final result. □

Corollary 22. *Under the model assumptions* (4.2) *with $d < n$ and assuming a sufficiently large n, it holds with probability at least $1 - \delta$ that*

$$L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 \lesssim \sigma^2 \left(\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(36/\delta)}{n}} \right)^2 \tag{3.27}$$

Proof. Recall from the proof of Theorem 24 that with probability at least $1 - \delta'$ we have

$$\langle w - w^*, x \rangle \leq \left(\sqrt{d} + 2\sqrt{\log(4/\delta')} \right) \left\| \Sigma^{1/2}(w^* - w) \right\|_2$$

where $\delta' = \delta/9$ so the result follows from Theorem 21 with $\mathcal{K} = \mathbb{R}^d$. □

Corollary 23. *Applying Theorem 21 with $\mathcal{K} = \{\|w\|_1 \leq \|w^*\|_1\}$ the rescaled ℓ_1-ball and under the sparsity and compatability condition assumptions of* (19)*, we have with probability at least $1 - \delta$ that the LASSO solution*

$$\hat{w}_{LASSO} = \operatorname*{arg\,min}_{w : \|w\|_1 \leq \|w^*\|_1} \hat{L}(w)$$

satisfies

$$L(\hat{w}_{LASSO}) - \sigma^2 \lesssim \frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{\sigma^2 k \log(16d/\delta)}{n} \tag{3.28}$$

provided n is sufficiently large that

$$\sqrt{\frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{8k \log(16d/\delta)}{n}} \leq 0.999.$$

Proof. Recall from the proof of Theorem 19, more specially (D.8), that with probability at least $1 - \delta/8$

$$\begin{aligned} \langle w - w^*, x \rangle &\leq \|w - w^*\|_1 \|x\|_\infty \\ &\leq 2\|(w - w^*)_S\|_1 \|x\|_\infty \leq \frac{2k^{1/2}}{\phi(\Sigma, S)} \|w - w^*\|_\Sigma \max_i \sqrt{2\Sigma_{ii} \log(16d/\delta)}. \end{aligned}$$

so the result follows from Theorem 21. □

D.4 Proofs for Section 3.5

The following training error bounds are standard, which we include for completeness.

Lemma 69. *Under the model assumptions in* (4.2) *with $d \leq n$, consider the ordinary least square estimator $\hat{w}_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$. With probability at least $1-\delta$, it holds that*

$$\sqrt{\hat{L}(\hat{w}_{\mathrm{OLS}})} \leq \sigma \left(\sqrt{1 - \frac{d}{n}} + 2\sqrt{\frac{\log(4/\delta)}{n}} \right) \tag{D.10}$$

Similarly, with probability at least $1-\delta$, it holds that

$$\|\hat{w}_{\mathrm{OLS}} - w^*\|_{\hat{\Sigma}} \leq \sigma \left(\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(4/\delta)}{n}} \right) \tag{D.11}$$

Proof. By our model assumptions, we can write $\hat{w}_{\mathrm{OLS}} = w^* + (X^T X)^{-1} X^T \xi$, and so $Y - X\hat{w}_{\mathrm{OLS}} = (I - X(X^T X)^{-1} X^T)\xi$. Since $(I - X(X^T X)^{-1} X^T)$ is almost surely an idempotent matrix with rank $n - d$, it follows that the distribution of

$$\frac{n\hat{L}(\hat{w}_{\mathrm{OLS}})}{\sigma^2} = \frac{1}{\sigma^2}\, \xi^T (I - X(X^T X)^{-1} X^T)\xi,$$

is a Chi-square distribution with $n - d$ degrees of freedom. By the same reasoning, the distribution of

$$\frac{n\, \|\hat{w}_{\mathrm{OLS}} - w^*\|_{\hat{\Sigma}}^2}{\sigma^2} = \frac{1}{\sigma^2}\, \xi^T X (X^T X)^{-1} X^T \xi$$

is a Chi-square distribution with d degrees of freedom. By Lemma 52, with probability at least $1 - \delta$, it holds that

$$\frac{\sqrt{n}}{\sigma} \sqrt{\hat{L}(\hat{w}_{\mathrm{OLS}})} \leq \sqrt{n - d} + 2\sqrt{\log(4/\delta)}.$$

Similarly, we have

$$\frac{\sqrt{n}}{\sigma}\left\|\hat{w}_{\mathrm{OLS}}-w^*\right\|_{\hat{\Sigma}} \leq \sqrt{d}+2\sqrt{\log(4/\delta)}.$$

Rearranging the terms conclude the proof. □

Theorem 24. *Under the model assumptions in* (3.22)*, let* $\gamma = d/n < 1$*. There exists some* $\epsilon \lesssim \left(\frac{\log(36/\delta)}{n}\right)^{1/2}$ *such that for all sufficiently large* n*, with probability* $1-\delta$ *it holds uniformly for all* $w \in \mathbb{R}^d$ *that*

$$\left|\sqrt{L(w)-\sigma^2}-\sqrt{\frac{\gamma\hat{L}(w)}{(1-\gamma)^2}}\right| \leq \epsilon\sqrt{\hat{L}(w)}+\sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma}-\sigma^2\right)+\epsilon\hat{L}(w)}. \tag{3.30}$$

For the empirical risk minimizer $\hat{w}_{\mathrm{OLS}} = (X^TX)^{-1}X^TY$*, the right hand side of* (3.30) *is approximately zero because we also have*

$$\hat{L}(\hat{w}_{\mathrm{OLS}}) \leq \sigma^2(1-\gamma)+\sigma^2\epsilon\sqrt{1-\gamma}. \tag{3.31}$$

Therefore, we obtain the following generalization bound:

$$L(\hat{w}_{\mathrm{OLS}})-\frac{\sigma^2}{1-\gamma} \lesssim \sigma^2\left(\frac{\log(36/\delta)}{n}\right)^{1/4}. \tag{3.32}$$

Proof. By Lemma 52, we can pick

$$\begin{aligned} F(w) &= \left(\sqrt{d}+2\sqrt{\log(4/\delta')}\right)\left\|\Sigma^{1/2}(w^*-w)\right\|_2 \\ &= \left(\sqrt{d}+2\sqrt{\log(4/\delta')}\right)\sqrt{L(w)-\sigma^2}. \end{aligned}$$

Let $\delta' = \delta/9$ and replace δ by $\delta/3$ in Theorem 20, plug in the estimates from Lemma 69 using confidence level $\delta/9$, then by a union bound with $\gamma = \frac{d}{n}$ and $\epsilon = \sqrt{\frac{\log(36/\delta)}{n}}$, we have

$$\sqrt{\hat{L}(\hat{w}_{\mathrm{OLS}})} \leq \sigma\sqrt{1-\gamma}+2\sigma\epsilon \tag{D.12}$$

and the bound (3.24) becomes

$$L(w) \leq (1+14\epsilon)\left(\sqrt{\hat{L}(w)} + (\sqrt{\gamma}+2\epsilon)\sqrt{L(w)-\sigma^2}\right)^2.$$

We can simplify this by expanding the square

$$(1+14\epsilon)^{-1}L(w) \leq \hat{L}(w) + (\sqrt{\gamma}+2\epsilon)^2(L(w)-\sigma^2) + 2(\sqrt{\gamma}+2\epsilon)\sqrt{\hat{L}(w)}\sqrt{L(w)-\sigma^2}.$$

Rearranging, we arrive at

$$\begin{aligned}&\left[(1+14\epsilon)^{-1} - (\sqrt{\gamma}+2\epsilon)^2\right](L(w)-\sigma^2)\\ &\leq \hat{L}(w) - (1+14\epsilon)^{-1}\sigma^2 + 2(\sqrt{\gamma}+2\epsilon)\sqrt{\hat{L}(w)}\sqrt{L(w)-\sigma^2}.\end{aligned}$$

Note that this is a quadratic equation in terms of $\sqrt{L(w)-\sigma^2}$

$$(L(w)-\sigma^2) - 2\frac{(\sqrt{\gamma}+2\epsilon)\sqrt{\hat{L}(w)}}{(1+14\epsilon)^{-1}-(\sqrt{\gamma}+2\epsilon)^2}\sqrt{L(w)-\sigma^2} \leq \frac{\hat{L}(w)-(1+14\epsilon)^{-1}\sigma^2}{(1+14\epsilon)^{-1}-(\sqrt{\gamma}+2\epsilon)^2}.$$

We can complete the square, which leads to the following

$$\begin{aligned}&\left[\sqrt{L(w)-\sigma^2} - \frac{(\sqrt{\gamma}+2\epsilon)\sqrt{\hat{L}(w)}}{(1+14\epsilon)^{-1}-(\sqrt{\gamma}+2\epsilon)^2}\right]^2\\ &\leq \frac{(1+14\epsilon)^{-1}}{(1+14\epsilon)^{-1}-(\sqrt{\gamma}+2\epsilon)^2}\left(\frac{\hat{L}(w)}{(1+14\epsilon)^{-1}-(\sqrt{\gamma}+2\epsilon)^2} - \sigma^2\right)\end{aligned}$$

Observe that $(1+14\epsilon)^{-1} - (\sqrt{\gamma}+2\epsilon)^2 = 1-\gamma-O(\epsilon)$ and so

$$\frac{\sqrt{\gamma}}{1-\gamma} \leq \frac{(\sqrt{\gamma}+2\epsilon)}{(1+14\epsilon)^{-1}-(\sqrt{\gamma}+2\epsilon)^2} \leq \frac{\sqrt{\gamma}}{1-\gamma} + O(\epsilon).$$

We can handle the other terms similarly. Plugging in (D.12) concludes the proof. □

Lemma 70. *Let w^*, w be arbitrary vectors with $w^* \neq 0$, let V be the (one-dimensional) span of w^*, and let P_V be the orthogonal projection onto V. Then for any vector x,*

$$\langle w - w^*, x\rangle \leq \|w - w^*\|_2 \cdot \|P_V x\|_2 + \|x\|_2 \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.$$

Proof. Observe that by expanding the square, we have

$$\|w - w^*\|_2^2 = \|w\|_2^2 + \|w^*\|_2^2 - 2\langle P_V w, w^*\rangle$$

and so rearranging gives the Parallelogram identity

$$\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2 = 2\langle P_V w, w^*\rangle.$$

Taking absolute value of both sides and using that $P_V w$ and w^* are colinear gives

$$\left|\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right| = 2\|P_V w\|_2 \|w^*\|_2.$$

Combining this with the Pythagorean Theorem, we find

$$\|P_{V^\perp} w\|_2^2 = \|w\|_2^2 - \|P_V w\|_2^2 = \|w\|_2^2 - \left(\frac{\left|\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right|}{2\|w^*\|}\right)^2.$$

Thus, applying the Cauchy-Schwarz inequality

$$\begin{aligned}\langle w - w^*, x\rangle &= \langle P_V(w - w^*), x\rangle + \langle P_{V^\perp} w, x\rangle \\ &\leq \langle P_V(w - w^*), x\rangle + \|P_{V^\perp} w\|_2 \|x\|_2\end{aligned}$$

and plugging in the identity for $\|P_{V^\perp} w\|_2$ gives

$$\begin{aligned}\langle w - w^*, x\rangle &\le \langle w - w^*, P_V x\rangle + \|x\|_2 \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}} \\ &\le \|w - w^*\|_2 \cdot \|P_V x\|_2 + \|x\|_2 \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.\end{aligned}$$

which is the desired inequality. □

Lemma 71. *Under the assumptions of Theorem 20 with $\gamma = d/n > 1$ and the further assumption that the data has isotropic covariance $\Sigma = I_d$, there exists $\epsilon \lesssim \sqrt{\frac{\log(18/\delta)}{n}}$ such that with probability at least $1 - \delta$, we have*

$$\|w - w^*\|_2^2 + \sigma^2 \le (1 + \epsilon) \left(\sqrt{\hat{L}(w)} + \sqrt{\gamma} \cdot \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}} \right)^2.$$

Proof. Observe that $\frac{\langle w^*, x\rangle}{\|w^*\|_2} \sim \mathcal{N}(0, 1)$ and so by a standard Gaussian tail bound, Lemma 52 and a union bound, with probability at least $1 - \delta$, it holds that

$$\|P_V x\|_2 = \left\| \frac{w^* (w^*)^T}{\|w^*\|_2^2} x \right\|_2 = \frac{|\langle w^*, x\rangle|}{\|w^*\|_2} \le \sqrt{2 \log(6/\delta)}$$

and $\|x\|_2 \le \sqrt{d} + 2\sqrt{\log(6/\delta)}$.

Combining Lemma 70 with Theorem 20 and another union bound gives

$$\begin{aligned}&\frac{1}{\sqrt{1 + \beta_1}} \sqrt{\|w - w^*\|_2^2 + \sigma^2} \\ &\le \sqrt{\hat{L}(w)} + \|w - w^*\|_2 \sqrt{\frac{2 \log(18/\delta)}{n}} \\ &\quad + \left(\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(18/\delta)}{n}} \right) \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.\end{aligned}$$

Using the fact that $\|w - w^*\|_2 \le \sqrt{\|w - w^*\|_2^2 + \sigma^2}$ and $d > n$, we have

$$\left(1 + 2\sqrt{\frac{\log(18/\delta)}{n}}\right)^{-1} \left(\frac{1}{\sqrt{1+\beta_1}} - \sqrt{\frac{2\log(18/\delta)}{n}}\right) \sqrt{\|w - w^*\|_2^2 + \sigma^2}$$
$$\le \sqrt{\hat{L}(w)} + \sqrt{\gamma} \cdot \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.$$

To simplify, there exists $\epsilon \lesssim \sqrt{\frac{\log(18/\delta)}{n}}$ such that

$$\frac{1}{\sqrt{1+\epsilon}} \sqrt{\|w - w^*\|_2^2 + \sigma^2} \le \sqrt{\hat{L}(w)} + \sqrt{\gamma} \cdot \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.$$

and rearranging concludes the proof. □

The generalization bound from Lemma 71 holds for all w; we now show what happens when we specialize it to interpolators.

Theorem 25. *Under the model assumptions in* (4.2) *with* $\gamma = d/n > 1$ *and* $\Sigma = I_d$*, there exists* $\epsilon \lesssim \left(\frac{\log(18/\delta)}{n}\right)^{1/2}$ *such that with probability at least* $1 - \delta$*, the following holds uniformly over all* w *such that* $\hat{L}(w) = 0$*:*

$$\begin{aligned} &\left| L(w) - \left[\sigma^2 + \|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w^*\|_2^2\right]\right| \\ \le &2\|w^*\|_2 \sqrt{\left(1 - \frac{1}{\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \frac{\sigma^2}{\gamma} + 3\epsilon\|w\|_2^2}. \end{aligned} \tag{3.33}$$

Proof. By Lemma 71, there exists some $\epsilon \lesssim \sqrt{\frac{\log(18/\delta)}{n}}$ such that with probability at least $1 - \delta$,

for all w such that $\hat{L}(w) = 0$ it holds that

$$\begin{aligned}
&\|w - w^*\|_2^2 + \sigma^2 \\
&\leq (1+\epsilon)\gamma \left(\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2} \right) \\
&= (1+\epsilon)\gamma \left(\|w\|_2^2 - \frac{(\|w\|_2^2 + \|w^*\|_2^2)^2 - 2(\|w\|_2^2 + \|w^*\|_2^2)\|w - w^*\|_2^2 + \|w - w^*\|_2^4}{4\|w^*\|_2^2} \right)
\end{aligned}$$

Rearranging, we have

$$\begin{aligned}
&4\|w^*\|_2^2 \cdot \frac{\|w - w^*\|_2^2 + \sigma^2}{(1+\epsilon)\gamma} \\
&\leq 4\|w^*\|_2^2 \cdot \|w\|^2 - (\|w\|_2^2 + \|w^*\|_2^2)^2 + 2(\|w\|_2^2 + \|w^*\|_2^2)\|w - w^*\|_2^2 - \|w - w^*\|_2^4
\end{aligned}$$

Grouping the terms with $\|w - w^*\|_2^2$, we see that

$$\begin{aligned}
&\|w - w^*\|_2^4 + \left(\frac{4\|w^*\|_2^2}{(1+\epsilon)\gamma} - 2(\|w\|_2^2 + \|w^*\|_2^2) \right) \cdot \|w - w^*\|_2^2 + 4\|w^*\|_2^2 \cdot \frac{\sigma^2}{(1+\epsilon)\gamma} \\
&\leq 4\|w^*\|_2^2 \cdot \|w\|^2 - (\|w\|_2^2 + \|w^*\|_2^2)^2
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
\|w - w^*\|_2^4 - 2\left(\|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w^*\|_2^2 \right) \cdot \|w - w^*\|_2^2 \\
+(\|w\|_2^2 - \|w^*\|_2^2)^2 + 4\|w^*\|_2^2 \cdot \frac{\sigma^2}{(1+\epsilon)\gamma} \leq 0.
\end{aligned}$$

To complete the square, we compute

$$\begin{aligned}
&\left(\|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w^*\|_2^2\right)^2 - (\|w\|_2^2 - \|w^*\|_2^2)^2 - 4\|w^*\|_2^2 \cdot \frac{\sigma^2}{(1+\epsilon)\gamma} \\
=&\left(\frac{4}{(1+\epsilon)^2\gamma^2} - \frac{4}{(1+\epsilon)\gamma}\right)\|w^*\|_2^4 + 4\left(1 - \frac{1}{(1+\epsilon)\gamma}\right)\|w\|_2^2\|w^*\|_2^2 - 4\|w^*\|_2^2 \cdot \frac{\sigma^2}{(1+\epsilon)\gamma} \\
=&4\|w^*\|_2^2\left[\left(1 - \frac{1}{(1+\epsilon)\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{(1+\epsilon)\gamma}\right) - \frac{\sigma^2}{(1+\epsilon)\gamma}\right] \\
\le&4\|w^*\|_2^2\left[\left(1 + \epsilon - \frac{1}{\gamma}\right)\left(\|w\|_2^2 + \epsilon\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \frac{\sigma^2}{\gamma}\right]
\end{aligned}$$

where in the last step we use $(1+\epsilon)^2 \ge 1$ and $\frac{\sigma^2(1+\epsilon)}{\gamma} \ge \frac{\sigma^2}{\gamma}$. To simplify, it is routine to check that

$$\left(1 + \epsilon - \frac{1}{\gamma}\right)\left(\|w\|_2^2 + \epsilon\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \left(1 - \frac{1}{\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) \le 3\epsilon\|w\|_2^2$$

and so we can conclude that

$$\begin{aligned}
&\left|\|w - w^*\|_2^2 - \left[\|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w^*\|_2^2\right]\right| \\
&\le 2\|w^*\|_2\sqrt{\left(1 - \frac{1}{\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \frac{\sigma^2}{\gamma} + 3\epsilon\|w\|_2^2}.
\end{aligned}$$

as desired. □

Theorem 26. *Under the model assumptions in* (4.2) *with* $\gamma = d/n > 1$ *and* $\Sigma = I_d$*, there exists* $\epsilon \lesssim \left(\frac{\log(40/\delta)}{n}\right)^{1/2}$ *such that with probability at least* $1 - \delta$*, it holds that*

$$\min_{w: Xw = Y} \|w\|_2^2 \le (1+\epsilon)\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma - 1}\right). \tag{3.34}$$

Thus, by Theorem 25, we have

$$\begin{aligned} & L(\hat{w}) - \left[\left(1-\frac{1}{\gamma}\right)\|w^*\|_2^2 + \sigma^2\frac{\gamma}{\gamma-1}\right] \\ & \leq \epsilon\left(\frac{\|w^*\|_2^2}{\gamma}+\frac{\sigma^2}{\gamma-1}\right) + \|w^*\|_2\sqrt{\epsilon\left(\frac{\|w^*\|_2^2}{\gamma}+\frac{\sigma^2}{\gamma-1}\right)} \end{aligned} \tag{3.35}$$

where $\hat{w}$ is the minimal-ℓ_2 norm interpolator. If we fix σ^2, γ and $\|w^*\|_2$, then as $n \to \infty$

$$L(\hat{w}) \to \left(1-\frac{1}{\gamma}\right)\|w^*\|_2^2 + \sigma^2\frac{\gamma}{\gamma-1} \quad \textit{in probability.} \tag{3.36}$$

Proof. The proof strategy here follows the same lines as in Theorem 2 of Koehler et al. [33], but handles the w^* term more carefully. First, we introduce the Lagrangian and apply a change of variable

$$\begin{aligned} \min_{Xw=Y}\|w\|^2 &= \min_w \max_\lambda \langle\lambda, Xw-Y\rangle + \|w\|^2 \\ &= \min_w \max_\lambda \langle\lambda, Xw-\xi\rangle + \|w+w^*\|^2 \end{aligned}$$

To apply CGMT (Theorem 45), we need a double truncation argument. For any $r, t > 0$, introduce the following problem:

$$\Phi_r(t) = \min_{\|w+w^*\|^2\leq 2t}\ \max_{\|\lambda\|\leq r}\ \langle\lambda, Xw-\xi\rangle + \|w+w^*\|^2. \tag{D.13}$$

We also introduce

$$\begin{aligned} \Phi(t) &= \min_{\|w+w^*\|^2\leq 2t}\ \max_\lambda \langle\lambda, Xw-\xi\rangle + \|w+w^*\|^2 \\ &= \min_{\substack{Xw=\xi \\ \|w+w^*\|^2\leq 2t}} \|w+w^*\|^2 \end{aligned} \tag{D.14}$$

and claim that $\Phi_r(t) \to \Phi(t)$ as $r \to \infty$. By definition, $\Phi_r(t) \leq \Phi_s(t)$ for $r \leq s$. We consider two

cases:

1. $\Phi(t) = \infty$, i.e. the minimization problem defining $\Phi(t)$ is infeasible. In this case, we know that for all $\|w + w^*\|^2 \leq 2t$

$$\|Xw - \xi\|_2 > 0.$$

   By compactness, there exists $\mu = \mu(X, \xi) > 0$ (in particular, independent of r) such that

$$\|Xw - \xi\|_2 \geq \mu.$$

   Therefore, considering λ along the direction of $Xw - \xi$ shows that

$$\Phi_r(t) = \min_{\|w+w^*\|^2 \leq 2t} \max_{\|\lambda\|_2 \leq r} \langle \lambda, Xw - \xi \rangle + \|w + w^*\|^2 \geq r\mu$$

   so $\Phi_r(t) \to \infty$ as $r \to \infty$.

