On The Performance Of The Maximum Likelihood Over Large Models

by

Gil Kur

B.Sc, Technion – Israel Institute of Technology (2015)
M.Sc, Weizmann Institute of Science (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

© 2023 Gil Kur. All rights reserved.
The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Gil Kur
Department of Electrical Engineering and Computer Science
August 31, 2023

Certified by: Alexander Rakhlin
Professor of Brain and Cognitive Sciences
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

On The Performance Of The Maximum Likelihood Over Large Models

by

Gil Kur

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This dissertation investigates non-parametric regression over large function classes, specifically, non-Donsker classes. We will present the concept of non-Donsker classes and study the statistical performance of Least Squares Estimator (LSE) — which also serves as the Maximum Likelihood Estimator (MLE) under Gaussian noise — over these classes. (1) We demonstrate the minimax sub-optimality of the LSE in the non-Donsker regime, extending traditional findings of Birgé and Massart 93' and resolving a longstanding conjecture of Gardner, Markus and Milanfar 06'. (2) We reveal that in the non-Donsker regime, the sub-optimality of LSE arises solely from its elevated bias error term (in terms of the bias and variance decomposition). (3) We introduce the first minimax optimal algorithm for multivariate convex regression with a polynomial runtime in the number of samples – showing that one can overcome the sub-optimality of the LSE in efficient runtime. (4) We study the minimal error of the LSE both in random and fixed design settings.

Thesis Supervisor: Alexander Rakhlin
Title: Professor of Brain and Cognitive Sciences

Acknowledgments

I'd like to begin by extending my profound appreciation to my advisor, Prof. Alexander (Sasha) Rakhlin. His steadfast guidance and encouragement have been pivotal throughout my PhD journey, affording me the liberty to explore my own research interests and define my life's trajectory. His teachings have profoundly shaped my approach to research, my capacity for conveying my thoughts in writing, and my understanding of the importance of transparency and professionalism. Sasha's considerable influence has undoubtedly honed my thinking, amplified my writing skills, and cultivated a strong sense of professionalism within me.

In addition, I'm tremendously grateful to my second (informal) advisor, mentor and true friend, Prof. Adityanand Guntobiyna. His constant emotional and professional support, along with his unfaltering patience during the more formidable stages of my PhD, have been instrumental.

I wish to express my deepest gratitude to Prof. Tomaso Poggio. His affirmation, philosophical guidance, and financial support during my second research year have been invaluable. His mentorship throughout my tenure at MIT has clearly demonstrated the essence of scientific reasoning, emphasizing the importance of perceiving learning theory through the lens of science, rather than viewing it merely as a field for proving algorithmic bounds. I am deeply grateful for the generous financial aid I obtained from the Eli and Dorothy Berman fellowship in the final stretch of my PhD journey. I also want to extend my appreciation for the assistance received through the Collaboration on the Theoretical Foundations of Deep Learning grant. My earnest recognition goes to Professors Peter Bartlett and Misha Belkin, whose active involvement brought a significant shift in my outlook on learning theory research. Next, my thanks go to Prof. Shiri Artstien-Avidan, who facilitated the introduction to my friend and collaborator, Eli Putterman.

Moreover, I extend my appreciation to my academic advisor, Professor Piotr Idnyk, and my graduate student officer, Professor Leslie Kolodziejski. Their invaluable time, support, and guidance have been pivotal throughout my Ph.D. journey. I also want to acknowledge my Master's advisor and friend, Professor Boaz Nadler, who encouraged me to undertake my Ph.D. abroad and supported me throughout both my Master's and Ph.D. journeys. I would next like to thank my mentor and friend, Professor Emanuel Milman, for his professional guidance and personal assistance from

undergraduate studies all the way through my doctoral work. Additionally, I extend my gratitude to the entire AGA group, including Professors Bo'az Klartag, Gideon Schechtman, Ronen Eldan, and Shiri Artstein for their valuable help over the years.

On a personal level, it is crucial for me to express my profound gratitude to my mother, Nomy Lavan-Kur, for her unwavering support and encouragement despite the numerous hurdles I faced in deciding to pursue a PhD at a highly esteemed university abroad. I deeply cherish her support. Additionally, I want to acknowledge Haim Strauss, who welcomed me as if I were his own son and granted me the freedom to pursue my dreams and undertake my PhD abroad. His support has been invaluable. Although my father, Yosi Kur, is no longer with us, I would like to honor him for believing in my skills and talent when no one else did, and for instilling in me valuable values and a unique life perspective during the early stages of my life.

I also wish to thank Yuval Dagan, a valued friend and colleague, for his consistent emotional support, friendship, and cooperation during my time at MIT. Similarly, my friend Tamir Hemo deserves special mention for his unwavering emotional support and companionship. Moreover, I want to recognize my friend Eli Putterman, for our fruitful collaboration and his assistance in writing and articulating ideas in our joint papers. Finally, I would like to acknowledge my dear friend and collaborator, Steven Hoehner, for his invaluable help throughout the years.

As I conclude, I'd like to extend my heartfelt thanks to all my collaborators and friends who journeyed with me throughout my PhD. My friends at MIT, including but not limited to Adam Block, Aviv Netayahu, and Dan Mikulincer, have enriched my experience by creating a lively and intellectually stimulating environment for work and collaboration.

In closing, I want to express my sincere gratitude to my thesis committee members, Prof. Martin Wainwright and Prof. Yury Polyanskiy.

Contents

Chapter 1

Introduction

This dissertation delves into the statistical performance of Empirical Risk Minimization (ERM) utilizing squared loss, commonly known as Least Squares Estimator (LSE), within the context of regression tasks. Notably, in the Gaussian noise model, LSE also serves as the Maximum Likelihood Estimator (MLE). ERM has emerged as a critical concept in the field of machine learning and statistical learning theory, shaping the way researchers and practitioners approach model development and evaluation. As the volume and complexity of data continue to grow exponentially, the need for effective and robust learning algorithms becomes increasingly pertinent. ERM procedures address this need by offering a principled and mathematically grounded framework for model selection and optimization. By minimizing the empirical risk or loss on a given dataset, these procedures provide a means to identify models that best capture the underlying patterns and relationships within the data, ultimately leading to improved generalization performance on unseen observations. The importance of ERM cannot be overstated, as it serves as the foundation for numerous widely-used learning algorithms and techniques, ranging from linear regression to deep learning. Additionally, ERM has proven instrumental in fostering a deeper understanding of the trade-offs between model complexity and generalization, thereby guiding the development of more efficient and effective learning algorithms.

Formally speaking, this dissertation studies the problem of regression where the goal is to estimate a function $f^* : \mathcal{X} \to \mathbb{R}$ from observations $\mathcal{D} := \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ drawn according to the model

$$Y_i = f^*(X_i) + \xi_i \quad \text{for } i = 1, \dots, n \tag{1.1}$$

where $X_1, \dots, X_n$ are n-design points (fixed or random) and $\xi_1, \dots, \xi_n$ are i.i.d random variables with mean zero and finite variance σ^2. In the *random design* setting, the design points are drawn i.i.d. from the same marginal distribution $\mathbb{P}$ over $\mathcal{X}$, i.e. $X_1, \dots, X_n \underset{i.i.d.}{\sim} \mathbb{P}$. In the *fixed design* setting, the design points $X_1 = x_1, \dots, X_n = x_n$ are arbitrary and fixed, and we denote the uniform measure over them by $\mathbb{P}^{(n)}$. Throughout this thesis, we assume the following:

Assumption 1. *The underlying function f^* lies in a known* convex *class of functions $\mathcal{F}$.*[1]

The convexity of $\mathcal{F}$ means that for any $f, g \in \mathcal{F}$ and $\lambda \in [0,1]$, $\lambda f + (1-\lambda)g \in \mathcal{F}$, and the assumption of $f^* \in \mathcal{F}$ is known as the well-specified model. $\mathcal{F}$ being a convex class of functions ensures that the least squares solution is uniquely defined on the data points, and that it is computationally tractable. Though, the convexity of $\mathcal{F}$ is a strong assumption, it is commonly used to for studying the statistical performance of ERM procedures (cf. [Cha14; BBM05; Men14]). Next, we assume that noise is isotropic Gaussian.

Assumption 2. *$\xi_1, \dots, \xi_n$ are independent $N(0,1)$ random variables.*

The assumption of normality is employed for the sake of simplicity in this presentation and due to the fact that minimax rates (in the regression model) are generally proven in the setting of normal noise [YB99]. Yet, our results remain applicable for other distributions of the noise $\bar{\xi}$.

Next, we define an estimator as a measurable (random) function $\bar{f}_n : (\mathcal{D}, \Omega) \mapsto \{\mathcal{X} \to \mathbb{R}\}$, i.e for any realization of the input $\mathcal{D}$ and some random parameter ("seed") $\omega \in \Omega$, $\bar{f}_n$ outputs some real-valued measurable function on $\mathcal{X}$. In many cases the estimator is a deterministic operator, i.e., a function $\mathcal{D} \to \{\mathcal{X} \to \mathbb{R}\}$, but we will also consider estimators that may depend on some random parameters.

[1] We also assume that $\mathcal{F}$ is also a closed set in terms of $L_2(\mathbb{Q})$, where $\mathbb{Q} \in \{\mathbb{P}^{(n)}, \mathbb{P}\}$, which ensures that ERM is well-defined. Closedness means that for any sequence $\{f_n\}_{n=1}^{\infty} \subset \mathcal{F}$ converging to f with respect to the norm of $L_2(\mathbb{Q})$, the limit f lies in $\mathcal{F}$.

The natural and classical estimator for any regression task is the LSE[2], which is defined by

$$\widehat{f}_n \in \mathrm{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2, \tag{1.2}$$

which is also the MLE under Assumption 2. Note that under Assumption 1, the LSE is *uniquely* defined on the data points, as LSE is simply the projection operator on the closed convex set [Cha14]

$$\mathcal{F}_{\overline{X}} := \{(f(X_1), \dots, f(X_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n.$$

That is, the vector $(\widehat{f}_n(X_1), \dots, \widehat{f}_n(X_n)) \in \mathbb{R}^n$ is uniquely defined as the closest point in $\mathcal{F}_{\overline{X}}$ to $(Y_1, \dots, Y_n)$.

Remark 1. $\widehat{f}_n$ may not be be unique on the entire $\mathcal{X}$, and therefore, we may equip it with a map Ψ that chooses a unique solution over the entire domain $\mathcal{X}$ from all possible solutions, i.e. $\widehat{f}_{n,\Psi} : \mathcal{D} \to \mathcal{F}$ is defined as

$$\mathcal{D} \underset{\widehat{f}_n}{\mapsto} \{f \in \mathcal{F} : \quad \forall 1 \le i \le n \; f(X_i) = \widehat{f}_n(X_i)\} \underset{\Psi}{\mapsto} \mathcal{F}.$$

For example, Ψ can be the minimal ℓ_2 norm solution in overparametrized linear regression.

As we mentioned earlier, this thesis mainly studies the statistical performance of the LSE in the task of regression.[3] There are many ways to measure the performance of an estimator, and we mainly focus on the maximum risk (see e.g., [Tsy03a]) with respect to mean-squared loss. In details, the maximum risk (or simply the risk) of an estimator $\bar{f}_n$ in the fixed design is defined via

$$\mathcal{R}(\bar{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) := \sup_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \|\widehat{f}_n - f^*\|_n^2, \tag{1.3}$$

where the expectation is taken over $\overline{\xi}$, and $\| \cdot \|_n$ denotes the $L_2(\mathbb{P}^{(n)})$ norm. Similarly, in the

[2] In the random design setting, the LSE may be also equipped with an additional map Ψ, see Remark 1 below.

[3] In Chapter 4 below, we construct and estimate the statistical guarantees of other estimators.

random design setting, we define

$$\mathcal{R}(\bar{f}_n, \mathcal{F}, \mathbb{P}) := \sup_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \|\widehat{f}_n - f^*\|^2, \tag{1.4}$$

where the expectation is taken over both ξ and X; and $\|\cdot\|$ denotes the $L_2(\mathbb{P})$ norm. [4] In words, for a given estimator $\bar{f}_n$, this measure chooses an underlying function $f^* \in \mathcal{F}$ that maximizes the estimator's error. In this fashion, the minimax rate of the function class $\mathcal{F}$ is defined via

$$\mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)}) := \inf_{\bar{f}_n} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \|\widehat{f}_n - f^*\|_n^2 \tag{1.5}$$

and

$$\mathcal{M}(n, \mathcal{F}, \mathbb{P}) := \inf_{\bar{f}_n} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \|\widehat{f}_n - f^*\|^2 \tag{1.6}$$

in the fixed and random designs, respectively. That is, the minimax risk is simply the minimal possible risk achievable by an estimator.

In this dissertation, we focus on non-parametric regression [Tsy03a], namely, when the function class $\mathcal{F}$ cannot be described by a *finite* number of parameters. For example, the class of all linear classifiers in $\mathbb{R}^d$ is parametric, since it can be described by d parameters. Whereas, the class of all α-Hölder functions (suppose on the unit cube in $\mathbb{R}^d$) cannot be described by finite number of parameters for all $d \geq 1$, and therefore it is a non-parametric class.

Non-parametric models motivate the following definition of minimax optimality:

Definition 1. *In the random design setting, an estimator $\bar{f}_n$ is minimax optimal if there exists $C_1 = C(\mathcal{F}, \mathbb{P}) > 0$ depending only on $\mathcal{F}$ and $\mathbb{P}$ such that*

$$\mathcal{R}(\bar{f}_n, \mathcal{F}, \mathbb{P}) \leq C_1 \cdot \mathcal{M}(n, \mathcal{F}, \mathbb{P}).$$

In the fixed design setting, an estimator $\bar{f}_n$ is called minimax optimal *(in the number of samples) if*

[4] If $\bar{f}_n$ is a random estimator, the expectation is taken over its random parameters as well.

there exists a constant $C_2 = C(\mathcal{F}) > 0$ depending only on $\mathcal{F}$ such that

$$\mathcal{R}(\bar{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) \leq C_2 \cdot \mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)}).$$

Remark 2. In models with relatively low number of parameters, such as under-parameterized linear regression, a refined notion of minimax optimality is used. One restrict the constants $C_1 > 0, C_2 > 0$ to absolute constants that are independent of $\mathcal{F}$ and $\mathbb{P}$.

The statistical effectiveness and minimax optimality of the LSE are among the most critical issues in the field of statistics (cf. [Gee00] and references within). This dissertation begins by posing the following question:

Question. *Given $(n, \mathcal{F}, \mathbb{P})$ (or $(\mathcal{F}, \mathbb{P}^{(n)})$), can the LSE be deemed as minimax optimal?*

This question has been thoroughly investigated, but remains unresolved in general. In this dissertation, we aim to bring further clarity to this issue and present new findings regarding the performance of ERM across diverse function classes. Our initial approach to addressing this question involves understanding the intricate link between the minimax rate and the metric entropy of the model $(\mathcal{F}, \mathbb{P})$. To explain this link, we first need to introduce a series of definitions and assumptions:

Definition 2 (Entropy numbers)**.** *The ϵ-metric entropy of a function class $\mathcal{G} \subset \{\mathcal{X} \to \mathbb{R}\}$ under the $L_2(\mathbb{Q})$-norm is the logarithm of the minimal cardinality of a finite set of functions $\mathcal{N}_\epsilon$ such that for any $f \in \mathcal{G}$ there exists a $f' \in \mathcal{N}_\epsilon$ with $\|f - f'\|_{\mathbb{Q}} \leq \epsilon$.*

We remark that the set $\mathcal{N}_\epsilon$ is called an ϵ-net or ϵ-covering. Throughout this manuscript, we use $\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P})$ to denote the ϵ-entropy of the model $(\mathcal{F}, \mathbb{P})$.

First, we introduce the notation of $\asymp, \gtrsim, \lesssim$ to denote equality/inequality up to an absolute constant. Next, we denote by $\mathrm{Diam}_{\mathbb{Q}}(\mathcal{G})$ to be the diameter of the class $\mathcal{G}$ with respect to the norm $L_2(\mathbb{Q})$, i.e $\mathrm{Diam}_{\mathbb{Q}}(\mathcal{G}) := \sup_{f,g \in \mathcal{G}} \|f - g\|_{\mathbb{Q}}$. Our next assumption is mainly used to ensure that the minimax can be described in terms of the ϵ-entropy numbers of the entire class $\mathcal{F}$.

Assumption 3. *The diameter of $\mathcal{F}$ in terms of $\mathbb{Q} \in \{\mathbb{P}, \mathbb{P}^{(n)}\}$ is of order one, in our notation* $\mathrm{Diam}_{\mathbb{Q}}(\mathcal{F}) \asymp 1$.

The following is a classical result in non-parametric statistics [YB99; LeC73]:

Theorem 1. *[Minimax rates in fixed design] Under Assumptions 1-3*[5]*, the following holds:*

$$\mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)}) \asymp \epsilon_*^2,$$

where ϵ_* is the solution of

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}^{(n)}) \asymp n\epsilon^2. \tag{1.7}$$

In the random design setting, we further assume that the class is uniformly bounded:

Assumption 4. *There exists an absolute constant $\Gamma \geq 0$ such that $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq \Gamma$, i.e. $\mathcal{F}$ is uniformly bounded by Γ.*

With this additional assumption, the same characterization of the minimax rate is available in this case as well:

Theorem 2. *[Minimax rates in random design [YB99]] Under Assumptions 1-4*[6]*, the following holds:*

$$\mathcal{M}(n, \mathcal{F}, \mathbb{P}) \asymp \epsilon_*^2,$$

where ϵ_* is the solution of

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}) \asymp n\epsilon^2. \tag{1.8}$$

The central insight from these findings is that the minimax rate, up to a multiplicative absolute constant, only depends on a straightforward and inherent feature of the model's geometry - its metric entropy.

Can the LSE be minimax optimal in the number of samples? Remarkably, as we will demonstrate, for many function classes known as Donsker classes, LSE is always minimax optimal. Initially associated with the uniform convergence of empirical processes, Donsker classes have since become a fundamental part of non-parametric statistics and statistical learning theory [VW96].

[5] And the additional assumption of $\log \mathcal{N}(\epsilon/2, \mathcal{F}, \mathbb{P}^{(n)}) / \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}^{(n)}) > 1$, for $\epsilon \gtrsim \mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)})$.

[6] And the additional assumption of $\liminf_{\epsilon > 0} \log \mathcal{N}(\epsilon/2, \mathcal{F}, \mathbb{P}) / \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}) > 1$.

Here, we divide to categorize existing results on the LSE into two groups based on whether the class of functions $\mathcal{F}$ belongs to the *Donsker* regime or the *non-Donsker* regime. As the full definition of Donsker classes involves the introduction of several notions we will not otherwise use, we will present a simplified definition [7], sufficient for our purposes, that is widely used (cf. [Gee00; CR06]). To this end, we introduce the definition of ϵ-entropy with bracketing:

Definition 3 (Entropy numbers with bracketing)**.** *The ϵ-metric entropy with bracketing of a function class $\mathcal{G} \subset \{\mathcal{X} \to \mathbb{R}\}$ under the $L_2(\mathbb{Q})$-norm is the logarithm of the minimal number of N pairs of functions $\mathcal{N}_{[],\epsilon} := \{(l_i, u_i)\}_{i=1}^{N}$ that satisfy the following:*

1. *Every $(l, u) \in \mathcal{N}_{[],\epsilon}$ satisfies $\|l - u\|_{\mathbb{Q}} \leq \epsilon$.*
2. *For every $f \in \mathcal{G}$ there exists $(l, u) \in \mathcal{N}_{[],\epsilon}$ such that $l \leq f \leq u$.*

We remark that the set $\mathcal{N}_{[],\epsilon}$ is called an ϵ-net with bracketing. Throughout this manuscript, we use $\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \mathbb{P})$ to denote the ϵ-entropy with bracketing of the model $(\mathcal{F}, \mathbb{P})$.

Note that by definition $\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}) \leq \log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \mathbb{P})$. In the asymptotic setting, where $(\mathcal{F}, \mathbb{P})$ is fixed and n grows, it is considered a very mild assumption to assume the converse, i.e. that $\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \mathbb{P}) \lesssim C(\mathcal{F}, \mathbb{P}) \cdot \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P})$ (cf. [BM93]). This mild assumption is used extensively in analyzing ERM on non-parametric classes in the random design setting [8] (cf. [Gee00]). In order to define our notions of $\mathbb{P}$-Donsker (or non-Donsker) classes, we shall assume it as well:

Assumption 5. *The following holds for all $\epsilon \in (0, \mathrm{Diam}_{\mathbb{P}}(\mathcal{F}))$:*

$$\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \mathbb{P}) \underset{\mathcal{F},\mathbb{P}}{\asymp} \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}),$$

where $\underset{\mathcal{F},\mathbb{P}}{\asymp}$ denotes equality up to a multiplicative constant depending on $\mathcal{F}, \mathbb{P}$.

Remark 3. Assumption 5 has many roles in the literature of empirical process theory (cf. [Pol02; Tal14] and references within). For one, this assumption implies the existence of a constant $C_3 =$

[7] The formal definition of Donsker classes appear in [VW96, Pg. 17] and [Gee00, Cpt. 6].

[8] The bracketing property does not have any significance in the setting of fixed design regression (cf. [Pol02]).

$C_3(\mathbb{P}, \mathcal{F}) > 0$ such that with high probability (over $X_1, \dots, X_n$) the following holds uniformly for all $f, g \in \mathcal{F}$ (cf. [Gee00, Lemma 5.16]):

$$4^{-1}\|f-g\|^2 - C_3 \cdot \mathcal{M}(n, \mathcal{F}, \mathbb{P}) \le \|f-g\|_n^2 \le 4\|f-g\|^2 + C_3 \cdot \mathcal{M}(n, \mathcal{F}, \mathbb{P}),$$

where $\|\cdot\|_n$, with some ambiguity in the notation, denotes the $L_2(\mathbb{P}_n)$ norm which is defined by the *random* (uniform) empirical measure over $X_1, \dots, X_n \underset{i.i.d.}{\sim} \mathbb{P}$, i.e. $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$. This holds true in particular when taking $f = f^*$, $g = \widehat{f}_n$, and hence under this additional assumption, the ERM is minimax optimal in random design if and only if it is minimax optimal in fixed design.

Now, as we promised above, we are finally ready to present the simplified definition of $\mathbb{P}$-Donsker classes [Gee00, Theorem 6.3]:

Definition 4 (Simplified Definition of Donsker classes)**.** *Under Assumptions 1-5, the $\mathcal{F}$ is called $\mathbb{P}-$Donsker if*

$$\int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P})} d\epsilon < \infty. \tag{1.9}$$

In particular, $\mathcal{F}$ is $\mathbb{P}$-Donsker, when there exists $p \in (0,2)$ such that

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}) \underset{\mathcal{F},\mathbb{P}}{\asymp} \epsilon^{-p} \quad \forall \epsilon \in (0, \mathrm{Diam}_{\mathbb{P}}(\mathcal{F})).$$

We remark that when $p \ge 2$, the integral of (1.9) diverges, and such a class does not satisfies the formal definition of Donsker class.

Remark 4. The $\mathbb{P}$-Donsker property is tightly connected to finite VC-dimension. For example, if $\mathcal{F} \subset \{\mathcal{X} \to \{0,1\}\}$, then it is $\mathbb{P}$-Donsker for any marginal $\mathbb{P}$ if and only if its VC-dimension is finite [Dud99, Theorem 10.1.4].