2. Otherwise $\Phi(t) < \infty$, i.e. the minimization problem defining $\Phi(t)$ is feasible. In this case, we can let $w(r)$ be an arbitrary minimizer achieving the objective $\Phi_r(t)$ for each $r \geq 0$ by compactness. By compactness again, the sequence $\{w(r)\}_{r=1}^{\infty}$ at positive integer values of r has a subsequential limit $w(\infty)$ such that $\|w(\infty) + w^*\| \leq 2t$. Equivalently, there exists an increasing sequence r_n such that $\lim_{n\to\infty} w(r_n) = w(\infty)$.

   Suppose for the sake of contradiction that $Xw(\infty) \neq \xi$, then by continuity, there exists $\mu > 0$ and a sufficiently small $\epsilon > 0$ such that for all $\|w - w(\infty)\|_2 \leq \epsilon$

$$\|Xw - \xi\|_2 \geq \mu.$$

   This implies that for sufficiently large n, we have

$$\|Xw(r_n) - \xi\|_2 \geq \mu$$

and by the same argument as in the previous case

$$\Phi_{r_n}(t) = \max_{\|\lambda\|_2 \le r} \langle \lambda, Xw(r_n) - \xi \rangle + \|w(r_n) + w^*\|^2 \ge r\mu$$

so $\Phi_{r_n} \to \infty$, but this is impossible since $\Phi_r(t) \le \Phi(t) < \infty$. By contradiction, it must be the case that $Xw(\infty) = \xi$. By taking $\lambda = 0$ in the definition of $\Phi_r(t)$, we have

$$\Phi_{r_n}(t) \ge \|w(r_n) + w^*\|^2.$$

By continuity, we show that

$$\liminf_{n\to\infty} \Phi_{r_n}(t) \ge \lim_{n\to\infty} \|w(r_n) + w^*\|^2 = \|w(\infty) + w^*\|^2 \ge \Phi(t).$$

Since $\Phi_{r_n}(t) \le \Phi(t)$, the limit of $\Phi_{r_n}(t)$ exists and equals $\Phi(t)$. We can conclude that $\lim_{r\to\infty} \Phi_r(t) = \Phi(t)$ because $\Phi_r(t)$ is an increasing function of r.

In both cases, we have $\Phi_r(t) \to \Phi(t)$ as $r \to \infty$. The auxiliary problem corresponding to $\Phi_r(t)$ is

$$\phi_r(t) = \min_{\|w+w^*\|^2 \le 2t} \max_{\|\lambda\|_2 \le r} \|\lambda\| \langle H, w \rangle + \|w\| \langle G, \lambda \rangle - \langle \lambda, \xi \rangle + \|w + w^*\|^2 \tag{D.15}$$

which is upper bounded by

$$\begin{aligned} \phi(t) &= \min_{\|w+w^*\|^2 \le 2t} \max_{\lambda} \|\lambda\| \langle H, w \rangle + \|w\| \langle G, \lambda \rangle - \langle \lambda, \xi \rangle + \|w + w^*\|^2 \\ &= \min_{\substack{\langle H, w \rangle + \|G\|w\| - \xi\| \le 0 \\ \|w+w^*\|^2 \le 2t}} \|w + w^*\|^2. \end{aligned} \tag{D.16}$$

Applying CGMT and the fact that $\Phi_r(t)$ monotonically increases to $\Phi(t)$ almost surely, we can

conclude

$$\begin{aligned}\Pr\left(\min_{Xw=Y}\|w\|^2 > t \,|\, \xi\right) &= \Pr\left(\Phi(t) > t \,|\, \xi\right) = \Pr\left(\lim_{r\to\infty}\Phi_r(t) > t \,|\, \xi\right)\\ &\le \lim_{r\to\infty}\Pr\left(\Phi_r(t) > t \,|\, \xi\right)\\ &\le 2\cdot\lim_{r\to\infty}\Pr\left(\phi_r(t) > t \,|\, \xi\right)\\ &\le 2\cdot\Pr\left(\phi(t) > t \,|\, \xi\right) = 2\cdot\Pr\left(\min_{\langle H,w\rangle+\|G\|w\|-\xi\|\le 0}\|w+w^*\|^2 > t \,|\, \xi\right)\end{aligned}$$

By tower law, we have shown that

$$\Pr\left(\min_{Xw=Y}\|w\|^2 > t\right) \le 2\cdot\Pr\left(\min_{\|G\|w\|-\xi\|\le\langle H,w\rangle}\|w+w^*\|^2 > t\right).$$

To upper bound the minimum, we consider w of the form $\alpha w^* + \beta PH$ where $P = I - \frac{w^*(w^*)^T}{\|w^*\|^2}$. For the simplicity of notation, define

$$\epsilon = 2\sqrt{\frac{\log(40/\delta)}{n}} \quad \text{and} \quad \rho = \sqrt{\frac{1}{n}} + 2\sqrt{\frac{\log(20/\delta)}{n}}.$$

By a union bound, the following collection of events occurs with probability at least $1-\delta/2$:

1. By Lemma 51, it holds that
$$|\langle \xi, G\rangle| \le \rho\|\xi\|\cdot\|G\|$$

2. By Lemma 52, it holds that

$$(1-\epsilon)\sigma\sqrt{n} \le \|\xi\| \le (1+\epsilon)\sigma\sqrt{n}$$

$$(1-\epsilon)\sqrt{n} \le \|G\| \le (1+\epsilon)\sqrt{n}$$

$$\left(\sqrt{\frac{d-1}{n}} - \epsilon\right)\sqrt{n} \le \|PH\| \le \left(\sqrt{\frac{d-1}{n}} + \epsilon\right)\sqrt{n}$$

3. By standard Gaussian tail bound, it holds that

$$|\langle H, w^* \rangle| \leq \|w^*\| \epsilon \sqrt{n}$$

The above bounds imply that

$$\begin{aligned} \|G\|w\| - \xi\|^2 &= \|G\|^2 \|w\|^2 + \|\xi\|^2 - 2\|w\|\langle G, \xi \rangle \\ &\leq (1+\rho)(\|G\|^2 \|w\|^2 + \|\xi\|^2) \\ &\leq (1+\rho)(1+\epsilon)^2 n(\|w\|^2 + \sigma^2). \end{aligned}$$

By orthogonality, observe that

$$\|w\|^2 = \alpha^2 \|w^*\|^2 + \beta^2 \|PH\|^2$$

$$\langle H, w \rangle = \alpha \langle H, w^* \rangle + \beta \|PH\|^2,$$

and so to ensure that $\|G\|w\| - \xi\| \leq \langle H, w \rangle$, we can choose β such that

$$(1+\rho)^{1/2}(1+\epsilon)\sqrt{n(\alpha^2\|w^*\|^2 + \beta^2\|PH\|^2 + \sigma^2)} + \alpha\|w^*\|\epsilon\sqrt{n} \leq \beta\|PH\|^2.$$

Note that it suffices to have

$$(1+\rho)^{1/2}(1+2\epsilon)\sqrt{n(\alpha^2\|w^*\|^2 + \beta^2\|PH\|^2 + \sigma^2)} \leq \beta\|PH\|^2$$

$$\iff \alpha^2 \frac{\|w^*\|^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n} - 1} + \frac{\sigma^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n} - 1} \leq \beta^2\|PH\|^2$$

Again, by orthogonality, we have

$$\|w + w^*\|^2 = (1+\alpha)^2\|w^*\|^2 + \beta^2\|PH\|^2$$

and so

$$\begin{aligned}
&\min_{\|G\|w\|-\xi\|\le\langle H,w\rangle} \|w+w^*\|^2 \\
\le&\frac{\sigma^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n}-1}+\min_{\alpha}(1+\alpha)^2\|w^*\|^2+\alpha^2\frac{\|w^*\|^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n}-1} \\
=&\frac{\sigma^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n}-1}+\frac{\|w^*\|^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n}}
\end{aligned}$$

Finally, we can plug in the high probability lower bound for $\|PH\|\sqrt{n}$ and the proof is complete after some routine calculations. □

Theorem 27. *Using the notation of Theorem 24, we have with probability at least $1-\delta$ that for all w with $\|w\|_1 \le \|w^*\|_1$,*

$$\left|\sqrt{L(w)-\sigma^2}-\sqrt{\frac{\gamma\hat{L}(w)}{(1-\gamma)^2}}\right| \le \epsilon\sqrt{\hat{L}(w)}+\sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma}-\sigma^2\right)+\epsilon\hat{L}(w)} \tag{3.37}$$

provided $\gamma+2\epsilon/\sqrt{n}<1$, where

$$\mathcal{K}' := \{u : \|w^*+u\|_1 \le \|w^*\|_1\} \quad \textit{and} \quad \gamma := \frac{1}{n}\cdot W(\mathcal{K}'\cap S^{n-1})^2.$$

Proof. We use that for $\mathcal{K}' := \{u : \|w^*+u\|_1 \le \|w^*\|_1\}$

$$\langle w^*-w, x\rangle \le \|w^*-w\| \sup_{u\in\mathcal{K}'\cap S^{n-1}}\langle u, x\rangle$$

where S^{n-1} is the unit sphere. Recall that $\omega := W(\mathcal{K}'\cap S^{n-1})$ denotes the Gaussian width of the intersection of the tangent cone $\mathcal{K}'$ with the unit sphere. Let $\epsilon = \Theta\left(\frac{\log(36/\delta)}{n}\right)^{1/2}$ as in

Theorem 24, then with this notation Theorem 20 gives

$$\sigma^2 + \|w^* - w\|_2^2 \leq (1+\beta)\left(\sqrt{\hat{L}(w)} + \|w^* - w\|_2(\omega + 2\epsilon)/\sqrt{n}\right)^2$$
$$\leq (1+14\epsilon)\left(\sqrt{\hat{L}(w)} + \|w^* - w\|_2(\omega + 2\epsilon)/\sqrt{n}\right)^2.$$

This is a quadratic equation in $\|w^* - w\|_2$ which is of exactly the same form as the quadratic equation that arose in the analysis of Ordinary Least Squares (proof of Theorem 24), if we define $\gamma = \omega^2/n$. So solving the quadratic equation in the exact same way, we find that under the assumption $\gamma + 2\epsilon/\sqrt{n} < 1$ that

$$\left|\sqrt{L(w) - \sigma^2} - \sqrt{\frac{\gamma\hat{L}(w)}{(1-\gamma)^2}}\right| \leq \epsilon\sqrt{\hat{L}(w)} + \sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma} - \sigma^2\right) + \epsilon\hat{L}(w)}. \tag{D.17}$$

□

Theorem 29. *Under the model assumptions, fix $\gamma = d/n$ to be some value in $(0,1)$ and pick any $c > 0$. Then there exists another absolute constant $c' > 0$ such that for all sufficiently large n, with probability at least $1 - \delta$, there exists a $w \in \mathbb{R}^d$ such that*

$$\hat{L}(w) - \hat{L}(\hat{w}_{\mathrm{OLS}}) \leq c \cdot \frac{\sigma^2}{n^{1/2}}, \tag{3.38}$$

but the population error satisfies

$$L(w) - L(\hat{w}_{\mathrm{OLS}}) \geq c' \cdot \frac{\sigma^2}{n^{1/4}}. \tag{3.39}$$

Proof. Consider the following estimator:

$$w_\alpha = w^* + \alpha\left(\hat{w}_{\mathrm{OLS}} - w^*\right)$$
$$= w^* + \alpha(X^TX)^{-1}X^T\xi$$

Then the training error is

$$\begin{aligned}\hat{L}(w_\alpha) &= \frac{1}{n}\|Y - Xw_\alpha\|^2 = \frac{1}{n}\|\xi - \alpha X(X^TX)^{-1}X^T\xi\|^2 \\ &= \frac{1}{n}\| \left(I - X(X^TX)^{-1}X^T\right)\xi + (1-\alpha)X(X^TX)^{-1}X^T\xi\|^2 \\ &= \frac{1}{n}\| \left(I - X(X^TX)^{-1}X^T\right)\xi\|^2 + (1-\alpha)^2\frac{1}{n}\|X(X^TX)^{-1}X^T\xi\|^2 \\ &= \hat{L}(\hat{w}_{\mathrm{OLS}}) + (1-\alpha)^2\left\|\hat{w}_{\mathrm{OLS}} - w^*\right\|_{\hat{\Sigma}}^2\end{aligned}$$

With probability at least $1-\delta$, it holds that

$$\|\hat{w}_{\mathrm{OLS}} - w^*\|_{\hat{\Sigma}}^2 \leq \sigma^2\left(\sqrt{\gamma} + 2\sqrt{\frac{\log(4/\delta)}{n}}\right)^2$$

which can again be upper bounded by, for example, $4\sigma^2\gamma$ for a sufficiently large n. Therefore, we can let

$$(1-\alpha)^2 4\sigma^2\gamma = c\cdot\frac{\sigma^2}{\sqrt{n}}$$

and it suffices to pick

$$\alpha = 1 + \sqrt{\frac{c}{4\gamma}}\cdot\frac{1}{n^{1/4}}.$$

So if we define $c' = 2\sqrt{\frac{c}{4\gamma}}$, then the excess error of w_α satisfies

$$\begin{aligned}L(w_\alpha) - \sigma^2 &= \|\Sigma^{1/2}(w_\alpha - w^*)\|^2 \\ &= \alpha^2\|\Sigma^{1/2}(\hat{w}_{\mathrm{OLS}} - w^*)\|^2 \\ &\geq \left(1 + \frac{c'}{n^{1/4}}\right)\cdot L(\hat{w}_{\mathrm{OLS}}).\end{aligned}$$

The last inequality follows from the fact that $L(\hat{w}_{\mathrm{OLS}}) \geq \sigma^2$. □

Theorem 30. *Under the model assumptions in* (4.2) *with* $d \leq n$*, consider the ordinary least square estimator* $\hat{w}_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$*. It holds that*

$$\begin{aligned} \mathbb{E}L(\hat{w}_{\mathrm{OLS}}) &= \sigma^2 \frac{n-1}{n-d-1} \\ \mathrm{Var}(L(\hat{w}_{\mathrm{OLS}})) &= 2\sigma^4 \frac{d(n-1)}{(n-d-1)^2(n-d-3)} \end{aligned} \tag{3.40}$$

Hence as $d/n \to \gamma$*, it holds that*

$$\mathbb{E}L(\hat{w}_{\mathrm{OLS}}) \to \frac{\sigma^2}{1-\gamma} \quad \textit{and} \quad \frac{n}{\sigma^4}\mathrm{Var}(L(\hat{w}_{\mathrm{OLS}})) \to \frac{2\gamma}{(1-\gamma)^3}. \tag{3.41}$$

If d *is held constant, as* $n \to \infty$*, we have*

$$n\mathbb{E}[L(\hat{w}_{\mathrm{OLS}}) - \sigma^2] \to \sigma^2 d \quad \textit{and} \quad \frac{n^2}{\sigma^4}\mathrm{Var}(L(\hat{w}_{\mathrm{OLS}})) \to 2d. \tag{3.42}$$

Proof. Write $X = Z\Sigma^{1/2}$ and recall that

$$\begin{aligned} L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 &= \|\hat{w}_{\mathrm{OLS}} - w^*\|_\Sigma^2 = \left\|\Sigma^{1/2}(X^T X)^{-1} X^T \xi\right\|_2^2 \\ &= \xi^T Z (Z^T Z)^{-2} Z^T \xi. \end{aligned}$$

First, we compute the expectation. By the tower law, we have

$$\begin{aligned} \mathbb{E}L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 &= \mathbb{E}\left[\mathbb{E}\left[\xi^T Z (Z^T Z)^{-2} Z^T \xi \mid Z\right]\right] \\ &= \sigma^2 \mathbb{E}\,\mathrm{Tr}((Z^T Z)^{-1}) \\ &= \sigma^2\,\mathrm{Tr}(\mathbb{E}\left[(Z^T Z)^{-1}\right]) \end{aligned}$$

Proposition 2.1 of von Rosen [72] shows that

$$\mathbb{E}[(Z^T Z)^{-1}] = \frac{1}{n-d-1} I_d,$$

and so

$$\mathbb{E}L(\hat{w}_{\mathrm{OLS}}) = \sigma^2 + \sigma^2 \frac{d}{n-d-1} = \sigma^2 \frac{n-1}{n-d-1}.$$

To compute the variance, by the law of total variance, we have

$$\begin{aligned}\mathrm{Var}(L(\hat{w}_{\mathrm{OLS}})) &= \mathrm{Var}(L(\hat{w}_{\mathrm{OLS}}) - \sigma^2) \\ &= \mathbb{E}\,\mathrm{Var}(\xi^T Z(Z^T Z)^{-2} Z^T \xi \mid Z) + \mathrm{Var}(\mathbb{E}(\xi^T Z(Z^T Z)^{-2} Z^T \xi \mid Z))\end{aligned}$$

By the variance formula of Gaussian quadratic form, we have

$$\mathrm{Var}(\xi^T Z(Z^T Z)^{-2} Z^T \xi \mid Z) = 2\sigma^4 \,\mathrm{Tr}((Z^T Z)^{-2})$$

Proposition 2.1 of von Rosen [72] shows that

$$\mathbb{E}[(Z^T Z)^{-2}] = \frac{n-1}{(n-d)(n-d-1)(n-d-3)} I_d,$$

and so

$$\mathbb{E}\,\mathrm{Var}(\xi^T Z(Z^T Z)^{-2} Z^T \xi \mid Z) = \frac{2\sigma^4 d(n-1)}{(n-d)(n-d-1)(n-d-3)}.$$

To compute the second term, observe that

$$\begin{aligned}\mathrm{Var}(\mathbb{E}(\xi^T Z(Z^T Z)^{-2} Z^T \xi \mid Z)) &= \sigma^4 \,\mathrm{Var}(\mathrm{Tr}((Z^T Z)^{-1})) \\ &= \sigma^4 \,\mathrm{Var}(\mathrm{vec}(I_d)^T \mathrm{vec}((Z^T Z)^{-1})) \\ &= \sigma^4 \mathrm{vec}(I_d)^T \,\mathrm{Var}(\mathrm{vec}((Z^T Z)^{-1}))\mathrm{vec}(I_d)\end{aligned}$$

Proposition 2.1 of von Rosen [72] shows that

$$\mathrm{Var}(\mathrm{vec}((Z^T Z)^{-1})) = \frac{I_{d^2} + \sum_{i,j}(e_i \otimes e_j)(e_j^T \otimes e_i^T)}{(n-d)(n-d-1)(n-d-3)} + 2\frac{\mathrm{vec}(I_d)\mathrm{vec}(I_d)^T}{(n-d)(n-d-1)^2(n-d-3)}$$

and so

$$\begin{aligned}
&\frac{1}{\sigma^4}\operatorname{Var}(\mathbb{E}(\xi^T Z(Z^T Z)^{-2} Z^T \xi \mid Z)) \\
=&\frac{2d}{(n-d)(n-d-1)(n-d-3)} + \frac{2d^2}{(n-d)(n-d-1)^2(n-d-3)} \\
=&\frac{2d(n-1)}{(n-d)(n-d-1)^2(n-d-3)}.
\end{aligned}$$

Finally, we have shown that

$$\operatorname{Var}(L(\hat{w}_{\mathrm{OLS}})) = 2\sigma^4 \frac{d(n-1)}{(n-d-1)^2(n-d-3)}. \qquad \square$$

Theorem 31. *Under the model assumptions in* (4.2) *with $d \le n$, consider the ordinary least square estimator $\hat{w}_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$ and denote $\gamma = d/n$. Assume that $\gamma \le 0.999$, then with probability at least $1-\delta$, it holds that*

$$L(\hat{w}_{\mathrm{OLS}}) - \frac{\sigma^2}{1-\gamma} \lesssim \sigma^2 \sqrt{\frac{\gamma \log(36/\delta)}{n}}.$$

Proof. We are interested in the excess risk:

$$L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 = \|\Sigma^{1/2}(\hat{w}_{\mathrm{OLS}} - w^*)\|^2 = \|(Z^T Z)^{-1} Z^T \xi\|^2.$$

Notice that

$$\|(Z^T Z)^{-1} Z^T \xi\|^2 = \left((Z^T Z)^{-1/2} Z^T \xi\right)^T (Z^T Z)^{-1} \left((Z^T Z)^{-1/2} Z^T \xi\right)$$

and we have the following equality:

$$\begin{aligned}
b^T (Z^T Z)^{-1} b &= \max_u -\|Zu\|^2 + 2\langle u, b\rangle \\
&= \max_u \min_v \|v\|^2 + 2\langle v, Zu\rangle + 2\langle u, b\rangle.
\end{aligned}$$

We can plug in $(Z^TZ)^{-1/2}Z^T\xi$ into b. The b term may seem a bit complicated, but the key observation is that conditioned on Z, the distribution of $(Z^TZ)^{-1/2}Z^T\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ actually does not depend on Z, and so they are independent. Therefore, we can condition on $b = (Z^TZ)^{-1/2}Z^T\xi$ and the law of Z remains unchanged. To apply Theorem 45, we need use a truncation argument. Define the truncated problem as

$$\Phi_r = \max_{\|u\| \le r} \min_v \|v\|^2 + 2\langle v, Zu\rangle + 2\langle u, b\rangle, \tag{D.18}$$

then by Theorem 47, we have

$$\begin{aligned}
&\Pr\left(L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 > t \,|\, (Z^TZ)^{-1/2}Z^T\xi = b\right) \\
=&\Pr\left(\lim_{r\to\infty} \Phi_r > t\right) \le \lim_{r\to\infty} \Pr\left(\Phi_r > t\right).
\end{aligned}$$

Given u, the minimizer $v = -Zu$ satisfies $\|v\| \le r\|Z\|$ and so for any $M > 0$, we have

$$\begin{aligned}
\Pr\left(\Phi_r > t\right) &\le \Pr\left(\max_{\|u\|\le r} \min_{\|v\| \le rM} \|v\|^2 + 2\langle v, Zu\rangle + 2\langle u, b\rangle > t\right) + \Pr(\|Z\| \ge M) \\
&\le 2\Pr\left(\max_{\|u\|\le r} \min_{\|v\| \le rM} \|v\|^2 + 2\|v\|\langle H, u\rangle + 2\|u\|\langle G, v\rangle + 2\langle u, b\rangle > t\right) \\
&\quad + \Pr(\|Z\| \ge M) \\
&= 2\Pr\left(\max_{\|u\|\le r} \min_{\|v\| \le rM} \|v\|^2 + 2\|v\|\left(\langle H, u\rangle - \|G\|\|u\|\right) + 2\langle u, b\rangle > t\right) \\
&\quad + \Pr(\|Z\| \ge M)
\end{aligned}$$

by Gaussian minimax theorem. On the event that $\|G\| \ge \|H\|$, the minimizer is

$$\|v\| = \|G\|\|u\| - \langle H, u\rangle \ge (\|G\| - \|H\|)\|u\| > 0.$$

At the same time, we have $\|v\| \le r(\|G\| + \|H\|)$ and so

$$\begin{aligned}\Pr\left(\Phi_r > t\right) \le & 2\Pr\left(\max_{\|u\|\le r} 2\langle u, b\rangle - \left(\langle H, u\rangle - \|G\|\|u\|\right)^2 > t, \|G\| > \|H\|\right) \\ & + 2\Pr(\|G\| \le \|H\|) + 2\Pr(\|G\| + \|H\| \ge M) + \Pr(\|Z\| \ge M).\end{aligned}$$

As the max over $\{u : \|u\| \le r\}$ is always smaller than the overall max, taking $M \to \infty$, we have

$$\Pr\left(\Phi_r > t\right) \le 2\Pr\left(\max_{u} 2\langle u, b\rangle - \left(\|G\|\|u\| - \langle H, u\rangle\right)^2 > t, \|G\| > \|H\|\right) + 2\Pr(\|G\| \le \|H\|)$$

Observe that any u can be decomposed into two parts: one part spanned by b and the other part in the orthogonal complement of b. Formally, we write $u = \alpha b + k$ where $\langle k, b\rangle = 0$, and the problem becomes

$$\max_{\alpha\in\mathbb{R}, \langle k,b\rangle = 0} 2\alpha\|b\|^2 - \left(\|G\| \cdot \sqrt{\alpha^2\|b\|^2 + \|k\|^2} - \langle H, k\rangle - \alpha\langle H, b\rangle\right)^2.$$

Define $P = I_d - \frac{bb^T}{\|b\|^2}$. On the event that $\|G\| > \|H\|$, the quantity inside the square is always positive and so we want to choose the direction of k that make $\langle H, k\rangle$ as large as possible:

$$\begin{aligned}& \max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \min_{\langle k,b\rangle = 0} \left(\|G\| \cdot \sqrt{\alpha^2\|b\|^2 + \|k\|^2} - \langle H, k\rangle - \alpha\langle H, b\rangle\right)^2 \\ = & \max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \left(\min_{\langle k,b\rangle = 0} \|G\| \cdot \sqrt{\alpha^2\|b\|^2 + \|k\|^2} - \langle H, k\rangle - \alpha\langle H, b\rangle\right)^2 \\ = & \max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \left(\min_{\beta\ge 0} \|G\| \cdot \sqrt{\alpha^2\|b\|^2 + \beta^2} - \beta\|PH\| - \alpha\langle H, b\rangle\right)^2 \\ = & \max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \left(|\alpha| \cdot \|b\|\sqrt{\|G\|^2 - \|PH\|^2} - \alpha\langle H, b\rangle\right)^2 \\ \le & \max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \alpha^2\|b\|^2 \left(\sqrt{\|G\|^2 - \|PH\|^2} - \frac{|\langle H, b\rangle|}{\|b\|}\right)^2 = \frac{\|b\|^2}{\left(\sqrt{\|G\|^2 - \|PH\|^2} - \frac{|\langle H,b\rangle|}{\|b\|}\right)^2}\end{aligned}$$

By the tower law, we have shown that

$$\begin{aligned}
&\Pr\left(L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 > \frac{\|b\|^2}{t}\right) \\
\leq& 2\Pr\left(\|G\| \leq \|H\|\right) + 2\Pr\left(\sqrt{\|G\|^2 - \|PH\|^2} - \frac{|\langle H, b\rangle|}{\|b\|} < \sqrt{t}, \|G\| > \|H\|\right) \\
=& 2\Pr\left(\|G\| \leq \|H\| \quad \text{or} \quad \sqrt{\|G\|^2 - \|PH\|^2} - \frac{|\langle H, b\rangle|}{\|b\|} < \sqrt{t}, \|G\| > \|H\|\right)
\end{aligned}$$

For the simplicity of notation, denote

$$\epsilon = 2\sqrt{\frac{\log(32/\delta)}{n}}.$$

By a union bound, with probability at least $1 - \delta/2$, the following occurs:

1. by Lemma 52 and the fact that $b \sim \mathcal{N}(0, \sigma^2 I_d)$, it holds that

$$\|G\|^2 \geq n(1-\epsilon)^2$$

$$\|PH\|^2 \leq n(\sqrt{\gamma} + \epsilon)^2 \quad \text{and} \quad \|b\|^2 \leq \sigma^2 n(\sqrt{\gamma} + \epsilon)^2$$

2. As $\frac{\langle H, b\rangle}{\|b\|} \sim \mathcal{N}(0, 1)$, by standard Gaussian concentration, it holds that

$$\frac{|\langle H, b\rangle|}{\|b\|} \leq \epsilon\sqrt{n}$$

Therefore, for sufficiently large n, we have $\|G\| > \|H\|$ and we can pick t by setting

$$\sqrt{t} = \sqrt{n(1-\epsilon)^2 - n(\sqrt{\gamma} + \epsilon)^2} - \epsilon\sqrt{n}$$

and so with probability at least $1-\delta$, we have

$$L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 \leq \frac{\sigma^2(\sqrt{\gamma}+\epsilon)^2}{\left(\sqrt{(1-\epsilon)^2 - (\sqrt{\gamma}+\epsilon)^2} - \epsilon\right)^2}.$$

It is then routine to check the desired bound. □

D.5 Proofs for Section 3.6

Theorem 32. *Suppose that $d_1 d_2 > n$, then there exists some $\epsilon \lesssim \sqrt{\frac{\log(32/\delta)}{n}} + \frac{n}{d_1 d_2}$ such that with probability at least $1-\delta$, it holds that*

$$\min_{\forall i \in [n], \langle A_i, X\rangle = y_i} \|X\|_* \leq \|X^*\|_* + (1+\epsilon)\sqrt{\frac{n\sigma^2}{d_1 \vee d_2}}. \tag{3.44}$$

Proof. Without loss of generality, we will assume that $d_1 \leq d_2$. We will vectorize the measurement matrices and estimator $A_1, ..., A_n, X \in \mathbb{R}^{d_1 \times d_2}$ as $a_1, ..., a_n, x \in \mathbb{R}^{d_1 d_2}$ and define $\|x\|_* = \|X\|_*$. Denote $A = [a_1, ..., a_n]^T \in \mathbb{R}^{n \times d_1 d_2}$. We define the primary problem Φ by

$$\begin{aligned} \min_{\forall i \in [n], \langle A_i, X\rangle = \xi} \|X\|_* &= \min_{Ax = \xi} \|x\|_* \\ &= \min_x \sup_\lambda \langle \lambda, Ax - \xi\rangle + \|x\|_* := \Phi. \end{aligned} \tag{D.19}$$

Next, we define the truncation problem by

$$\Phi_r := \min_{\|x\|_2 \leq r} \sup_\lambda \langle \lambda, Ax - \xi\rangle + \|x\|_* \tag{D.20}$$

$$\Phi_{r,s} := \min_{\|x\|_2 \leq r} \sup_{\|\lambda\|_2 \leq s} \langle \lambda, Ax - \xi\rangle + \|x\|_*. \tag{D.21}$$

The corresponding auxiliary problems are

$$\begin{aligned}\Psi_{r,s} &:= \min_{\|x\|_2 \le r} \sup_{\|\lambda\|_2 \le s} \|\lambda\|_2 \langle H, x\rangle + \|x\|_2 \langle G, \lambda\rangle - \langle \lambda, \xi\rangle + \|x\|_* \\ &= \min_{\|x\|_2 \le r} \sup_{0 \le \lambda \le s} \lambda \left(\langle H, x\rangle + \|G\|x\|_2 - \xi\|_2\right) + \|x\|_*\end{aligned} \tag{D.22}$$

and the limit as $s \to \infty$ and then the limit as $r \to \infty$

$$\Psi_r := \min_{\substack{\|x\|_2 \le r \\ \|G\|x\|_2 - \xi\|_2 \le -\langle H, x\rangle}} \|x\|_* \tag{D.23}$$

$$\Psi := \min_{\|G\|x\|_2 - \xi\|_2 \le -\langle H, x\rangle} \|x\|_*. \tag{D.24}$$

By the same argument used in the proof of Theorem 26, it holds that $\Pr(\Phi > t) \le 2\Pr(\Psi \ge t)$ and so it suffices to analyze the auxiliary problem. We will pick x of the form $x = -\alpha H$ for some $\alpha \ge 0$, which needs to satisfy $\alpha \|H\|_2^2 \ge \|\alpha G\|H\|_2 - \xi\|_2$. By a union bound, the following events occur simultaneously with probability at least $1 - \delta/2$:

1. by Lemma 52, it holds that

$$\begin{aligned}\|G\|_2 &\le \sqrt{n} + 2\sqrt{\log(32/\delta)} \\ \frac{\|\xi\|_2}{\sigma} &\le \sqrt{n} + 2\sqrt{\log(32/\delta)} \\ \|H\|_2 &\le \sqrt{d_1 d_2} + 2\sqrt{\log(32/\delta)}\end{aligned}$$

2. Condition on ξ, we have $\frac{1}{\|\xi\|}\langle G, \xi\rangle \sim \mathcal{N}(0, 1)$ and so by standard Gaussian tail bound $\Pr(|Z| > t) \le 2e^{-t^2/2}$

$$\frac{|\langle G, \xi\rangle|}{\|\xi\|} \le \sqrt{2\log(16/\delta)}$$

Then we can use AM-GM inequality to show for sufficiently large n

$$\begin{aligned}
&\|\alpha G\|H\|_2 - \xi\|_2^2 \\
=&\alpha^2\|G\|_2^2\|H\|_2^2 + \|\xi\|^2 - 2\alpha\|H\|_2\langle G, \xi\rangle \\
\leq& n\alpha^2\|H\|_2^2\left(1 + 2\sqrt{\frac{\log(32/\delta)}{n}}\right)^2 + \|\xi\|^2 + 2\sqrt{n}\alpha\|H\|_2\|\xi\|_2\sqrt{\frac{2\log(16/\delta)}{n}} \\
\leq& n\alpha^2\|H\|_2^2\left(1 + 10\sqrt{\frac{\log(32/\delta)}{n}}\right) + \left(1 + \sqrt{\frac{2\log(16/\delta)}{n}}\right)\|\xi\|_2^2
\end{aligned}$$

and it suffices to let

$$\alpha^2\|H\|_2^4 \geq n\alpha^2\|H\|_2^2\left(1 + 10\sqrt{\frac{\log(32/\delta)}{n}}\right) + \left(1 + \sqrt{\frac{2\log(16/\delta)}{n}}\right)\|\xi\|_2^2.$$

Rearranging the above inequality, we can choose

$$\alpha = \left(\frac{1 + 10\sqrt{\frac{\log(32/\delta)}{n}}}{1 - \frac{n}{d_1 d_2}\left(1 + 10\sqrt{\frac{\log(32/\delta)}{n}}\right)\left(1 + 2\sqrt{\frac{\log(32/\delta)}{d_1 d_2}}\right)^2}\right)^{1/2} \frac{\sqrt{n\sigma^2}}{\|H\|_2^2}$$

and since H as a matrix can have at most rank d_1, by Cauchy-Schwarz inequality on the singular values of H, we have $\|H\|_* \leq \sqrt{d_1}\|H\|_2$ and

$$\|x\|_* = \alpha\|H\|_* \leq \alpha\sqrt{d_1}\|H\|_2 \leq (1+\epsilon)\sqrt{\frac{d_1(n\sigma^2)}{d_1 d_2}} = (1+\epsilon)\sqrt{\frac{n\sigma^2}{d_2}}$$

for some $\epsilon \lesssim \sqrt{\frac{\log(32/\delta)}{n}} + \frac{n}{d_1 d_2}$. □

Theorem 33. *Fix any $\delta \in (0,1)$. There exist constants $c_1, c_2, c_3 > 0$ such that if $d_1 d_2 > c_1 n$, $d_2 > c_2 d_1$, $n > c_3 r(d_1+d_2)$, then with probability at least $1-\delta$ that*

$$\frac{\|\hat{X} - X^*\|_F^2}{\|X^*\|_F^2} \lesssim \frac{r(d_1+d_2)}{n} + \sqrt{\frac{r(d_1+d_2)}{n}}\frac{\sigma}{\|X^*\|_F} + \left(\sqrt{\frac{d_1}{d_2}} + \frac{n}{d_1 d_2}\right)\frac{\sigma^2}{\|X^*\|_F^2}. \tag{3.45}$$

Proof. Note that $\langle A, X^*\rangle \sim \mathcal{N}(0, \|X^*\|_F^2)$ and so by the standard Gaussian tail bound $\Pr(|Z| \geq t) \leq 2e^{-t^2/2}$, Theorem 53 and a union bound, it holds with probability at least $1 - \delta/8$ that

$$\begin{aligned} |\langle A, X^*\rangle| &\leq \sqrt{2\log(32/\delta)}\|X^*\|_F \\ \|A\|_{op} &\leq \sqrt{d_1} + \sqrt{d_2} + \sqrt{2\log(32/\delta)}. \end{aligned}$$

Therefore, we can choose F in Theorem 20 by

$$\begin{aligned} \langle X - X^*, A\rangle &\leq \|A\|_{op}\|X\|_* + |\langle A, X^*\rangle| \\ &\leq \left(\sqrt{d_1} + \sqrt{d_2} + \sqrt{2\log(32/\delta)}\right)\|X\|_* + \sqrt{2\log(32/\delta)}\|X^*\|_F := F(X). \end{aligned}$$

Applying Theorem 20, we have

$$\sigma^2 + \|\hat{X} - X^*\|_F^2 = L(\hat{X}) \leq (1+\beta_1)\frac{F(\hat{X})^2}{n}$$

Moreover, since X^* has rank r and so $\|X^*\|_* \leq \sqrt{r}\|X^*\|_F$, we use Theorem 32 to show

$$\|\hat{X}\|_* \leq \sqrt{r}\|X^*\|_F + (1+\epsilon)\sqrt{\frac{n\sigma^2}{d_2}}.$$

To control $F(\hat{X})$, observe that

$$\begin{aligned} &\left(\sqrt{d_1} + \sqrt{d_2} + \sqrt{2\log(32/\delta)}\right)\sqrt{r}\|X^*\|_F + \sqrt{2\log(32/\delta)}\|X^*\|_F \\ \leq &\left(2\sqrt{r(d_1+d_2)} + (1+\sqrt{r})\sqrt{2\log(32/\delta)}\right)\|X^*\|_F \end{aligned}$$

and so we have

$$\frac{F(\hat{X})}{\sqrt{n}} \leq \left(2\sqrt{\frac{r(d_1+d_2)}{n}} + (1+\sqrt{r})\sqrt{\frac{2\log(32/\delta)}{n}}\right)\|X^*\|_F + \left(1+\sqrt{\frac{d_1}{d_2}} + \sqrt{\frac{2\log(32/\delta)}{d_2}}\right)(1+\epsilon)\sigma.$$

The desired conclusion follows by plugging in the above estimates and some rearrangements. □

APPENDIX E

PROOFS FOR SECTION 4

In this appendix, we will first prove Theorem 41, which immediately implies Theorem 36 as a special case, and then Theorem 37 and its applications.

Notation. We consider the general setting introduced in Section 4.5. Following the tradition in statistics, we denote $X = (x_1, ..., x_n)^T \in \mathbb{R}^{n\times d}$ as the design matrix. In the proof section, we slightly abuse the notation of η_i to mean Xw_i^* and ξ to mean the n-dimensional random vector whose i-th component satisfies $y_i = g(\eta_{1,i}, ..., \eta_{k,i}, \xi_i)$. We will write $X = Z\Sigma^{1/2}$ where Z is a random matrix with i.i.d. standard normal entries if $\mu = 0$.

Throughout this section, we can first assume $\mu = 0$ in Assumption (A) without loss of generality because if we define $\tilde{f} : \mathbb{R} \times \mathcal{Y} \times \Theta \to \mathbb{R}$ by

$$\tilde{f}(\hat{y}, y, \theta) := f(\hat{y} + \langle w(\theta), \mu\rangle, y, \theta), \tag{E.1}$$

then by definition, it holds that

$$f(\langle w(\theta), x\rangle, y, \theta) = \tilde{f}(\langle w(\theta), x - \mu\rangle, y, \theta)$$

and so we can apply the theory on $\tilde{f}$ first and then translate to the problem on f. Similarly, we can also assume $\Sigma^{1/2}w_1^*, ..., \Sigma^{1/2}w_k^*$ are orthonormal without loss of generality. This is because we can denote $W \in \mathbb{R}^{d\times k}$ by $W = [w_1^*, ..., w_k^*]$ and let $\tilde{W} = W(W^T\Sigma W)^{-1/2}$. By definition, it holds that $\tilde{W}^T\Sigma\tilde{W} = I$ and so the columns of $\tilde{W} = [\tilde{w}_1^*, ..., \tilde{w}_k^*]$ satisfy $\Sigma^{1/2}\tilde{w}_1^*, ..., \Sigma^{1/2}\tilde{w}_k^*$ are orthonormal. If we define $\tilde{g} : \mathbb{R}^{k+1} \to \mathbb{R}$ by

$$\tilde{g}(\eta_1, ..., \eta_k, \xi) = g([\eta_1, ..., \eta_k](W^T\Sigma W)^{1/2} + \mu^T W, \xi), \tag{E.2}$$

then $y = \tilde{g}(x^T \tilde{W}, \xi)$ and so we can apply the theory on $\tilde{g}$.