In this thesis, we will mainly consider a *special* case (and yet very natural) of the general notion of non-Donsker classes [9], which is nonetheless very widely applicable (as will be discussed below), and it is defined as follows:

[9] In Chapter 3, we will also study properties of ERM on additional settings.

Definition 5 (Non-Donsker Classes)**.** *Under Assumptions 1-5, we say that $\mathcal{F}$ is non $\mathbb{P}$-Donsker when there exists $p > 2$ such that*

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}) \underset{\mathcal{F},\mathbb{P}}{\asymp} \epsilon^{-p} \quad \forall \epsilon \in (0, \mathrm{Diam}_{\mathbb{P}}(\mathcal{F})).$$

Throughout this thesis, we will refer to the class of models that satisfy Definition 4 as the "Donsker regime", and for models that satisfy Definition 5 as the "non-Donsker regime".

Note that under the above two definitions, Theorem 2 above implies that the minimax rate of $\mathcal{F}$ will be of order $n^{-\frac{2}{2+p}}$, whether the class is Donsker or non-Donsker. Furthermore, it should be noted that different variants of the same general problem may fall in different regimes. For example, the class of support functions of compact convex subsets of $\mathbb{R}^d$ is Donsker for $d \leq 5$ and non-Donsker for $d \geq 6$. We revisit this example in detail in Chapter 2 below.

Finally, we reiterate that Definition 5 is not the general definition, but a special case. Yet our definition covers many natural non-parametric classes such Isotonic, convex, α-Hölder functions and more (cf. [Dud99]). We now describe the state of knowledge as it stood before our contributions:

On the Optimality of the LSE: In the Donsker regime, it is well-known that (under Definition 4), the LSE achieves the rate $n^{-2/(2+p)}$ [BM93]. Namely, according to our definitions, the LSE is minimax optimal when the class $\mathcal{F}$ is $\mathbb{P}$-Donsker.

In contrast, the minimax optimality of the LSE in the non-Donsker regime remains unresolved. In a fundamental paper, [BM93] proved that in the non-Donsker regime, the rate of convergence of the LSE is *always* at least as fast as $n^{-1/p}$, and they also showed by example that this bound may be tight. Namely, they observed that it is possible to design "unnatrual" function classes $\mathcal{F}$ where the LSE provably achieves a risk of order $n^{-1/p}$ (up to logarithmic factors). By unnatrual, we mean classes there were designed in adversarial way to maximize the error of the LSE, and do not appear in real applications or even in theory. Later, [Bir06] provided additional unnatrual examples of such classes. As it was mentioned earlier, the minimax rate of estimation is still of order $n^{-2/(2+p)}$ and therefore the LSE *may* be suboptimal in the non-Donsker regime.

Some important progress on this open problem has been made in the recent papers [KDR19;

Han19; Car+18; Han+19] (including a paper of the author of this thesis). Specifically, these papers have shown that there exist “natural” non-Donsker families of functions where the LSE achieves the minimax rate of order $n^{-2/(2+p)}$, among them: the class of bounded convex functions supported on “smooth" domains in $\mathbb{R}^d$ for $d \geq 3$, the class of multivariate isotonic functions over $\mathbb{R}^d$ for $d \geq 2$, and some other examples. Prior to our work, all known non-Donsker classes for which the LSE was proved minimax suboptimal were somehow “pathological” in their nature.

We summarize this short discussion in the following observations:

Observation I: For a general non-Donsker class $(\mathcal{F}, \mathbb{P})$, the risk of the LSE can be $n^{-1/p}$ or $n^{-2/(2+p)}$ (or some intermediate rate); i.e. the LSE can be minimax optimal or sub-optimal in the non-Donsker regime

Observation II: Unlike the Donsker regime, the minimax optimality of the LSE in the non-Donsker regime is influenced by additional characteristics of the class besides its entropy numbers alone.

This naturally leads to the following question:

Question. *Are there “natural” non-Donsker classes in which the LSE attains a sub-optimal error rate?*

The first part of this thesis answers this question. We show that there exist “natural” function classes for which the LSE attains the this sub-optimal rate. We provide geometric conditions on the geometry of the model $(\mathcal{F}, \mathbb{P})$ which implies the sub-optimality of the LSE, and show that these are satisfied by the class of convex Lipschitz functions when $d \geq 5$ and the class of support functions of convex sets in $\mathbb{R}^d$ when $d \geq 6$, hence implying that the LSE is minimax suboptimal on these classes. We discuss this result in §1.1 below.

This result resolves a long-standing open problem [GKM06] on the sub-optimality of the LSE on the problem of estimating a convex set in dimension $d \geq 6$ from noisy support function measurements; see for example [GKM06; Gun12; SC19a].

A natural follow-up question is what may be the *cause* of potential suboptimality of the LSE (or ERM in general) in the non-Donsker regime.

Question. *What is the reason for the potential sub-optimality of LSE in the non-Donsker regime?*

The second part of this thesis gives insight into this question. First, we remind the classical bias-variance decomposition of the mean squared error of an estiamtor $\bar{f}_n$:

$$\mathbb{E}_{\mathcal{D}} \int (\bar{f}_n - f^*)^2 d\mathbb{Q} = \underbrace{\mathbb{E}_{\mathcal{D}} \int (\bar{f}_n - \mathbb{E}_{\mathcal{D}}\bar{f}_n)^2 d\mathbb{Q}}_{V(\bar{f}_n)} + \underbrace{\int (\mathbb{E}_{\mathcal{D}}\bar{f}_n - f^*)^2 d\mathbb{Q}}_{B^2(\bar{f}_n)}, \tag{1.10}$$

where $\mathbb{Q} = \mathbb{P}^{(n)}$ in the fixed design setting and $\mathbb{Q} = \mathbb{P}$ in the random design setting. We refer to the first term as the variance error term and the second term as the (squared) bias error term. Next, we also define the maximal variance of an estimator $\bar{f}_n$ as

$$\mathcal{V}(\bar{f}_n, \mathcal{F}, \mathbb{Q}) := \sup_{f^* \in \mathcal{F}} V(\bar{f}_n) = \sup_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \int (\bar{f}_n - \mathbb{E}\bar{f}_n)^2 d\mathbb{Q}, \tag{1.11}$$

where $\mathbb{Q} \in \{\mathbb{P}^{(n)}, \mathbb{P}\}$. In §1.2 and Chapter 3 below, we will show that in the fixed design setting (under Assumptions 1-3), ERM has a minimax optimal variance error term up to a multiplicative absolute constant, i.e.

$$\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) \lesssim \mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)}). \tag{1.12}$$

Furthermore, we will bound $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P})$ for various models in the random design setting. In particular, we will show that if $\mathcal{F}$ is a *non*-$\mathbb{P}$-Donsker class (see Definition 5), then

$$\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim \mathcal{M}(n, \mathcal{F}, \mathbb{P}). \tag{1.13}$$

Note that for $\mathbb{P}$-Donsker classes (see Definition 4), the LSE is always minimax optimal; therefore its variance error term is minimax optimal as well. This yields the following corollary:

Corollary (Informal)**.** *In the non-Donsker regime, the suboptimality of the LSE can only be due to the bias error term.*

Now is an appropriate time to inquire: Could our fixation on LSE and ERM procedures be unnecessary? Specifically, are there generic estimators which can be applied to any model $(\mathcal{F}, \mathbb{P})$

and achieve minimax optimality? Alternatively, are there methods that can reduce the (potentially) significant bias of LSE or ERM procedures?

Indeed, such estimators exist, including aggregation procedures, the star algorithm, and more (cf. [Yan04] and the discussion sections in [RST17])[10]. But these estimators have a significant limitation: all of them are based on constructing a minimal ϵ-net with respect to $L_2(\mathbb{P})$. As an example, we present the simplest of these estimators, known as the sieve estimator [Gee00, Pgs. 184-185]. The main idea behind it is to restrict the LSE to choose elements in the ϵ-net instead of the entire class $\mathcal{F}$. Formally, for each $\epsilon \in (0,1)$, we define

$$\widehat{f}_n^{\epsilon} := \text{argmin}_{f \in N_\epsilon} n^{-1} \sum_{i=1}^{n} (Y_i - f(X_i))^2, \tag{1.14}$$

where N_ϵ is defined as the minimal ϵ-net with respect to $L_2(\mathbb{P})$.

Intuitively, as $\epsilon \to 0$, the approximation error (the distance between the true function f^* and the best approximation f_ϵ^* to f^* in N_ϵ) decreases, while the estimation error (the average distance between $\widehat{f}_n^{\epsilon}$ and f_ϵ^*) increases. It turns out that there exists a choice of ϵ that balances the two terms, which is precisely the minimax rate ϵ_*, and $\widehat{f}_n^{\epsilon_*}$ is a minimax optimal estimator.

However, this approach has the fundamental limitation that constructing such an ϵ-net is in general computationally hard; in most situations, the size of an ϵ_*-net is itself superpolynomial in the number of samples n, and lack of convexity in the optimization problem (1.14) means that it usually cannot be solved in polynomial time. This motivates the following question:

Question. *Does a minimax optimal algorithm exist for non-Donsker classes that is both minimax optimal and has a polynomial runtime in the number of samples?*

If such an algorithm exists and is relatively simple to compute, it could potentially replace ERM procedures in practice. Settling this question could have significant practical implications. Regrettably, despite numerous efforts and attempts, we have not been able to fully resolve this question. However, we have achieved one advance which indicates that there might be hope for a

[10] In an forthcoming publication titled "Local Risk Bounds for Entropy Regularized Aggregation", Prof. N. Zhivotovskiy, and his colleagues provided an aggregation method that does not use ϵ-nets. Nevertheless, it is important to note that their procedure has a super-polynomial runtime in the number of samples.

positive resolution in the future.

In the third part of this thesis, we introduce the first *efficient* minimax optimal algorithm for the task of multivariate convex Lipschitz regression when the dimension $d \geq 5$ – which is a non-Donsker class. In this task, the LSE is provably minimax sub-optimal (a result of the author that was discussed above) and exhibits poor performance when tested in practice. Moreover, all previously known algorithms (approximately a dozen) are either provably minimax suboptimal or do not come with any statistical guarantees on maximum risk. We believe that all of them can be proven to be minimax sub-optimal. We remark that our algorithm has a runtime of order $n^{O(d)}$ which is polynomial in the number of samples for fixed dimension d; prior to our work all known minimax optimal algorithms had an exponential runtime in n.

In the final part of this thesis, we move away from the notion of maximum risk, aiming to estimate the statistical performance of LSE under the most optimistic manner, that is its minimal error. Namely,

$$\inf_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{Q},$$

where $\mathbb{Q} \in \{\mathbb{P}^{(n)}, \mathbb{P}\}$.

Naturally, we aim to understand the following question: Is the minimal error of the LSE is affected by the richness of the entire class $\mathcal{F}$ both in the fixed or random design setting?. In this introduction, we only state our result in the fixed design, as in the random design it is more complicated and requires more terminology.

In the fixed design setting, the answer to the aforementioned question is affirmative. Specifically, for any fixed design measure $\mathbb{P}^{(n)}$, under Assumptions 1-3, we demonstrate the following:

$$\inf_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}^{(n)} \gtrsim \mathcal{W}_{\mathbf{x}}(\mathcal{F})^2,$$

where $\mathcal{W}_{\mathbf{x}}(\mathcal{F}) := \mathbb{E}_{\overline{\xi}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(x_i)$ (and remember that $\xi_1, \ldots, \xi_n \underset{i.i.d.}{\sim} N(0,1)$).

The quantity $\mathcal{W}_{\mathbf{x}}(\mathcal{F})$ is known as the Gaussian complexity, which is a fundamental measure in many fields, including mathematical statistics and high-dimensional geometry [Wai19; AAGM15], and will be used throughout this thesis. Roughly speaking, the Gaussian complexity measures the richness of the class in the fixed design setting. Therefore, the latter implies that the minimal error (or the adaptive rate) of the LSE is lower bounded by the square of the Gaussian complexity of the entire class.

Before we delve into the detailed results, let us outline the organization of this thesis and introduce the common notation used throughout:

Organization: Chapter 2 and §1.1 are based on [KRG20]. Chapter 3 and §1.2 are based on [KPR23]. Chapter 4 and §1.3 are based on [KP22]. Chapter 5 and §1.4 are based on [KR21].

Notation: The following notational conventions are used throughout this document:

- C, c with subscripts represent positive absolute constants. When these constants depend on parameters such as $p, p_1, \ldots$, they will be denoted as $c(p), C(p), \ldots$ or $c(p, p_1), C(p, p_1), \ldots$. Note that the values of these constants may change from line to line and from section to section.

- $\mathbb{P}_n$ signifies the uniform measure over $X_1, \ldots, X_n$, where $X_1, \ldots, X_n \underset{\text{i.i.d.}}{\sim} \mathbb{P}$. The notation $\| \cdot \|_n$ is used to represent the $L_2(\mathbb{P}^{(n)})$ norm in the fixed design setting, and $L_2(\mathbb{P}_n)$ in the random design setting.

- Standard big-O, Ω, and Θ notation is used, along with their parameterized versions $O_{p_1}, \Omega_{p_1}, \Theta_{p_1}$ for some parameters p_1. For example, $O(n^2)$ implies bounded above by Cn^2 for some absolute constant $C \geq 0$, whereas $O_{p_1}(n^2)$ implies bounded by $C(p_1)n^2$ for some constant $C(p_1)$ depending only on p_1.

- The design points and the noise vector are represented in vector form as $\overline{X} := (X_1, \ldots, X_n)$ and $\overline{\xi} := (\xi_1, \ldots, \xi_n)$, respectively. For any finite-dimensional vector v, $\|v\|_2$ denotes its Euclidean norm, and $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$, i.e. $\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d : \|v\|_2 = 1\}$.

1.1 On the sub-optimality of least squares

Here, we give a new recipe for establishing a lower bound for the LSE's risk in the *non*-Donsker regime, both in fixed and random design. As an application, we complement the aforementioned recent results by proving that there also exist "natural" non-Donsker families of functions, where the LSE cannot achieve a rate faster than $n^{-1/p}$, up to logarithmic multiplicative factors (such as the class of support functions of convex bodies in $\mathbb{R}^d$ for $d \geq 6$). In other words, for these non-Donsker classes, the LSE is provably *suboptimal*. In this introduction, we only define the class of support functions on convex bodies in $\mathbb{R}^d$ and presenting our sub optimality result on the LSE when $d \geq 6$. In Chapter 2 below, we present our approach, which involves a technical result giving general conditions on a function class under which one can prove a lower bound for the risk of the LSE. This approach is quite technical in its nature and necessitates the use of additional statistical notions.

The sub-optimality of least squares for estimating a convex set

The task of estimating a convex set from noisy support function measurements is a classical and known task in non-parametric statistics and geometric tomography [GKM06; Bru16; Bru13; Gun12; Fis+97; BGS15; SC19a]. First, let us define the support function of a compact convex set $K \subset \mathbb{R}^d$. For any vector $u \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$, the support function $h_K : \mathbb{S}^{d-1} \to \mathbb{R}$ is defined via

$$h_K(u) = \max_{x \in K} \langle x, u \rangle .$$

The support function uniquely determines the compact convex set K and is a fundamental object in convex geometry (see, for example, [Sch14, §1.7] or [Roc70, §13]). Consider now the problem of estimating an unknown compact, convex set K^* from observations $(X_1, Y_1), \dots, (X_n, Y_n)$ drawn according to the model:

$$Y_i = h_{K^*}(X_i) + \xi_i \quad \text{for } i = 1, \dots n$$

where $X_1, \dots, X_n$ are design points (fixed or random) and $\xi_1, \dots, \xi_n \underset{i.i.d.}{\sim} N(0,1)$. Recovery of K^* is a fundamental problem in geometric tomography [PW90; Gar95]. The natural estimator in this problem is the LSE, here it is defined by

$$\widehat{K}_n \in \mathrm{argmin}_{K \in \mathcal{C}_d} \sum_{i=1}^{n} (Y_i - h_K(X_i))^2$$

where $\mathcal{C}_d$ denotes the class of all compact, convex sets in $\mathbb{R}^d$. Basic properties and algorithms for computing $\widehat{K}_n$ can be found in [PW90; LKW92] and [Kid+08]. Rigorous accuracy results for $\widehat{K}_n$ as an estimator of $K^* \subset B_d$ (here B_d denotes the unit-Euclidean ball) were proved in [GKM06]. Specifically, [GKM06, Corollary 5.7] proved that, under the fixed design setting of any $X_1, \dots, X_n$ (fixed points) that form a well-separated set (i.e. a set of n points on the unit-sphere that are at least $c(d)n^{-\frac{1}{d-1}}$ far from each other),

$$\mathcal{R}(h_{\widehat{K}_n}, \mathcal{C}_d(1), \mathbb{P}^{(n)}) \leq O_d(\beta_n) \quad \text{where } \beta_n = \begin{cases} n^{-4/(d+3)} & \text{for } d = 2,3,4 \\ n^{-4/(d+3)} \cdot \log n & \text{for } d = 5 \\ n^{-2/(d-1)} & \text{for } d \geq 6. \end{cases} \tag{1.15}$$

where $\mathbb{P}^{(n)}$ denotes the uniform measure on these n-well separated points.

Complementarily, [Gun12] proved that the minimax rate of estimation in this problem equals $\Theta_d(n^{-4/(d+3)})$ for all $d \geq 2$. These two results combined imply that the least squares estimator $\widehat{K}_n$ is minimax optimal for $d = 2,3,4$ and nearly minimax optimal (up to the logarithmic multiplicative factor $\log n$) for $d = 5$. However, there is a gap between the upper bound $O_d(n^{-2/(d-1)})$ on the rate of convergence of the least squares estimator (1.15) and the minimax rate $\Theta_d(n^{-4/(d+3)})$ for $d \geq 6$. This gap has remained open since the paper [GKM06] where it was suggested that the upper bound is accurate and that the least squares estimator is indeed minimax *suboptimal* for $d \geq 6$. The goal of this section is to confirm the long-standing conjecture of [GKM06].

Specifically, we will use our recipe that appears below, to prove that the LSE is suboptimal for $d \geq 6$, in the sense that there exist sets for which the rate of convergence for LSE is bounded from below by $\Omega_d(n^{-2/(d-1)})$ up to a logarithmic multiplicative factor. Before, we state our theorem, we denote by $\mathcal{C}_d(1) := \{K \in \mathcal{C}_d : K \subset B_d\}$, and by $\mathrm{U}(\mathbb{S}^{d-1})$ to be the uniform distribution on $\mathbb{S}^{d-1}$.

Theorem 3. *Let $d \geq 6$ and $n \geq d+1$. Suppose that $X_1, \ldots, X_n \underset{i.i.d.}{\sim} \mathrm{U}(\mathbb{S}^{d-1})$, then there exist a positive constants γ_d depending only on d such that*

$$\mathcal{R}(h_{\widehat{K}_n}, \mathcal{C}_d(1), \mathrm{U}(\mathbb{S}^{d-1})) = \Omega_d(n^{-2/(d-1)} \log(n)^{-\gamma_d}).$$

Note that $h_{\widehat{K}_n}$ in not restricted to $\mathcal{C}_d(1)$, i.e. it can return any convex set in $\mathbb{R}^d$. We emphasize the our proof implies that the same bound will be valid for the restricted LSE over $\mathcal{C}_d(1)$, hence this result is consistent with our assumptions on non-Donsker classes that appeared above. The proof of the corollary can be modified to hold for any fixed design of n well-separated points. The modification will follow from the fact that n well-separated points are a discrete approximation to the uniform measure on the sphere. Therefore, we settle the question in [GKM06].

Remark 5. In the random design setting, our result is valid for any density $g(x)$ on the sphere, such that $g(x) \geq c_1(d)$ for all $x \in \mathbb{S}^{d-1}$.

1.2 On the Variance, Admissibility and Stability of Empirical Risk Minimization

The study of asymptotic consistency of Maximum Likelihood has been central to the field for almost a century [Wal49]. Along with consistency, failures of Maximum Likelihood have been thoroughly investigated for nearly as long [NS48; Bah58; Fer82]. In the context of estimation of non-parametric models, the seminal work of [BM93] provided *sufficient* conditions for minimax optimality (in a non-asymptotic sense) of Least Squares while also presenting an example of a model class where this basic procedure is sub-optimal. Three decades later, we still do not have necessary and sufficient conditions for minimax optimality of Least Squares when the model class is large. While the present work does not resolve this question, it makes several steps towards understanding the behavior of Least Squares — equivalently, Empirical Risk Minimization (ERM) with square loss — in large models.

Beyond intellectual curiosity, the question of minimax optimality of Least Squares is driven by the desire to understand the current practice of fitting large or overparametrized models, such as neural networks, to data (cf. [BRT19; Bar+20]). At the present moment, there is little theoretical understanding of whether such unregularized data-fitting procedures are optimal, and the study of their statistical properties may lead to new methods with improved performance.

In addition to minimax optimality, many other important properties of Least Squares with large models are yet to be understood. For instance, little is known about its *stability* with respect to perturbations of the data. It is also unclear whether *approximate* minimizers of empirical loss enjoy similar statistical properties as the exact solution. Conversely, one may ask whether in the landscape of possible solutions, a small perturbation of Least Squares itself is a near-optimizer of empirical loss.

In this part of the thesis, we provide novel insights into the aforementioned questions for convex function classes in various settings. In details, we show the following:

1. We prove that in the fixed design setting, the variance term of ERM agrees with the minimax rate of estimation; thus, if the ERM is minimax suboptimal, this must be due to the bias term in the bias-variance decomposition (see Theorem 8 in §3.4.1 below). We remark that this result implies Eq. (1.12) that appeared above.

2. In the random design, using tools from empirical processes theory, we provide an upper bound for the variance error term under a uniform boundedness assumption on the class (see Theorem 13 in §3.2.1 below). This bound also implies that under classical assumptions in empirical process theory, the variance error term is minimax optimal (see Corollary 4 in §3.2.1 below). In particular, it also covers our definition of the "non-Donsker" regime (see Def. 5 above), and therefore it implies Eq. (1.13) above.

3. In the random design setting, we prove that under an isoperimetry assumption on the noise, the expected conditional variance error term (in terms of the total law of variance) of ERM is upper bounded by the "lower isometry" remainder, a parameter we introduce following [BBM05; Men14] (see Theorem 14 in §3.2.2 below).

   Furthermore, under an additional isoperimetry assumption on the covariates, we prove that the variance of ERM is upper bounded by the lower isometry remainder on any robust learning architecture (namely, a class that all its functions are $O(1)$-Lipschitz) that almost interpolates the observations. (cf. [BS23]) (see Theorem 15 in §3.2.2 below).

4. It is known that ERM is always admissible in the fixed design [Cha14; CGZ17], in the sense that for any convex class $\mathcal{F}$, there is no estimator that can have a lower error than ERM (up to a multiplicative absolute constant) on *every* regression function. We provide a short proof of this result for fixed design via a fixed-point theorem (see Theorem 11 and Corollary 2 below). Using the same method, we also prove a somewhat weaker result in the random design case, generalizing the main result of [Cha14] (see Theorem 16 below).

5. We show that ERM is stable, in the sense that all almost-minimizers (up to the minimax rate) of square loss are close in the space of functions. This result is a non-asymptotic analogue of the asymptotic analysis in [CR06], and extends its scope to non-Donsker classes (see Theorems 8 and 13 below).