E.1 Proof of Theorem 41

Lemma 72. *Fix any integer $k < d$ and any k vectors $w_1^*, ..., w_k^*$ in $\mathbb{R}^d$ such that $\Sigma^{1/2} w_1^*, ..., \Sigma^{1/2} w_k^*$ are orthonormal. Denoting*

$$P = I_d - \sum_{i=1}^{k} (\Sigma^{1/2} w_i^*)(\Sigma^{1/2} w_i^*)^T, \tag{E.3}$$

the distribution of X conditional on $Xw_1^* = \eta_1, ..., Xw_k^* = \eta_k$ is the same as that of

$$\sum_{i=1}^{k} \eta_i (\Sigma w_i^*)^T + ZP\Sigma^{1/2}. \tag{E.4}$$

Proof. We can write $X = Z\Sigma^{1/2}$. The key observation is that ZP, $Z\Sigma^{1/2} w_1^*, ..., Z\Sigma^{1/2} w_k^*$ are independent. To see why this is the case, we can vectorize each term:

$$\begin{pmatrix} \text{vec}(ZP) \\ \text{vec}(Z\Sigma^{1/2} w_1^*) \\ ... \\ \text{vec}(Z\Sigma^{1/2} w_k^*) \end{pmatrix} = \begin{pmatrix} P \otimes I_n \\ (\Sigma^{1/2} w_1^*)^T \otimes I_n \\ ... \\ (\Sigma^{1/2} w_k^*)^T \otimes I_n \end{pmatrix} \text{vec}(Z)$$

From the above representation, we see that the joint distribution is multivariate Gaussian and the covariance matrix is

$$\begin{pmatrix} P \otimes I_n \\ (\Sigma^{1/2} w_1^*)^T \otimes I_n \\ ... \\ (\Sigma^{1/2} w_k^*)^T \otimes I_n \end{pmatrix} \begin{pmatrix} P \otimes I_n \\ (\Sigma^{1/2} w_1^*)^T \otimes I_n \\ ... \\ (\Sigma^{1/2} w_k^*)^T \otimes I_n \end{pmatrix}^T = \text{diag}\left(P \otimes I_n, I_n, ..., I_n\right)$$

Therefore, the distribution of ZP remains unchanged after conditioning on $Z\Sigma^{1/2}w_1^*, ..., Z\Sigma^{1/2}w_k^*$, and we can write

$$\begin{aligned} Z &= Z\left(\sum_{i=1}^{k}(\Sigma^{1/2}w_i^*)(\Sigma^{1/2}w_i^*)^T\right) + ZP \\ &= \sum_{i=1}^{k}\eta_i(\Sigma^{1/2}w_i^*)^T + ZP. \end{aligned}$$

The proof is concluded by the fact that $X = Z\Sigma^{1/2}$. □

The above lemma allows us to ignore the multi-index model (B) and simply treat y as deterministic by a conditioning argument, while preserving the Gaussianity of the design matrix. Next, we will write the generalization problem as a Primary Optimization problem in Gaussian Minimax Theorem (see Theorem 46 in Appendix A). For generality, we will let F be any deterministic function and then choose it in the end.

Lemma 73. *Fix an arbitrary set $\Theta \subseteq \mathbb{R}^p$ and let $F : \Theta \to \mathbb{R}$ be any deterministic and continuous function. Consider dataset (X, Y) drawn i.i.d. from the data distribution $\mathcal{D}$ according to (A) and (B) with $\mu = 0$ and orthonormal $\Sigma^{1/2}w_1^*, ..., \Sigma^{1/2}w_k^*$. Then conditioned on $Xw_1^* = \eta_1, ..., Xw_k^* = \eta_k$ and ξ, if we define*

$$\Phi := \sup_{\substack{(w,u,\theta)\in\mathbb{R}^d\times\mathbb{R}^n\times\Theta \\ w = P\Sigma^{1/2}w(\theta)}} \inf_{\lambda\in\mathbb{R}^n} \langle\lambda, Zw\rangle + \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) \tag{E.5}$$

where P is defined in (E.3) *and ψ is a deterministic and continuous function given by*

$$\begin{aligned} \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) = F(\theta) &- \frac{1}{n}\sum_{i=1}^{n} f(u_i, g(\eta_{1,i}, ..., \eta_{k,i}, \xi_i), \theta) \\ &+ \langle\lambda, \left(\sum_{i=1}^{k}\eta_i(\Sigma w_i^*)^T\right) w(\theta) - u\rangle, \end{aligned} \tag{E.6}$$

then it holds that for any $t \in \mathbb{R}$, we have

$$\Pr\left(\sup_{\theta\in\Theta} F(\theta) - \hat{L}(\theta) > t \,\middle|\, \eta_1, ..., \eta_k, \xi\right) = \Pr(\Phi > t). \tag{E.7}$$

Proof. By introducing a variable $u = Xw(\theta)$, we have

$$\begin{aligned}\sup_{\theta\in\Theta} F(\theta) - \hat{L}(\theta) &= \sup_{\theta\in\Theta} F(\theta) - \frac{1}{n}\sum_{i=1}^{n} f(\langle w(\theta), x_i\rangle, y_i, \theta) \\ &= \sup_{\theta\in\Theta, u\in\mathbb{R}^n} \inf_{\lambda\in\mathbb{R}^n} \langle \lambda, Xw(\theta) - u\rangle + F(\theta) - \frac{1}{n}\sum_{i=1}^{n} f(u_i, y_i, \theta).\end{aligned}$$

Conditioned on $Xw_1^* = \eta_1, ..., Xw_k^* = \eta_k$ and ξ, the above is only random in X by our multi-index model assumption on y. By Lemma 72, the above is equal in law to

$$\begin{aligned}&\sup_{\theta\in\Theta, u\in\mathbb{R}^n} \inf_{\lambda\in\mathbb{R}^n} \langle \lambda, \left(\sum_{i=1}^{k} \eta_i(\Sigma w_i^*)^T + ZP\Sigma^{1/2}\right) w(\theta) - u\rangle + F(\theta) - \frac{1}{n}\sum_{i=1}^{n} f(u_i, y_i, \theta) \\ =& \sup_{\theta\in\Theta, u\in\mathbb{R}^n} \inf_{\lambda\in\mathbb{R}^n} \langle \lambda, \left(ZP\Sigma^{1/2}\right) w(\theta)\rangle + \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) \\ =& \sup_{\substack{(w,u,\theta)\in\mathbb{R}^d\times\mathbb{R}^n\times\Theta \\ w = P\Sigma^{1/2} w(\theta)}} \inf_{\lambda\in\mathbb{R}^n} \langle \lambda, Zw\rangle + \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) \\ =& \Phi.\end{aligned}$$

The function ψ is continuous because we require F, f and w to be continuous in the definitions. □

Next, we are ready to apply Gaussian Minimax Theorem. Although the domains in (E.5) are not compact, we can use the truncation lemmas 47 and 48 in Appendix A.

Lemma 74. *In the same setting as Lemma 73, define the auxiliary problem as*

$$\Psi := \sup_{\substack{(u,\theta)\in\mathbb{R}^n\times\Theta \\ \langle H, P\Sigma^{1/2} w(\theta)\rangle \geq \left\| \|P\Sigma^{1/2} w(\theta)\|_2 G + \sum_{i=1}^{k} \langle w(\theta), \Sigma w_i^*\rangle \eta_i - u \right\|_2}} F(\theta) - \frac{1}{n}\sum_{i=1}^{n} f(u_i, y_i, \theta) \tag{E.8}$$

then for any $t \in \mathbb{R}$, it holds that

$$\Pr\left(\sup_{\theta\in\mathcal{K}} F(\theta) - \hat{L}(\theta) > t\right) \leq 2\Pr(\Psi \geq t). \tag{E.9}$$

where the randomness in the second probability is taken over $G, H, \eta_1, ..., \eta_k$ and ξ.

Proof. Denote $\mathcal{S}_r = \{(w, u, \theta) \in \mathbb{R}^d \times \mathbb{R}^n \times \Theta : w = P\Sigma^{1/2}w(\theta) \text{ and } \|w\|_2 + \|u\|_2 + \|\theta\|_2 \leq r\}$. The set $\mathcal{S}_r$ is bounded by definition and closed by the continuity of w. Hence, it is compact. Next, we denote the truncated problems:

$$\Phi_r := \sup_{(w,u,\theta)\in\mathcal{S}_r} \inf_{\lambda\in\mathbb{R}^n} \langle\lambda, Zw\rangle + \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) \tag{E.10}$$

$$\Phi_{r,s} := \sup_{(w,u,\theta)\in\mathcal{S}_r} \inf_{\|\lambda\|_2\leq s} \langle\lambda, Zw\rangle + \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi). \tag{E.11}$$

By definition, we have $\Phi_r \leq \Phi_{r,s}$ and so

$$\Pr(\Phi_r > t) \leq \Pr(\Phi_{r,s} > t).$$

The corresponding auxiliary problems are

$$\begin{aligned}
\Psi_{r,s} &:= \sup_{(w,u,\theta)\in\mathcal{S}_r} \inf_{\|\lambda\|_2\leq s} \|\lambda\|_2\langle H, w\rangle + \|w\|_2\langle G, \lambda\rangle + \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) \\
&= \sup_{(w,u,\theta)\in\mathcal{S}_r} \inf_{\|\lambda\|_2\leq s} \|\lambda\|_2\langle H, w\rangle + \langle\lambda, \|w\|_2 G + \sum_{i=1}^{k} \eta_i\langle w(\theta), \Sigma w_i^*\rangle - u\rangle \\
&\qquad + F(\theta) - \frac{1}{n}\sum_{i=1}^{n} f(u_i, g(\eta_{1,i}, ..., \eta_{k,i}, \xi_i), \theta) \\
&= \sup_{(w,u,\theta)\in\mathcal{S}_r} \inf_{0\leq\lambda\leq s} \lambda\left(\langle H, w\rangle - \left\|\|w\|_2 G + \sum_{i=1}^{k} \eta_i\langle w(\theta), \Sigma w_i^*\rangle - u\right\|_2\right) \\
&\qquad + F(\theta) - \frac{1}{n}\sum_{i=1}^{n} f(u_i, g(\eta_{1,i}, ..., \eta_{k,i}, \xi_i), \theta)
\end{aligned}$$

and the limit of $s \to \infty$:

$$\Psi_r := \sup_{\substack{(w,u,\theta)\in\mathcal{S}_r \\ \langle H,w\rangle \geq \left\| \|w\|_2 G + \sum_{i=1}^k \eta_i \langle w(\theta), \Sigma w_i^* \rangle - u \right\|_2}} F(\theta) - \frac{1}{n}\sum_{i=1}^n f(u_i, g(\eta_{1,i}, ..., \eta_{k,i}, \xi_i), \theta)$$

By definition, it holds that $\Psi_r \leq \Psi$ and so

$$\Pr(\Psi_r \geq t) \leq \Pr(\Psi \geq t).$$

Thus, it holds that

$$\begin{aligned}
\Pr(\Phi > t) &= \lim_{r\to\infty} \Pr(\Phi_r > t) && \text{by Lemma 47} \\
&\leq \lim_{r\to\infty} \lim_{s\to\infty} \Pr(\Phi_{r,s} > t) \\
&\leq 2 \lim_{r\to\infty} \lim_{s\to\infty} \Pr(\Psi_{r,s} \geq t) && \text{by Theorem 46} \\
&= 2 \lim_{r\to\infty} \Pr(\Psi_r \geq t) && \text{by Lemma 48} \\
&\leq 2 \Pr(\Psi \geq t).
\end{aligned}$$

The proof concludes by applying Lemma 73 and the tower law. □

The following two simple lemmas will be useful to analyze the auxiliary problem.

Lemma 75. *For $a, b, H > 0$, we have*

$$\sup_{\lambda\geq 0} -\lambda a + \frac{\lambda}{H+\lambda} b = (\sqrt{b} - \sqrt{Ha})_+^2.$$

Proof. Observe that

$$\sup_{\lambda\geq 0} -\lambda a + \frac{\lambda}{H+\lambda} b = b - \inf_{\lambda\geq 0} \lambda a + \frac{H}{H+\lambda} b.$$

Define $f(\lambda) = \lambda a + \frac{H}{H+\lambda} b$, then

$$f'(\lambda) = a - \frac{Hb}{(H+\lambda)^2} \leq 0 \iff (H+\lambda)^2 \leq \frac{Hb}{a}$$
$$\iff -\sqrt{\frac{Hb}{a}} - H \leq \lambda \leq \sqrt{\frac{Hb}{a}} - H$$

Since we require $\lambda \geq 0$, we only need to consider whether $\sqrt{\frac{Hb}{a}} - H \geq 0 \iff b \geq Ha$. If $b < Ha$, the infimum is attained at $\lambda = 0$. Otherwise, the infimum is attained at $\lambda^* = \sqrt{\frac{Hb}{a}} - H$, at which point

$$f(\lambda^*) = 2\sqrt{Hba} - Ha.$$

Plugging in, we see that the expression is equivalent to $(\sqrt{b} - \sqrt{Ha})_+^2$ in both cases. □

Lemma 76. *For $a, b \geq 0$, we have*

$$\sup_{\lambda \geq 0} -\lambda a - \frac{b}{\lambda} = -\sqrt{4ab}$$

Proof. Define $f(\lambda) = -\lambda a - \frac{b}{\lambda}$, then

$$f'(\lambda) = -a + \frac{b}{\lambda^2} \geq 0 \iff \frac{b}{a} \geq \lambda^2$$

and so in the domain $\lambda \geq 0$, the optimum is attained at $\lambda^* = \sqrt{b/a}$ at which point $f(\lambda^*) = -2\sqrt{ab}$. □

We are now ready to analyze the auxiliary problem.

Lemma 77. *In the same setting as in Lemma 73, assume that for every $\delta > 0$*

(A) $C_\delta : \mathbb{R}^d \to [0, \infty]$ is a continuous function such that with probability at least $1 - \delta/4$ over $H \sim \mathcal{N}(0, I_d)$, uniformly over all $w \in \mathbb{R}^d$, we have that

$$\langle \Sigma^{1/2} PH, w \rangle \leq C_\delta(w) \tag{E.12}$$

(B) ϵ_δ *is a positive real number such that with probability at least* $1-\delta/4$ *over* $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$ *drawn i.i.d. from* $\tilde{D}$*, it holds uniformly over all* $\theta \in \Theta$ *that*

$$\frac{1}{n}\sum_{i=1}^{n} f(\langle \phi(w(\theta)), \tilde{x}_i\rangle, \tilde{y}_i, \theta) \geq \frac{1}{1+\epsilon_\delta}\mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{D}}[f(\langle \phi(w(\theta)), \tilde{x}\rangle, \tilde{y}, \theta)]. \tag{E.13}$$

where the distribution $\tilde{D}$ *over* $(\tilde{x}, \tilde{y})$ *is given by*

$$\tilde{x} \sim \mathcal{N}(0, I_{k+1}), \quad \tilde{\xi} \sim \mathcal{D}_\xi, \quad \tilde{y} = g(\tilde{x}_1, ..., \tilde{x}_k, \tilde{\xi})$$

and the mapping $\phi : \mathbb{R}^d \to \mathbb{R}^{k+1}$ *is defined as*

$$\phi(w) = (\langle w, \Sigma w_1^*\rangle, ..., \langle w, \Sigma w_k^*\rangle, \|P\Sigma^{1/2}w\|_2)^T.$$

Then the following is true:

(i) suppose for some choice of M_θ *that is continuous in* θ*, it holds for every* $y \in \mathcal{Y}$ *and* $\theta \in \Theta$*,* f *is* M_θ*-Lipschitz with respect to the first argument, then with probability at least* $1-\delta$*, uniformly over all* $\theta \in \Theta$*, we have*

$$L(\theta) \leq (1+\epsilon_\delta)\left(\hat{L}(\theta) + M_\theta\sqrt{\frac{C_\delta(w(\theta))^2}{n}}\right). \tag{E.14}$$

(ii) suppose for some choice of H_θ *that is continuous in* θ*, it holds for every* $y \in \mathcal{Y}$ *and* $\theta \in \Theta$*,* f *is non-negative and* $\sqrt{f}$ *is* $\sqrt{H_\theta}$*-Lipschitz with respect to the first argument, then with probability at least* $1-\delta$*, uniformly over all* $\theta \in \Theta$*, we have*

$$L(\theta) \leq (1+\epsilon_\delta)\left(\sqrt{\hat{L}(\theta)} + \sqrt{\frac{H_\theta C_\delta(w(\theta))^2}{n}}\right)^2. \tag{E.15}$$

Proof. First, let's simplify the auxiliary problem (E.8). Changing variables to subtract the quantity

$G_i \left\| P\Sigma^{1/2} w(\theta) \right\|_2 + \sum_{l=1}^k \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i}$ from each of the former u_i, we have that

$$\Psi = \sup_{\substack{(u,\theta)\in\mathbb{R}^n\times\Theta \\ \|u\|_2 \le \langle H, P\Sigma^{1/2} w(\theta)\rangle}} F(\theta) - \frac{1}{n}\sum_{i=1}^n f\left(u_i + G_i \left\| P\Sigma^{1/2} w(\theta) \right\|_2 + \sum_{l=1}^k \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i}, y_i, \theta \right)$$

and separating the optimization problem in u and θ, we obtain

$$\begin{aligned}\Psi = \sup_{\theta\in\Theta}\ & F(\theta) \\ & - \frac{1}{n} \inf_{\substack{u\in\mathbb{R}^n: \\ \|u\|_2 \le \langle H, P\Sigma^{1/2} w(\theta)\rangle}} \sum_{i=1}^n f\left(u_i + G_i \left\| P\Sigma^{1/2} w(\theta) \right\|_2 + \sum_{l=1}^k \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i}, y_i, \theta \right).\end{aligned}$$

Next, we will lower bound the infimum term by weak duality to obtain upper bound on Ψ:

$$\begin{aligned}& \inf_{\substack{u\in\mathbb{R}^n: \\ \|u\|_2 \le \langle H, P\Sigma^{1/2} w(\theta)\rangle}} \sum_{i=1}^n f\left(u_i + G_i \left\| P\Sigma^{1/2} w(\theta) \right\|_2 + \sum_{l=1}^k \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i}, y_i, \theta \right) \\ &= \inf_{u\in\mathbb{R}^n} \sup_{\lambda\ge 0} \lambda(\|u\|_2^2 - \langle \Sigma^{1/2} PH, w(\theta)\rangle^2) \\ &\qquad + \sum_{i=1}^n f\left(u_i + G_i \left\| P\Sigma^{1/2} w(\theta) \right\|_2 + \sum_{l=1}^k \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i}, y_i, \theta \right) \\ &\ge \sup_{\lambda\ge 0} -\lambda \langle \Sigma^{1/2} PH, w(\theta)\rangle^2 \\ &\qquad + \inf_{u\in\mathbb{R}^n} \sum_{i=1}^n f\left(u_i + G_i \left\| P\Sigma^{1/2} w(\theta) \right\|_2 + \sum_{l=1}^k \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i}, y_i, \theta \right) + \lambda\|u\|_2^2 \\ &= \sup_{\lambda\ge 0} -\lambda \langle \Sigma^{1/2} PH, w(\theta)\rangle^2 \\ &\qquad + \sum_{i=1}^n \inf_{u_i\in\mathbb{R}} f\left(u_i + G_i \left\| P\Sigma^{1/2} w(\theta) \right\|_2 + \sum_{l=1}^k \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i}, y_i, \theta \right) + \lambda u_i^2 \\ &= \sup_{\lambda\ge 0} -\lambda \langle \Sigma^{1/2} PH, w(\theta)\rangle^2 + \sum_{i=1}^n f_\lambda\left(G_i \left\| P\Sigma^{1/2} w(\theta) \right\|_2 + \sum_{l=1}^k \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i}, y_i, \theta \right).\end{aligned}$$

Suppose that for every $y \in \mathcal{Y}$ and $\theta \in \Theta$, f is M_θ-Lipschitz with respect to the first argument, then by the same argument in the proof of equation (4.7), the above can be further lower bounded by the following quantity:

$$\sup_{\lambda \geq 0} -\lambda \langle \Sigma^{1/2} PH, w(\theta) \rangle^2 - \frac{nM_\theta^2}{4\lambda} + \sum_{i=1}^{n} f\left(\sum_{l=1}^{k} \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i} + \left\| P\Sigma^{1/2} w(\theta) \right\|_2 G_i, y_i, \theta \right).$$

On the other hand, suppose that for every $y \in \mathcal{Y}$ and $\theta \in \Theta$, f is non-negative and $\sqrt{f}$ is $\sqrt{H_\theta}$-Lipschitz with respect to the first argument, then by the same argument in the proof of equation (4.8), the above can be further lower bounded by:

$$\sup_{\lambda \geq 0} -\lambda \langle \Sigma^{1/2} PH, w(\theta) \rangle^2 + \frac{\lambda}{H_\theta + \lambda} \left[\sum_{i=1}^{n} f\left(\sum_{l=1}^{k} \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i} + \left\| P\Sigma^{1/2} w(\theta) \right\|_2 G_i, y_i, \theta \right) \right].$$

Notice that if we write $\tilde{x}_i = (\eta_{1,i}, ..., \eta_{k,i}, G_i)$, then $(\tilde{x}_i, y_i)$ are independent with distribution exactly equal to $\tilde{\mathcal{D}}$. Moreover, we have

$$f\left(\sum_{l=1}^{k} \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i} + \left\| P\Sigma^{1/2} w(\theta) \right\|_2 G_i, y_i, \theta \right) = f(\langle \phi(w(\theta)), \tilde{x}_i \rangle, y_i, \theta)$$

and it is easy to see that the joint distribution of $(\langle \phi(w(\theta)), \tilde{x} \rangle, y)$ with $(\tilde{x}, y) \sim \tilde{\mathcal{D}}$ is exactly the same as $(\langle w(\theta), x \rangle, y)$ with $(x, y) \sim \mathcal{D}$. As a result, we have that

$$\mathbb{E}_{(\tilde{x},y)\sim\tilde{D}}[f(\langle \phi(w(\theta)), \tilde{x} \rangle, y, \theta)] = L(\theta).$$