6. While any almost-minimizer of the square loss is close to the minimizer with respect to the underlying population distribution, the converse is not true. We prove that for any non-Donsker class of functions, there exists a target regression function such that, with high probability, there exists a function with high empirical error near the ERM solution; this means that the landscape of near-solutions is, in some sense, irregular (see Theorem 12 below).

1.3 Efficient Minimax Optimal Estimators For Multivariate Convex Regression

Let Ω be a convex set in $\mathbb{R}^d$ that lies in the (Euclidean) ball with radius one, denoted by B_d, i.e. $\Omega \subset B_d$; and let $\mathcal{F}(\Omega)$ be the class of all convex functions that their domain is Ω. In this part, we consider the following sub-classes of $\mathcal{F}(\Omega)$:

1. $\mathcal{F}_L(\Omega)$ – the class of convex L-Lipschitz functions supported on a polytope Ω.

2. $\mathcal{F}^{\Gamma}(\Omega)$ – the class of Γ-uniformly bounded convex functions supported on a polytope Ω.

These tasks are known as L-Lipschitz convex regression [SS11] and Γ-bounded convex regression [HW16], respectively. For reasons that will become apparent later, we always take $d \geq 5$ (which, unsurprisingly, corresponds to the non-Donsker regime of these tasks, see Definition 5). In our work, we further assume that $\mathbb{P}$ satisfies the following condition:

Assumption 6. $\mathbb{P}$ *is uniformly bounded on its support by some constants* $c(d), C(d) > 0$*, namely,* $c(d) \leq \frac{d\mathbb{P}}{dx}(x) \leq C(d)$*, for all* $x \in \Omega$*. Furthermore,* $\mathbb{P}$ *is* known *to us (i.e. we know the density of* $\mathbb{P}$ *on each point in* Ω*).*

Convex regression tasks have been a central concern in the “shape-constrained" statistics literature [DL12], and have innumerable applications in a variety of disciplines, from economic theory [Var82] to operations research [PT03] and more [Bal16]. In general, convexity is extensively studied in pure mathematics [AAGM15], computer science [LV07], and optimization [BBV04]. We remark that there is a density-estimation counterpart of the convex regression problem, known as log-concave density estimation [Sam18; CSS10], and these two tasks are closely related [KDR19; KS16].

Due to the appearance of convex regression in various fields, it has been studied from many perspectives and by many different communities. For example, in the mathematical statistics literature the minimax rates of convex regression tasks and the risk of the maximum likelihood estimator (MLE) are the main areas of interest; an incomplete sample of works treating this problem is [Gun12; GS13; Gar95; GW17; KRG20; Han19; Bru13; DSS18; DKS16; Car+18; KDR19; BGS15]. In operations research, work has focused on the algorithmic aspects of convex regression, i.e., finding scalable and efficient algorithms; see, e.g., [Gho+21; Bru16; OC21; SC21; Bal16; Maz+19; CM20; BM21; Sia+21; Sim+18; HD12; Bla+19; LST20; CMS21; Bal22]. Initially, convex regression was mostly studied in the univariate case, which is now considered to be well-understood. Multivariate convex regression has only begun to be explored in recent years, and is still an area of active research.

The naive algorithm for any variant of the convex regression task is the LSE, which is defined via

$$\widehat{f}_n \in \operatorname{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

where $\mathcal{F} = \mathcal{F}_L(\Omega)$ or $\mathcal{F} = \mathcal{F}^{\Gamma}(\Omega)$ in our convex regression tasks. From a computational point of view, the LSE can be formulated as a quadratic programming problem with $O(n^2)$ constraints, and therefore can be computed in polynomial runtime in n [SS11; HW16]; however, the LSE has been seen empirically not to be scalable for large number of samples [CM20].

From a statistical point of view, the minimax rates of both of our convex regression tasks are $\Theta_{d,L,\sigma}(n^{-\frac{4}{d+4}})$ and $\Theta_{d,\Gamma,\sigma,\Omega}(n^{-\frac{4}{d+4}})$ (respectively) for all $d \geq 1$ [GW17; Bro76; YB99]. The LSE is minimax optimal *only* in low dimension, when $d \leq 4$ [BM93], while for $d \geq 5$ it attains a suboptimal risk of $\widetilde{\Theta}_d(n^{-\frac{2}{d}})$ [Kur+20]. The poor statistical performance of the LSE for $d \geq 5$ has also been verified empirically [GKM06; Gho+21]. There are known minimax optimal estimators when $d \geq 5$, yet all of them are computationally inefficient. Moreover, all of them are based on some sort of discretization of the relevant function classes, i.e., they consider some ϵ-nets (see Definition 2 above). In our tasks, these algorithms require examining nets of cardinality that is exponential in the number of samples, and are thus perforce inefficient [RST17; Gun12].

The empirically-observed poor performance of the LSE and the computational intractability of known minimax optimal estimators have motivated the study of efficient algorithms for convex regression with better statistical properties than the LSE; an incomplete list of relevant works appears above. However, previously studied algorithms are either provably minimax suboptimal or do not provide any statistical guarantees at all with respect to the minimax risk. We would however like to mention the “adaptive partitioning” estimator constructed in [HD13], which is the first provable computationally efficient estimator for convex regression which has been shown to be *consistent* in the L_∞ norm. The authors’ approach is somewhat related to our proposed algorithm, but it is unknown whether their algorithm is minimax optimal.

Our main results are the existence of computationally efficient minimax optimal estimators for the task of multivariate Lipschitz convex regression and bounded convex regression under polytopal support. Specifically, we prove the following results:

Theorem 4. *Let $d \geq 5$ and $n \geq d+1$. Then, under Assumption 6, for the task of L-Lipschitz convex* *regression on a convex polytope $\Omega \subset B_d$, there exists an efficient estimator, $\widehat{f}_{L,n}$, with runtime of at most $n^{O(d)}$ such that*

$$\mathcal{R}(\widehat{f}_{L,n}, \mathcal{F}_L(\Omega), \mathbb{P}) \leq (\sigma + L)^2 n^{-\frac{4}{d+4}} \log(n)^{h(d)} + C(\Omega) n^{-\frac{d}{d+4}} \log(n)^{2\cdot h(d)}, \tag{1.16}$$

where $h(d) \leq 3d$ and $C(\Omega)$ is a constant that only depends on the polytope Ω.[11]

Theorem 4 gives a minimax optimal estimator in many natural cases; e.g., it applies when the polytope Ω is assumed to have only $C(d)$ vertices or facets, where $C(d)$ is a constant that only depends on d; this class includes, for instance, the unit cube, the simplex, and the ℓ_1 ball. Then by [McM70], the second term in our bound is of order $\widetilde{O}_d(n^{-\frac{d}{d+4}})$, which is strictly smaller than the minimax rate $\Theta_d(n^{-\frac{4}{d+4}})$.

Remark 6. In the future extended version of the manuscript that this part is based on, we remove the redundant term that depends on Ω. Specifically, we show that

$$\mathcal{R}(\widehat{f}_{L,n}, \mathcal{F}_L(\Omega), \mathbb{P}) \leq (\sigma + L)^2 n^{-\frac{4}{d+4}} \log(n)^{h(d)}$$

for any convex domain $\Omega \subset B_d$. The proof for an arbitrary convex body Ω involves a technical (and quite standard) detour through tedious techniques in stochastic geometry, and is therefore omitted in this version.

Our second main result, concerning bounded convex regression, is proved in the same way as Theorem 4, using the entropy bounds of [GW17].

Theorem 5. *Let $d \geq 5$ and $n \geq d+1$. Then, under Assumption 6, for the task of Γ-bounded convex regression on the polytope $\Omega \subset B_d$, there exists an efficient estimator, $\widehat{f}^{\Gamma,n}$, with runtime of at most $O_d(n^{O(d)})$ such that*

$$\mathcal{R}(\widehat{f}^{\Gamma,n}, \mathcal{F}^{\Gamma}(\Omega), \mathbb{P}) \leq C_1(\Omega)(\sigma + \Gamma)^2 n^{-\frac{4}{d+4}} \log(n)^{h(d)},$$

where $h(d) \leq 3d$ and $C_1(\Omega)$ is a constant that only depends on Ω.

[11] Specifically, it depends on the flags number of polytope Ω (cf. [RSW19]).

As we mentioned earlier, for both of these two tasks the minimax rate is of order $n^{-\frac{4}{d+4}}$, so up to polylogarithmic factors in n, the above estimators are minimax optimal. We note that in Theorem 5, the dependence of the constants on the polytope Ω is unavoidable (differently from Theorem 4). This follows from the results of [GW17; HW16], in which the authors showed the geometry of the support of the measure $\mathbb{P}$ affects the minimax rate of bounded convex regression. For example, in the extreme case $\Omega = B_d$, the minimax rate is of order $n^{-\frac{2}{d+1}}$, which is asymptotically larger than the error rate for polytopes; thus, if we take a sequence of polytopes Ω_n which approaches B_d, the sequence of constants $C(\Omega_n)$ will necessarily blow up.

We consider our results as mainly a proof-of-concept for the existence of efficient estimators for the task of convex regression when $d \geq 5$. Due to their high polynomial runtime, in practice our estimators would probably not work well. However, as we mentioned above, the other minimax optimal estimators in the literature are computationally inefficient, and they all require consideration of some net of exponential size in n; our estimator is conceptually quite different. We hope that insights from our algorithm can be used to construct a practical estimator with the same desirable statistical properties. From a purely theoretical point of view, our estimators are the first known minimax optimal efficient estimators for non-Donsker classes for which their LSE are provably minimax suboptimal in $L_2(\mathbb{P})$. Prior to this work, there were efficient optimal estimators for non-Donsker classes such that their corresponding LSE is provably efficient and optimal (such as log-concave density estimation and isotonic regression, cf. [KDR19; Han+19; Han19; PS22]). Our work should be contrasted with these earlier works. We show that it is possible to overcome the suboptimality of the LSE with an efficient optimal algorithm in the non-Donsker regime - a result that was unknown before this work.

We prove Theorems 4 and 5 in Chapter 4 below, we remark that the proof of Theorem 5 uses the same method as that of Theorem 4, along with the main result of [GW17, Thm 1.1]. We conclude this part with the following remarks:

Remark 7.

1. We conjecture that the estimators of Theorems 4 and 5 are minimax-optimal up to constants that only depend on d, σ, i.e the. $\log(n)$ factors are unnecessary.

2. Our estimators' runtime is of order $n^{O(d)}$, which is significantly higher than the $O_d(n^2)$ runtime of the suboptimal Least Squares.

3. When $d \geq 5$, one can show that when f^* is a max k-affine function (restricted to $P \subset B_d$), i.e. $f^* \mathbb{1}_P(x) = \max_{1 \leq i \leq k} a_i^\top x + b_i$, our estimator attains a parametric rate, i.e.

$$\mathbb{E} \int (\widehat{f}^{\mathrm{T}} - f^*)^2 d\mathbb{P} \leq \widetilde{O}_d \left(\frac{C(P,k)}{n} \right).$$

   When $d \leq 4$, [HW16] showed that LSE attains a parametric rate as well. However, when $d \geq 5$, the LSE attains a non-parametric error of $\widetilde{\Theta}_d(C(P,k)n^{-4/d})$ ([Kur+20]; for a more general result see [KR21]). Therefore, our algorithm has the proper adaptive rates when $d \geq 5$; see [Gho+21] for more details.

4. An interesting property of our estimator is that the random design setting, i.e. the fact that data points $X_1, \ldots, X_n$ are drawn from $\mathbb{P}$ rather than fixed, is essential to its success, a phenomenon not often observed when studying shape-constrained estimators. Usually these estimators also perform well on a "nice enough" fixed design set, for example when $\Omega = [-1/2, 1/2]^d$ and $X_1, \ldots, X_n$ are the regular grid points.

1.4 On the Minimal Error of Empirical Risk Minimization

Suppose a "simple" class $\mathcal{H}$ of models captures the relationship between the covariates X and the response variable Y. Inspired by the use of overparameterized models, we may take a much larger class $\mathcal{H} \subset \mathcal{F}$ for computational or other purposes (such as lack of explicit description of $\mathcal{H}$) and minimize training loss over this larger class.

It is natural to ask whether the learning procedure can *adapt* to the fact that data comes from a simple model $f^* \in \mathcal{H}$, in the sense that the prediction error depends on the statistical complexity of $\mathcal{H}$ rather than $\mathcal{F}$. We do have positive examples of this type: the least squares solution over the class of all convex functions $\mathcal{F}$ on a convex compact subset of $\mathbb{R}^d$ (with $d \leq 4$) automatically enjoys the faster "parametric" rate $\widetilde{O}(1/n)$ of convergence to the true regression function $f^* \in \mathcal{H}$, where $\mathcal{H}$ constitutes the k-max affine convex functions with $k = O(1)$ pieces. This rate should be contrasted with the slow non-parametric rate $\Theta(n^{-4/(d+4)})$ when the true regression function is 'complex' and cannot be approximated well by a piece-wise linear convex function.

How generic is this phenomenon of automatic adaptivity of empirical minimizers to simplicity of the true model? An affirmative answer would lend credibility to the practice of taking large models, whereas a negative answer would necessitate the study of conditions that can make such adaptivity possible.

This papt of this thesis, studies the fundamental limits of adaptivitiy of LSE in the setting of nonparametric regression, in both random and fixed design. In contrast with the standard minimax approach to lower bounds, which may hide the true performance of LSE on simple models, we focus on lower bounds that hold for *any* (rather than the worst-case) regression function in a given class. In the fixed design setting, we show that—informally speaking—for rich classes $\mathcal{F}$, dependence on the global statistical complexity of the class is unavoidable, as it controls the error of LSE for any true regression function f^*, no matter how 'simple' it is. In contrast, in the random design case, the situation is more subtle. Somewhat counter-intuitively, we show that for rich classes $\mathcal{F}$, adaptation to the simplicity of f^* may only be possible if the local neighborhood of f^* in $\mathcal{F}$ is nearly as rich as the class $\mathcal{F}$. This finding can be viewed through the lens of recent results on interpolation [BRT19; BHM18; Bar+20; LR20]. In these papers, the solutions can be seen as 'simple-plus-spiky' [Wyn+17] with spikes responsible for fitting the training data without affecting the error with respect to the population. Since in these models there are enough degrees of freedom to fit any noisy data, the effective function classes have rich local neighborhoods. In such cases, it is still possible that "overfitting" to the training data does not result it large out-of-sample error. Conversely, we show that—again, informally speaking—if f^* is embedded in a local neighborhood in $\mathcal{F}$ with low complexity, the empirical minimizer will necessarily be attracted to a solution far away from f^* with respect to the out-of-sample loss. This finding initially appeared counter-intuitive to the author of this thesis. We discuss and provide the main results (and their proofs) in Chapter 5 below.

Chapter 2

On the Sub-Optimality of Least Squares in the non-Donsker Regime

In this chapter, we will provide our general recipe to prove a lower bound of order $n^{-1/p}$ on the risk of the LSE (as we promised in §1.1). In this chapter, $\asymp, \gtrsim, \lesssim$, denote equality/inequality up to a multiplicative constant that only depends on p (or d in the case the support function of convex functions).

In order to provide a good intuition for this recipe, we begin with describing our approach which motivated our novel technique:

Sketch to our approach: First, as we discussed in the preceding chapter, the minimax rate for a function class $(\mathcal{F}, \mathbb{P})$ is determined by its ϵ-entropy numbers, which is a measure of the *global* geometry of $(\mathcal{F}, \mathbb{P})$. However, the error rate of the LSE on a function f^* is determined by the *local* "complexity" (which will be the Gaussian complexity) of $\mathcal{F}$ in the neighborhood of f^* (cf. [Cha14]). The possible disparity between this local "complexity" for certain choices of f^* and between the global geometry is the engine behind our result. To explain what we mean requires the introduction of an additional notion, the *Gaussian complexity*, which is tightly linked to the metric entropy and to the error of the LSE. As a warm up, we first provide a rough sketch of our recipe.

First, recall that definition of the Gaussian complexity, that is defined for a class $\mathcal{G} \subset \{\mathcal{X} \to \mathbb{R}\}$

and n data points $x_1, \dots, x_n \in \mathcal{X}$, as

$$\mathcal{W}_x(\mathcal{G}) = \mathbb{E} \sup_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(x_i), \tag{2.1}$$

where $\xi_1, \dots, \xi_n \underset{i.i.d.}{\sim} N(0,1)$. We denote the t-neighbourhood around f_0 (in terms of $L_2(\mathbb{P}^{(n)})$) by

$$B_n(f_0, t) := \{f \in \mathcal{F} : \|f - f_0\|_n \leq t\}.$$

Our main tool for lower-bounding the risk of the LSE is a result of [Cha14] which states that the error of the LSE on an underlying function f^* is given by ϵ_{LSE}^2, where

$$\epsilon_{LSE} := \operatorname{argmax}_{\epsilon \geq 0} H_{f^*}(\epsilon), \tag{2.2}$$

where

$$H_{f^*}(\epsilon) = \mathcal{W}_x(B_n(f^*, \epsilon)) - \frac{\epsilon^2}{2}. \tag{2.3}$$

It turns out that the function $H_{f^*}(\epsilon)$ is strictly concave – indeed, it is easy to see that $\mathcal{W}_x(B_n(f^*, \epsilon))$ is concave by the convexity of $\mathcal{F}$, and $-\frac{\epsilon^2}{2}$ is certainly concave.

Recall that the entropy numbers of a non-Donsker class (see Definition 5 above) behaves as $\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}^{(n)}) \asymp \epsilon^{-p}$ for some $p > 2$; and the minimax rate is given by $\epsilon_*^2 \asymp n^{-\frac{2}{p+2}}$. Hence, in order to show that the LSE is suboptimal on a given class $\mathcal{F}$, we must find some function $f^* \in \mathcal{F}$ such that the ϵ maximizing (2.2) is asymptotically larger than $n^{-\frac{1}{p+2}}$.

Our strategy for doing this is as follows: we find a function $f_0 \in \mathcal{F}$ whose local complexity is "large" at the scale $n^{-1/p}$, i.e., $\mathcal{W}_x(B_n(f_0, n^{-1/p})) \gtrsim n^{-1/p}$, and another function $f_{(n)}$ at distance of order $n^{-1/2p}$ from f_0 such that $\mathcal{W}_x(B_n(f_{(n)}, t)) \lesssim n^{-1/2p} t$ for $t \leq c_1 n^{-1/2p}$. (This sketch is valid up to polylogarithmic factors in the number of samples.)

For large enough $C > 0$, the $C \cdot n^{-1/2p}$-neighborhood of $f_{(n)}$ contains an $n^{-1/p}$-neighborhood of f_0 and hence its width is bounded below by $\mathcal{W}_x(B_n(f, n^{-1/p})) \gtrsim n^{-1/p}$. At first glance, this does not seem to control the behavior of the function

$$H_{f_{(n)}}(\epsilon) = \mathcal{W}_x(B_n(f_{(n)}, \epsilon)) - \frac{\epsilon^2}{2}$$

for $\epsilon \asymp n^{-1/2p}$, since $n^{-1/p}$ and $(n^{-1/2p})^2$ have the same asymptotic order. However, using the fact that the function $t \mapsto \mathcal{W}_x(B_n(f_{(n)}, t))/t$ is decreasing (under Assumption 1), we are able to obtain that $H_{f_{(n)}}(C_1 n^{-1/2p}) \geq C_2 n^{-1/p}$ for some other constants C_1, C_2.

On the other hand, using the lower bound $\mathcal{W}_x(B_n(f_{(n)}, t)) \leq C_3 n^{-1/2p} t$ for $t \leq c n^{-1/2p}$ one obtains

$$H_{f_{(n)}}(c_2 n^{-1/2p}) \leq \mathcal{W}_x(B_n(f_{(n)}, C \cdot n^{-1/2p})) \leq C_3 n^{-1/2p} \cdot c_2 n^{-1/2p} = c_2 C_3 n^{-1/p}$$

and choosing c_2 small enough one obtains that $H_{f_{(n)}}(c_2 n^{-1/2p}) \leq H_{f_{(n)}}(C_2 n^{-1/2p})$, which implies by the concavity of $H_{f_{(n)}}$ that $\operatorname{argmax}_\epsilon H_{f_{(n)}}(\epsilon) \geq c_2 n^{-1/2p} \gg n^{-1/(p+2)}$, and hence that the squared error of the LSE on $f_{(n)} \in \mathcal{F}$ is asymptotically larger than the minimax rate, as claimed.

We make two further remarks to guide the reader before proceeding to the formal statement of our assumptions and results. First, in order to study the Gaussian complexity of different subsets of $\mathcal{F}$, we use standard inequalities – Sudakov's and Dudley's inequalities – which connect the Gaussian complexity of a set to its entropy numbers; it is often much easier to compute the entropy numbers of a set than to bound the Gaussian complexity directly. Thus, the assumptions under which our general theorem (Theorem 6) holds are phrased in terms of entropy numbers.

Second, in the application of our general results to the class of support functions of convex bodies, the choice of f_0 and $f_{(n)}$ is quite natural – f_0 is simply the support function of the Euclidean ball, while $f_{(n)}$ will be the best polytopal approximation to the ball with $\sqrt{n}$-vertices. This concludes the sketch of our approach.

2.1 The Sub-Optimality Recipe

We start by stating assumptions that imply sub-optimality of LSE in a fixed design. Then, we extend these assumption to the random design setting. Later, these assumptions will be verified in the case of support functions of convex sets in $\mathbb{R}^d$.

Assumption 7. *[S.O. Assumptions: Fixed design] Let $p > 2$ and $n \gtrsim 1$. Assume there are two functions $f_0, f_{(n)} \in \mathcal{F}$ such that for some $\beta, \gamma \geq 0$,*

$$\log \mathcal{N}(\epsilon, B_n(f_0, 2\epsilon), \mathbb{P}^{(n)}) \gtrsim \epsilon^{-p} \quad \forall\, \epsilon \gtrsim n^{-1/p} \log(n)^{\gamma}. \tag{2.4}$$

Furthermore, suppose

$$\|f_{(n)} - f_0\|_n \lesssim n^{-\frac{1}{2p}} \log(n)^{-s_{p,\gamma}}, \tag{2.5}$$

where $s_{p,\gamma} = \frac{\gamma}{2}(\frac{p}{2} - 1)$, and either

$$\log \mathcal{N}(\epsilon, B_n(f_{(n)}, t), \mathbb{P}^{(n)}) \lesssim \sqrt{n} \log(n)^{p \cdot s_{p,\gamma}} \left(\log \frac{1}{\epsilon}\right)^{\beta} \left(\frac{t}{\epsilon}\right)^{p} \tag{2.6}$$

for every $n^{-\frac{1}{p}} \log(n)^{-s_{p,\gamma}} \lesssim \epsilon \leq t \lesssim n^{-\frac{1}{2p}} \log(n)^{-\frac{\beta}{p} - \frac{3s_{p,\gamma}}{2}}$ or a weaker condition

$$\mathcal{W}_x(B_n(f_{(n)}, t)) \lesssim n^{-\frac{1}{2p}} \log(n)^{\frac{\beta}{p} + s_{\gamma,p}} t, \tag{2.7}$$

for some $t \lesssim n^{-\frac{1}{2p}} \log(n)^{-\frac{\beta}{p} - \frac{3s_{p,\gamma}}{2}}$ holds.

The following theorem establishes that under the aforementioned assumptions, the rate of convergence of the LSE is bounded below by $n^{-1/p}$, up to a poly-logarithmic multiplicative factor.

Theorem 6. *Let $n \geq 1$ and $(\mathcal{F}, \mathbb{P}^{(n)})$. Then, under Assumptions 1,2,3, and Assumption 7 for some $p > 2$, the following holds:*

$$\mathcal{R}(\widehat{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) \gtrsim n^{-\frac{1}{p}} \log(n)^{-\frac{2\beta}{p} - 3s_{p,\gamma}}.$$

We shall next extend Theorem 6 to the random design setting. Recall that in this setting, the design points $X_1, \dots, X_n \underset{i.i.d.}{\sim} \mathbb{P}$ and we denote by $\mathbb{P}_n$ the random empirical measure of $X_1, \dots, X_n$ and $\|\cdot\|_n$ denotes the $L_2(\mathbb{P}_n)$ and $\|\cdot\|$ denotes the $L_2(\mathbb{P})$ norm.

It is sufficient for our purposes to establish a quasi-isometry with high probability, i.e.

$$\forall f, g \in \mathcal{F}, \quad 2^{-1}\|f-g\|_n - d_n \leq \|f-g\| \leq 2\|f-g\|_n + d_n, \tag{2.8}$$

for some remainder d_n that decays to zero with increasing n. In the case of uniformly bounded functions, sufficient conditions for the two-sided inequality (2.8) can be found in the literature on local Rademacher averages (see e.g., [Bou02; BBM05]), while the right-hand side of (2.8) holds under weaker conditions [KM13; Men14; Men17]. For the purpose of proving lower bounds for random design, however, we need the more demanding left-hand side of (2.8). Hence, we shall assume that functions in $\mathcal{F}$ are uniformly bounded.

In addition, we assume the following:

Assumption 8. *The growth of Koltchinskii-Pollard entropy for all $\epsilon \in (0,1)$ satisfies the following:*

$$\sup_{n \in \mathbb{N}} \sup_{\mathbb{Q} \in \mathcal{P}_n} \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{Q}) \underset{(\mathcal{F}, \mathbb{P})}{\asymp} \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}) \tag{2.9}$$

where $\mathcal{P}_n$ denotes the set of all probability measures supported on finite subsets of $\mathcal{X}$ of cardinality at most n. Under this assumption, with high probability $d_n \lesssim (\log n)^2 n^{-1/p}$ [RST17]. This will allow us to reduce the random design setting to a fixed design and use Theorem 6, since the remainder $(d_n)^2$ is of the lower order than $n^{-1/p}$.

To establish a lower bound for LSE in random design, we may verify that above assumptions (2.4), (2.5), (2.6) hold with high probability for random measures $\mathbb{P}_n$ and employ near-isometry for the distance between LSE and the regression function. Alternatively, it may be easier to verify the corresponding assumptions in the population. We now state this latter approach.

Assumption 9. *[S.O. Assumptions: Random Design] Let $p > 2$. Assume there exists a function $f_0 \in \mathcal{F}$ such that for every integer $m \gtrsim 1$, there exists $f_{(m)} \in \mathcal{F}$ such that*

$$\log \mathcal{N}(\epsilon, B(f_0, 2\epsilon), \mathbb{P}) \gtrsim \epsilon^{-p} \quad \forall \epsilon \in (0,1) \tag{2.10}$$

and

$$\|f_{(m)} - f_0\| \lesssim m^{-\frac{1}{p}}. \tag{2.11}$$

Furthermore, for some $\beta \geq 0$

$$\log \mathcal{N}_{[]}(\epsilon, B(f_{(m)}, t), \mathbb{P}) \lesssim m \left(\log \frac{1}{\epsilon}\right)^{\beta} \left(\frac{t}{\epsilon}\right)^{p}, \tag{2.12}$$

for every $0 < \epsilon \leq t \lesssim \log(m)^{-\beta/p} m^{-1/p}$.

We remark that Assumption (2.10) is satisfied for non-Donsker classes of functions (cf. [YB99, Lemma 3] for the proof of existence). The following is our lower bound for the random-design setting:

Corollary 1. *Let* $n \geq 1$ *and* $(\mathcal{F}, \mathbb{P})$*. Then, under Assumptions 1,2,4,8 and Assumption 9 for some* $p > 2$*, the following holds:*

$$\mathcal{R}(\widehat{f}_n, \mathcal{F}, \mathbb{P}) \gtrsim n^{-\frac{1}{p}} \log(n)^{-\frac{2\beta}{p} - 3(\frac{p}{2} - 1)}.$$

Remark 8. In some cases it may be possible to relax Assumptions 8 and 4 in Corollary 1. For example, when $(\mathcal{F}, \mathbb{P})$ satisfies a small-ball condition or L_2-L_4 entropy condition [Men14; Kur+20]).

The ideas behind our assumptions: Equation (2.10) says that the neighborhood around f_0 has high "complexity". Next, Equation (2.11) states that f_0 is approximated by $f_{(n)}$ up to the accuracy $n^{-1/p}$ for every n. Finally, Equation (2.12) captures the "simplicity" of the functions $f_{(n)}$ in relation to the complex function f_0 satisfying (2.10). Note that when $t = O(\epsilon)$, the right hand side of (2.12) is logarithmic in ϵ (assuming that n is not too large) while the right hand side of (2.10) is polynomial in $1/\epsilon$. Thus the local neighborhood of $f_{(n)}$ is smaller than that of f_0 and in this sense $f_{(n)}$ is simpler than f_0. Note also that there is a factor of m on the right hand side of (2.12) which means that the complexity of the functions $f_{(n)}$ increases with n. There exist natural non-Donsker function classes $\mathcal{F}$ which satisfy (2.10), (2.11) and (2.12). For example, we show that the class of support functions of compact convex sets in $\mathbb{R}^d$ satisfies these properties with f_0 being the support function of the unit ball, and $f_{(m)}$ the support function of a regular polytopal approximation to the unit ball with m vertices. In a twin paper of the author of this thesis [Kur+20], provide more classes that satisfy these properties.

The proof of Theorem 6 will reveal that $\widehat{f}_n$ achieves the rate $n^{-1/p}$ (up to logarithmic factors) when the regression function is in the class $\{f_{(m)}, m \geq 1\}$. In fact, the specific function achieving the rate $n^{-1/p}$ is $f^* = f_{(m)}$ for $m \asymp \sqrt{n}$. It is interesting to note that for $m \asymp \sqrt{n}$ and $t \asymp n^{-1/2p}$, the local neighborhood $B_n(f_{(m)}, t)$ has the same metric entropy as the right hand side of (2.10). Our main technical insight is that the sub-optimality occurs at a function $f^* = f_{(m)}$ instead of perhaps a more natural candidate function such as f_0 (we actually believe that the rate at f_0 may be equal to $\Theta_p(n^{-2/(2+p)})$).

2.1.1 The non-Donsker regime – revisited

The recent results on families in the non-Donsker regime [KDR19; Han19; Car+18; Kur+20; Han+19] and the our results, the LSE may be optimal or sub-optimal in the non-Donsker regime for natural classes of functions, even if uniformly bounded. These results also indicate that LSE achieves a risk that equals to the Gaussian complexity of the class, up to a constant that depends on dimension. Namely, in contrast to the Donsker regime, there is no localization. In the problem of convex regression with uniformly bounded functions and *Euclidean ball* as domain, as well as in multiple isotopic regression, the minimax rate equals to the Gaussian complexity of the family. In contrast, by Theorem 3, for support function regression and for convex uniformly bounded regression (or Lipshitz-convex regression) with support on the *cube* [Kur+20], the Gaussian complexity differs from the minimax rate.

2.2 Proofs

Notation Throughout this text, c, C with subscripts are positive absolute constants that do not depend on the dimension d. Additionally, positive constants that only depend on the dimension or on p are explicitly denoted, respectively, by $c(d), C(d), c(p), C(p)$. These constants may change from line to line, and from section to section $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot)$ denotes behavior up to logarithmic factors in n or ϵ. Also, remember the notation $\|\cdot\|_2$, which represents the Euclidean norm in $\mathbb{R}^d$.

2.2.1 Proof of Theorem 6

Our main technical tool is the following important result of [Cha14] which gives sharp upper and lower bounds for the accuracy of an LSE over a convex family of functions in the fixed-design setting. We use the following notation in this result.

Theorem 7. *[[Cha14, Theorem 1.1]] Let $\mathcal{F}$ be a convex family of functions and consider the LSE for the fixed design setting. Let*

$$t_f := \text{argmax}_{t \geq 0} H_f(t) \quad \textit{where } H_f(t) := \mathcal{W}_x(B_n(f,t)) - \frac{t^2}{2}. \tag{2.13}$$

Then, t_f is unique, $H_f(\cdot)$ is a concave function, and we have

$$\Pr\left\{0.5t_f^2 \leq \|\widehat{f}_n - f\|_n^2 \leq 2t_f^2\right\} \geq 1 - 3p_n, \tag{2.14}$$

where $p_n = \exp\left(-cnt_f^2\right)$.

The proof of Theorem 6 is reduced to the following "two points" lemma that is mainly based on Theorem 7. The notation $f_{(n)}$ is to emphasize the fact that $f_{(n)}$ is a function that is chosen based on the number of samples.

Lemma 1. *[Lower bound for fixed design] Let $f_0, f_{(n)} \in \mathcal{F}$, and $r_n \geq Cn^{-1/4}, \delta_n, w_n, s_n$ be positive constants. Also, let $w_n' = \max\{\frac{r_n^2}{8\delta_n}, w_n\}$, and assume the following:*

- $\mathcal{W}_x(B_n(f_0, s_n)) \geq r_n^2$
- $\|f_{(n)} - f_0\|_n \leq \delta_n$*, and* $r_n \leq 2\delta_n$, $s_n \leq \delta_n$.
- $\mathcal{W}_x(B_n(f_{(n)}, t)) \leq w_n' t$ *for some* $t \leq \frac{r_n^4}{16\delta_n^2 w_n'}$.

Then, when $Y_i = f_{(n)}(x_i) + \xi_i$, the following holds with high probability

$$\|\widehat{f}_n - f_{(n)}\|_n^2 \geq c \cdot \min\left\{ \frac{r_n^8}{\delta_n^4 w_n^2}, \frac{r_n^4}{\delta_n^2} \right\}.$$

Proof of Lemma 1. By Theorem 7, it is enough to show that $t_{f_{(n)}}$ is greater than $\frac{r_n^4}{16\delta_n^2 w_n}$. Namely, the functional $H_{f_{(n)}}(\cdot)$ attains its unique maximum on a $t_{f_{(n)}} \geq \frac{r_n^4}{16\delta_n^2 w_n}$. First, using the second assumption and the upper bound on δ_n, we see that

$$B_n(f_0, s_n) \subseteq B_n(f_{(n)}, \|f_{(n)} - f_0\|_n + s_n) \subseteq B_n(f_{(n)}, 2\delta_n).$$

By the convexity of $\mathcal{F}$, the function $t \mapsto \mathcal{W}_x(B_n(f_{(n)}, t))/t$ is nonincreasing. Therefore, by the first assumption and the above inclusion, for all $t \leq 2\delta_n$

$$\mathcal{W}_x(B_n(f_{(n)}, t)) \geq \frac{r_n^2}{2\delta_n} t.$$

Now, by the last equation and the second assumption, we know that for $t_1 = \frac{r_n^2}{2\delta_n} \leq 2\delta_n$

$$H_{f_{(n)}}(\frac{r_n^2}{2\delta_n}) = \mathcal{W}_x(B_n(f_{(n)}, \frac{r_n^2}{2\delta_n})) - \frac{\left(\frac{r_n^2}{2\delta_n}\right)^2}{2} \geq \frac{r_n^2}{2\delta_n} \cdot \left(\frac{r_n^2}{2\delta_n}\right) - \frac{\left(\frac{r_n^2}{2\delta_n}\right)^2}{2} \geq \frac{r_n^4}{8\delta_n^2}.$$

Finally, by the last assumption, we know that for $t_2 = \frac{r_n^4}{16\delta_n^2 w_n'} < t_1$ the following holds:

$$H_{f_{(n)}}(\frac{r_n^4}{16\delta_n^2 w_n'}) \leq \mathcal{W}_x(B_n(f_{(n)}, \frac{r_n^4}{16\delta_n^2 w_n'})) \leq w_n' \frac{r_n^4}{16\delta_n^2 w_n'} = \frac{r_n^4}{16\delta_n^2}.$$

Since $H_{f_{(n)}}(\cdot)$ is concave in t, we conclude by the last two equations that $t_{f_{(n)}} \geq \frac{r_n^4}{16\delta_n^2 w_n'}$ and the claim follows from Theorem 7. □

In addition to Lemma 1, we need the following two standard facts (which can be found, for example, in [Kol11]) to prove Theorem 6.

Lemma 2 (Sudakov Minoration)**.** *There exists a universal positive constant* $c \geq 0$ *such that the following holds for any class* $\mathcal{F}$ *of real-valued functions:*

$$\mathcal{W}_x(\mathcal{F}) \geq c \cdot \sup_{\epsilon > 0} \epsilon \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}^{(n)})/n}.$$

Lemma 3 (Dudley Integral)**.** *There exists a universal positive constant* $C \geq 0$ *such that the following holds for any class* $\mathcal{F}$ *of real-valued functions:*

$$\mathcal{W}_x(\mathcal{F}) \leq C \cdot \inf_{\epsilon > 0} \left(\epsilon + \frac{1}{\sqrt{n}} \int_{\epsilon}^{\mathrm{Diam}_{\mathbb{P}^{(n)}}(\mathcal{F})/2} \sqrt{\log \mathcal{N}(u, \mathcal{F}, \mathbb{P}^{(n)})} du \right),$$

where $\mathrm{diam}_{\mathbb{P}^{(n)}}(\mathcal{F})$ *is the diameter of* $\mathcal{F}$ *with respect to* $L^2(\mathbb{P}^{(n)})$.

Proof of Theorem 6. We aim to apply Lemma 1. In order to satisfy its first assumption, we need to estimate the Gaussian width of $B_n(f_0, C_3 n^{-1/p} \log(n)^{\gamma})$. By Lemma 2 and Eq. (2.10)

$$\begin{aligned} \mathcal{W}_x(B_n(f_0, C_3 n^{-1/p} \log(n)^{\gamma})) &\geq \frac{C}{\sqrt{n}} \sup_{\epsilon \geq C_1 n^{-1/p} \log(n)^{\gamma}} \left\{ \epsilon \sqrt{\log \mathcal{N}(\epsilon, B_n(f_0, C_3 n^{-1/p} \log(n)^{\gamma}), \mathbb{P}^{(n)})} \right\} \\ &\geq c_3(p) n^{-\frac{1}{p}} \log(n)^{-\gamma(\frac{p}{2}-1)}, \end{aligned} \tag{2.15}$$

where we chose $\epsilon = C_1(p) n^{-1/p} \log(n)^{\gamma}$. Now, we aim to upper bound $\mathcal{W}_x(B_n(f_{(n)}, t))$ when

$$t = c_1(p) n^{-\frac{1}{2p}} \log(n)^{-\frac{\beta}{p} - \frac{3\gamma}{2}(\frac{p}{2}-1)},$$

for some suitable $c_1(p)$. By using Lemma 3, and Eq. (2.11), we see that

$$\begin{aligned}\mathcal{W}_x(B_n(f_{(n)},t)) &\leq C \inf_{\epsilon>0}\left(\epsilon + \frac{1}{\sqrt{n}}\int_{\epsilon}^{t}\sqrt{\log\mathcal{N}(u, B_n(f_{(n)},t), \mathbb{P}^{(n)})}du\right)\\ &\leq C \inf_{\epsilon>0}\left(\epsilon + \frac{C_2(p)}{\sqrt{n}}\int_{\epsilon}^{t}\left(\frac{tn^{\frac{1}{2p}}\log(n)^{\frac{\gamma}{2}(\frac{p}{2}-1)}\log(u^{-1})^{\frac{\beta}{p}}}{u}\right)^{\frac{p}{2}}du\right)\\ &\underset{(*)}{\leq} C \inf_{\epsilon>0}\left(\epsilon + \frac{C_3(p)}{\sqrt{n}}\left(\frac{tn^{\frac{1}{2p}}\log(\epsilon^{-1})^{\frac{\beta}{p}}\log(n)^{\frac{\gamma}{2}(\frac{p}{2}-1)}}{\epsilon}\right)^{\frac{p}{2}}\epsilon\right)\\ &\leq C_4(p)n^{-\frac{1}{2p}}\log(n)^{\frac{\beta}{p}+\frac{\gamma}{2}(\frac{p}{2}-1)}t,\end{aligned} \tag{2.16}$$

where in $(*)$ we used the fact that $p > 2$, which implies that $\int_{\epsilon}^{t}\left(\frac{tn^{\frac{1}{2p}}\log(n)^{\frac{\gamma}{2}(\frac{p}{2}-1)}\log(u^{-1})^{\frac{\beta}{p}}}{u}\right)^{\frac{p}{2}}du \leq C_1\int_{\epsilon}^{2\epsilon}\left(\frac{tn^{\frac{1}{2p}}\log(n)^{\frac{\gamma}{2}(\frac{p}{2}-1)}\log(u^{-1})^{\frac{\beta}{p}}}{u}\right)^{\frac{p}{2}}du$; and set

$$\epsilon = C_1(p)n^{-\frac{1}{2p}}\log(n)^{\frac{\beta}{p}+\frac{\gamma}{2}(\frac{p}{2}-1)}t = C_2(p)n^{-\frac{1}{p}}\log(n)^{-\gamma(\frac{p}{2}-1)}.$$

Hence, it is enough to assume (2.7) in place of (2.6). Finally, we can apply Lemma 1 with the following parameters:

$$\|f_{(n)} - f_0\|_n \leq c_1(p)n^{-\frac{1}{2p}}\log(n)^{-\frac{\gamma}{2}(\frac{p}{2}-1)} =: \delta_n,$$

where we used Eq. (2.6). Also, we set $s_n = C_5(p)n^{-1/p}\log(n)^{\gamma}$,

$$r_n^2 = \min\{c_3(p), c_1(p)/8\}n^{-\frac{1}{p}}\log(n)^{-\gamma(\frac{p}{2}-1)}$$

and

$$w_n = C_4(p)n^{-\frac{1}{2p}}\log(n)^{\frac{\beta}{p}+\frac{\gamma}{2}(\frac{p}{2}-1)}.$$

Therefore, Lemma 1 gives that

$$\mathcal{R}(\widehat{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) \geq c_5(p) \left(\frac{n^{-\frac{1}{p}} \log(n)^{-\gamma(\frac{p}{2}-1)}}{n^{-\frac{1}{2p}} \log^{\frac{\beta}{p}+\frac{\gamma}{2}(\frac{p}{2}-1)}(n)} \right)^2 \gtrsim n^{-\frac{1}{p}} \log(n)^{-\frac{2\beta}{p}-3\gamma(\frac{p}{2}-1)},$$

and the claim follows. □

2.2.2 Proof of Corollary 1

The following is an almost immediate consequence of [Gee00, Thm 5.11] (see §2.3 below).

Lemma 4. *Let $\mathcal{G}$ be a family of functions uniformly bounded by one, and $\mathbb{P}$ be some probability measure. For any $f \in \mathcal{G}$ and $t \geq 0$, let $\varepsilon(f,t)$ be the stationary point of*

$$\frac{1}{\sqrt{n}} \int_{\frac{\delta}{c}}^{t} \sqrt{\log \mathcal{N}_{[]}(u, B(f,t), \mathbb{P})} du = \delta$$

for some absolute constant $c > 0$. Then, for any $\varepsilon \geq \varepsilon(f,t)$ the following holds with probability of at least $1 - C\exp(-cn\varepsilon^2)$:

$$\mathcal{W}_x(B(f,t)) \lesssim \varepsilon + t \cdot n^{-1/2}.$$

Proof of Corollary 1. First, we take $m = C(p)\sqrt{n}\log(n)^{\frac{p\gamma}{2}(\frac{p}{2}-1)}$ in (2.11) and set $f_{(n)} := f_{(m)}$. By using Eq. (2.8) and the estimate $d_n \lesssim (\log n)^2 n^{-1/p}$ from [RST17], we see that with high probability,

$$\|f_{(n)} - f_0\|_n \leq 2\|f_{(n)} - f_0\| + C\log^2(n) n^{-\frac{1}{p}} \lesssim n^{-\frac{1}{2p}} \log(n)^{\frac{\gamma}{2}(\frac{p}{2}-1)}.$$

Thus Eq. (2.5) holds for $f_{(n)}$. Next, by using Eq. (2.8), it is easy to see that Eq. (2.4) hold with $\gamma = 2$. It remains to show that (2.7) holds.

First, by (2.8), we know that for $t \geq Cd_n$ the following holds:

$$\mathcal{W}_x(B_n(f_{(n)}, t)) \leq \mathcal{W}_x(B(f_{(n)}, 2t + d_n)) \leq \mathcal{W}_x(B(f_{(n)}, 4t)).$$

To estimate the last term, we use our bracketing assumption along with the uniform boundedness of functions in $\mathcal{F}$. These assumptions let us apply Lemma 4 that yields an upper bound on this term. In order to upper bound $\varepsilon(f_{(n)}, 4t)$ (the fixed point of Lemma 4), we use Eq. (2.12) and derive that

$$\frac{1}{\sqrt{n}} \int_{\frac{\varepsilon}{c}}^{4t} \left(\frac{t n^{\frac{1}{2p}} \log(n)^{\frac{\gamma}{2}(\frac{p}{2}-1)} \log(u^{-1})^{\frac{\beta}{p}}}{u} \right)^{\frac{p}{2}} du \lesssim \frac{1}{\sqrt{n}} \left(\frac{t n^{\frac{1}{2p}} \log(n)^{\frac{\beta}{p}} \log(n)^{\frac{\gamma}{2}(\frac{p}{2}-1)}}{\varepsilon} \right)^{\frac{p}{2}} \varepsilon$$

where in the last inequality we used the fact that $p > 2$ and $\varepsilon \geq 1/n$. Setting the right-hand side equal to ε, we find an upper bound on the fixed point

$$\varepsilon(f_{(n)}, 4t) \lesssim n^{-\frac{1}{2p}} \log^{\frac{\beta}{p} + \frac{\gamma}{2}(\frac{p}{2}-1)}(n) t.$$

We now set $t \asymp n^{-\frac{1}{2p}} \log(n)^{-\frac{\beta}{p} - \frac{3 s_{p,\gamma}}{2}}$ and apply Lemma 4 with $\varepsilon \asymp n^{-1/p} \log(n)^{-\gamma(p/2-1)}$, concluding that with probability of at least $1 - C\exp(-\widetilde{\Omega}(n^{1-2/p}))$

$$\mathcal{W}_x(B(f_{(n)}, 4t)) \lesssim \varepsilon \asymp n^{-\frac{1}{2p}} \log^{\frac{\beta}{p} + \frac{\gamma}{2}(\frac{p}{2}-1)}(n) t.$$

Hence, all the assumptions of Theorem 6 hold with high probability. Therefore, by Eq. (2.8) and Theorem 6, we conclude that

$$\begin{aligned} \|\widehat{f}_n - f^*\|^2 &\geq \frac{1}{4}\|\widehat{f}_n - f^*\|_n^2 - 2d_n^2 \geq \frac{1}{8}\|\widehat{f}_n - f^*\|_n^2 \\ &\gtrsim n^{-\frac{1}{p}} \log(n)^{-\frac{2\beta}{p} - 3\gamma(\frac{p}{2}-1)}, \end{aligned}$$

and the claim follows. □

2.2.3 Proof of Theorem 3

Throughout this subsection $\mathbb{P} := \mathrm{U}(\mathbb{S}^{d-1})$. In this proof we will use the following auxiliary lemmas:

Lemma 5 ([Bro76; Dud99]). *The following holds for all* $0 < \epsilon < 1$,

$$c_1 2^{-d} \epsilon^{-\frac{d-1}{2}} \leq \log \mathcal{N}_2 \left(\epsilon, \mathcal{C}_d(1), \mathbb{P}\right) \leq \log \mathcal{N}_\infty \left(\epsilon, \mathcal{C}_d(1)\right) \leq C d^{5/2} \epsilon^{-\frac{d-1}{2}},$$

where $\log \mathcal{N}_\infty \left(\epsilon, \mathcal{C}_d(1)\right)$ *denotes covering with respect to* $\| \cdot \|_\infty$.