By our assumption (E.12), (E.13) and a union bound, we have with probability at least $1 - \delta/2$

$$|\langle \Sigma^{1/2} PH, w(\theta) \rangle| \leq C_\delta(w(\theta))$$

$$\frac{1}{n} \sum_{i=1}^{n} f\left(\sum_{l=1}^{k} \langle w(\theta), \Sigma w_l^* \rangle \eta_{l,i} + \left\| P\Sigma^{1/2} w(\theta) \right\|_2 G_i, y_i, \theta \right) \geq \frac{1}{1 + \epsilon_\delta} L(\theta).$$

Therefore, if f is M_θ-Lipschitz, then by by Lemma 76, we have

$$\begin{aligned}\Psi &\leq \sup_{\theta\in\Theta} F(\theta) - \sup_{\lambda\geq 0} -\lambda\frac{C_\delta(w(\theta))^2}{n} - \frac{M_\theta^2}{4\lambda} + \frac{1}{1+\epsilon_\delta}L(\theta)\\ &= \sup_{\theta\in\Theta} F(\theta) + \sqrt{M_\theta^2\frac{C_\delta(w(\theta))^2}{n}} - \frac{1}{1+\epsilon_\delta}L(\theta)\end{aligned}$$

Consequently, by taking $F(\theta) = \frac{1}{1+\epsilon_\delta}L(\theta) - M_\theta\sqrt{\frac{C_\delta(w(\theta))^2}{n}}$ and Lemma 74, we have shown that with probability at least $1-\delta$, we have

$$\sup_{\theta\in\mathcal{K}} F(\theta) - \hat{L}(\theta) \leq 0 \implies \frac{1}{1+\epsilon_\delta}L(\theta) \leq \hat{L}(\theta) + M_\theta\sqrt{\frac{C_\delta(w(\theta))^2}{n}}.$$

If $\sqrt{f}$ is $\sqrt{H_\theta}$-Lipschitz, then by Lemma 75

$$\begin{aligned}\Psi &\leq \sup_{\theta\in\mathcal{K}} F(\theta) - \sup_{\lambda\geq 0} -\lambda\frac{C_\delta(w(\theta))^2}{n} + \frac{\lambda}{H_\theta+\lambda}\frac{1}{1+\epsilon_\delta}L(\theta)\\ &= \sup_{\theta\in\mathcal{K}} F(\theta) - \left(\sqrt{\frac{L(\theta)}{1+\epsilon_\delta}} - \sqrt{\frac{H_\theta C_\delta(w(\theta))^2}{n}}\right)_+^2.\end{aligned}$$

Consequently, by taking $F(\theta) = \left(\sqrt{\frac{L(\theta)}{1+\epsilon_\delta}} - \sqrt{\frac{H_\theta C_\delta(w(\theta))^2}{n}}\right)_+^2$ and Lemma 74, we have shown that with probability at least $1-\delta$, we have

$$\sup_{\theta\in\mathcal{K}} F(\theta) - \hat{L}(\theta) \leq 0.$$

Rearranging, either we have

$$\sqrt{\frac{L(\theta)}{1+\epsilon_\delta}} - \sqrt{\frac{H_\theta C_\delta(w(\theta))^2}{n}} < 0 \implies L(\theta) < (1+\epsilon_\delta)\frac{H_\theta C_\delta(w(\theta))^2}{n}$$

or we have

$$\sqrt{\frac{L(\theta)}{1+\epsilon_\delta}} - \sqrt{\frac{H_\theta C_\delta(w(\theta))^2}{n}} \geq 0 \implies \left(\sqrt{\frac{L(\theta)}{1+\epsilon_\delta}} - \sqrt{\frac{H_\theta C_\delta(w(\theta))^2}{n}}\right)^2 \leq \hat{L}(\theta)$$
$$\implies L(\theta) \leq (1+\epsilon_\delta)\left(\sqrt{\hat{L}(\theta)} + \sqrt{\frac{H_\theta C_\delta(w(\theta))^2}{n}}\right)^2.$$

In either case, the desired bound holds. □

Finally, we are ready to prove Theorem 41, which is restated below for convenience.

Theorem 41. *Suppose that assumptions (A), (B), (E) and (F) hold. Denote $W \in \mathbb{R}^{d\times k}$ by $W = [w_1^*, ..., w_k^*]$ and let $Q = I - W(W^T\Sigma W)^{-1}W^T\Sigma$. For any $\delta \in (0,1)$, let $C_\delta : \mathbb{R}^d \to [0,\infty]$ be a continuous function such that with probability at least $1-\delta/4$ over $x \sim \mathcal{N}(0,\Sigma)$, uniformly over all $\theta \in \Theta$,*

$$\langle Qw(\theta), x\rangle \leq C_\delta(w(\theta)). \tag{4.34}$$

Then it holds that

(i) if for each $\theta \in \Theta$ and $y \in \mathcal{Y}$, f is M_θ-Lipschitz with respect to the first argument and M_θ is continuous in θ, then with probability at least $1-\delta$, it holds that uniformly over all $\theta \in \Theta$, we have

$$(1-\epsilon)\, L(\theta) \leq \hat{L}(\theta) + M_\theta\sqrt{\frac{C_\delta(w(\theta))^2}{n}} \tag{4.35}$$

(ii) if for each $\theta \in \Theta$ and $y \in \mathcal{Y}$, f is non-negative and $\sqrt{f}$ is $\sqrt{H_\theta}$-Lipschitz with respect to the first argument, and H_θ is continuous in θ, then with probability at least $1-\delta$, it holds that uniformly over all $\theta \in \Theta$, we have

$$(1-\epsilon)\, L(\theta) \leq \left(\sqrt{\hat{L}(\theta)} + \sqrt{\frac{H_\theta\, C_\delta(w(\theta))^2}{n}}\right)^2 \tag{4.36}$$

where $\epsilon = O\left(\tau\sqrt{\frac{h\log(n/h)+\log(1/\delta)}{n}}\right)$.

Proof. We apply the reduction argument at the beginning of the appendix. Given $\mathcal{D}$ that satisfies assumptions (A) and (B), we define $[\tilde{w}_1^*, ..., \tilde{w}_k^*] = \tilde{W} = W(W^T\Sigma W)^{-1/2}$ and $\tilde{f}, \tilde{g}$ as in (E.1) and (E.2). For $\{(x_i, y_i)\}_{i=1}^n$ sampled independently from $\mathcal{D}$, we observe that the joint distribution of $(x_i - \mu, y_i)$ can also be described by $\mathcal{D}'$ as follows:

(A') $x \sim \mathcal{N}(0, \Sigma)$

(B') $y = \tilde{g}(\eta_1, ..., \eta_k, \xi)$ where $\eta_i = \langle x, \tilde{w}_i \rangle$.

Indeed, we can check that

$$\begin{aligned} y &= g(x^T W, \xi) \\ &= g((x-\mu)^T \tilde{W}(W^T\Sigma W)^{1/2} + \mu^T W, \xi) \\ &= \tilde{g}((x-\mu)^T \tilde{W}, \xi). \end{aligned}$$

Moreover, by construction, we have

$$\begin{aligned} \hat{L}(\theta) &= \frac{1}{n}\sum_{i=1}^n \tilde{f}(\langle w(\theta), x_i - \mu \rangle, y_i, \theta) \\ L(\theta) &= \mathbb{E}_{\mathcal{D}'} \tilde{f}(\langle w(\theta), x_i \rangle, y_i, \theta) \end{aligned}$$

and $\mathcal{D}'$ satisfies assumptions (A) and (B) with $\mu = 0$ and orthonormal $\Sigma^{1/2}\tilde{w}_1^*, ..., \Sigma^{1/2}\tilde{w}_1^*$ and falls into the setting in Lemma 73. We see that f being Lipschitz or square-root Lipschitz is equivalent to $\tilde{f}$ being Lipschitz or square-root Lipschitz. It remains to check assumptions (E.12) and (E.13) and then apply Lemma 77. Observe that

$$\begin{aligned} \Sigma^{-1/2} P \Sigma^{1/2} &= \Sigma^{-1/2}\left(I_d - \Sigma^{1/2}\tilde{W}\tilde{W}^T\Sigma^{1/2}\right)\Sigma^{1/2} \\ &= I_d - \tilde{W}\tilde{W}^T\Sigma = I - W(W^T\Sigma W)^{-1}W^T\Sigma \\ &= Q \end{aligned} \tag{E.16}$$

and so $\Sigma^{1/2}P = Q^T\Sigma^{1/2}$.

To check that (E.12) holds, observe that $\langle \Sigma^{1/2} PH, w \rangle$ has the same distribution as $\langle Qw, x \rangle$. To check that (E.13) holds, we will apply Theorem 57. Note that the joint distribution of $(\langle \phi(w(\theta)), \tilde{x} \rangle, \tilde{y})$ with $(\tilde{x}, \tilde{y}) \sim \tilde{\mathcal{D}}$ is exactly the same as $(\langle w(\theta), x \rangle, y)$ with $(x, y) \sim \mathcal{D}'$ and so

$$\frac{\mathbb{E}_{\tilde{\mathcal{D}}}[\tilde{f}(\langle \phi(w(\theta)), x \rangle, y, \theta)^4]^{1/4}}{\mathbb{E}_{\tilde{\mathcal{D}}}[\tilde{f}(\langle \phi(w(\theta)), x \rangle, y, \theta)]} = \frac{\mathbb{E}_{\mathcal{D}'}[\tilde{f}(\langle w(\theta), x \rangle, y, \theta)^4]^{1/4}}{\mathbb{E}_{\mathcal{D}'}[\tilde{f}(\langle w(\theta), x \rangle, y, \theta)]} = \frac{\mathbb{E}_{\mathcal{D}}[f(\langle w(\theta), x \rangle, y, \theta)^4]^{1/4}}{\mathbb{E}_{\mathcal{D}}[f(\langle w(\theta), x \rangle, y, \theta)]}.$$

Therefore, the assumption (E) is equivalent to the condition in Theorem 57. Note that $\{(x, y) \mapsto \mathbb{1}\{\tilde{f}(\langle \phi(w(\theta)), x \rangle, y, \theta) > t\} : (\theta, t) \in \Theta \times \mathbb{R}\}$ is a subclass of $\{(x, y) \mapsto \mathbb{1}\{f(\langle w, x \rangle + b, y, \theta) > t\} : (w, b, t, \theta) \in \mathbb{R}^{k+1} \times \mathbb{R} \times \mathbb{R} \times \Theta\}$. Therefore, by assumption (F), we can apply Theorem 57 and (E.13) holds. □

E.2 Proof of Theorem 37

Lemma 78. *Consider $f : \mathbb{R} \times \mathcal{Y} \times \Theta \to \mathbb{R}$ and dataset (X, Y) drawn i.i.d. from the data distribution $\mathcal{D}$ according to (A) and (B) with $\mu = 0$ and orthonormal $\Sigma^{1/2} w_1^*, ..., \Sigma^{1/2} w_k^*$. Define the random quantity Ψ over $x \sim \mathcal{N}(0, \Sigma)$ and $(\tilde{x}_i, \tilde{y}_i)$ drawn i.i.d. from the surrogate distribution $\tilde{\mathcal{D}}$ in Lemma 77 by*

$$\Psi := \inf_{\substack{(u,\theta) \in \mathbb{R}^n \times \Theta \\ \|u\|_2 \leq \langle Qw(\theta), x \rangle}} \frac{1}{n} \sum_{i=1}^{n} f(u_i + \langle \phi(w(\theta)), \tilde{x}_i \rangle, \tilde{y}_i, \theta). \tag{E.17}$$

Suppose that we can interchange the supremum and infimum in the following problem:

$$\begin{aligned} &\inf_{\theta \in \Theta, u \in \mathbb{R}^n} \sup_{\lambda \in \mathbb{R}^n} \langle \lambda, Xw(\theta) - u \rangle + \frac{1}{n} \sum_{i=1}^{n} f(u_i, y_i, \theta) \\ = &\sup_{\lambda \in \mathbb{R}^n} \inf_{\theta \in \Theta, u \in \mathbb{R}^n} \langle \lambda, Xw(\theta) - u \rangle + \frac{1}{n} \sum_{i=1}^{n} f(u_i, y_i, \theta) \end{aligned} \tag{E.18}$$

then it holds that for any $t \in \mathbb{R}$, we have

$$\Pr\left(\inf_{\theta\in\Theta} \hat{L}(\theta) > t\right) \leq 2\Pr(\Psi \geq t). \tag{E.19}$$

Proof. By introducing a variable $u = Xw(\theta)$ and our assumption (E.18), we have

$$\begin{aligned}
\inf_{\theta\in\Theta} \hat{L}(\theta) &= \inf_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} f(\langle w(\theta), x_i\rangle, y_i, \theta) \\
&= \inf_{\theta\in\Theta, u\in\mathbb{R}^n} \sup_{\lambda\in\mathbb{R}^n} \langle \lambda, u - Xw(\theta)\rangle + \frac{1}{n}\sum_{i=1}^{n} f(u_i, y_i, \theta) \\
&= \sup_{\lambda\in\mathbb{R}^n} \inf_{\theta\in\Theta, u\in\mathbb{R}^n} \langle \lambda, u - Xw(\theta)\rangle + \frac{1}{n}\sum_{i=1}^{n} f(u_i, y_i, \theta)
\end{aligned}$$

By Lemma 72, conditioned on $Xw_1^* = \eta_1, ..., Xw_k^* = \eta_k$ and ξ, the above is equal in law to

$$\begin{aligned}
&\sup_{\lambda\in\mathbb{R}^n} \inf_{\theta\in\Theta, u\in\mathbb{R}^n} \langle \lambda, u - \left(\sum_{i=1}^{k} \eta_i(\Sigma w_i^*)^T + ZP\Sigma^{1/2}\right) w(\theta)\rangle + \frac{1}{n}\sum_{i=1}^{n} f(u_i, y_i, \theta) \\
=& \sup_{\lambda\in\mathbb{R}^n} \inf_{\theta\in\Theta, u\in\mathbb{R}^n} -\langle \lambda, ZP\Sigma^{1/2} w(\theta)\rangle - \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) \\
=& -\inf_{\lambda\in\mathbb{R}^n} \sup_{\theta\in\Theta, u\in\mathbb{R}^n} \langle \lambda, ZP\Sigma^{1/2} w(\theta)\rangle + \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) \\
=& -\inf_{\lambda\in\mathbb{R}^n} \sup_{\substack{(w,u,\theta)\in\mathbb{R}^d\times\mathbb{R}^n\times\Theta \\ w = P\Sigma^{1/2} w(\theta)}} \langle \lambda, Zw\rangle + \psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) \\
:=& -\Phi
\end{aligned}$$

where ψ is defined as

$$\begin{aligned}
\psi(u, \theta, \lambda \,|\, \eta_1, ..., \eta_k, \xi) =& \langle \lambda, \sum_{i=1}^{k} \langle w(\theta), \Sigma w_i^*\rangle \eta_i - u\rangle \\
&- \frac{1}{n}\sum_{i=1}^{n} f(u_i, g(\eta_{1,i}, ..., \eta_{k,i}, \xi_i), \theta).
\end{aligned} \tag{E.20}$$

Denote the compact set $\mathcal{S}_r = \{(w,u,\theta) \in \mathbb{R}^d \times \mathbb{R}^n \times \Theta : w = P\Sigma^{1/2}w(\theta) \text{ and } \|w\|_2 + \|u\|_2 + \|\theta\|_2 \leq r\}$. Next, we consider the truncated problems:

$$\Phi_r := \inf_{\lambda \in \mathbb{R}^n} \sup_{(w,u,\theta)\in\mathcal{S}_r} \langle \lambda, Zw\rangle + \psi(u,\theta,\lambda \,|\, \eta_1, ..., \eta_k, \xi) \tag{E.21}$$

$$\Phi_{r,s} := \inf_{\|\lambda\|_2 \leq s} \sup_{(w,u,\theta)\in\mathcal{S}_r} \langle \lambda, Zw\rangle + \psi(u,\theta,\lambda \,|\, \eta_1, ..., \eta_k, \xi). \tag{E.22}$$

The corresponding auxiliary problems are

$$\Psi_{r,s} := \inf_{\|\lambda\|_2 \leq s} \sup_{(w,u,\theta)\in\mathcal{S}_r} \|\lambda\|_2\langle H, w\rangle + \|w\|_2\langle G, \lambda\rangle + \psi(u,\theta,\lambda \,|\, \eta_1, ..., \eta_k, \xi) \tag{E.23}$$

$$\Psi_r := \inf_{\lambda \in \mathbb{R}^n} \sup_{(w,u,\theta)\in\mathcal{S}_r} \|\lambda\|_2\langle H, w\rangle + \|w\|_2\langle G, \lambda\rangle + \psi(u,\theta,\lambda \,|\, \eta_1, ..., \eta_k, \xi). \tag{E.24}$$

We can simplify

$$\begin{aligned}
\Psi_r &= \inf_{\lambda \in \mathbb{R}^n} \sup_{(w,u,\theta)\in\mathcal{S}_r} \|\lambda\|_2\langle H, w\rangle + \langle \lambda, G\|w\|_2 + \sum_{i=1}^k \langle w(\theta), \Sigma w_i^*\rangle \eta_i - u\rangle - \frac{1}{n}\sum_{i=1}^n f(u_i, y_i, \theta) \\
&\geq \sup_{(w,u,\theta)\in\mathcal{S}_r} \inf_{\lambda \in \mathbb{R}^n} \|\lambda\|_2\langle H, w\rangle + \langle \lambda, G\|w\|_2 + \sum_{i=1}^k \langle w(\theta), \Sigma w_i^*\rangle \eta_i - u\rangle - \frac{1}{n}\sum_{i=1}^n f(u_i, y_i, \theta) \\
&\geq \sup_{(w,u,\theta)\in\mathcal{S}_r} \inf_{\lambda \geq 0} \lambda \left(\langle H, w\rangle - \left\| G\|w\|_2 + \sum_{i=1}^k \langle w(\theta), \Sigma w_i^*\rangle \eta_i - u \right\|_2 \right) - \frac{1}{n}\sum_{i=1}^n f(u_i, y_i, \theta) \\
&= \sup_{\substack{(w,u,\theta)\in\mathcal{S}_r \\ \langle H, w\rangle \geq \left\| G\|w\|_2 + \sum_{i=1}^k \langle w(\theta), \Sigma w_i^*\rangle \eta_i - u \right\|_2}} -\frac{1}{n}\sum_{i=1}^n f(u_i, g(\eta_{1,i}, ..., \eta_{k,i}, \xi_i), \theta) \\
&= - \inf_{\substack{(w,u,\theta)\in\mathcal{S}_r \\ \langle H, w\rangle \geq \left\| G\|w\|_2 + \sum_{i=1}^k \langle w(\theta), \Sigma w_i^*\rangle \eta_i - u \right\|_2}} \frac{1}{n}\sum_{i=1}^n f(u_i, g(\eta_{1,i}, ..., \eta_{k,i}, \xi_i), \theta)
\end{aligned}$$

Therefore, we can define the limit as $r \to \infty$

$$\begin{aligned}
\Psi := & \inf_{\substack{(u,\theta)\in\mathbb{R}^n\times\Theta \\ \left\|G\|P\Sigma^{1/2}w(\theta)\|_2+\left(\sum_{i=1}^k \eta_i(\Sigma w_i^*)^T\right)w(\theta)-u\right\|_2\le\langle H,P\Sigma^{1/2}w(\theta)\rangle}} \frac{1}{n}\sum_{i=1}^n f(u_i,y_i,\theta) \\
= & \inf_{\substack{(u,\theta)\in\mathbb{R}^n\times\Theta \\ \|u\|_2\le\langle H,P\Sigma^{1/2}w(\theta)\rangle}} \frac{1}{n}\sum_{i=1}^n f(u_i+\sum_{l=1}^k\langle w(\theta),\Sigma w_l^*\rangle\eta_{l,i}+\|P\Sigma^{1/2}w(\theta)\|_2 G_i,y_i,\theta) \\
= & \inf_{\substack{(u,\theta)\in\mathbb{R}^n\times\Theta \\ \|u\|_2\le\langle x,Qw(\theta)\rangle}} \frac{1}{n}\sum_{i=1}^n f(u_i+\langle\phi(w(\theta)),\tilde{x}_i\rangle,y_i,\theta)
\end{aligned}$$

where we write $\tilde{x}_i = (\eta_{1,i},...,\eta_{k,i},G_i)$ and $x = \Sigma^{1/2}H$ because $P\Sigma^{1/2} = \Sigma^{1/2}Q$. Finally, we conclude the proof by showing that

$$\begin{aligned}
\Pr\left(\inf_{\theta\in\Theta}\hat{L}(\theta) > t \,\middle|\, \eta_1,...,\eta_k,\xi\right) &= \Pr(-\Phi > t) = \Pr(\Phi < -t) \\
&\le \lim_{r\to\infty}\Pr(\Phi_r < -t) && \text{since } \Phi \ge \Phi_r \\
&= \lim_{r\to\infty}\lim_{s\to\infty}\Pr(\Phi_{r,s} < -t) && \text{by Lemma } 47 \\
&\le 2\lim_{r\to\infty}\lim_{s\to\infty}\Pr(\Psi_{r,s} \le -t) && \text{by Theorem } 46 \\
&\le 2\lim_{r\to\infty}\Pr(\Psi_r \le -t) && \text{since } \Psi_{r,s} \ge \Psi_r \\
&= 2\Pr(\cap_r\{\Psi_r \le -t\}) \\
&\le 2\Pr(\Psi \ge t) && \text{by Lemma } 47. \qquad \square
\end{aligned}$$