Lemma 6 ([Bro76; Dud99]). *Let* $h_{B_d} : \mathbb{S}^{d-1} \to \mathbb{R}$ *be the support function of the unit ball in* $\mathbb{R}^d$, *namely the constant function* $h_{B_d} \equiv 1$. *Then for any* $0 \leq \epsilon \leq c$, *there exists a set of cardinality* $M(\epsilon, d) \geq 2^{10^{-d} \epsilon^{-(d-1)/2}}$ *of convex functions on* $\mathbb{S}^{d-1}$, *denoted by* $h_1, \ldots, h_{K_{M(\epsilon,d)}}$, *that has the following properties:*

- $\forall\, 1 \leq i < j \leq M(\epsilon, d)$: $\|h_{K_i} - h_{K_j}\| \geq \epsilon$.
- $\forall\, 1 \leq i \leq M(\epsilon, d)$: $\|h_{K_i} - h_{B_d}\| \leq 8\epsilon$.

Definition 6. *We say that* $f : \mathbb{R}^d \to \mathbb{R}$ *is a* k-piecewise simplicial linear *if* f *is a convex piecewise linear function and its support can be written as a union of k-simplicials, and* f *is linear in each of them.*

For the following lemma, let $\mathcal{P}_d$ be the regular simplex with $d+1$ faces that contains the unit Euclidean ball, with vertices at distance d from the origin, and let

$$\mathcal{F}(\Gamma, \mathcal{P}_d) := \{ f : \mathcal{P}_d \to \mathbb{R} : \ f \text{ is convex function and } \|f\|_\infty \leq \Gamma \ \}$$

Lemma 7 (Theorem 4.4 in [Kur+20]). *Let* $f_k \in \mathcal{F}(\Gamma, \mathcal{P}_d)$ *be k piecewise simplicial linear. Then, the following bound holds for* $0 \leq \epsilon \leq t$:

$$\mathcal{N}_{[]}(\epsilon, B_{\mathrm{Unif}}(f_k, t), \mathrm{Unif}) \leq C(d) \left(t/\epsilon\right)^{d/2} \log^{d+1}\left(\epsilon^{-1}\right) \log(\Gamma),$$

where $B_{\mathrm{Unif}}(f_k, t) \subset \mathcal{F}(\Gamma, \mathcal{P}_d)$ and Unif *denotes the normalized Lebesgue volume measure on $\mathcal{P}_d$.*

In order to prove Corollary 3, we first prove that when we restrict the LSE to $\mathcal{C}(C_1 \cdot d)$ (here C_1 is some absolute constant) we achieve the desired bound:

Lemma 8. *For every $d \geq 6, n \geq C(d)$, the following holds*

$$\operatorname{argmin}_{K \in \mathcal{C}(C_1 \cdot d)} \sum_{i=1}^{n} (Y_i - h_K(X_i))^2 \geq c(d) \log^{\gamma_d}(n) n^{-\frac{2}{d-1}}, \tag{2.17}$$

whenever $K^* \in \mathcal{C}_d(1)$.

Then, in Subsection 2.3.2, we prove that with high probability,

$$\operatorname{argmin}_{K \in \mathcal{C}} \sum_{i=1}^{n} (Y_i - h_K(X_i))^2 = \operatorname{argmin}_{K \in \mathcal{C}(C_1 \cdot d)} \sum_{i=1}^{n} (Y_i - h_K(X_i))^2$$

whenever $K^* \in \mathcal{C}_d(1)$, and therefore Corollary 3 follows.

Proof of Lemma 8. Let $h_K \in \mathcal{C}_d(1)$ and $t > 0$, denote by

$$B(h_K, t) := \{L \in \mathcal{C}(C_1 \cdot d) : \|h_K - h_L\| \leq t\}.$$

We will need the following Lemma (its proof appears in Section 2.3.1).

Lemma 9 (The regular polytope lemma)**.** *For every $m \geq 4^d, d \geq 2$ there exists a polytope $P_{m,d} \subset B_d$ with m vertices that satisfies the following:*

1. $\|h_{P_{m,d}}(x) - 1\|_\infty \leq C_1 m^{-\frac{2}{d-1}}$
2. $\log \mathcal{N}_{[]}(\epsilon, B_{\mathbb{P}}(h_{P_{m,d}}, t), \mathbb{P}) \leq C(d) m \log(\epsilon^{-1})^d (t/\epsilon)^{(d-1)/2}$ *for all* $0 < \epsilon \leq t$.

Now, observe that the family is uniformly bounded by $C_1 \cdot d$. Therefore, we can invoke Corollary 1 with $f_{(m)} = h_{P_{m,d}}$, $\beta = d$, $f_0 = h_{B_d}$, and $p = \frac{d-1}{2}$. By Lemmas 5, 6, and 9, we satisfy all the corollary's assumptions (observe that an upper bound on ∞-covering implies that Koltchinski-Pollard entropy satisfies the same upper bound). Thus $\mathcal{R}(\hat{h}_{\hat{K}_n}, \mathcal{C}_d(1), \mathbb{P}) \geq \widetilde{\Omega}(n^{-\frac{2}{d-1}})$, and the claim follows. □

2.3 Loose Ends and Missing Parts

Proof of Lemma 4. Under the conditions of the Lemma, [Gee00, Thm 5.11] implies that for the fixed point ε,

$$\sup_{g\in B(f,t)} \left|\mathbb{P}_n(g) - \mathbb{P}(g)\right| \lesssim \varepsilon,$$

with probability of at least $1 - C\exp(-c(n(\varepsilon/t)^2)) \geq 1 - C\exp(-cn\varepsilon^2)$ since $t \leq 1$ due to uniform boundedness. Now, using de-symetrization argument (e.g. [Kol11]), we know that

$$\mathbb{E}_{\overline{X}}\mathcal{W}_x(B(f,t)) - \frac{t}{\sqrt{n}} \lesssim \mathbb{E}\sup_{g\in B(f,t)} \left|\mathbb{P}_n(g) - \mathbb{P}(g)\right| \lesssim \varepsilon.$$

Therefore, in expectation we have the desired bound. Since the class in uniformly bounded by one, we can apply Adamzacks's inequality (e.g. [Kol11]) which gives

$$\Pr\left(\mathcal{W}_x(B(f,t)) - C_1(\varepsilon + t/\sqrt{n}) \geq u\right) \leq \exp(-cnu^2)$$

and by setting $u = C(\varepsilon + (t/n)^{1/2})$ the claim follows. □

Next, we will use the following lemma:

Lemma 10 (Jacobian of the radial function, see for example [SW08])**.** *Let $H = \{x \in \mathbb{R}^d \,:\, x^T u = h\}$ be a $d-1$ hyper-plane and any integrable $f : H \to \mathbb{R}$, the following holds for the radial function $R(x) = x/\|x\|_2$:*

$$\int_H f(x)\frac{u^T x}{\|x\|_2^d}dx = \int_{R(H)} f(R^{-1}(x))dS(x).$$

Lemma 11 (Basic facts on the support function, see for example [AAGM15])**.**

- *For every $K \in \mathcal{C}_d(1)$, the function h_K can be extended to the whole of $\mathbb{R}^d$ via $h_K(x) = \|x\|_2 h_K(x/\|x\|_2)$ and this extension makes h_K a convex and 1-Lipschitz function on $\mathbb{R}^d$.*

- *For any $\lambda > 0$, $K \in \mathcal{C}$ the following holds for all $x \in \mathbb{R}^d$: $h_{\lambda K}(x) = \lambda h_K(x)$.*

2.3.1 Proof of Lemma 9

In this proof, we denote by $B_G(x,r)$ a geodesic ball on the sphere with center x and radius r. Let $\mathcal{V} := \{v_i\}_{i=1}^{s}$ be a set of vertices of size $m \leq s \leq 400d\ln(d)m$ with norm one. This set is a $m^{-\frac{1}{d-1}}$-net on $\mathbb{S}^{d-1}$ with the following property: In any geodesic ball with radius of $m^{-\frac{1}{d-1}}$ on the sphere there are at most $400d\ln(d)$ points. When $m \geq C^d$ such a set exists by [BW03].

Now, we define a polytope $P_{m,d} := \operatorname{conv}\{\mathcal{V}\}$. First, observe that for each $x \in \mathbb{R}^d$ the following holds:

$$h_{P_{m,d}}(x) = \max_{v\in\mathcal{V}} \|x\|_2\|v\|_2\cos(\angle(x,v_\pi)) = \|x\|_2\cos(\angle(x,v_\pi)), \tag{2.18}$$

where v_π is the closest point in the net to $x/\|x\|_2$. Then, using the fact that $\cos(t) = 1 - t^2/2 + O(t^4)$, we know that when $m \geq C_1^d$, for all $u \in \mathbb{S}^{d-1}$

$$0 \leq 1 - h_{P_{m,d}}(u) \leq Cm^{-\frac{2}{d-1}}$$

and the first part of the Lemma follows.

Now, we prove the last part of Lemma 9, that is an upper bound on the entropy numbers. Let us consider the support function when it is restricted to the facets of regular simplex $\mathcal{P}_d$, denoted by S_k, $1 \leq k \leq d+1$. Now we prove the following:

Lemma 12. *For any $k \in \{1,\dots,d+1\}$, $h_{P_{m,d}} : S_k \to \mathbb{R}$ is at most $C\ln(d)m$ piecewise linear, and every piece has at most $C(d\ln(d))^2 4^{d-1}$ $(d-2)$-facets.*

Proof. First, we use the Voronoi cells of the vertices $\mathcal{V}$ when they are restricted to $\mathbb{S}^{d-1}$, that is each cell, denoted by $C(v_i)$, $1 \leq i \leq |\mathcal{V}|$, is defined by

$$C(v_i) := \{x \in \mathbb{S}^{d-1} : \|v_i - x\|_2 \leq \|v_j - x\|_2 \ \ \forall j \neq i\}.$$

By Eq. (2.18) we know that for each $x \in \mathbb{R}^d$ the value of $h_{P_{m,d}}$ is attained on the closest $v_i \in \mathcal{V}$ to

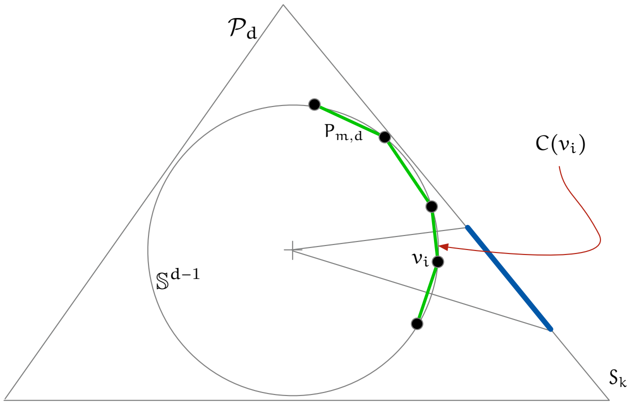

Figure 2-1: Illustration of the proof. The function $h_{P_{m,d}}$ is linear on the set $R^{-1}(C(v_i))$ (in blue) and convex piece-wise linear on S_k.

$x/\|x\|$ (since all the vertices of $P_{n,d}$ have the same norm). Thus, we can write

$$S_k = \bigcup_{i=1}^{|\mathcal{V}|} R^{-1}(C(v_i)) \cap S_k$$

where R denotes the radial function from $\mathcal{P}_d$ to $\mathbb{S}^{d-1}$. Moreover, observe that

$$h_{P_{m,d}} : R^{-1}(C(v_i))(x) \equiv v_i^T x.$$

Since $h_{P_{m,d}} : S_k \to \mathbb{R}$ is a convex function, we conclude that it i also piecewise linear. Due to the regularity of the net $\mathcal{V}$, we know that number of pieces can be bounded by $2|\mathcal{V}|/(d+1) \leq C\ln(d)m$.

Next, observe that the number of $d-2$ facets of each piece, which corresponds to some $v_i \in \mathcal{V}$, is determined by the number of neighbors of the Voronoi cell $C(v_i)$. By the construction of $\mathcal{V}$, it can be bounded by $C(d\ln(d))^2 4^{d-1}$. To see this, since $\mathcal{V}$ is a $m^{-1/(d-1)}$-net, clearly all the neighbors of $v_i \in \mathcal{V}$ lie in $B_G(v_i, 4\epsilon)$. Moreover, we can bound the number of vertices in this ball by

$$\begin{aligned} |\{v \in \mathcal{V} : v \in B_G(v_i, 4\epsilon)\}| &\leq 400 d\ln(d) \cdot |\mathcal{N}(\epsilon, B_G(v_i, 4\epsilon), d_G)| \\ &\leq C(d\ln(d))^2 4^{d-1}, \end{aligned}$$

and the claim follows. □

Now, we are ready to prove Lemma 9. It is easy see that each $h_{P_{m,d}} : S_k \to \mathbb{R}$ is uniformly bounded by d (since $K \subset B_d$ implies that h_K restricted to S_k is bounded by d). Moreover, since we assume that all $h_L \in B(h_K, t)$ are bounded in $C_1 d$, we also know that our family is uniformly bounded by Cd^2. Now, recall that all of $h_{P_{m,d}}$ $C\sqrt{m}/(d-1)$-pieces have at most $(Cd\ln(d))^2 4^{(d-1)}$ facets, hence, it is also $C(d)\sqrt{m}$ piecewise simplicial linear. Thus, we can apply Lemma 7 and find a $(d+1)^{-1/2}\epsilon$-net, denoted by $\{f_{k,j}\}_{j=1}^{S_\epsilon}$, with respect to $\text{Unif}(S_k)$, where

$$S_\epsilon := \log \mathcal{N}_{[]}((d+1)^{-1/2}\epsilon, B_{\text{Unif}(S_k)}(h_{P_{m,d}}, t), \text{Unif}(S_k)) \leq C(d) m \log(\epsilon^{-1})^d \left(\frac{t}{\epsilon}\right)^{(d-1)/2}.$$

Recall that we aim to bound the entropy numbers of the the support function on the sphere. Therefore, we use the radial function to project each facet S_k onto the sphere, and show that the set of functions

$$\{\|R^{-1}(x)\|^{-1} f_{k,j}(R^{-1}(x))\}_{j=1}^{S_\epsilon}$$

forms an $(d+1)^{-1/2}\epsilon$-net on $B_{\text{Unif}(R(S_k))}(h_{P_{m,d}}, t)$ with respect to uniform measure.

Let $h_K \in \mathcal{C}(C_1 \cdot d)$ (which is convex and bounded by one) that is at least $(d+1)^{-1/2}\epsilon$ far from $h_{P_{m,d}}$, when the support functions are restricted to $R(S_k)$. Denote by $h_{\pi,K}$ the closest member to the h_K with respect to $\text{Unif}(S_k)$ in the aforementioned set. Then,

$$\begin{aligned}
&\sqrt{\int_{R(S_k)} (h_K(x) - \|R^{-1}(x)\|^{-1} h_{\pi,K}(R^{-1}(x))^2 dS(x)} \\
&= \sqrt{\int_{R(S_k)} \|R^{-1}(x)\|^{-2} (h_K(R^{-1}(x)) - h_{\pi,K}(R^{-1}(x)))^2 dS(x)} \\
&= \sqrt{\int_{S_k} (h_K(x) - h_{\pi,K}(x))^2 \frac{u_i^T x}{\|x\|^{d+2}} dx} \leq \sqrt{\int_{S_k} (h_K(x) - h_{\pi,K}(x))^2 dx} \leq (d+1)^{-1/2}\epsilon,
\end{aligned}$$

where we used the homogeneity of the support function, and Lemma 10. Therefore, we found our desired net. Now, we can use [GW17], and conclude that for all $0 \leq \epsilon \leq t$

$$\begin{aligned}\log \mathcal{N}_{[]}(\epsilon, B_{\mathrm{Unif}}(h_{P_{m,d}}, t), \mathbb{P}) &\leq \sum_{k=1}^{d+1} \log \mathcal{N}_{[]}(\frac{\epsilon}{\sqrt{d+1}}, B_{S_k}(h_{P_{m,d}}, t), \mathrm{Unif}(S_k)) \\ &\leq \sum_{k=1}^{d+1} \log \mathcal{N}_{[]}(\frac{\epsilon}{\sqrt{d+1}}, B_{\mathrm{Unif}(R(S_k))}(h_{P_{m,d}}, t), \mathrm{Unif}(R(S_k))) \\ &\lesssim m \log(\epsilon^{-1})^d \, (t/\epsilon)^{(d-1)/2},\end{aligned}$$

and the claim follows.

2.3.2 Reduction to Lemma 8

We will show that when $K^* \in \mathcal{C}_d(1)$, the LSE only considers convex sets in $\mathcal{C}(C_1 \cdot d)$. First, observe that the score of h_{K^*} is bounded by 5 when n is large enough. To see this,

$$\begin{aligned}n^{-1} \sum_{i=1}^{n} (Y_i - h_{K^*}(X_i))^2 &\leq 2n^{-1} \left(\sum_{i=1}^{n} \xi_i^2 + h_{K^*}(X_i)^2 \right) \\ &\leq 2 \left(\int_{\mathbb{S}^{d-1}} h_K^2(x) d\mathbb{P}(x) + 1 + O(\frac{1}{\sqrt{n}}) \right) \leq 5.\end{aligned} \tag{2.19}$$

where we used the fact that $h_{K^*}(x) \leq 1$ when $x \in \mathbb{S}^{d-1}$.

Now, let $K \notin \mathcal{C}(C_1 \cdot d)$. Then, K has a vertex (denoted by v_1) with $\|v_1\|_2 > C_1 d$. In this case for C_1 that is large enough, there exists a spherical cap of $\mathbb{S}^{d-1}$, denoted by $C(v_1)$, with center $v_1/\|v_1\|_2$ and normalized surface area of $1/3$, such that

$$h_K(u) \geq 100 \ \ \forall u \in C(v_1).$$

To see this, observe that if $C(v_1)$ has a normalized surface of $1/3$, then it implies that the geodesic distance between $v_1/\|v_1\|_2$ and its boundary is at most $\pi/2 - c_2/(d-1)$ (for some fixed $c_2 \geq 0$). Thus, for any $u \in C(v_1)$ the following holds:

$$\begin{aligned} h_K(u) &\geq \|v_1\|_2 \cos(\angle(v_1, u)) \geq C_1 d \cos(\pi/2 - c_2/(d-1)) \\ &= C_1 d \sin(c_2/(d-1)) \geq C_1 c_2/2, \end{aligned}$$

therefore if we choose C_1 to be large enough then $h_K(u) \geq 100$ for all $u \in C(v_1)$.

Now, using the fact that X_i are i.i.d uniform on the sphere, we know that with probability of at least $1 - e^{-cn}$, there are at least $n/4$ points in this cap (using concentration for Bernoulli random variables). Moreover, since $\xi_i \sim N(0,1)$, we know that with high probability, at least 0.49 of the points that lie in this cap, the ξ_i are negative. Thus, we conclude that with high probability

$$n^{-1} \sum_{i=1}^{n} (Y_i - h_K(X_i))^2 \geq n^{-1} \sum_{X_i \in C(v_1),\ \xi_i \leq 0} (h_K(X_i) - 1 - \xi_i)^2 \geq 11.$$

Using the last equation and Eq. (2.19), only sets in $\mathcal{C}(C_1 \cdot d)$ will be considered, and the claim follows.

Chapter 3

On the Variance, Admissibility, and Stability of Empirical Risk Minimization

We begin with some notation, in the fixed design, we will abuse notation and treat $\widehat{f}_n$ as a vector in $\mathbb{R}^n$. In addition to $\widehat{f}_n$, we analyze properties of the set of δ-approximate minimizers of empirical loss, defined for $\delta > 0$ via

$$\mathcal{O}_\delta := \left\{ f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{f}_n(X_i))^2 + \delta \right\}. \tag{3.1}$$

Note that $\mathcal{O}_\delta$ is a random set, in both fixed and random designs. Recall the bias-variance decomposition of the mean squared error of an estimator $\bar{f}_n$ (see Eq. (1.10) above), and the definition of $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{Q})$ as the maximal variance error of an estimator $\bar{f}_n$ (see Eq. (1.11) above). Finally, we remind the total law of variance, that is

$$V(\widehat{f}_n) = \underbrace{\mathbb{E}_{\overline{X}}\mathbb{E}_{\overline{\xi}}\left[\int \left(\widehat{f}_n - \mathbb{E}_{\overline{\xi}}\left[\widehat{f}_n|\overline{X}\right]\right)^2 d\mathbb{P}\right]}_{\mathbb{E}V(\widehat{f}_n|\overline{X})} + \underbrace{\mathbb{E}_{\overline{X}}\left[\int \left(\mathbb{E}_{\overline{\xi}}\left[\widehat{f}_n|\overline{X}\right] - \mathbb{E}_{\overline{X},\overline{\xi}}\widehat{f}_n\right)^2 d\mathbb{P}\right]}_{V(\mathbb{E}(\widehat{f}_n|\overline{X}))}. \tag{3.2}$$

We refer to the two terms as the expected conditional variance and the variance of the conditional expectation, respectively. Thus, in the random design setting, the mean squared error decomposes into three terms:

$$\mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{P} = \mathbb{E}_{\overline{\mathrm{X}}} V(\widehat{f}_n | \overline{\mathrm{X}}) + V(\mathbb{E}_{\overline{\xi}}(\widehat{f}_n | \overline{\mathrm{X}})) + B^2(\widehat{f}_n). \tag{3.3}$$

3.1 Main Results: Fixed Design

Before we present our results, we remind Assumptions 1,2,3 from the first chapter. The first assumption states that $\mathcal{F}$ is convex, the second assumption is that the noise is an isotropic Gaussian, and the final assumption states that the diameter of $\mathcal{F}$ (in terms of $L_2(\mathbb{P}^{(n)})$) is of order one.

Under these three assumptions, Theorem 1 ensures that the minimax $\mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)})$ is of order ϵ_*^2, where ϵ_* is the stationary point of $\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}^{(n)}) \asymp n\epsilon^2$.

Now, we can state our results. The first result, uses the notion of the set $\mathcal{O}_\delta$ of δ-approximate minimizers from (3.1).

Theorem 8. *Let $n \geq 1$ and $\delta_n \lesssim \epsilon_*^2$. Then, under Assumptions 1,2,3, there exists an absolute constant $c > 0$, such that for every $f^* \in \mathcal{F}$ the event*

$$\sup_{f \in \mathcal{O}_{\delta_n}} \int (f - \mathbb{E}_{\mathcal{D}} \widehat{f}_n)^2 d\mathbb{P}^{(n)} \lesssim \epsilon_*^2 \tag{3.4}$$

holds with probability at least $1 - 2\exp(-cn\epsilon_*^2)$; and in particular, $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) \lesssim \mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)})$.

Theorem 8 establishes our promised result form first chapter, and in particular it implies Equation (1.12) above. It shows the variance of ERM necessarily enjoys the minimax rate of estimation, and, hence, any sub-optimality of ERM must arise from the bias term in the error. The theorem also incorporates a stability result: not only is the ERM close to its expected value $\mathbb{E}_{\mathcal{D}} \widehat{f}_n$ with high probability, but any approximate minimizer (up to an excess error of δ_n^2) is close to $\mathbb{E}_{\mathcal{D}} \widehat{f}_n$ as well. We remark that we prove this bound for any $\overline{\xi}$ that satisfies the LCP property (see Equation (3.10) below) which holds for i.i.d. $N(0,1)$ noise as well.

Remark 9. We remark that $\mathcal{V}(\bar{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) \lesssim \epsilon_*^2$ holds for any estimator $\bar{f}_n$ for which the map $\overline{\xi} \mapsto \bar{f}_n(\overline{\xi})$ is $O(1)$-Lipschitz. Furthermore, Theorem 8 also holds for the misspecified model (i.e., even if f^* does not lie in $\mathcal{F}$).

Our next result complements Theorem 8 above, providing a lower bound on $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}^{(n)})$.

Theorem 9. *Let $n \geq 1$. Then, under Assumptions 1,2, the following holds:*

$$\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) \gtrsim \mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)})^2.$$

We remark that the proof of Theorem 9 is specific to $\widehat{f}_n$ and cannot be extended immediately to other estimators. Next, we refine the result of Theorem 8, by both obtaining the exact characterization (up to a multiplicative absolute constant) of the variance and relaxing Assumption 3. First, for every $f^* \in \mathcal{F}$, we define the following set:

$$\mathcal{H}_* := \{f \in \mathcal{F} : \|f - \mathbb{E}_{\mathcal{D}}\widehat{f}_n\|_n^2 \leq 4 \cdot V(\widehat{f}_n)\}. \tag{3.5}$$

In simple words, we take a neighborhood around the $\mathbb{E}_{\mathcal{D}}\widehat{f}_n$ with a radius of order the square root of the variance error term of $\widehat{f}_n$, when the underlying function is $f^* \in \mathcal{F}$. In the statement below, recall the notation of $\mathcal{M}(\mathcal{H}_*, \mathbb{P}^{(n)})$ denotes the minimax risk over $\mathcal{H}_*$.

Theorem 10. *Let $n \geq 1$. Then, under Assumptions 1,2, for any underlying $f^* \in \mathcal{F}$, the following holds:*

$$V(\widehat{f}_n) \asymp \mathcal{M}(\mathcal{H}_*, \mathbb{P}^{(n)}),$$

where $\mathcal{H}_*$ is defined in (3.5).

This theorem shows that the variance of $\widehat{f}_n$ (for a given $f^* \in \mathcal{F}$) equals (up to a multiplicative absolute constant) to the minimax rate of a subclass of $\mathcal{H}_* \subset \mathcal{F}$ (which depends on f^*). Clearly, it implies the optimality of the variance error term of $\widehat{f}_n$ without Assumption 3, since $\mathcal{M}(\mathcal{H}_*, \mathbb{P}^{(n)}) \lesssim \mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)})$ for any $f^* \in \mathcal{F}$.

Our next result, which almost follows immediately from the proof of the last theorem, regards $\widehat{f}_n$ when it is "nearly" unbiased for some target function $f^* \in \mathcal{F}$. Specifically, we show that for $f^* \in \mathcal{F}$ such that $B^2(\widehat{f}_n) \leq 4 \cdot V(\widehat{f}_n)$ (where B^2 is the squared bias as defined in (1.10)), $\widehat{f}_n$ is "locally" minimax optimal on the set

$$\mathcal{G}_* := \{f \in \mathcal{F} : \|f - f^*\|_n^2 \leq 4 \cdot \mathrm{E}_{f^*}^2 + S/n\}, \tag{3.6}$$

where $\mathrm{E}_{f^*}^2 := \mathbb{E}_{\mathcal{D}}\|\widehat{f}_n - f^*\|_n^2$ and $S > 0$ is an absolute constant (that will defined in §3.4.2 below). Note that if $B^2(\widehat{f}_n) \leq 4 \cdot V(\widehat{f}_n)$, then $\mathrm{E}_{f^*}^2 \lesssim \mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)})$ by Theorem 10; but $\mathrm{E}_{f^*}^2$ may be much smaller than the minimax rate. In the statement of the theorem below, recall the definition of the risk .

Theorem 11. *Let $n \geq 1$. Then, under Assumptions 1,2, for any $f^* \in \mathcal{F}$ such that $B^2(\widehat{f}_n) \leq 4 \cdot V(\widehat{f}_n)$, the following holds:*

$$\mathcal{R}(\widehat{f}_n, \mathcal{G}_*, \mathbb{P}^{(n)}) \asymp \mathcal{M}(\mathcal{G}_*, \mathbb{P}^{(n)}),$$

where $\mathcal{G}_*$ is defined in (3.6).

In other words, even if we restrict our class to a $\Theta(\max\{\mathrm{E}_{f^*}, \sqrt{S/n}\})$-neighborhood of f^*, no estimator on the restricted class $\mathcal{G}_*$ performs (up to a multiplicative absolute constant) better on every regression function $f \in \mathcal{G}_*$ than $\widehat{f}_n$ (with respect to the original class $\mathcal{F}$) does on f^*.

Remark 10. Both Theorems 10,11 hold with $\widehat{f}_n$ replaced by any estimator $\widehat{g}_n$ for which the map $\overline{\xi} \mapsto \widehat{g}_n(\overline{\xi})$ is $O(1)$-Lipschitz. Furthermore, their proofs are almost identical to each other as we will see in §3.4.2 below. In both $\mathcal{H}_*$ and $\mathcal{G}_*$ there is nothing special about the constant 2; it may be replaced by any absolute constant.

Next, we recall the notion of *admissibility* of $\widehat{f}_n$, first established by [Cha14], with a simplified proof given by [CGZ17]. The result states that for any estimator $\bar{f}_n$, there *exists* a regression function $f^* \in \mathcal{F}$ such that ERM over the data drawn according to (1.1) has error which is no worse (up to an absolute constant) than that of $\bar{f}_n$. Hence, while ERM may be suboptimal for some models $f^* \in \mathcal{F}$, it cannot be ruled out completely as a learning procedure.

It is fairly straightforward to see that Theorem 11 implies $\widehat{f}_n$ is admissible if at least one $f^* \in \mathcal{F}$ exists with a "small bias". Our next result offers an alternative proof for Chatterjee's admissibility theorem. In this proof, under Assumptions 1 and 2, we demonstrate that there always exists an $f^* \in \mathcal{F}$ such that $\widehat{f}_n$ exhibits a small bias.

Corollary 2. *[Chatterjee's Admissibility Theorem] Let $n \geq 1$ and $\bar{f}_n$ be any estimator. Then, under Assumptions 1,2, there exists an underlying function $f^* \in \mathcal{F}$ (that depends on $\bar{f}_n$) such that*

$$\mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}^{(n)} \leq C \cdot \mathbb{E}_{\mathcal{D}} \int (\bar{f}_n - f^*)^2 d\mathbb{P}^{(n)}, \tag{3.7}$$

where $C > 0$ is a universal constant that is independent of $\mathcal{F}$, $\mathbb{P}^{(n)}$, and n.

Remark 11. Under Assumptions 1,2,3, our proof demonstrates that the admissibility property is valid for any $\widehat{g}_n$ such that $\overline{\xi} \mapsto \widehat{g}_n(\overline{\xi})$ is $O(1)$-Lipschitz. Furthermore, it offers a new and simplified perspective on this profound theorem. Specifically, we show that admissibility hinges on the existence of a target regression function $f^* \in \mathcal{F}$ such that the estimator not only has a 'small bias' but is also "stable" around it. Remarkably, the existence of such a target function is ensured by a purely topological argument—the Brouwer's fixed-point theorem. From a statistical perspective, this has a simple interpretation: a "stable" estimator cannot have a "high" bias on every target function within a compact function class.

Remark 12. [CGZ17] also gave an explicit upper bound of $1.65 \cdot 10^5$ for the constant $C > 0$ in (3.7). Making the constants explicit in our proof yields a bound of the order 10^3 rather than 10^5; we have not attempted to optimize the constants, so we believe this can be improved further.

Note that if we place $\bar{f}_n$ in Corollary 2 to be a minimax optimal estimator immediately yields the following:

Corollary 3. *[Weak admissibility] Let $(\mathcal{F}, \mathbb{P}^{(n)})$ that satisfies Assumptions 1,2. Then, there exists an underlying function $f^* \in \mathcal{F}$ such that*

$$\mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}^{(n)} \leq C_1 \cdot \mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)}), \tag{3.8}$$

where $C_1 > 0$ is a universal constant that is independent of $\mathcal{F}$, $\mathbb{P}^{(n)}$, and n.

That is, ERM cannot have minimax suboptimal error for every $f^* \in \mathcal{F}$; we name this result *weak* admissibility. Namely, in the fixed design setting, the *minimal* error of ERM is at most the minimax rate[1]

Finally, we establish the following counter-intuitive behavior of the landscape around $\widehat{f}_n$ when $(\mathcal{F}, \mathbb{P}^{(n)})$ is a non-Donsker class (a notion which will be defined formally in §3.4.8 below).

Theorem 12. *Assume that $(\mathcal{F}, \mathbb{P}^{(n)})$ is a non-Donsker class satisfying Assumptions 1,2 for all n that are large enough. Then, there exist a sequence of functions ${f^*}_n \in \mathcal{F}$ and a sequence $C_n = \omega(1)$, such that for each n, $\mathbb{E}\|\widehat{f}_n - {f^*}_n\|_n^2 \lesssim \epsilon_*(n)^2$ and*

$$\left\{ f \in \mathcal{F} : \int (f - \widehat{f}_n)^2 d\mathbb{P}^{(n)} \lesssim \epsilon_*^2 \right\} \not\subset \mathcal{O}_{C_n \cdot \epsilon_*(n)^2}, \tag{3.9}$$

with probability of at least $1 - 2\exp(-c_2 n \epsilon_*(n)^2)$.

Theorem 12 says that when the underlying function is ${f^*}_n$, the LSE solution $\widehat{f}_n$ displays counter-intuitive behavior: on the one hand, $\widehat{f}_n$ estimates ${f^*}_n$ optimally, but on the other hand, for most $\overline{\xi}$ there exist functions which are very close to $\widehat{f}_n$ in $L_2(\mathbb{P}^{(n)})$, and yet far from being minimizers of the squared error.

Remark 13. Is is easy to verify that for $\delta_n = \omega(\mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)}))$, the inequality (3.4) cannot be true in the generality of our assumptions. Therefore, the stability threshold of $\mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)})$ is tight, up to a multiplicative absolute constant.

We conclude this part with the following remark, which we will return to in §3.2.2. Theorems 8, and 12 and Corollary 3 hold for any $\overline{\xi}$ that satisfies a Lipschitz Concentration Property (LCP) with an absolute constant $c_L > 0$: this means that if $F : \mathbb{R}^n \to \mathbb{R}$ is 1-Lipschitz (with respect to the Euclidean norm on $\mathbb{R}^n$), then

$$\Pr(|F(\overline{\xi}) - \mathbb{E}F(\overline{\xi})| \geq t) \leq 2\exp(-c_L t^2). \tag{3.10}$$

[1] In the paper of [KR21], they provided a lower bound to the minimal error of ERM in the fixed design setting. Later, we shall state and prove our new weak admissibility result for $\widehat{f}_n$, analogous to Corollary 3, in the random design setting.

The LCP condition is stronger than the sub-Gaussian random vector assumption [BLM13], but it is significantly less restrictive than requiring normal noise (in which case $c_L = 1/2$ [Led01]). The LCP condition is also known as the isoperimetry condition (cf. [BS23, Sec. 1.3]).

Remark 14. Herbst's argument [Wai19] implies that the Lipschitz concentration property holds for any random vector $\overline{\xi}$ satisfying a log-Sobolev inequality; the converse is not true in general. However, in the seminal work of [Mil09], it was shown that if $\overline{\xi}$ is assumed to be log-concave, then $\overline{\xi}$ which satisfies a LCP with constant c_L also satisfies a log-Sobolev inequality with constant $\Theta(c_L)$.

3.2 Main Results: Random Design

We now turn our attention to the random design setting. Here, we establish similar results, albeit under additional assumptions, and with significantly more effort. We shall use two different approaches: the first uses the classical tools of empirical process theory [Gee00], and the second, inspired by our fixed-design methods, relies heavily on isoperimetry and concentration of measure (cf. [Led01]) to understand the behaviour of $\widehat{f}_n$ when $\overline{X}, \overline{\xi}$ satisfy the LCP property.

For technical reasons (see Remark 17 below), we assume in this section that the error rate of the class $\mathcal{F}$ is more than parametric:

Assumption 10. *$\epsilon_*^2 \gtrsim \log(n)/n$, where ϵ_* is the stationary point of $n\epsilon^2 \asymp \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P})$.*

3.2.1 The Empirical Processes Approach

First, we assume that the function class and the noise are uniformly bounded.

Assumption 11. *There exist universal constants $\Gamma_1, \Gamma_2 > 0$ such that $\mathcal{F}$ is uniformly upper-bounded by Γ_1, i.e. $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq \Gamma_1$; and the components of $\overline{\xi} = (\xi_1, \ldots, \xi_n)$ are i.i.d. zero mean and bounded almost surely by Γ_2.*

Remark 15. The uniform boundedness assumption is taken to simplify the presentation when appealing to Talagrand's inequality, and can be relaxed to i.i.d. sub-Gaussian, at the a price of a multiplicative factor of $\log n$ in the error term in Theorem 13 below.

Next, we follow definitions of [BBM05] of the lower and upper isometry remainders of $(\mathcal{F}, \mathbb{P}, n)$, denoted by $\mathcal{I}_L(n, \mathbb{P})$ and $\mathcal{I}_U(n, \mathbb{P})$ respectively. First, for each realization of the input $\overline{\mathrm{X}} := (X_1, \ldots, X_n)$, define

$$\mathcal{I}_L(\overline{\mathrm{X}}) := \min \left\{ A > 0 \,\middle|\, \forall f, g \in \mathcal{F} : \quad 2^{-1} \int (f-g)^2 d\mathbb{P} - A \leq \int (f-g)^2 d\mathbb{P}_n \right\}$$

and

$$\mathcal{I}_U(\overline{\mathrm{X}}) := \min \left\{ A > 0 \,\middle|\, \forall f, g \in \mathcal{F} : \quad 2^{-1} \int (f-g)^2 d\mathbb{P}_n - A \leq \int (f-g)^2 d\mathbb{P} \right\}$$

where $\mathbb{P}_n$ is the uniform (random) measure over $\overline{\mathrm{X}}$. In words, a bound on the lower isometry remainder ensures that the distance with respect to the underlying distribution $\mathbb{P}$ between any two functions in $\mathcal{F}$ is small if the distance between them with respect to the empirical measure is small, up to a multiplicative constant and an additive remainder, and conversely for the upper isometry remainder.

Definition 7. *$\mathcal{I}_L(n, \mathbb{P})$ and $\mathcal{I}_U(n, \mathbb{P})$ are defined as the minimal constants $\delta_1(n), \delta_2(n) \geq 0$ such that*

$$\Pr\left(\mathcal{I}_L(\overline{\mathrm{X}}) \leq \delta_1(n)\right) \geq 1 - n^{-1} \quad \textit{and} \quad \Pr\left(\mathcal{I}_U(\overline{\mathrm{X}}) \leq \delta_2(n)\right) \geq 1 - n^{-1},$$

respectively.

Definition 8. *Set $\epsilon_U := \max\{\epsilon_*, \widetilde{\epsilon}\}$, where $\widetilde{\epsilon}$ is the solution of*

$$\mathcal{I}_U(n, \mathbb{P}) \cdot \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}) \asymp n\epsilon^4. \tag{3.11}$$

Note that when $\mathcal{I}_U(n, \mathbb{P}) \lesssim \epsilon_*^2$, then $\epsilon_U \asymp \epsilon_*$, while if $\mathcal{I}_U(n, \mathbb{P}) \gg \epsilon_*^2$ then $\epsilon_U \gg \epsilon_*$. The following is our main result in this approach to the random design setting:

Theorem 13. *Let $n \geq 1$ and $\epsilon_V^2 := \max\{\epsilon_U^2, \mathcal{I}_L(n, \mathbb{P})\}$. Then, under Assumptions 1,10,11, for* *every $f^* \in \mathcal{F}$ the following holds with probability of at least $1 - 2n^{-1}$:*

$$\sup_{f \in \mathcal{O}_{\delta_n}} \int (f - \mathbb{E}_{\mathcal{D}} \widehat{f}_n)^2 d\mathbb{P} \lesssim \epsilon_V^2,$$

where $\delta_n = O(\epsilon_V^2)$; and in particular $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim \epsilon_V^2$.

Theorem 13 is a generalization of Theorem 8 to the random design case, and its proof uses the strong convexity of the loss and Talagrand's inequality. In §3.2.5 below, we discuss this bound in the context of "distribution unaware" estimators.

Remark 16. Note that a bound similar to that of Theorem 13 cannot hold for the bias error term. Indeed, one can construct a class $\mathcal{F}$ with $\mathcal{I}_L(n, \mathbb{P}) \lesssim \epsilon_*^2$ and $\mathcal{V}(n, \mathbb{P}, \mathcal{F}) \asymp \epsilon_*^2$ for which the bias error term $\sup_{f^* \in \mathcal{F}} B^2(\widehat{f}_n)$ does not decrease to 0 with n; moreover, for this class one has

$$\mathbb{E}_{\overline{X}, \overline{\xi}} \|\widehat{f}_n - \mathbb{E}_{\overline{\xi}} \widehat{f}_n | \overline{X} \|_n^2 \asymp 1.$$

That is, neither the bias nor the *empirical* variance converge to zero. A remarkable consequence of our results is that even though the ERM only observes the random empirical measure $\mathbb{P}_n$, its variance, measured in terms of $\mathbb{P}$, converges to zero when $\mathcal{I}_L(n, \mathbb{P}) \to 0$.

Remark 17. When $\epsilon_*^2 \lesssim \log(n)/n$, the proof of Theorem 13 shows that there exists an $f_c \in \hat{f}_n$ such that $\|f - f_c\| \leq C_1 \epsilon_V$ outside an event $\mathcal{E}$ of probability $\exp(-C_2 n \epsilon_*^2)$ for absolute constants $C_1, C_2 > 0$. For instance, in the parametric case $\epsilon_*^2 \asymp n^{-1}$, this bound only yields that $\Pr(\mathcal{E}) \geq 1 - c_1$, for some $c_1 \in (0, 1)$. In particular, we cannot conclude that $\mathcal{V}(\hat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim \epsilon_*^2$, because we have no control over the behavior of $\hat{f}_n$ on $\mathcal{E}^c$.

An immediate and useful corollary of this result is that if we have sufficient control of the upper and lower isometry remainders, the variance will be minimax optimal:

Corollary 4. *Let $n \geq 1$. Then, under Assumptions 1,10,11 and the additional assumption of $\max\{\mathcal{I}_L(n, \mathbb{P}), \mathcal{I}_U(n, \mathbb{P})\} \lesssim \epsilon_*^2$, then with probability of at least $1 - 2n^{-1}$*

$$\sup_{f \in \mathcal{O}_{\delta_n}} \int (f - \mathbb{E}_{\mathcal{D}} \widehat{f}_n)^2 \, d\mathbb{P} \lesssim \epsilon_*^2$$

if $\delta_n \lesssim \epsilon_*^2$; in particular, $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim \epsilon_*^2$.

The assumption that $\max\{\mathcal{I}_L(n,\mathbb{P}), \mathcal{I}_U(n,\mathbb{P})\} \lesssim \epsilon_*^2$ is considered very mild in the empirical process and shape constraints literature ([Gee00, Chp. 5] and references therein), and it holds for many classical models. In the context of this thesis, the last corollary covers our definition of non-Donsker classes (see Definition 5 above); and in particular it implies Equation (1.13) above. Finally, we remark that Corollary 4 extends the scope of [CR06] to non-Donsker classes.

Remark 18. The assumption of $\max\{\mathcal{I}_L(n,\mathbb{P}), \mathcal{I}_U(n,\mathbb{P})\} \lesssim \epsilon_*^2$ holds for classes that satisfy the Koltchinskii-Pollard condition [RST17] or the $L_2 - L_{2+\delta}$ entropy equivalence condition (see [LM13] and references therein). In the classical regime, i.e. when $\mathcal{F}$ is fixed and n grows, it is hard to construct function classes that does not satisfy this assumption for n that is large enough [BM93]

Remark 19. Corollary 4 may also be derived directly from Theorem 8 if the noise is assumed to satisfy the LCP property. But the corollary holds for any sub-Gaussian noise – which is significantly more general.

3.2.2 Random Design: The Isoperimetry Approach

In order to motivate this part, we point out that just requiring that $\mathcal{I}_L(n,\mathbb{P}) \lesssim \epsilon_*^2$ is considered to be a very mild assumption (cf. the influential work of [Men14]). However, the upper bound of Theorem 13 depends on the upper isometry remainder as well; we would like to find some sufficient conditions under which this dependency can be removed. Moreover, note that the isometry remainders are connected to the geometry of $(\mathcal{F}, \mathbb{P})$ and not directly to the stability properties of the *estimator*. Using a different approach, based on isoperimetry, we will upper-bound the variance of ERM based on some "interpretable" stability parameters of the estimator itself. These stability parameters will be data-dependent relatives of the lower isometry remainder.

We remind that $\widehat{f}_n$ is uniquely defined on the data points $\overline{\mathrm{X}}$ when $\mathcal{F}$ is a convex closed function class, but it may not be unique over the entire $\mathcal{X}$ (as multiple functions in $\mathcal{F}$ may take the same values at $X_1, \dots, X_n$). Here, we (implicitly) assume that $\widehat{f}_n$ is equipped with a selection rule such that it is also unique over the entire $\mathcal{X}$ (e.g., choosing the minimal norm solution [Has+22; Bar+20]); thus, in all the results below, $\widehat{f}_n$ denotes an element of $\mathcal{F}$.

Now, we are ready to present our results. First, as we mentioned at the end of §3.1, here we shall assume that $\overline{\xi}$ is isotropic noise that satisfies the LCP condition (which includes Gaussian noise).

Assumption 12. *$\overline{\xi}$ is an isotropic random vector satisfying the LCP condition of Equation* (3.10) *with parameter $c_L = \Theta(1)$.*

Recall that ϵ_* is defined as the stationary point of $n\epsilon^2 \asymp \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P})$, and that the conditional variance of $\widehat{f}_n$, which is a function of the realization $\overline{X}$ of the input, is defined as

$$V(\widehat{f}_n|\overline{X}) := \mathbb{E}_{\overline{\xi}}[\|\widehat{f}_n - \mathbb{E}_{\overline{\xi}}[\widehat{f}_n|\overline{X}]\|^2];$$

that is, we fix the data points $\overline{X}$ and take expectation over the noise. The formulation of the following definition involves a yet-to-be-defined (large) absolute constant $M > 0$, which will be specified in the proof of Theorem 14 (see §3.3.3 below).