Theorem 37. *Under assumptions (A) and (B), let $f : \mathbb{R}\times\mathcal{Y}\to\mathbb{R}$ be convex with respect to the first argument for any $y\in\mathcal{Y}$. Fix $\mathcal{K}\subset\mathbb{R}^d, \mathcal{B}\subset\mathbb{R}$ to be any bounded convex sets. Suppose that τ is such that with probability at least $1-\delta/2$ over $x\sim\mathcal{N}(0,\Sigma)$ and $(\tilde{x}_i,\tilde{y}_i)_{i=1}^n$ sampled i.i.d. from*

$\tilde{\mathcal{D}}$ defined in (4.16)*, it holds that*

$$\min_{(w,b)\in\mathcal{K}\times\mathcal{B}} \sup_{\lambda\geq 0} \frac{1}{n}\sum_{i=1}^{n} f_\lambda(\langle\phi(w),\tilde{x}_i\rangle + w^T\mu + b, \tilde{y}_i) - \frac{\lambda}{n}\langle Qw, x\rangle^2 \tag{4.18}$$

Then with probability at least $1-\delta$*, it holds that*

$$\min_{(w,b)\in\mathcal{K}\times\mathcal{B}} \hat{L}_f(w,b) \leq \tau. \tag{4.19}$$

Proof. As in the proof of Theorem 41, we can assume without loss of generality that $\mu = 0$ and $\Sigma^{1/2}w_1^*, ..., \Sigma^{1/2}w_k^*$ are orthonormal, and we can apply Lemma 78. Let $\theta = (w,b)$ and $\Theta = \mathcal{K}\times\mathcal{B}$. Define $w(\theta) = w$ and $f' : \mathbb{R}\times\mathcal{Y}\times\Theta$ by $f(\hat{y}, y, \theta) = f(\hat{y}+b, y)$ and we can check

$$\begin{aligned} &\inf_{(w,b,u)\in\mathcal{K}\times\mathcal{B}\times\mathbb{R}^n} \sup_{\lambda\in\mathbb{R}^n} \langle\lambda, Xw-u\rangle + \frac{1}{n}\sum_{i=1}^{n} f(u_i+b, y_i) \\ =\ &\sup_{\lambda\in\mathbb{R}^n} \inf_{(w,b,u)\in\mathcal{K}\times\mathcal{B}\times\mathbb{R}^n} \langle\lambda, Xw-u\rangle + \frac{1}{n}\sum_{i=1}^{n} f(u_i+b, y_i) \end{aligned} \tag{E.25}$$

because $\mathcal{K}\times\mathcal{B}\times\mathbb{R}^n$ is convex and f is convex with respect to the first argument and we can apply the minimax theorem [61]. Then by applying Lemma 78 to f', it suffices to analyze

$$\begin{aligned} \Psi &= \inf_{\substack{(u,w,b)\in\mathbb{R}^n\times\mathcal{K}\times\mathcal{B} \\ \|u\|_2\leq\langle Qw,x\rangle}} \frac{1}{n}\sum_{i=1}^{n} f(u_i + \langle\phi(w),\tilde{x}_i\rangle + b, \tilde{y}_i) \\ &= \inf_{(w,b)\in\mathcal{K}\times\mathcal{B}} \inf_{u\in\mathbb{R}^n} \sup_{\lambda\geq 0} \frac{\lambda}{n}\left(\|u\|_2^2 - \langle Qw,x\rangle^2\right) + \frac{1}{n}\sum_{i=1}^{n} f(u_i + \langle\phi(w),\tilde{x}_i\rangle + b, \tilde{y}_i) \\ &= \inf_{(w,b)\in\mathcal{K}\times\mathcal{B}} \sup_{\lambda\geq 0} \inf_{u\in\mathbb{R}^n} \frac{\lambda}{n}\left(\|u\|_2^2 - \langle Qw,x\rangle^2\right) + \frac{1}{n}\sum_{i=1}^{n} f(u_i + \langle\phi(w),\tilde{x}_i\rangle + b, \tilde{y}_i) \\ &= \inf_{(w,b)\in\mathcal{K}\times\mathcal{B}} \sup_{\lambda\geq 0} -\frac{\lambda}{n}\langle Qw,x\rangle^2 + \frac{1}{n}\sum_{i=1}^{n} \inf_{u_i\in\mathbb{R}} f(u_i + \langle\phi(w),\tilde{x}_i\rangle + b, \tilde{y}_i) + \lambda u_i^2 \\ &= \inf_{(w,b)\in\mathcal{K}\times\mathcal{B}} \sup_{\lambda\geq 0} \frac{1}{n}\sum_{i=1}^{n} f_\lambda(\langle\phi(w),\tilde{x}_i\rangle + b, \tilde{y}_i) - \frac{\lambda}{n}\langle Qw,x\rangle^2 \end{aligned} \tag{E.26}$$

where in the third inequality we use the convexity of f and the minimax theorem [61] again. $\square$

E.3 Proofs for Section 4.3 and 4.4

First, we show norm bounds for the minimal ℓ_2 norm interpolant and the max-margin classifier by applying Theorem 37 to the square loss and the squared hinge loss. Finally, we will use similar arguments to calculate the minmimal norm required to interpolate for phase retrieval and ReLU regression. We will use the following lemmas.

Lemma 79. *Suppose that $a, b > 0$. Then if $a/b > 1$, we have*

$$\max_{\lambda \geq 0} \left[\frac{\lambda}{1+\lambda} a - \lambda b \right] = (\sqrt{a} - \sqrt{b})^2,$$

and if $a/b \leq 1$ then

$$\max_{\lambda \geq 0} \left[\frac{\lambda}{1+\lambda} a - \lambda b \right] = 0.$$

Proof. Observe that the objective can be rewritten as

$$g(\lambda) := a - \frac{1}{1+\lambda} a - \lambda b$$

and the derivative of this expression with respect to λ is

$$g'(\lambda) = \frac{1}{(1+\lambda)^2} a - b.$$

Therefore the unique critical point of g on the domain $(-1, \infty)$ is at $1 + \lambda = \sqrt{a/b}$. This is the global maximum of g on this domain because g goes to $-\infty$ as $\lambda \to -1$ and as $\lambda \to \infty$. At this point, we have that

$$g(\lambda) = a - \sqrt{ab} - (\sqrt{a/b} - 1)b = a + b - 2\sqrt{ab} = (\sqrt{a} - \sqrt{b})^2.$$

If $a/b > 1$ this is the global maximum on $[0, \infty)$. Otherwise, the maximum is at the boundary at $\lambda = 0$. □

Lemma 80. *For any $\hat{y}, \epsilon \in \mathbb{R}$, it holds that if $y \in \mathbb{R}$, then*

$$(y - (\hat{y} + \epsilon))^2 = (y - \hat{y})^2 - 2(y - \hat{y})\epsilon + \epsilon^2, \tag{E.27}$$

and if $y \in \{-1, 1\}$, then

$$(1 - y(\hat{y} + \epsilon))_+^2 \leq (1 - y\hat{y})_+^2 - 2y(1 - y\hat{y})_+\epsilon + \epsilon^2. \tag{E.28}$$

Proof. The first equality is straightforward to check. For the second inequality, we claim that

$$(1 - y(\hat{y} + \epsilon))_+ \leq |(1 - y\hat{y})_+ - \epsilon y|.$$

Indeed, if $1 - y(\hat{y} + \epsilon) \leq 0$, then there is nothing to prove. Otherwise, by monotonicity of $x \to |x|$ and $x \to x_+$, it is clear that $|(1 - y\hat{y})_+ - \epsilon y| \geq (1 - y\hat{y}) - \epsilon y = (1 - y(\hat{y} + \epsilon))_+$. Taking the square of both hand sides concludes the proof. □

Lemma 81. *Consider $Q = I - \sum_{i=1}^k w_i^*(w_i^*)^T\Sigma$ where $\Sigma^{1/2}w_1^*, ..., \Sigma^{1/2}w_k^*$ are orthonormal and we let R be the orthogonal projection matrix onto the image of Q. Then it holds that $\mathrm{rank}(R) = d - k$ and*

$$R\Sigma w_i^* = 0 \quad \textit{for any } \ i = 1, ..., k.$$

Moreover, we have $QR = R$ and $RQ = Q$, and so

$$\frac{1}{\mathrm{Tr}(R\Sigma R)} \leq \left(1 - \frac{k}{n} - \frac{n}{R(Q^T\Sigma Q)}\right)^{-1} \frac{1}{\mathrm{Tr}(Q^T\Sigma Q)}$$
$$\frac{n}{R(R\Sigma R)} \leq \left(1 - \frac{k}{n} - \frac{n}{R(Q^T\Sigma Q)}\right)^{-2} \frac{n}{R(Q^T\Sigma Q)}.$$

Proof. It is obvious that $\mathrm{rank}(R) = \mathrm{rank}(Q)$ and by the rank-nullity theorem, it suffices to show the nullity of Q is k. To this end, we observe that

$$\begin{aligned} Qw = 0 &\iff \Sigma^{-1/2}\left(I - \sum_{i=1}^{k}(\Sigma^{1/2}w_i^*)(\Sigma^{1/2}w_i^*)^T\right)\Sigma^{1/2}w = 0 \\ &\iff \left(I - \sum_{i=1}^{k}(\Sigma^{1/2}w_i^*)(\Sigma^{1/2}w_i^*)^T\right)\Sigma^{1/2}w = 0 \\ &\iff \Sigma^{1/2}w \in \text{span}\{\Sigma^{1/2}w_1^*, ..., \Sigma^{1/2}w_k^*\} \\ &\iff w \in \text{span}\{w_1^*, ..., w_k^*\}. \end{aligned}$$

It is also straightforward to verify that $Q^2 = Q$ and $Q^T\Sigma w_i^* = 0$ for $i = 1, ..., k$. For any $v \in \mathbb{R}^d$, Rv lies in the image of Q and so there exists w such that $Rv = Qw$. Then we can check that

$$\begin{aligned} v^T R\Sigma w_i^* &= \langle Rv, \Sigma w_i^*\rangle \\ &= \langle Qw, \Sigma w_i^*\rangle = \langle w, Q^T\Sigma w_i^*\rangle = 0 \end{aligned}$$

and

$$\begin{aligned} (QR)v &= Q(Rv) \\ &= Q(Qw) = Q^2 w \\ &= Qw = Rv. \end{aligned}$$

Since the choice of v is arbitrary, it must be the case that $R\Sigma w_i^* = 0$ and $QR = R$. For any $v \in \mathbb{R}^d$, we can check

$$(RQ)v = R(Qv) = Qv$$

by the definition of orthogonal projection. Therefore, it must be the case that $RQ = Q$. Finally,

we use $R = QR = RQ^T$ to show that

$$\begin{aligned}
\operatorname{Tr}(R\Sigma R) &= \operatorname{Tr}(RQ^T\Sigma QR) = \operatorname{Tr}(Q^T\Sigma QR) \\
&= \operatorname{Tr}(Q^T\Sigma Q) - \operatorname{Tr}(Q^T\Sigma Q(I-R)) \\
&\geq \operatorname{Tr}(Q^T\Sigma Q) - \sqrt{\operatorname{Tr}((Q^T\Sigma Q)^2)\operatorname{Tr}((I-R)^2)} \\
&= \operatorname{Tr}(Q^T\Sigma Q)\left(1 - \sqrt{\frac{k}{R(Q^T\Sigma Q)}}\right) \\
&= \operatorname{Tr}(Q^T\Sigma Q)\left(1 - \frac{k}{n} - \frac{n}{R(Q^T\Sigma Q)}\right)
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{Tr}((R\Sigma R)^2) &= \operatorname{Tr}(\Sigma R\Sigma R) \\
&= \operatorname{Tr}(\Sigma QRQ^T\Sigma QRQ^T) \\
&= \operatorname{Tr}((RQ^T\Sigma Q)R(Q^T\Sigma QR)) \\
&\leq \operatorname{Tr}((RQ^T\Sigma Q)(Q^T\Sigma QR)) = \operatorname{Tr}((Q^T\Sigma Q)^2R) \\
&\leq \operatorname{Tr}((Q^T\Sigma Q)^2).
\end{aligned}$$

Rearranging concludes the proof. □

We will prove Theorem 38. The proof for theorem 14 is exactly the same.

Theorem 38. *Under assumptions (A) and (B), let $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be the squared hinge loss $f(\hat{y}, y) = (1 - \hat{y}y)_+^2$ with $\mathcal{Y} = \{-1, 1\}$. Let Q be the same as in Theorem 36 and $\Sigma^\perp = Q^T\Sigma Q$. Fix any $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$ such that $Qw^\sharp = 0$ and for some $\rho \in (0, 1)$, it holds that*

$$\hat{L}_f(w^\sharp, b^\sharp) \leq (1+\rho)L_f(w^\sharp, b^\sharp). \tag{4.25}$$

Then with probability at least $1 - \delta$, for some $\epsilon \lesssim \rho + \log\left(\frac{1}{\delta}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma^\perp)}} + \frac{k}{n} + \frac{n}{R(\Sigma^\perp)}\right)$,

it holds that

$$\min_{(w,b)\in\mathbb{R}^{d+1}:\hat{L}_f(w,b)=0} \|w\|_2 \leq \|w^\sharp\|_2 + (1+\epsilon)\sqrt{\frac{nL_f(w^\sharp,b^\sharp)}{\mathrm{Tr}(\Sigma^\perp)}}. \tag{4.26}$$

Proof. It suffices to show that for the desired choice of $B = \|w^\sharp\|_2 + (1+\epsilon)\sqrt{\frac{nL_f(w^\sharp,b^\sharp)}{\mathrm{Tr}(\Sigma^\perp)}}$, we have

$$\min_{\|w\|\leq B} \hat{L}_f(w, b^\sharp) = 0.$$

To this end, we will apply Theorem 37. For the squared hinge loss (as well as the square loss), it holds that $f_\lambda = \frac{\lambda}{\lambda+1} f$ and so we want

$$\min_{\|w\|_2\leq B} \max_{\lambda\geq 0} \frac{\lambda}{\lambda+1}\left[\frac{1}{n}\sum_{i=1}^n f(\langle\phi(w),\tilde{x}_i\rangle + \langle w,\mu\rangle + b^\sharp, \tilde{y}_i)\right] - \frac{\lambda}{n}\langle Qw, x\rangle^2 \leq 0.$$

By Lemma 79, it suffices to find a w such that $\|w\|_2 \leq B$ and

$$\frac{1}{n}\sum_{i=1}^n f(\langle\phi(w),\tilde{x}_i\rangle + \langle w,\mu\rangle + b^\sharp, \tilde{y}_i) \leq \frac{1}{n}\langle Qw, x\rangle^2. \tag{E.29}$$

To this end, we let R to the orthogonal projection matrix onto the image of Q and consider w of the form $w^\sharp + \alpha\frac{Rx}{\|Rx\|_2}$. Note that Lemma 81 can be applied here because we can consider $[\tilde{w}_1^*, ..., \tilde{w}_k^*] = \tilde{W} = W(W^T\Sigma W)^{-1/2}$ and $\Sigma^{1/2}\tilde{w}_1^*, ..., \Sigma^{1/2}\tilde{w}_k^*$ are orthonormal by construction. It is easy to check that $I - \tilde{W}\tilde{W}^T\Sigma = I - W(W^T\Sigma W)^{-1}W^T\Sigma = Q$. Then we check that

$$Qw = Qw^\sharp + \alpha\frac{QRx}{\|Rx\|_2} = \alpha\frac{Rx}{\|Rx\|_2} \quad \text{and} \quad \langle Qw, x\rangle = \alpha\|Rx\|_2.$$

Since $R\Sigma\tilde{W} = 0$, we have

$$\phi(w) = \begin{pmatrix} \tilde{W}^T\Sigma w \\ \|\Sigma^{1/2}Qw\|_2 \end{pmatrix} = \begin{pmatrix} \tilde{W}^T\Sigma w^\sharp \\ \frac{\alpha\|\Sigma^{1/2}Rx\|_2}{\|Rx\|_2} \end{pmatrix}$$

and so

$$\langle \phi(w), \tilde{x}_i \rangle = \langle \phi(w^\sharp), \tilde{x}_i \rangle + \frac{\alpha \|\Sigma^{1/2} R x\|_2}{\|Rx\|_2} \tilde{x}_{i,k+1}.$$

Moreover, we can assume[1] μ to be in the span of ΣW. Then we have $\langle w, \mu \rangle = \langle w^\sharp, \mu \rangle$. Next, we use Lemma 80 to show

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n f(\langle \phi(w), \tilde{x}_i \rangle + \langle w, \mu \rangle + b^\sharp, \tilde{y}_i) \\
\leq &\frac{1}{n} \sum_{i=1}^n f(\langle \phi(w^\sharp), \tilde{x}_i \rangle + \langle w^\sharp, \mu \rangle + b^\sharp, \tilde{y}_i) + \frac{\alpha^2 \|\Sigma^{1/2} R x\|_2^2}{\|Rx\|_2^2} \left(\frac{1}{n} \sum_{i=1}^n \tilde{x}_{i,k+1}^2 \right) \\
&- 2 \frac{\alpha \|\Sigma^{1/2} R x\|_2}{\|Rx\|_2} \left(\frac{1}{n} \sum_{i=1}^n \tilde{y}_i \sqrt{f(\langle \phi(w^\sharp), \tilde{x}_i \rangle + \langle w^\sharp, \mu \rangle + b^\sharp, \tilde{y}_i)} \tilde{x}_{i,k+1} \right).
\end{aligned}$$

Since $Qw^\sharp = 0$ and $\tilde{y}_i$ only depends on $\tilde{x}_{i,1}, ..., \tilde{x}_{i,k}$, we have $\tilde{y}_i \sqrt{f(\langle \phi(w^\sharp), \tilde{x}_i \rangle + \langle w^\sharp, \mu \rangle + b^\sharp, \tilde{y}_i)}$ is independent of $\tilde{x}_{i,k+1}$ and so $\frac{1}{n} \sum_{i=1}^n \tilde{y}_i \sqrt{f(\langle \phi(w^\sharp), \tilde{x}_i \rangle + \langle w^\sharp, \mu \rangle + b^\sharp, \tilde{y}_i)} \tilde{x}_{i,k+1}$ has the same distribution as

$$\sqrt{\frac{1}{n^2} \sum_{i=1}^n f(\langle \phi(w^\sharp), \tilde{x}_i \rangle + \langle w^\sharp, \mu \rangle + b^\sharp, \tilde{y}_i)} \cdot \mathcal{N}(0, 1).$$

Moreover, we have $Rx \sim \mathcal{N}(0, R\Sigma R)$ and $\Sigma^{1/2} Rx \sim \mathcal{N}(0, \Sigma^{1/2} R \Sigma R \Sigma^{1/2})$, and it is easy to check that

$$\mathrm{Tr}(\Sigma^{1/2} R \Sigma R \Sigma^{1/2}) = \mathrm{Tr}((R\Sigma R)^2).$$

By a union bound, the following occur together with probability at least $1 - \delta/2$ for some absolute constant $C > 0$:

1. Using the first part of Lemma 64, we have

$$\|Rx\|_2^2 \geq \mathrm{Tr}(R\Sigma R) \left(1 - C \frac{\log(32/\delta)}{\sqrt{R(R\Sigma R)}} \right)$$

1. If not, we can simply increase k by 1 and let $w_{k+1}^* = \Sigma^{-1} \mu$.

2. Using the last part of Lemma 64, requiring $R(R\Sigma R) \gtrsim \log(32/\delta)^2$

$$\frac{\|\Sigma^{1/2}Rx\|_2^2}{\|Rx\|_2^2} \leq C\log(32/\delta)\frac{\mathrm{Tr}((R\Sigma R)^2)}{\mathrm{Tr}(R\Sigma R)}$$

3. Using subexponential Bernstein's inequality (Theorem 2.8.1 of Vershynin [71]), requiring $n = \Omega(\log(1/\delta))$,

$$\frac{1}{n}\sum_{i=1}^n \tilde{x}_{i,k+1}^2 \leq 2$$

4. Using standard Gaussian tail bound $\Pr(|Z| \geq t) \leq 2e^{-t^2/2}$, we have

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n \tilde{y}_i\sqrt{f(\langle\phi(w^\sharp),\tilde{x}_i\rangle + \langle w^\sharp,\mu\rangle + b^\sharp,\tilde{y}_i)}\tilde{x}_{i,k+1}\right| \\ \leq&\sqrt{\frac{1}{n}\sum_{i=1}^n f(\langle\phi(w^\sharp),\tilde{x}_i\rangle + \langle w^\sharp,\mu\rangle + b^\sharp,\tilde{y}_i)}\sqrt{\frac{2\log(32/\delta)}{n}}\end{aligned}$$