Definition 9. *For each realization $\overline{X}$ and $f^* \in \mathcal{F}$, let $\rho_S(\overline{X}, f^*)$ be defined as the minimal constant $\delta(n)$ such that*

$$\Pr_{\overline{\xi}}\left\{\overline{\xi} \in \mathbb{R}^n : \forall \overline{\xi}' \in B_n(\overline{\xi}, M\epsilon_*) : \|\widehat{f}_n(\overline{X}, \overline{\xi}') - \widehat{f}_n(\overline{X}, \overline{\xi})\|^2 \leq \delta(n)\right\} \geq \exp(-c_2 n\epsilon_*^2). \tag{3.12}$$

where $B_n(\overline{\xi}, r) = \{\overline{\xi}' \in \mathbb{R}^n : \|\overline{\xi} - \overline{\xi}'\|_n \leq r\}$, and $c_2 > 0$ is an absolute constant.

We also set $\rho_S(\overline{X}) := \sup_{f^* \in \mathcal{F}} \rho_S(\overline{X}, f^*)$. Note that $\rho_S(\overline{X})$ measures the optimal radius of stability (or "robustness") of $\widehat{f}_n$ to perturbations of the noise *when the underlying function and data points $\overline{X}$ are fixed.* This is a weaker notion than the lower isometry remainder; in fact, one can verify that $\rho_S(\overline{X}) \lesssim \max\{\mathcal{I}_L(\overline{X}), \epsilon_*^2\}$ for every realization $\overline{X}$ (see Lemma 21 below for completeness). Now, we are ready to present our first theorem:

Theorem 14. *Let $n \geq 1$. Then, for any realization of the input $\overline{X}$ satisfying Assumptions 1,3,10,12, and any underlying $f^* \in \mathcal{F}$, the following holds:*

$$V(\widehat{f}_n|\overline{X}) \lesssim \max\{\rho_S(\overline{X}, f^*), \epsilon_*^2\}.$$

In particular, $\sup_{f^* \in \mathcal{F}} \mathbb{E}_{\overline{X}} V(\widehat{f}_n|\overline{X}) \lesssim \max\{\mathcal{I}_L(n,\mathbb{P}), \epsilon_*^2\}$.

Note that if $\mathcal{I}_L(n,\mathbb{P}) \lesssim \epsilon_*^2$ – a very mild assumption – then we obtain that the expected conditional variance is minimax optimal. We believe that it is impossible to bound the total variance via the lower isometry remainder alone. The intuition is $\widehat{f}_n$ only observes a given realization $\overline{X}$, and in general, the geometry of the function class "looks different" under different realizations (see the discussion in §3.2.5 for further details).

In our next result, we identify models under which we can bound the total variance of $\widehat{f}_n$ by the lower isometry remainder. Inspired by the recent paper of [BS23], we assume that $X \sim \mathbb{P}$ satisfies an isoperimetry condition and that $\mathcal{F}$ is a robust function class, i.e. $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R} : \|f\|_{Lip} = O(1)\}$. (A function class all of whose functions are $O(1)$-Lipschitz is often referred to the literature as a "robust learning architecture.")

We fix a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on $\mathcal{X}$, and denote by d_n the metric on $\mathcal{X}^n$ given by $d_n(\overline{X}, \overline{X}')^2 = \sum d(X_i, X_i')^2$.

Assumption 13. *The law $\mathbb{P}^{\otimes n}$ of $(X_1, \ldots, X_n)$ satisfies the LCP condition with respect to d_n, i.e. for all 1-Lipschitz functions $F : \mathcal{X}^n \to \mathbb{R}$,*

$$\Pr(|F(X) - \mathbb{E}F(X)| \geq t) \leq 2\exp(-c_X t^2), \tag{3.13}$$

where c_X is a constant depending only on $(\mathcal{X}, \mathbb{P})$. Furthermore, we assume that $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R} : \|f\|_{Lip} \leq C\}$, for some $C > 0$.

In general, it is insufficient to assume that $(\mathcal{X}, \mathbb{P})$ satisfies an LCP, as this does not imply that $(\mathcal{X}^n, \mathbb{P}^{\otimes n})$ satisfies an LCP with a constant independent of n. However, if $(\mathcal{X}, \mathbb{P})$ satisfies a concentration inequality which tensorizes "nicely," such as a log-Sobolev or W_2-transportation cost inequality (cf. [Led01, §5.2, §6.2]) then $(\mathcal{X}^n, \mathbb{P}^n)$ does satisfy such an LCP.

Next, we assume that with high probability, $\widehat{f}_n$ nearly interpolates the observations.

Assumption 14. *There exist absolute constants $c_I, C_I > 0$, such that the following holds:*

$$\Pr_{\mathcal{D}}\left(n^{-1}\textstyle\sum_{i=1}^n (\widehat{f}_n(X_i) - Y_i)^2 \leq C_I \epsilon_*^2\right) \geq 1 - \exp(-c_I n \epsilon_*^2).$$

Note that for this assumption to hold, the dimension of $\mathcal{X}$ must be $\Omega(\log(n))$, as interpolation with $O(1)$-Lipschitz functions is impossible in when the dimensionality is sub-logarithmic in the number of samples n (this follows, e.g., from the behaviour of the entropy numbers of the class of Lipschitz functions; cf. [Dud99]).

We now introduce another stability notion, which involves the random set $\mathcal{O}_\delta \subset \mathcal{F}$ of almost-minimizers of the empirical loss, as defined in (3.1); note that $\mathcal{O}_\delta$ depends on both $\overline{X}$ and $\overline{\xi}$. The random variable $\mathrm{Diam}_{\mathbb{P}}(\mathcal{O}_\delta)$ measures the stability of the ERM with respect to small errors in the minimization procedure. The formulation of the following definition involves another yet-to-be-defined (large) absolute constant $M' > 0$, which will be specified in the proof of Theorem 15 (see §3.4.6 below), as well as the constant c_I from Assumption 14.

Definition 10. $\rho_{\mathcal{O}}(n, \mathbb{P}, f^*)$ *is defined as the smallest* $\delta(n) \geq 0$ *such that*

$$\Pr_{\mathcal{D}}(\mathrm{Diam}_{\mathbb{P}}(\mathcal{O}_{M'\epsilon_*^2}) \leq \sqrt{\delta(n)}) \geq 2\exp(-c_I n \epsilon_*^2), \tag{3.14}$$

where $c_I \geq 0$ *is* the same *absolute constant defined in Assumption 14.*

In order to understand the relation between this and the previous stability notions, note that under Assumption 14 and the event $\mathcal{E}$ of Definition 10, we have that on an event of nonnegligible probability, $\rho_S(\overline{X}, f^*) \leq \rho_{\mathcal{O}}(n, \mathbb{P}, f^*)$; in addition, $\rho_{\mathcal{O}}(n, \mathbb{P}, f^*) \lesssim \max\{\mathcal{I}_L(n, \mathbb{P}), \epsilon_*^2\}$ (see Lemma 22 below). Under these additional two assumptions, we state our bound for the variance of $\widehat{f}_n$:

Theorem 15. *Let* $n \geq 1$. *Then, under Assumptions 1,3,10,12-14, the following holds for any target regression function* $f^* \in \mathcal{F}$:

$$V(\widehat{f}_n) \lesssim c_X^{-1} \cdot \max\{\epsilon_*^2, \rho_{\mathcal{O}}(n, \mathbb{P}, f^*)\};$$

and in particular, we have that $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim c_X^{-1} \cdot \max\{\epsilon_*^2, \mathcal{I}_L(n, \mathbb{P})\}$.

This result connects the total variance of $\widehat{f}_n$ to a "probabilistic" threshold for the $L_2(\mathbb{P})$-diameter of the *data-dependent* set of $\Theta(\epsilon_*^2)-$approximating solutions of $\widehat{f}_n$ – two parameters which at first sight are unrelated.

Remark 20. If we remove Assumption 14, our proof only implies that $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim \epsilon_V^2$, where ϵ_V is defined in Theorem 13 above. Also, under Assumptions 1,3,10,12,13, one may arrive the conclusion of Theorem 13 – $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim \epsilon_V^2$, by using the LCP for X and $\overline{\xi}$ and the robustness of $\mathcal{F}$ in place of Talagrand's inequality.

3.2.3 Weak Admissibility

In order to present our weak admissibility result, we assume the following:

Assumption 15. *For every $x \in \mathcal{X}$, the evaluation functional $f \mapsto f(x)$ is continuous in the $L_2(\mathbb{P})$ norm when restricted to $\mathcal{F}$. Furthermore, $\mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P})$ is finite for every $\epsilon \in (0, 1)$.*

This regularity condition will be used in order to apply a fixed-point theorem for continuous functions on a compact convex set in a Banach space. (Note that we have assumed that $\mathcal{F}$ is closed and convex in Assumption 1, and the finite ϵ-entropy ensures that $\mathcal{F}$ is totally bounded, hence compact.) The following is our weak admissibility result:

Finally, we state our new weak admissibility result, which holds under the following additional assumption:

Theorem 16. *[Weak admissibility of ERM in random design] Let $n \geq 1$. Then, under Assumptions 1 and 15, there exists $f^* \in \mathcal{F}$ (depending only on n) such that*

$$\mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{P} \lesssim \max\{\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}), \mathcal{I}_L(n, \mathbb{P})\}.$$

We believe that our bound cannot be improved in general (see §3.2.5 below more details). It implies that when $\max\{\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}), \mathcal{I}_L(n, \mathbb{P})\} \lesssim \epsilon_*^2$ ERM is weakly admissible. Differently from the fixed design setting, we conjecture that there are models $(n, \mathcal{F}, \mathbb{P})$ in which ERM is not weakly admissible.

Conclusions

Our results show that the variance term of ERM enjoys optimal rate in two distinct regimes: the classical regime (where the function class is fixed and the number of samples is increasing), and the benign overfitting regime (in which capacity of $\mathcal{F}$ is large compared to the number of samples). It follows that in both regimes the optimality of ERM is determined by its bias (or its implicit bias). This work raises the natural question of finding computationally efficient debiasing methods for ERM, with the hope that it may improve the statistical performance of ERM on large models.

3.2.4 Discussion and Prior Work

Stability of ERM Stability of learning procedures has been an active area of research in the early 2000's, and the topic came to spotlight again recently because of the connections to differential privacy and to robustness of learning methods with respect to adversarial perturbations. In the interest of space, we only compare present results to those of [CR06]. In the latter paper, the authors showed that the $L_1(\mathbb{P})$-diameter of the set of almost-minimizers of empirical error (with respect to any loss function) asymptotically shrinks to zero as long as the perturbation is $o(n^{-1/2})$ and the class $\mathcal{F}$ is Donsker. The analysis there relies on passing from the empirical process to the associated Gaussian process in the limit, and studying uniqueness of its maximum using anti-concentration properties. While the result there holds without assumptions of convexity of the class, it is limited by (a) its asymptotic nature and (b) the assumption that the class is not too complex. In contrast, the present work uses more refined non-asymptotic concentration results, at the expense of assumptions such as convexity and minimax optimal lower and upper lower isometry remainders. Crucially, the present result holds for non-Donsker classes — those for which the empirical process does not converge to the Gaussian process.

Shape-constrained regression In shape-constrained regression, the class $\mathcal{F}$ is chosen to be a set of functions with a certain "shape" property, such as convexity or monotonicity [SS18]. In these problems, a common theme is that $\mathcal{F}$ becomes too large (or, non-Donsker) when the dimension d of the input space surpasses a certain value. For instance, in convex (Lipschitz) regression, the transition happens at $d = 5$. It has been shown that below the threshold, the ERM procedure is minimax optimal [SS11; HW16; KS16; SS11; Gun12], but in higher dimensions, for most shape-constrained tasks, it is unknown whether ERM is minimax optimal. The few known cases of minimax optimality of ERM in high dimensions include the tasks of *isotonic regression* and *log-concave density estimation* [Han+19; KDR19; Car+18], and suboptimality of ERM above the threshold in convex regression and estimation of convex bodies has been established in [Kur+20; KRG20].

Our results imply that in all these high-dimensional shape-constrained regression tasks, the sub-optimality of ERM can only be due to the high bias of ERM. These results also align with the empirical observation that for the problem of estimation of convex sets, the ERM has a bias towards “smooth" convex sets [SC19a; Gho+21].

High Dimensional Statistics In classical statistics, the MLE is typically unbiased, and the standard approach is to introduce bias into the procedure in order to reduce the variance, overall achieving a better trade-off (see [SC19b, §1] and references within). In contrast, in high-dimensional models, the MLE may suffer from high bias even in tasks such as logistic regression and sparse linear regression (cf. [CS20; JM18]). Our results align with this line of work.

Regularization It is statistical folklore that the main benefit of incorporating a regularization term, such as ridge penalties, into an ERM procedure is that such regularized estimators have reduced variance. However, our findings seem to indicate that the ERM exhibits much lower conditional variance than variance of conditional expectation. This suggests that regularization primarily reduces the variance of the conditional expectation as opposed to the full variance error term.

3.2.5 Is the bound of Theorem 13 optimal?

When $\max\{\mathcal{I}_L(n,\mathbb{P}), \mathcal{I}_U(n,\mathbb{P})\} \gg \epsilon_*^2$, we have that $\epsilon_V^2 \gg \epsilon_*^2$ – our upper bound on the variance is much larger than the minimax error. Therefore, this bound seems at first glance to be suboptimal.

However, to the best of our knowledge, all estimators that attain the minimax rate, such as aggregation and the star algorithm (cf. [Yan04]), depend on the marginal distribution $\mathbb{P}$ of the covariates. In many cases, though, we do not know or have an oracle access to the marginal distribution $\mathbb{P}$, and the estimator only estimator only has access to $\mathbb{P}_n$ and to $\mathcal{F}$. It is natural to ask what is the minimax rate of "distribution-unaware" estimators that only depend on $\mathbb{P}_n$ and the function class $\mathcal{F}$, when the underlying distribution $\mathbb{P}$ is allowed to vary over some family of distributions.

To this end, given some family of probability distributions $\mathcal{P}$ on a domain $\mathcal{X}$, consider the following measurement of optimality of an estimator:[2]

$$\Delta_{(du)}(n,\mathcal{F},\mathcal{P}) = \inf_{\bar{f}_n} \sup_{\mathbb{Q}\in\mathcal{P}} \sup_{f^*\in\mathcal{F}} \frac{\mathbb{E}_{\mathcal{D}} \int (\bar{f}_n - f^*)^2 d\mathbb{Q}}{\mathcal{M}(n,\mathcal{F},\mathbb{Q})}.$$

We say that a distribution-unaware estimator is minimax optimal on $(n,\mathcal{F},\mathcal{P})$ when

$$\Delta_{(du)}(n,\mathcal{F},\mathcal{P}) = \Theta(1).$$

Unsurprisingly, if we do not place additional assumptions on $\mathcal{F}$ and $\mathcal{P}$ (beyond convexity), then $\Delta_{(du)}(n,\mathcal{F},\mathcal{P})$ may be $\omega(1)$ – no single estimator attains the minimax error on every distribution $\mathbb{Q} \in \mathcal{P}$, or, in other words, a minimax optimal estimator for $\mathbb{Q}$ must "know" $\mathbb{Q}$. In fact, one may construct a set of probability distributions $\mathcal{P}$ on a domain $\mathcal{X}$ and a function class $\mathcal{F}$ such that for any $\mathbb{Q} \in \mathcal{P}$, $\mathcal{M}(n,\mathcal{F},\mathbb{Q}) = O(n^{-1})$ (the parametric rate), and for any estimator $\bar{f}_n$, one may find $\mathbb{Q} \in \mathcal{P}$ such that $\mathcal{R}(\bar{f}_n,\mathcal{F},\mathbb{Q}) = \Omega(1)$; and in particular $\Delta_{(du)}(n,\mathcal{F}) = \omega(1)$ (see Example 1 below). The example also demonstrates (see Remark 22 below) that the "generalization diameter"

$$\Psi(n,\mathbb{P}) := \min_{f^*\in\mathcal{F}} \mathbb{E}[\mathrm{Diam}_{\mathbb{P}}(\{f \in \mathcal{F} : \forall i \in 1,\ldots,n \quad f^*(X_i) = f(X_i)\})]$$

should appear in the error of any "distribution-unaware" estimator in terms of $L_2(\mathbb{P})$ (though we do not know how to show this in complete generality). In the above example, $\Psi(n,\mathbb{P}) \asymp 1$, while the

[2] Note that here $\mathbb{E}_{\mathcal{D}}$ is the expected error when the noise $\overline{\xi}$ is drawn from some fixed distribution, and $X_1,\ldots,X_n$ are drawn i.i.d. from $\mathbb{Q}$.

minimax rate is $O(n^{-1})$; on the other hand, for every model, $\Psi(n, \mathbb{P}) \le \mathcal{I}_L(n, \mathbb{P})$. Therefore, it is not surprising that the bound of Theorem 13 includes the lower isometry remainder.

The upper isometry remainder $\mathcal{I}_U(n, \mathbb{P})$, though, is unrelated to $\Psi(n, \mathbb{P})$. Nonetheless, we conjecture that it cannot be removed from the bound of Theorem 13:

Conjecture 1. *The bound of Theorem 13 is optimal in the following way: There exist models $(n, \mathcal{F}, \mathbb{P})$ for which $\widehat{f}_n$ incurs an error $\epsilon_U^2 \gg \mathcal{I}_L(n, \mathbb{P}) \gtrsim \epsilon_*^2$ or an error $\mathcal{I}_L(n, \mathbb{P}) \gg \epsilon_U^2 \gg \epsilon_*^2$; and therefore our bound cannot be improved in general.*

The intuition behind this conjecture is as follows: the ERM sees the geometry of $\mathcal{F}$ with respect to $\mathbb{P}_n$ rather than $\mathbb{P}$, and perturbing the data points $X_1, \ldots, X_n$ by $\delta_1, \ldots, \delta_n$ can change the finite-dimensional geometry of the projected function class, reducing the "stability" of $\widehat{f}_n$. This is because $\widehat{f}_n$ is not a Lipschitz function in the data $\overline{X}$ with respect to the $L_2(\mathbb{P})$-norm, in contrast to its Lipschitz properties in $\overline{\xi}$ with respect to the $L_2(\mathbb{P}^{(n)})$-norm. Only if the upper and lower isometry constants are small, say of the order ϵ_*^2 does the metric geometry of $\mathcal{F}_n$ not depend too much on $\overline{X}$, up to ϵ_*^2, in which case we expect the variance to be bounded by ϵ_*^2.

Our confidence that this is the correct explanation for the appearance of $\mathcal{I}_U(n, \mathbb{P})$, rather than some other phenomenon, derives from Theorem 14, which precisely states that the expected conditional variance of $\widehat{f}_n$ is upper bounded by the lower isometry radius, i.e.

$$\mathbb{E}_{\overline{X}} V(\widehat{f}_n | \overline{X}) \lesssim \mathcal{I}_L(n, \mathbb{P}). \tag{3.15}$$

Therefore, if Conjecture 1 is correct and there are models in which $\mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}) \gtrsim \epsilon_U^2 \gg \mathcal{I}_L(n, \mathbb{P})$, this must be due to the variance of conditional expectations:

$$\sup_{f^* \in \mathcal{F}} V\left(\mathbb{E}_{\overline{\xi}}\left[\widehat{f}_n | \overline{X}\right]\right) \gtrsim \epsilon_U^2, \tag{3.16}$$

and $V\left(\mathbb{E}_{\overline{\xi}}\left[\widehat{f}_n | \overline{X}\right]\right)$ is precisely the error term which captures how the geometry of $\mathcal{F}$ varies under different realizations.

3.3 Proof sketches

In this section we sketch the proofs of less-technical results to give the reader a flavor of our methods. Detailed proofs are given in the next section.

3.3.1 Sketch of proof of Theorem 8

The proof uses the probabilistic method [AS16]. Let $f_1, \dots, f_N$ be centers of a minimal ϵ_*-cover of $\mathcal{F}$ with respect to $L_2(\mathbb{P}^{(n)})$, per Definition 2. First, since for any $i \in [N]$, the map $\overline{\xi} \mapsto \sqrt{n}\|\widehat{f}_n(\overline{\xi}) - f_i\|_n$ is 1-Lipschitz, Eq. (3.10) and a union bound ensure that with probability at least $1 - \frac{1}{2N}$, for all $i \in [N]$,

$$\mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - f_i\|_n - \|\widehat{f}_n - f_i\|_n \lesssim \sqrt{\frac{\log N}{n}}.$$

On the other hand, by the pigeonhole principle, there exists at least one $i^* \in [N]$ such that with probability at least $1/N$, $\|\widehat{f}_n - f_{i^*}\|_n \leq \epsilon_*$. Hence, there exists at least one realization of $\overline{\xi}$ for which both bounds hold, and thus, *deterministically*,

$$\mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - f_{i^*}\|_n \lesssim \epsilon_* + \sqrt{\frac{\log N}{n}} \lesssim \epsilon_*$$

where we used the balancing equation (1.7). Another application of (3.10) and integration of tails yields

$$V(\widehat{f}_n) = \mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - \mathbb{E}_{\overline{\xi}}\widehat{f}_n\|_n^2 \leq \mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - f_{i^*}\|_n^2 \lesssim \epsilon_*^2,$$

implying that the variance of ERM is minimax optimal.

3.3.2 Proof of Theorem 9

By the definition of the minimax risk, there exists some $f^* \in \mathcal{F}$ with risk at least $\delta^2 := \mathcal{M}(\mathcal{F}, \mathbb{P}^{(n)})$. By translating $\mathcal{F}$, we may assume $f^* = 0$ without loss of generality, so that $\mathbb{E}_{\overline{\xi}}[\|\widehat{f}_n\|_n^2] \gtrsim \delta^2$.

Write $\widehat{f}_n(\overline{\xi})$ for the ERM computed when the target function is $f^* = 0$ and the noise is $\overline{\xi}$, namely, the projection of $\overline{\xi}$ onto $\mathcal{F}$. We wish to show that $\mathbb{E}_{\overline{\xi}}[\|\hat{f}_n - \mathbb{E}\widehat{f}_n\|_n^2] \gtrsim \delta^4$.

The fact that $\widehat{f}_n(\overline{\xi})$ is the projection of the observation vector $\overline{\xi}$ on the convex set $\mathcal{F}$ implies, by convexity, that $\langle f - \widehat{f}_n, \widehat{f}_n - \overline{\xi}\rangle_n \geq 0$ for any $f \in \mathcal{F}$ (see §3.4.1 for the easy argument). Substituting $f = f^* = 0$ and rearranging immediately yields that for any $\overline{\xi}$,

$$\langle \widehat{f}_n(\overline{\xi}), \overline{\xi}\rangle_n \geq \|\widehat{f}_n(\overline{\xi})\|_n^2.$$

Write $f_e = \mathbb{E}_{\overline{\xi}}\hat{f}_n(\overline{\xi})$. Since $\mathbb{E}_{\overline{\xi}}\overline{\xi} = 0$, we may take expectations and insert $\hat{f}_e$ to obtain

$$\mathbb{E}_{\overline{\xi}}\langle \widehat{f}_n(\overline{\xi}) - f_e, \overline{\xi}\rangle_n \geq \mathbb{E}_{\overline{\xi}}\|\hat{f}_n(\overline{\xi})\|_n^2 \geq \delta^2.$$

Applying Cauchy-Schwarz, we obtain

$$\mathbb{E}_{\overline{\xi}}\|\widehat{f}_n(\overline{\xi}) - f_e\|_n^2 \cdot \mathbb{E}_{\overline{\xi}}\|\overline{\xi}\|_n^2 \geq \delta^4,$$

and because the noise is isotropic we immediately obtain

$$V(\widehat{f}_n) = \mathbb{E}_{\overline{\xi}}[\|\hat{f}_n(\overline{\xi}) - f_e\|_n^2] \geq \delta^4,$$

as desired.