5. By assumption, it holds that

$$\frac{1}{n}\sum_{i=1}^n f(\langle\phi(w^\sharp),\tilde{x}_i\rangle + \langle w^\sharp,\mu\rangle + b^\sharp,\tilde{y}_i) \leq (1+\rho)L_f(w^\sharp,b^\sharp).$$

Therefore, we can use AM-GM inequality to show that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n f(\langle\phi(w),\tilde{x}_i\rangle + b^\sharp,\tilde{y}_i) \leq&\frac{1}{n}\sum_{i=1}^n f(\langle\phi(w^\sharp),\tilde{x}_i\rangle + b^\sharp,\tilde{y}_i) + 2\alpha^2\frac{\|\Sigma^{1/2}Rx\|_2^2}{\|Rx\|_2^2} \\ &+ 2\alpha\frac{\|\Sigma^{1/2}Rx\|_2}{\|Rx\|_2}\sqrt{\frac{1}{n}\sum_{i=1}^n f(\langle\phi(w^\sharp),\tilde{x}_i\rangle + b^\sharp,\tilde{y}_i)}\sqrt{\frac{2\log(32/\delta)}{n}} \\ \leq&\left(1+\sqrt{\frac{2\log(32/\delta)}{n}}\right)(1+\rho)L_f(w^\sharp,b^\sharp) \\ &+ C\log(32/\delta)\left(2+\sqrt{\frac{2\log(32/\delta)}{n}}\right)\frac{\mathrm{Tr}((R\Sigma R)^2)}{\mathrm{Tr}(R\Sigma R)}\alpha^2\end{aligned}$$

and it suffices to pick α such that

$$\left(1+\sqrt{\frac{2\log(32/\delta)}{n}}\right)(1+\rho)L_f(w^\sharp,b^\sharp)+C\log(32/\delta)\left(2+\sqrt{\frac{2\log(32/\delta)}{n}}\right)\frac{\mathrm{Tr}((R\Sigma R)^2)}{\mathrm{Tr}(R\Sigma R)}\alpha^2$$
$$\leq\alpha^2\frac{\mathrm{Tr}(R\Sigma R)}{n}\left(1-C\frac{\log(32/\delta)}{\sqrt{R(R\Sigma R)}}\right).$$

Rearranging, we can set

$$\alpha^2=\frac{\left(1+\sqrt{\frac{2\log(32/\delta)}{n}}\right)(1+\rho)}{1-\frac{C\log(32/\delta)}{\sqrt{R(R\Sigma R)}}-C\log(32/\delta)\left(2+\sqrt{\frac{2\log(32/\delta)}{n}}\right)\frac{n}{R(R\Sigma R)}}\frac{nL_f(w^\sharp,b^\sharp)}{\mathrm{Tr}(R\Sigma R)}.$$

Note that $\|w\|_2\leq\|w^\sharp\|_2+\alpha$. The proof concludes by applying Lemma 81 to replace $R\Sigma R$ with $\Sigma^\perp$. □

Theorem 40. *Under assumptions (A) and (B), let $f:\mathbb{R}\times\mathcal{Y}\to\mathbb{R}$ be the loss defined in* (4.28) *with $\mathcal{Y}=\mathbb{R}_{\geq 0}$. Let Q be the same as in Theorem 36 and $\Sigma^\perp=Q^T\Sigma Q$. Fix any $(w^\sharp,b^\sharp)\in\mathbb{R}^{d+1}$ such that $Qw^\sharp=0$ and for some $\rho\in(0,1)$, it holds that*

$$\hat{L}_f(w^\sharp,b^\sharp)\leq(1+\rho)L_f(w^\sharp,b^\sharp). \tag{4.29}$$

Then with probability at least $1-\delta$, for some $\epsilon\lesssim\rho+\log\left(\frac{1}{\delta}\right)\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{R(\Sigma^\perp)}}+\frac{k}{n}+\frac{n}{R(\Sigma^\perp)}\right)$, it holds that

$$\min_{(w,b)\in\mathbb{R}^{d+1}:\hat{L}_f(w,b)=0}\|w\|_2\leq\|w^\sharp\|_2+(1+\epsilon)\sqrt{\frac{nL_f(w^\sharp,b^\sharp)}{\mathrm{Tr}(\Sigma^\perp)}}. \tag{4.30}$$

Proof. Let $I = \{i \in [n] : y_i > 0\}$ and observe that

$$\begin{aligned}\min_{\substack{w\in\mathbb{R}^d \\ \forall i\in[n],\sigma(\langle w,x_i\rangle+b^\sharp)=y_i}} \|w\|_2 &= \min_{\substack{w\in\mathbb{R}^d \\ \forall i\in I,\langle w,x_i\rangle+b^\sharp=y_i \\ \forall i\notin I,\langle w,x_i\rangle+b^\sharp\le 0}} \|w\|_2 \\ &= \min_{w\in\mathbb{R}^d} \sup_{\substack{\lambda\in\mathbb{R}^n \\ \forall i\notin I,\lambda_i\ge 0}} \sum_{i=1}^n \lambda_i(\langle w,x_i\rangle + b^\sharp - y_i) + \|w\|_2 \\ &= \min_{w\in\mathbb{R}^d} \sup_{\substack{\lambda\in\mathbb{R}^n \\ \forall i\notin I,\lambda_i\ge 0}} \langle \lambda, Z\Sigma^{1/2}w + (b^\sharp + \langle w,\mu\rangle)\vec{1} - y\rangle + \|w\|_2\end{aligned}$$

Note that if we condition on $Z\Sigma^{1/2}w_1^* = \eta_1, ..., Z\Sigma^{1/2}w_k^* = \eta_k$ and ξ, then both y and I are non-random. Without loss of generality, we can assume that $\Sigma^{1/2}w_1^*, ..., \Sigma^{1/2}w_k^*$ are orthonormal because otherwise we can define $\tilde{W} = W(W^T\Sigma W)^{-1/2}$ and conditioning on $Z\Sigma^{1/2}W$ is the same as conditioning on $Z\Sigma^{1/2}\tilde{W}$. Then by Lemma 72, the above quantity has the same distribution as

$$\Phi := \min_{w\in\mathbb{R}^d} \sup_{\substack{\lambda\in\mathbb{R}^n \\ \forall i\notin I,\lambda_i\ge 0}} \langle \lambda, ZP\Sigma^{1/2}w\rangle + \langle \lambda, \sum_{i=1}^k \langle w, \Sigma w_i^*\rangle \eta_i + (b^\sharp + \langle w,\mu\rangle)\vec{1} - y\rangle + \|w\|_2. \quad \text{(E.30)}$$

Observe that the minimax problem is convex in w and linear in λ. Moreover, the set $\{\lambda \in \mathbb{R}^n : \forall i \notin I, \lambda_i \ge 0\}$ is convex. Therefore, by the same truncation argument and CGMT, we can consider the auxiliary problem

$$\begin{aligned}\Psi := \min_{w\in\mathbb{R}^d} \sup_{\substack{\lambda\in\mathbb{R}^n \\ \forall i\notin I,\lambda_i\ge 0}} &\langle \lambda, \sum_{i=1}^k \langle w, \Sigma w_i^*\rangle \eta_i + \|P\Sigma^{1/2}w\|_2 G + (b^\sharp + \langle w,\mu\rangle)\vec{1} - y\rangle \\ &+ \|\lambda\|_2\langle H, P\Sigma^{1/2}w\rangle + \|w\|_2\end{aligned} \quad \text{(E.31)}$$

and it holds that $\Pr(\Phi > t) \leq 2\Pr(\Psi \geq t)$. If we define

$$f(\hat{y}, y) = \begin{cases} (\hat{y} - y)^2 & \text{if} \quad y > 0 \\ \sigma(\hat{y})^2 & \text{if} \quad y = 0 \end{cases} \tag{E.32}$$

then by Lemma 82, we have

$$\begin{aligned}
&\sup_{\substack{\lambda \in \mathbb{R}^n \\ \forall i \notin I, \lambda_i \geq 0}} \langle \lambda, \sum_{i=1}^{k} \langle w, \Sigma w_i^* \rangle \eta_i + \|P\Sigma^{1/2} w\|_2 G + (b^\sharp + \langle w, \mu \rangle)\vec{1} - y \rangle + \|\lambda\|_2 \langle H, P\Sigma^{1/2} w \rangle \\
&= \sup_{\lambda \geq 0} \lambda \left(\langle H, P\Sigma^{1/2} w \rangle + \sup_{\substack{u \in \mathbb{R}^n: \\ \|u\|_2 = 1 \\ \forall i \notin I, u_i \geq 0}} \langle u, \sum_{i=1}^{k} \langle w, \Sigma w_i^* \rangle \eta_i + \|P\Sigma^{1/2} w\|_2 G + (b^\sharp + \langle w, \mu \rangle)\vec{1} - y \rangle \right) \\
&= \sup_{\lambda \geq 0} \lambda \left(\langle H, P\Sigma^{1/2} w \rangle + \sqrt{\sum_{i=1}^{n} f \left(\sum_{l=1}^{k} \langle w, \Sigma w_l^* \rangle \eta_{l,i} + \|P\Sigma^{1/2} w\|_2 G_i + b^\sharp + \langle w, \mu \rangle, y_i \right)} \right).
\end{aligned}$$

Therefore, we have shown

$$\Psi = \min_{\substack{w \in \mathbb{R}^d: \\ \left(\sum_{i=1}^{n} f \left(\sum_{l=1}^{k} \langle w, \Sigma w_l^* \rangle \eta_{l,i} + \|P\Sigma^{1/2} w\|_2 G_i + b^\sharp + \langle w, \mu \rangle, y_i \right) \right)^{1/2} \leq -\langle H, P\Sigma^{1/2} w \rangle}} \|w\|_2$$

Note that Lemma 83 is the analog of Lemma 80 for ReLU regression. The rest of the proof is completely the same as proving equation (E.29) in Theorem 38 and so we omit the details. □

The following two lemmas are used in the proof of Theorem 40.

Lemma 82. *For any $v \in \mathbb{R}^n$ and $I \subseteq [n]$, it holds that*

$$\sup_{\substack{u \in \mathbb{R}^n : \|u\|_2 = 1 \\ \forall i \notin I, u_i \geq 0}} \langle u, v \rangle = \left(\sum_{i \in I \text{ or } i \notin I : v_i > 0} v_i^2 \right)^{1/2}.$$

Proof. Observe that

$$\begin{aligned}
\sup_{\substack{u\in\mathbb{R}^n:\|u\|_2=1\\ \forall i\notin I, u_i\geq 0}} \langle u, v\rangle &= \sup_{\substack{u\in\mathbb{R}^n:\|u\|_2=1\\ \forall i\notin I, u_i\geq 0}} \sum_{i\in I} u_i v_i + \sum_{i\notin I: v_i>0} u_i v_i + \sum_{i\notin I: v_i\leq 0} u_i v_i \\
&\leq \sup_{\substack{u\in\mathbb{R}^n:\|u\|_2=1\\ \forall i\notin I, u_i\geq 0}} \sum_{i\in I \text{ or } i\notin I: v_i>0} u_i v_i \\
&\leq \sup_{\substack{u\in\mathbb{R}^n:\|u\|_2=1\\ \forall i\notin I, u_i\geq 0}} \left(\sum_{i\in I \text{ or } i\notin I: v_i>0} u_i^2\right)^{1/2} \left(\sum_{i\in I \text{ or } i\notin I: v_i>0} v_i^2\right)^{1/2} \\
&\leq \left(\sum_{i\in I \text{ or } i\notin I: v_i>0} v_i^2\right)^{1/2}
\end{aligned}$$

and this upper bound is attainable by setting $u_i = v_i$ for $i \in I$ or $i \notin I : v_i > 0$ and 0 otherwise and then scale it to have unit ℓ_2 norm. □

Lemma 83. *Define*

$$f'(\hat{y}, y) = \begin{cases} \hat{y} - y & \text{if} \quad y > 0 \\ \sigma(\hat{y}) & \text{if} \quad y = 0 \end{cases} \tag{E.33}$$

then for any $\hat{y}, y, \epsilon$

$$f(\hat{y} + \epsilon, y) \leq f(\hat{y}, y) + \epsilon^2 + 2\epsilon f'(\hat{y}, y)$$

Proof. If $y > 0$, then

$$\begin{aligned}
f(\hat{y} + \epsilon, y) &= (\hat{y} + \epsilon - y)^2 = (\hat{y} - y)^2 + \epsilon^2 + 2\epsilon(\hat{y} - y) \\
&= f(\hat{y}, y) + \epsilon^2 + 2\epsilon f'(\hat{y}, y).
\end{aligned}$$

If $y = 0$, first observe that

$$\sigma(\hat{y} + \epsilon) \leq |\sigma(\hat{y}) + \epsilon|.$$

Indeed, if $\hat{y}+\epsilon \leq 0$, then there is nothing to prove. Otherwise, the proof follows by the monotonicty

of $x \to |x|$ and σ. Therefore, we can check

$$\begin{aligned} f(\hat{y}+\epsilon, y) &= \sigma(\hat{y}+\epsilon)^2 \\ &\leq (\sigma(\hat{y})+\epsilon)^2 = \sigma(\hat{y})^2 + \epsilon^2 + 2\epsilon\sigma(\hat{y}) \\ &= f(\hat{y}, y) + \epsilon^2 + 2\epsilon f'(\hat{y}, y). \end{aligned}$$

□

APPENDIX F

PROOFS FOR SECTION 5

F.1 Experimental Details

Linear Regression. Since we are considering quite high-dimensional settings and we need many repeated experiments for different regularization strengths, we generally want to avoid drawing a large test set to estimate the prediction error when it is possible. In the case of square loss, we can always write the population loss as

$$L_f(w) = L_f(\tilde{w}) + \|w - \tilde{w}\|_\Sigma^2$$

where $\tilde{w}$ is the optimal linear predictor satisfying the first order condition:

$$\mathbb{E}[x(x^T\tilde{w} - y)] = 0.$$

In the well-specified case, by the independence between x and ξ, the above becomes

$$\Sigma\tilde{w} = \Sigma w^* \implies \tilde{w} = w^*.$$

Therefore, we have $L_f(\tilde{w}) = \mathbb{E}[(y - \langle w^*, x\rangle)^2] = \sigma^2$. To determine the optimal linear predictor in the mis-specified case, we want to set

$$\begin{aligned}\Sigma\tilde{w} &= \mathbb{E}[xy] \\ &= \mathbb{E}[x(\langle w^*, x\rangle + |x_1| \cdot \cos x_2)] \\ &= \Sigma w^* + \mathbb{E}[x_1 \cdot |x_1|]\mathbb{E}[\cos x_2]e_1 + \mathbb{E}[|x_1|]\mathbb{E}[x_2 \cos x_2]e_2\end{aligned}$$

and so

$$\tilde{w} = w^* + \mathbb{E}[x_1 \cdot |x_1|]\mathbb{E}[\cos x_2]\Sigma^{-1}e_1 + \mathbb{E}[|x_1|]\mathbb{E}[x_2 \cos x_2]\Sigma^{-1}e_2.$$

At the same time, it is routine to check that the optimal error is given by

$$L_f(\tilde{w}) = \mathbb{E}[y^2] - \langle \mathbb{E}[xy], \Sigma^{-1}\mathbb{E}[xy] \rangle.$$

It remains to compute the null risk

$$\begin{aligned}
\mathbb{E}[y^2] &= \mathbb{E}[(\langle w^*, x\rangle + |x_1| \cdot \cos x_2 + x_3\xi)^2] \\
&= \mathbb{E}[(\langle w^*, x\rangle + |x_1| \cdot \cos x_2)^2] + \Sigma_{33}\sigma^2 \\
&= \langle w^*, \Sigma w^*\rangle + \mathbb{E}[x_1^2]\mathbb{E}[\cos^2 x_2] + 2\mathbb{E}[\langle w^*, x\rangle(|x_1| \cdot \cos x_2)] + \Sigma_{33}\sigma^2 \\
&= \langle w^*, \Sigma w^*\rangle + \mathbb{E}[x_1^2]\mathbb{E}[\cos^2 x_2] + \Sigma_{33}\sigma^2 \\
&\quad + 2\left(\mathbb{E}[x_1 \cdot |x_1|]\mathbb{E}[\cos x_2]w_1^* + \mathbb{E}[|x_1|]\mathbb{E}[x_2 \cos x_2]w_2^*\right)
\end{aligned}$$

and

$$\begin{aligned}
\langle \mathbb{E}[xy], \Sigma^{-1}\mathbb{E}[xy]\rangle &= \langle \Sigma w^* + \mathbb{E}[|x_1|\cos(x_2)x], w^* + \Sigma^{-1}\mathbb{E}[|x_1|\cos(x_2)x]\rangle \\
&= \langle \Sigma w^*, w^*\rangle + 2\langle w^*, \mathbb{E}[|x_1|\cos(x_2)x]\rangle \\
&\quad + \langle \mathbb{E}[|x_1|\cos(x_2)x], \Sigma^{-1}\mathbb{E}[|x_1|\cos(x_2)x]\rangle.
\end{aligned}$$

Therefore, we have

$$L_f(\tilde{w}) = \mathbb{E}[x_1^2]\mathbb{E}[\cos^2 x_2] + \Sigma_{33}\sigma^2 - \mathbb{E}[x_1 \cdot |x_1|]^2\mathbb{E}[\cos x_2]^2\Sigma_{11}^{-1} - \mathbb{E}[|x_1|]^2\mathbb{E}[x_2\cos(x_2)]^2\Sigma_{22}^{-1}$$

It remains to compute quantities like $\mathbb{E}[|x|], \mathbb{E}[x \cdot |x|], \mathbb{E}[\cos x], \mathbb{E}[x \cos x]$ for each of the eight feature distributions. Since they are one dimensional quantities, we can afford to draw a very large number of samples to estimate them.

Linear Classification. When the feature distribution is Gaussian, we can estimate

$$L_f(w, b) = \mathbb{E}\left[\max(0, 1 - y(\langle w, x\rangle + b))^2\right]$$

without drawing a new high-dimensional dataset from $\mathcal{D}$. First, we can write $x = \Sigma^{1/2} z$. Note that conditioning on η is the same as conditioning on $\langle w^*, x\rangle = \langle \Sigma^{1/2} w^*, z\rangle \sim \mathcal{N}(0, \|w^*\|_\Sigma^2)$ and the conditional distribution of z is

$$\frac{\eta - b^*}{\|w^*\|_\Sigma^2} \Sigma^{1/2} w^* + Pz$$

where $P = I - \frac{(\Sigma^{1/2} w^*)(\Sigma^{1/2} w^*)^T}{\|w^*\|_\Sigma^2}$ and so the conditional distribution of $\langle w, x\rangle + b$ is

$$\begin{aligned}
&\left\langle w, \Sigma^{1/2}\left(\frac{\eta - b^*}{\|w^*\|_\Sigma^2}\Sigma^{1/2} w^* + Pz\right)\right\rangle + b \\
=& b + \frac{\langle w, \Sigma w^*\rangle}{\|w^*\|_\Sigma^2}(\eta - b^*) + \langle P\Sigma^{1/2} w, z\rangle \sim \mathcal{N}\left(\mu(\eta),\, \sigma^2\right)
\end{aligned}$$

where $\mu(\eta) = b + \frac{\langle w, \Sigma w^*\rangle}{\|w^*\|_\Sigma^2}(\eta - b^*)$ and

$$\sigma^2 = w^T(\Sigma^{1/2} P \Sigma^{1/2}) w = w^T \Sigma w - \frac{\langle w, \Sigma w^*\rangle^2}{\|w^*\|_\Sigma^2}.$$

Since x is independent of y conditioned on η, we have that

$$\begin{aligned}
L(w, b) &= \mathbb{E}\left[\mathbb{E}\left[\max(0, 1 - y(\langle w, x\rangle + b))^2 \mid \eta\right]\right] \\
&= \mathbb{E}\left[g(\eta) \cdot \max(0, 1 - \mu(\eta) - \sigma z)^2 + (1 - g(\eta)) \cdot \max(0, 1 + \mu(\eta) + \sigma z)^2\right]
\end{aligned}$$

We can then estimate the population error by drawing samples from a two-dimensional distribution. In addition, the linear predictor that minimizes the population squared hinge loss generally does not have a simple closed-form expression, but we can run SGD on the population objective

in order to find the optimal linear predictor $\tilde{w}, \tilde{b}$. For simplicity, we choose

$$w^* = (5, 0, ..., 0) \quad \text{and} \quad b^* = 3.$$

In this case, we can simplify the optimization problem to an one-dimensional problem by observing that $\tilde{w}_i = 0$ for $i \neq 1$. Indeed, we can check the first order condition holds

$$\begin{aligned}\frac{\partial}{\partial w_i} L_f(\tilde{w}, \tilde{b}) &= -2\mathbb{E}\left[y \max(0, 1 - y(\langle \tilde{w}, x\rangle + \tilde{b}))x_i\right] \\ &= -2\mathbb{E}\left[y \max(0, 1 - y(\tilde{w}_1 x_1 + \tilde{b}))\right] \mathbb{E}\left[x_i\right] = 0\end{aligned}$$

because y is independent of x_i with $i \neq 1$. Therefore, we can just generate $\{x_{i,1}, y_i\}$ from $\mathcal{D}$ and perform one-pass SGD (theorem 6.1 of Bubeck [16]) to find $\tilde{w}_1, \tilde{b}$. In the experiments, we find choosing the initial step size to be 0.1 works well.

F.2 Proofs for Section 5.2

By assumption (A), we can write $x_{i|d-k} = h(x_{i|k}) \cdot \Sigma_{|d-k}^{1/2} z_i$ where $z_i \sim \mathcal{N}(0, I_{d-k})$. We will denote the matrix $Z = [z_1, ..., z_n]^T \in \mathbb{R}^{n\times(d-k)}$. Following the notation in section 5.2, we will also write $X = [X_{|k}, X_{|d-k}]$ where $X_{|k} \in \mathbb{R}^{n\times k}$ and $X_{|d-k} \in \mathbb{R}^{n\times(d-k)}$. The proofs in this section closely follows the proof of Theorem 41.

Theorem 43. *Consider dataset (X, Y) drawn i.i.d. from the data distribution $\mathcal{D}$ according to (A) and (B), and fix any $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ such that $\sqrt{f}$ is 1-Lipschitz for any $y \in \mathcal{Y}$. Fix any $\delta > 0$ and suppose there exists $\epsilon_\delta < 1$ and $C_\delta : \mathbb{R}^{d-k} \to [0, \infty]$ such that*

(i) with probability at least $1 - \delta/2$ over (X, Y) and $G \sim \mathcal{N}(0, I_n)$, it holds uniformly over all $w_{|k} \in \mathbb{R}^k$ and $\|w_{|d-k}\|_{\Sigma_{|d-k}} \in \mathbb{R}_{\geq 0}$ that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x_{i|k})^2} f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})\|w_{|d-k}\|_{\Sigma_{|d-k}} G_i, y_i) \geq (1-\epsilon_\delta)\mathbb{E}_{\mathcal{D}}\left[\frac{1}{h(x_{|k})^2} f(\langle w, x\rangle, y)\right]$$

(ii) with probability at least $1-\delta/2$ *over* $z_{|d-k} \sim \mathcal{N}(0, \Sigma_{|d-k})$*, it holds uniformly over all* $w_{|d-k} \in \mathbb{R}^{d-k}$ *that*

$$\langle w_{|d-k}, z_{|d-k}\rangle \leq C_\delta(w_{|d-k}) \tag{5.8}$$

then with probability at least $1-\delta$*, it holds uniformly over all* $w \in \mathbb{R}^d$ *that*

$$(1-\epsilon_\delta)\mathbb{E}\left[\frac{1}{h(x_{|k})^2} f(\langle w, x\rangle, y)\right] \leq \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{h(x_{i|k})^2} f(\langle w, x_i\rangle, y_i) + \frac{C_\delta(w_{|d-k})}{\sqrt{n}}\right)^2. \tag{5.9}$$

Proof. Note that

$$\langle w_{|d-k}, x_{i|d-k}\rangle = h(x_{i|k}) \cdot \langle w_{|d-k}, \Sigma_{|d-k}^{1/2} z_i\rangle$$

and so for any $f : \mathbb{R} \times \mathcal{Y} \times \mathbb{R}^k \to \mathbb{R}$, we can write

$$\begin{aligned}
\Phi &:= \sup_{w\in\mathbb{R}^d} F(w) - \frac{1}{n}\sum_{i=1}^n f(\langle w, x_i\rangle, y_i, x_{i|k}) \\
&= \sup_{\substack{w\in\mathbb{R}^d, u\in\mathbb{R}^n \\ u = Z\Sigma_{|d-k}^{1/2} w_{|d-k}}} F(w) - \frac{1}{n}\sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})u_i, y_i, x_{i|k}) \\
&= \sup_{w\in\mathbb{R}^d, u\in\mathbb{R}^n} \inf_{\lambda\in\mathbb{R}^n} \langle \lambda, Z\Sigma_{|d-k}^{1/2} w_{|d-k} - u\rangle + F(w) - \frac{1}{n}\sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})u_i, y_i, x_{i|k}).
\end{aligned}$$

By the same truncation argument used in Lemma 74, it suffices to consider the auxiliary problem:

$$\begin{aligned}
\Psi := \sup_{w\in\mathbb{R}^d, u\in\mathbb{R}^n} \inf_{\lambda\in\mathbb{R}^n} &\|\lambda\|_2 \langle H, \Sigma_{|d-k}^{1/2} w_{|d-k}\rangle + \langle G\|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 - u, \lambda\rangle \\
&+ F(w) - \frac{1}{n}\sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})u_i, y_i, x_{i|k}) \\
= \sup_{w\in\mathbb{R}^d, u\in\mathbb{R}^n} \inf_{\lambda\geq 0} &\lambda\left(\langle H, \Sigma_{|d-k}^{1/2} w_{|d-k}\rangle - \left\|G\|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 - u\right\|_2\right) \\
&+ F(w) - \frac{1}{n}\sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})u_i, y_i, x_{i|k})
\end{aligned}$$

Therefore, it holds that

$$\Psi = \sup_{\substack{w\in\mathbb{R}^d, u\in\mathbb{R}^n \\ \langle H,\Sigma_{|d-k}^{1/2} w_{|d-k}\rangle \geq \left\| G\|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 - u \right\|_2}} F(w) - \frac{1}{n}\sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})u_i, y_i, x_{i|k})$$

$$= \sup_{w\in\mathbb{R}^d} F(w) - \frac{1}{n} \inf_{\substack{u\in\mathbb{R}^n \\ \langle H,\Sigma_{|d-k}^{1/2} w_{|d-k}\rangle \geq \left\| G\|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 - u \right\|_2}} \sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})u_i, y_i, x_{i|k}).$$

Next, we analyze the infimum term:

$$\inf_{\substack{u\in\mathbb{R}^n \\ \langle H,\Sigma_{|d-k}^{1/2} w_{|d-k}\rangle \geq \left\| G\|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 - u \right\|_2}} \sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})u_i, y_i, x_{i|k})$$

$$= \inf_{\substack{u\in\mathbb{R}^n \\ \|u\|_2 \leq \langle H,\Sigma_{|d-k}^{1/2} w_{|d-k}\rangle}} \sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})\left(u_i + \|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 G_i\right), y_i, x_{i|k})$$

$$= \inf_{u\in\mathbb{R}^n} \sup_{\lambda\geq 0} \lambda(\|u\|^2 - \langle H, \Sigma_{|d-k}^{1/2} w_{|d-k}\rangle^2) + \sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})\left(u_i + \|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 G_i\right), y_i, x_{i|k})$$

$$\geq \sup_{\lambda\geq 0} \inf_{u\in\mathbb{R}^n} \lambda(\|u\|^2 - \langle H, \Sigma_{|d-k}^{1/2} w_{|d-k}\rangle^2) + \sum_{i=1}^n f(\langle w_{|k}, x_{i|k}\rangle + h(x_{i|k})\left(u_i + \|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 G_i\right), y_i, x_{i|k})$$

$$= \sup_{\lambda\geq 0} -\lambda\langle H, \Sigma_{|d-k}^{1/2} w_{|d-k}\rangle^2 + \sum_{i=1}^n \inf_{u_i\in\mathbb{R}} f(\langle w_{|k}, x_{i|k}\rangle + u_i + \|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 h(x_{i|k}) G_i, y_i, x_{i|k}) + \frac{\lambda}{h(x_{i|k})^2} u_i^2.$$

Now suppose that f takes the form $f(\hat{y}, y, x_{|k}) = \frac{1}{h(x_{|k})^2}\tilde{f}(\hat{y}, y)$ for some 1 square-root Lipschitz $\tilde{f}$ and by a union bound, it holds with probability at least $1-\delta$ that

$$\langle \Sigma_{|d-k}^{1/2} H, w_{|d-k} \rangle^2 \leq C_\delta (w_{|d-k})^2$$

$$\frac{1}{n} \sum_{i=1}^n \frac{1}{h(x_{i|k})^2} \tilde{f}(\langle w_{|k}, x_{i|k} \rangle + \|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 h(x_{i|k}) G_i, y_i) \geq (1-\epsilon_\delta) \mathbb{E}\left[\frac{1}{h(x_{|k})^2} \tilde{f}(\langle w, x \rangle, y) \right],$$

then the above becomes

$$\begin{aligned}
&\sup_{\lambda \geq 0} -\lambda \langle \Sigma_{|d-k}^{1/2} H, w_{|d-k} \rangle^2 + \sum_{i=1}^n \frac{1}{h(x_{i|k})^2} \tilde{f}_\lambda(\langle w_{|k}, x_{i|k} \rangle + \|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 h(x_{i|k}) G_i, y_i) \\
\geq &\sup_{\lambda \geq 0} -\lambda \langle \Sigma_{|d-k}^{1/2} H, w_{|d-k} \rangle^2 + \frac{\lambda}{\lambda+1} \sum_{i=1}^n \frac{1}{h(x_{i|k})^2} \tilde{f}(\langle w_{|k}, x_{i|k} \rangle + \|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 h(x_{i|k}) G_i, y_i) \\
\geq &\sup_{\lambda \geq 0} -\lambda C_\delta (w_{|d-k})^2 + \frac{\lambda}{\lambda+1} (1-\epsilon) n \mathbb{E}\left[\frac{1}{h(x_{|k})^2} \tilde{f}(\langle w, x \rangle, y) \right] \\
\geq &n \left(\sqrt{(1-\epsilon_\delta) \mathbb{E}\left[\frac{1}{h(x_{|k})^2} \tilde{f}(\langle w, x \rangle, y) \right]} - \frac{C_\delta(w_{|d-k})}{\sqrt{n}} \right)_+^2
\end{aligned}$$

where we apply Lemma 75 in the last step. Then if we take

$$F(w) = \left(\sqrt{(1-\epsilon_\delta) \mathbb{E}\left[\frac{1}{h(x_{|k})^2} \tilde{f}(\langle w, x \rangle, y) \right]} - \frac{C_\delta(w_{|d-k})}{\sqrt{n}} \right)_+^2$$

then we have $\Psi \leq 0$. To summarize, we have shown

$$\left(\sqrt{(1-\epsilon_\delta) \mathbb{E}\left[\frac{1}{h(x_{|k})^2} \tilde{f}(\langle w, x \rangle, y) \right]} - \frac{C_\delta(w_{|d-k})}{\sqrt{n}} \right)_+^2 - \frac{1}{n} \sum_{i=1}^n \frac{1}{h(x_{i|k})^2} \tilde{f}(\langle w, x_i \rangle, y_i) \leq 0$$

which implies

$$\mathbb{E}\left[\frac{1}{h(x_{|k})^2} \tilde{f}(\langle w, x \rangle, y) \right] \leq (1-\epsilon_\delta)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{h(x_{i|k})^2} \tilde{f}(\langle w, x_i \rangle, y_i) + \frac{C_\delta(w_{|d-k})}{\sqrt{n}} \right)^2$$

and we are done. □

Theorem 44. *Under assumptions (A) and (B), fix any $w^*_{|k} \in \mathbb{R}^k$ and suppose for some $\rho \in (0,1)$, it holds with probability at least $1-\delta/8$*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \langle w^*_{|k}, x_{i|k}\rangle}{h(x_{i|k})}\right)^2 \le (1+\rho)\cdot \mathbb{E}\left[\left(\frac{y - \langle w^*_{|k}, x_{|k}\rangle}{h(x_{|k})}\right)^2\right]. \tag{5.10}$$

Then with probability at least $1-\delta$, for some $\epsilon \lesssim \rho + \log\left(\frac{1}{\delta}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma_{|d-k})}} + \frac{n}{R(\Sigma_{|d-k})}\right)$, it holds that

$$\min_{w\in\mathbb{R}^d:\forall i,\langle w,x_i\rangle = y_i} \|w\|_2^2 \le \|w^*_{|k}\|_2^2 + (1+\epsilon)\frac{n\mathbb{E}\left[\left(\frac{y-\langle w^*_{|k},x_{|k}\rangle}{h(x_{|k})}\right)^2\right]}{\mathrm{Tr}(\Sigma_{|d-k})} \tag{5.11}$$

Proof. Fix any $w^*_{|k} \in \mathbb{R}^k$, we observe that

$$\begin{aligned}\min_{w\in\mathbb{R}^d:\forall i,\langle w,x_i\rangle = y_i} \|w\|_2^2 &= \min_{w\in\mathbb{R}^d:\forall i,\langle w_{|k},x_{i|k}\rangle + \langle w_{|d-k},x_{i|d-k}\rangle = y_i} \|w_{|k}\|_2^2 + \|w_{|d-k}\|_2^2 \\ &\le \|w^*_{|k}\|_2^2 + \min_{\substack{w_{|d-k}\in\mathbb{R}^{d-k}:\\ \forall i,\langle w_{|d-k},x_{i|d-k}\rangle = y_i - \langle w^*_{|k},x_{i|k}\rangle}} \|w_{|d-k}\|_2^2.\end{aligned}$$

Therefore, it is enough analyze

$$\begin{aligned}\Phi := &\min_{\substack{w_{|d-k}\in\mathbb{R}^{d-k}:\\ \forall i,\langle w_{|d-k},x_{i|d-k}\rangle = y_i - \langle w^*_{|k},x_{i|k}\rangle}} \|w_{|d-k}\|_2 \\ = &\min_{\substack{w_{|d-k}\in\mathbb{R}^{d-k}:\\ \forall i,\langle w_{|d-k},\Sigma^{1/2}_{|d-k}z_i\rangle = \frac{y_i - \langle w^*_{|k},x_{i|k}\rangle}{h(x_{i|k})}}} \|w_{|d-k}\|_2.\end{aligned}$$

By introducing the Lagrangian, we have

$$\begin{aligned}\Phi &= \min_{w_{|d-k}\in\mathbb{R}^{d-k}} \max_{\lambda\in\mathbb{R}^n} \sum_{i=1}^n \lambda_i \left(\langle \Sigma_{|d-k}^{1/2} w_{|d-k}, z_i\rangle - \frac{y_i - \langle w_{|k}^*, x_{i|k}\rangle}{h(x_{i|k})} \right) + \|w_{|d-k}\|_2 \\ &= \min_{w_{|d-k}\in\mathbb{R}^{d-k}} \max_{\lambda\in\mathbb{R}^n} \langle \lambda, Z\Sigma_{|d-k}^{1/2} w_{|d-k}\rangle - \sum_{i=1}^n \lambda_i \left(\frac{y_i - \langle w_{|k}^*, x_{i|k}\rangle}{h(x_{i|k})} \right) + \|w_{|d-k}\|_2.\end{aligned}$$

Similarly, the above is only random in Z after conditioning on $X_{|k} w_{|k}^*$ and ξ and the distribution of Z remains unchanged after conditioning because of the independence. By the same truncation argument as before and CGMT, it suffices to consider the auxiliary problem:

$$\begin{aligned}&\min_{w_{|d-k}\in\mathbb{R}^{d-k}} \max_{\lambda\in\mathbb{R}^n} \|\lambda\|_2 \langle H, \Sigma_{|d-k}^{1/2} w_{|d-k}\rangle + \sum_{i=1}^n \lambda_i \left(\|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 G_i - \frac{y_i - \langle w_{|k}^*, x_{i|k}\rangle}{h(x_{i|k})} \right) \\ &\qquad + \|w_{|d-k}\|_2 \\ = &\min_{w_{|d-k}\in\mathbb{R}^{d-k}} \max_{\lambda\in\mathbb{R}^n} \|\lambda\|_2 \left(\langle H, \Sigma_{|d-k}^{1/2} w_{|d-k}\rangle + \sqrt{\sum_{i=1}^n \left(\|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 G_i - \frac{y_i - \langle w_{|k}^*, x_{i|k}\rangle}{h(x_{i|k})} \right)^2} \right) \\ &\qquad + \|w_{|d-k}\|_2\end{aligned}$$

and so we can define

$$\Psi := \min_{\substack{w_{|d-k}\in\mathbb{R}^{d-k}: \\ \sqrt{\sum_{i=1}^n \left(\|\Sigma_{|d-k}^{1/2} w_{|d-k}\|_2 G_i - \frac{y_i - \langle w_{|k}^*, x_{i|k}\rangle}{h(x_{i|k})} \right)^2} \le \langle -\Sigma_{|d-k}^{1/2} H, w_{|d-k}\rangle}} \|w_{|d-k}\|_2.$$

To upper bound Ψ, we consider $w_{|d-k}$ of the form $-\alpha \frac{\Sigma_{|d-k}^{1/2} H}{\|\Sigma_{|d-k}^{1/2} H\|_2}$, then we just need

$$\sum_{i=1}^n \left(\alpha \frac{\|\Sigma_{|d-k} H\|_2}{\|\Sigma_{|d-k}^{1/2} H\|_2} G_i - \frac{y_i - \langle w_{|k}^*, x_{i|k}\rangle}{h(x_{i|k})} \right)^2 \le \alpha^2 \|\Sigma_{|d-k}^{1/2} H\|_2^2.$$

By a union bound, the following occur together with probability at least $1 - \delta/2$ for some absolute constant $C > 0$:

1. Using the first part of Lemma 64, we have

$$\|\Sigma_{|d-k}^{1/2} H\|_2^2 \geq \mathrm{Tr}(\Sigma_{|d-k}) \left(1 - C \frac{\log(32/\delta)}{\sqrt{R(\Sigma_{|d-k})}}\right)$$

2. Using the last part of Lemma 64, requiring $R(\Sigma_{|d-k}) \gtrsim \log(32/\delta)^2$

$$\frac{\|\Sigma_{|d-k} H\|_2^2}{\|\Sigma_{|d-k}^{1/2} H\|_2^2} \leq C \log(32/\delta) \frac{\mathrm{Tr}(\Sigma_{|d-k}^2)}{\mathrm{Tr}(\Sigma_{|d-k})}$$

3. Using subexponential Bernstein's inequality (Theorem 2.8.1 of Vershynin [71]), requiring $n = \Omega(\log(1/\delta))$,

$$\frac{1}{n} \sum_{i=1}^n G_i^2 \leq 2$$

4. Using standard Gaussian tail bound $\Pr(|Z| \geq t) \leq 2e^{-t^2/2}$, we have

$$\left| \frac{1}{n} \sum_{i=1}^n \frac{G_i (y_i - \langle w_{|k}^*, x_{i|k} \rangle)}{h(x_{i|k})} \right| \leq \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \langle w_{|k}^*, x_{i|k} \rangle}{h(x_{i|k})} \right)^2} \sqrt{\frac{2 \log(32/\delta)}{n}}$$

5. By assumption, it holds that

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \langle w_{|k}^*, x_{i|k} \rangle}{h(x_{i|k})} \right)^2 \leq (1 + \rho) \cdot \mathbb{E} \left[\left(\frac{y - \langle w_{|k}^*, x_{|k} \rangle}{h(x_{|k})} \right)^2 \right].$$

Then we use the above and the AM-GM inequality to show that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\alpha\frac{\|\Sigma_{|d-k}H\|_2}{\|\Sigma_{|d-k}^{1/2}H\|_2}G_i-\frac{y_i-\langle w_{|k}^*,x_{i|k}\rangle}{h(x_{i|k})}\right)^2\\
&\leq 2\alpha^2\frac{\|\Sigma_{|d-k}H\|_2^2}{\|\Sigma_{|d-k}^{1/2}H\|_2^2}+(1+\rho)\cdot\mathbb{E}\left[\left(\frac{y-\langle w_{|k}^*,x_{|k}\rangle}{h(x_{|k})}\right)^2\right]\\
&\quad+2\frac{\alpha\|\Sigma_{|d-k}H\|_2}{\|\Sigma_{|d-k}^{1/2}H\|_2}\sqrt{(1+\rho)\cdot\mathbb{E}\left[\left(\frac{y-\langle w_{|k}^*,x_{|k}\rangle}{h(x_{|k})}\right)^2\right]}\sqrt{\frac{2\log(32/\delta)}{n}}\\
&\leq C\log(32/\delta)\left(2+\sqrt{\frac{2\log(32/\delta)}{n}}\right)\alpha^2\frac{\mathrm{Tr}(\Sigma_{|d-k}^2)}{\mathrm{Tr}(\Sigma_{|d-k})}\\
&\quad+\left(1+\sqrt{\frac{2\log(32/\delta)}{n}}\right)(1+\rho)\cdot\mathbb{E}\left[\left(\frac{y-\langle w_{|k}^*,x_{|k}\rangle}{h(x_{|k})}\right)^2\right].
\end{aligned}
$$

After some rearrangements, it is easy to see that we can choose

$$
\alpha^2=\frac{\left(1+\sqrt{\frac{2\log(32/\delta)}{n}}\right)(1+\rho)}{1-C\frac{\log(32/\delta)}{\sqrt{R(\Sigma_{|d-k})}}-C\log(32/\delta)\left(2+\sqrt{\frac{2\log(32/\delta)}{n}}\right)\frac{n}{R(\Sigma_{|d-k})}}\frac{n\mathbb{E}\left[\left(\frac{y-\langle w_{|k}^*,x_{|k}\rangle}{h(x_{|k})}\right)^2\right]}{\mathrm{Tr}(\Sigma_{|d-k})}.
$$

and the proof is complete. □