3.3.3 Sketch of proof of Theorem 14

As is well-known, the Lipschitz concentration condition (3.10) is equivalent to an isoperimetric phenomenon: for any set $A \subset \mathbb{R}^n$ with $\Pr_{\overline{\xi}}(A) \geq 1/2$, its t-neighborhood $A_t = \{\overline{\xi} \in \mathbb{R}^n : \inf_{x \in A} \|x - \overline{\xi}\|_n \leq t\}$ satisfies

$$\Pr_{\overline{\xi}}(A_t) \geq 1 - 2\exp(-nt^2/2). \tag{3.17}$$

One sees quickly that this implies that if A has measure at least $2\exp(-nt^2/2)$, then A_{2t} has measure $1-2\exp(-nt^2/2)$.

Let $\mathcal{E}$ be the event of Definition 9. As in sub-section 3.3.1, one obtains via the pigeonhole principle and the definition of ϵ_* that there exists some $f_c \in \mathcal{F}$ such that

$$\Pr_{\overline{\xi}}(\underbrace{\{\widehat{f}_n \in B(f_c,\epsilon_*)\} \cap \mathcal{E}}_{A}\,|\overline{\mathrm{X}}) \geq \frac{\Pr_{\overline{\xi}}(\mathcal{E})}{\mathcal{N}(\epsilon_*,\mathcal{F},\mathbb{P})} \geq \exp(-C_3 n\epsilon_*^2).$$

By isoperimetry, $\Pr_{\overline{\xi}}(A_{2t}) \geq 1-2\exp(-nt^2/2)$, where $t = M\epsilon_*/2$ and M is chosen such that $(M/2)^2 \geq 2C_3$; this fixes the value of the absolute constant M used in (3.12).

Applying (3.12) yields that if $\overline{\xi} \in A \subset \mathcal{E}$ and $\|\overline{\xi}' - \overline{\xi}\|_n \leq M\epsilon_* = 2t$, $\|\widehat{f}_n(\overline{\xi}) - \widehat{f}_n(\overline{\xi}')\| \leq \rho_S(\overline{\mathrm{X}}, f^*)$ and so $\|\widehat{f}_n(\overline{\xi}') - f_c\| \leq \epsilon_* + \rho_S(\overline{\mathrm{X}}, f^*)$. This implies

$$\Pr_{\overline{\xi}}(\{\widehat{f}_n \in B(f_c,\epsilon_* + \rho_S(\overline{\mathrm{X}}, f^*))\}|\overline{\mathrm{X}}) \geq \Pr_{\overline{\xi}}(A_{2t}|\overline{\mathrm{X}}) \geq 1-2\exp(-nt^2/2),$$

which implies via conditional expectation that $V(\widehat{f}_n|\overline{\mathrm{X}}) \lesssim \max\{\rho_S(\overline{\mathrm{X}}), \epsilon_*^2\}$, as desired (where we used that $\epsilon_*^2 \gtrsim \log(n)/n$, and therefore $\exp(-nt^2) = O(\epsilon_*^2)$).

3.4 Remaining Proofs

We begin with additional notation: Given $x_1,\ldots,x_n$ and $h : \mathcal{X} \to \mathbb{R}$, we denote

$$G_h = n^{-1}\sum_{i=1}^{n} \xi_i h(x_i).$$

For $g \in \mathcal{F}$, we set $B_n(g,t) := \{f \in \mathcal{F} : \|f-g\|_n \leq t\}$, and $B(g,t) := \{f \in \mathcal{F} : \|f-g\| \leq t\}$. Throughout the proof, we denote by $c, c_1, c_2, \ldots \in (0,1)$, an $C, C_1, \ldots \geq 0$ absolute constants (not depending on $\mathcal{F}$ or on n) that may change from line to line.

3.4.1 Proof of Theorem 8

First, we show that for all $t \geq 0$ and for any fixed $f \in \mathcal{F}$, the following holds:

$$\Pr_{\overline{\xi}}\Big\{\Big|\|\widehat{f}_n - f\|_n - \mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - f\|_n\Big| \geq t\Big\} \leq 2\exp(-c_L n t^2), \tag{3.18}$$

Indeed, this will follow immediately from the LCP condition (3.10) with $L = n^{-1/2}$ if we prove that $h(\overline{\xi}) = \|\widehat{f}_n(\overline{\xi}) - f^*\|_n$ is a $n^{-1/2}$-Lipschitz function.

To prove this claim, observe that $\widehat{f}_n(\overline{\xi})$ is the projection of $\overline{Y} = f^* + \overline{\xi}$ onto the convex set

$$\mathcal{F}_n := \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n.$$

Therefore, we obtain

$$\begin{aligned} |h(\overline{\xi}_1) - h(\overline{\xi}_2)| = ||\|\widehat{f}_n(\overline{\xi}_1) - f^*\|_n - \|\widehat{f}_n(\overline{\xi}_2) - f^*\|_n| &\leq \|\widehat{f}_n(\overline{\xi}_1) - \widehat{f}_n(\overline{\xi}_2)\|_n \\ &\leq \|\overline{\xi}_1 - \overline{\xi}_2\|_n = n^{-1/2}\|\overline{\xi}_1 - \overline{\xi}_2\|_2, \end{aligned}$$

where we have used the fact that the projection to a convex set is a contracting operator. This concludes the proof of (3.18).

Next, fix $\epsilon > 0$ (to be chosen later), let $\mathcal{N}(\epsilon) := \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}^{(n)})$, and let $A = \{f_1, \dots, f_{\mathcal{N}}\}$ be a minimal ϵ-net of $\mathcal{F}$. By the pigeonhole principle, there exists at least one element $f_\epsilon \in A$ such that

$$\Pr(\|\widehat{f}_n - f_\epsilon\|_n \leq \epsilon) \geq 1/\mathcal{N}(\epsilon). \tag{3.19}$$

Also, setting $f = f_\epsilon$ in (3.18) we have

$$\Pr\Big(|\|\widehat{f}_n - f_\epsilon\|_n - \mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - f_\epsilon\|_n)| \geq t\Big) \leq 2\exp(-c_L n t^2). \tag{3.20}$$

Taking $t = \log(4)\sqrt{\frac{\log \mathcal{N}(\epsilon)}{c_L n}}$ in (3.20) yields

$$\Pr\left(\left| \|\widehat{f}_n - f_\epsilon\|_n - \mathbb{E}_{\overline{\xi}} \|\widehat{f}_n - f_\epsilon\|_n \right| \geq \log(4)\sqrt{\frac{\log \mathcal{N}(\epsilon)}{c_L n}} \right) \leq \frac{1}{2\mathcal{N}(\epsilon)}. \tag{3.21}$$

Combining (3.19) and (3.21) via the union bound we obtain

$$\Pr\left(\|\widehat{f}_n - f_\epsilon\|_n \leq \epsilon, \left| \|\widehat{f}_n - f_\epsilon\|_n - \mathbb{E}_{\overline{\xi}} \|\widehat{f}_n - f_\epsilon\|_n \right| \leq \log(4)\sqrt{\frac{\log \mathcal{N}(\epsilon)}{c_L n}} \right) \geq \frac{1}{2|\mathcal{N}(\epsilon)|} > 0$$

Since the event of the last equation holds with positive probability, we must have

$$\mathbb{E}_{\overline{\xi}} \|\widehat{f}_n - f_\epsilon\|_n \leq \epsilon + \log(4)\sqrt{\frac{\log \mathcal{N}(\epsilon)}{c_L n}}.$$

To optimize the RHS over ϵ, we take ϵ such that $\epsilon = \sqrt{\log \mathcal{N}(\epsilon)/(c_L n)}$ — i.e., $\epsilon = \epsilon_*$ — and get

$$\mathbb{E}_{\overline{\xi}} \|\widehat{f}_n - f_\epsilon\|_n \leq C\epsilon_* / \sqrt{c_L}.$$

Substituting in (3.20) and taking $t = \epsilon_*$, we obtain

$$\Pr(|\|\widehat{f}_n - f_\epsilon\|_n \geq C\epsilon_* / \sqrt{c_L}) \leq 2\exp(-cn\epsilon_*^2).$$

This easily implies that $\mathbb{E}[\|\widehat{f}_n - f_\epsilon\|_n^2] \leq C_1 \epsilon_*^2 / c_L$, and therefore also

$$\mathbb{E}[\|\widehat{f}_n - \mathbb{E}\widehat{f}_n\|_n] \leq (\mathbb{E}[\|\widehat{f}_n - \mathbb{E}\widehat{f}_n\|_n^2])^{1/2} \leq (\mathbb{E}[\|\widehat{f}_n - f_\epsilon\|_n^2])^{1/2} \leq C_2 \epsilon_* / \sqrt{c_L}.$$

Applying (3.18) once again, now with $f = \mathbb{E}\widehat{f}_n$, we obtain

$$\Pr(\|\widehat{f}_n - \mathbb{E}\widehat{f}_n\|_n^2 \geq C_2 \epsilon_*^2 / c_L) \leq 2\exp(-cn\epsilon_*^2). \tag{3.22}$$

Therefore, the claim of the theorem follows for $\delta_n = 0$ (i.e. all the exact minimizes).

Next, condition on the high-probability event of (3.22) above, and consider some $f \in \mathcal{O}_{\delta_n}$. Since

$$\|f - \mathbb{E}\widehat{f}_n\|_n^2 \leq 2(\|f - \widehat{f}_n\|_n^2 + \|\widehat{f}_n - \mathbb{E}\widehat{f}_n\|_n^2),$$

to obtain the theorem it suffices to show that for any $f \in \mathcal{O}_{\delta_n}$, we have

$$\|f - \widehat{f}_n\|_n^2 \leq \delta_n^2$$

deterministically.

This is a matter of elementary convex geometry: we know that $\widehat{f}_n$ is the closest point in the convex set $\mathcal{F}_n$ to the point $\overline{Y}$, which implies that the ball $B = B(\overline{Y}, \|\widehat{f}_n - \overline{Y}\|_n)$ is tangent to $\mathcal{F}_n$ at $\widehat{f}_n$. This implies that $\mathcal{F}_n$ is contained within the positive half-space H^+ defined by the supporting hyperplane of B at $\widehat{f}_n$, i.e.,

$$\mathcal{F}_n \subset H^+ = \{f : \langle \widehat{f}_n - \overline{Y}, f - \overline{Y} \rangle \geq \|\widehat{f}_n - \overline{Y}\|^2\}. \tag{3.23}$$

We now compute:

$$\|f - \overline{Y}\|_n^2 = \|f - \widehat{f}_n\|_n^2 + \|\widehat{f}_n - \overline{Y}\|_n^2 + 2\langle \widehat{f}_n - \overline{Y}, f - \widehat{f}_n \rangle_n.$$

Since $f \in \mathcal{F}_n$, (3.23) implies that $\langle f - \overline{Y}, \widehat{f}_n - \overline{Y} \rangle_n \geq \langle \widehat{f}_n - \overline{Y}, \widehat{f}_n - \overline{Y} \rangle_n$, or equivalently, $\langle f - \widehat{f}_n, \widehat{f}_n - \overline{Y} \rangle_n \geq 0$. Hence we obtain

$$\|f - \widehat{f}_n\|_n^2 \leq \|f - \overline{Y}\|_n^2 - \|\widehat{f}_n - \overline{Y}\|_n^2,$$

but the RHS is at most δ_n by the definition of $\mathcal{O}_{\delta_n}$. This concludes the proof.

3.4.2 Proof of Theorems 10 and 11

Here we only prove Theorem 11, as the proof will reveal, almost identical arguments imply Theorem 10. For completeness, the proof of Theorem 10 appears in §3.5 below. For simplicity, we prove

Theorem 11 assuming that the ERM is unbiased, in which case we have $V(\widehat{f}_n) = \mathrm{E}_{f^*}^2$. In general, if $B^2(\widehat{f}_n) \leq 2V(\widehat{f}_n)$ we have $V(\widehat{f}_n) \geq \frac{1}{3}\mathrm{E}_{f^*}^2$ as $B^2(\widehat{f}_n) + V(\widehat{f}_n) = \mathrm{E}_{f^*}^2$, and the proof goes along the same lines.

To estimate the minimax risk of the class $\mathcal{G}_* = \{f \in \mathcal{F} : \|f - f^*\|_n^2 \leq 4 \cdot \mathrm{E}_{f^*}^2 + S/n\}$, where $S > 0$ will be defined below. We use the following classical characterization [YB99] (that holds for Gaussian noise and for any class whose diameter is at most 1):

$$\mathcal{M}(\mathcal{G}_*, \mathbb{P}^{(n)}) \sim \min_{\epsilon} \left[\frac{\log \mathcal{N}(\epsilon, \mathcal{G}_*, \mathbb{P}^{(n)})}{n} + \epsilon^2 \right]. \tag{3.24}$$

Case I: $\mathrm{E}_{f^*}^2 \leq Sn^{-1}$. It is well known (see, e.g., [Wai19, Example 15.4]) that the minimax rate is at least $\Omega(\mathrm{Diam}_{\mathbb{P}^{(n)}}(\mathcal{G}_*)^2)$ on the restricted class $\mathcal{G}_*$. Namely,

$$\mathcal{M}(\mathcal{G}_*, \mathbb{P}^{(n)}) \geq c \cdot \mathrm{Diam}_{\mathbb{P}^{(n)}}(\mathcal{G}_*)^2 \geq \frac{c_1}{S}\mathrm{E}_{f^*}^2,$$

for some $c, c_1 \in (0,1)$. Next, recall that $\widehat{f}_n$ is a 1-Lipschitz map in $\overline{\xi}$ (and hence in $\overline{Y}$). Therefore, we conclude that

$$\mathcal{R}(\widehat{f}_n, \mathcal{G}_*, \mathbb{P}^{(n)}) \lesssim \mathrm{Diam}_{\mathbb{P}^{(n)}}(\mathcal{G}_*)^2 \lesssim \mathcal{M}(\mathcal{G}_*, \mathbb{P}^{(n)}),$$

and the claim follows for this case. We remark that the assumption of $4 \cdot B^2(\widehat{f}_n) \leq V(\widehat{f}_n)$ is not used our argument. Hence, this result holds for any f^* such that $\mathrm{E}_{f^*}^2 \leq Sn^{-1}$.

Case II: $\mathrm{E}_{f^*}^2 \geq Sn^{-1}$. We will prove that the covering number of $\mathcal{G}_*$ by balls of radius $\mathrm{E}_{f^*}/2$ is not too small:

$$\log \mathcal{N}(\mathrm{E}_{f^*}/2, \mathcal{G}_*, \mathbb{P}^{(n)}) \geq c_1 n \cdot \mathrm{E}_{f^*}^2, \tag{3.25}$$

for some universal constant $c_1 > 0$, to be chosen later.

Since $\mathcal{N}(\epsilon, \mathcal{G}_*, \mathbb{P}^{(n)})$ increases when ϵ decreases, this easily implies that the RHS of (3.24) is at least $c_2\mathrm{E}_{f^*}^2$ and hence, by Eq (3.24), that the minimax risk of $\mathcal{G}_*$ is at least by $c_2\mathrm{E}_{f^*}^2$. On the other hand, $\mathcal{M}(\mathcal{G}_*, \mathbb{P}^{(n)})$ is certainly bounded above by $4\mathrm{E}_{f^*}^2$, as this is an upper bound on the risk of the trivial estimator which returns the fixed element f^* on any sample. Hence, one obtains

$\mathcal{M}(\mathcal{G}_*, \mathbb{P}^{(n)}) \asymp \mathrm{E}_{f^*}^2$. Similarly, $\mathcal{R}(\widehat{f}_n, \mathcal{G}_*, \mathbb{P}^{(n)}) \asymp \mathrm{E}_{f^*}^2$, as on the one hand, the risk of $\widehat{f}_n$ cannot be smaller than the minimax risk, and on the other, the risk of any proper estimator on a class is bounded by the squared diameter of the class.

So it suffices to prove that (3.25) holds for an appropriate $c_1 > 0$. The proof strategy is very similar to that of Theorem 8. Suppose for the sake of contradiction that

$$\log \mathcal{N}(\mathrm{E}_{f^*}/2, B_n(f^*, 2\mathrm{E}_{f^*}), \mathbb{P}^{(n)}) < c_1 n \cdot \mathrm{E}_{f^*}^2.$$

We consider the distribution of the ERM $\widehat{f}_n$ when the true function is f^*. First, note that as $\mathbb{E}\|\widehat{f}_n - f^*\|^2 = \mathrm{E}_{f^*}^2$, and $\mathbb{E}\widehat{f}_n = f^*$, we have that

$$\Pr(\widehat{f}_n \in \mathcal{G}_*) = \Pr(\widehat{f}_n \in B_n(f^*, 2\mathrm{E}_{f^*})) = 1 - \Pr(\|f^* - \mathbb{E}\widehat{f}_n\|_n \geq 2\mathrm{E}_{f^*}) \geq 3/4$$

by Chebyshev's inequality. Let $A = \{f_1, \ldots, f_{\mathcal{N}}\}$ be a minimal $\mathrm{E}_{f^*}/2$- net in $\mathcal{G}_*$; by the pigeonhole principle, there exists at least one element $g \in A$ such that

$$\Pr(\|\widehat{f}_n - g\|_n \leq \mathrm{E}_{f^*}/2) \geq \frac{3}{4\mathcal{N}} \geq \frac{3}{4}\exp(-c_1 n \mathrm{E}_{f^*}^2). \tag{3.26}$$

Next, we apply (3.18) with $f = g$ and $t = \mathrm{E}_{f^*}/3$ to obtain

$$\Pr(\left| \|\widehat{f}_n - g\|_n - \mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - g\|_n \right| < \mathrm{E}_{f^*}/9) \geq 1 - 2\exp(-c_L n \mathrm{E}_{f^*}^2/9). \tag{3.27}$$

Recalling that we are in the case $\mathrm{E}_{f^*}^2 \geq S n^{-1}$, by choosing $c_1 > 0$ small enough and $S > 0$ large enough we can ensure that $\exp(n\mathrm{E}_{f^*}^2(1/18 - c_1)) > 8/3$, or equivalently

$$\frac{3}{4}\exp(-c_1 n \mathrm{E}_{f^*}^2) - 2\exp(-n\mathrm{E}_{f^*}^2/18) > 0. \tag{3.28}$$

Combining (3.26), (3.27), and (3.28) yields

$$\Pr(\|\widehat{f}_n - g\|_n \leq \mathrm{E}_{f^*}/2) + \Pr\left(\left| \|\widehat{f}_n - g\|_n - \mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - g\|_n \right| < \mathrm{E}_{f^*}/6\right) > 1,$$

so the two events

$$\left\{ \left| \|\widehat{f}_n - g\|_n - \mathbb{E}_{\overline{\xi}} \|\widehat{f}_n - g\|_n \right| < \mathrm{E}_{f^*}/6 \right\}, \{ \|\widehat{f}_n - g\|_n \leq \mathrm{E}_{f^*}/2 \}$$

have nonempty intersection, which implies that $\mathbb{E}_{\overline{\xi}} \|\widehat{f}_n - g\|_n < 2\mathrm{E}_{f^*}/3$.

Let $h(\xi) = \|\widehat{f}_n - g\|_n$. We have

$$\mathbb{E}h^2 = (\mathbb{E}h)^2 + \mathbb{E}(h - \mathbb{E}h)^2 < 4\mathrm{E}_{f^*}^2/9.$$

As h is $\frac{1}{\sqrt{n}}$-Lipschitz, the LCP implies that h is $\frac{1}{\sqrt{n}}$-subgaussian. Thus $h - \mathbb{E}h$ is a centered $\frac{1}{\sqrt{n}}$-subgaussian random variable, so $\mathbb{E}(h - \mathbb{E}h)^2 \leq \frac{2}{n}$ [Ver18, Proposition 2.5.2], and hence

$$\mathbb{E}_{\xi} \|\widehat{f}_n - g\|_n^2 < \frac{4}{9}\mathrm{E}_{f^*}^2 + \frac{2}{n}. \tag{3.29}$$

Again recalling that $\mathrm{E}_{f^*}^2 > Sn^{-1}$, by taking S large enough we can ensure that $\mathbb{E}_{\xi} \|\widehat{f}_n - g\|_n^2 < \mathrm{E}_{f^*}^2$, contradicting the assumption $V(\widehat{f}_n) = \mathrm{E}_{f^*}^2$.

3.4.3 Proof of Corollary 2

Here, we prove this result under the additional Assumption 3. For completeness, we provide a proof without this assumption in §3.5.1 below. We remark that it is very similar to the argument that we present here and does not provide any new insights, and therefore omitted.

Consider the map $F : \mathcal{F}_n \to \mathbb{R}^n$ defined via

$$f^* \to \mathbb{E}_{\overline{\xi}} \widehat{f}_n,$$

i.e., f^* maps to the expectation of the ERM $\widehat{f}_n$ when the underlying function is $f^* \in \mathcal{F}_n$. One verifies easily that F is continuous, since projection to a convex set is a 1-Lipschitz function; in addition, the convexity of $\mathcal{F}_n$ implies that $F(f^*) \in \mathcal{F}_n$ for all $f^* \in \mathcal{F}_n$. Thus F is a continuous map from the compact convex set $\mathcal{F}_n$ to itself, so by the Brouwer fixed point theorem, there exists an $f^* \in \mathcal{F}_n$ such that

$$f^* = \mathbb{E}_{\overline{\xi}} \widehat{f}_n.$$

Let $\mathrm{E}_{f^*}^2 = \mathbb{E}_{\overline{\xi}} \|\widehat{f}_n - f^*\|_n^2 = V(\widehat{f}_n)$, and let P denote the projection to $\mathcal{F}$, which is is 1-Lipschitz. For any $f \in \mathcal{G}_* = B(f^*, 2\mathrm{E}_{f^*}) \cap \mathcal{F}$ and any ξ we have

$$\begin{aligned} \|P(f+\overline{\xi}) - f\|_n^2 &\leq 3(\|P(f+\overline{\xi}) - P(f^*+\overline{\xi})\|_n^2 + \|P(f^*+\overline{\xi}) - f^*\|_n^2 + \|f^* - f\|_n^2) \\ &\leq 3(\|P(f^*+\overline{\xi}) - f^*)\|_n^2 + 8\mathrm{E}_{f^*}^2) = 3(\|\widehat{f}_n - f^*)\|_n^2 + 8\mathrm{E}_{f^*}^2), \end{aligned}$$

and taking the expectation over $\overline{\xi}$ we see that $\mathbb{E}_{\overline{\xi}} \|P(f+\overline{\xi}) - f\|_n^2$, the squared error of the ERM when the underlying function is f, is at most $27\mathrm{E}_{f^*}^2$.

Now let $\hat{h} : \mathbb{R}^n \to \mathcal{F}$ be any estimator. By Theorem 11, we have

$$\mathrm{E}_{f^*}^2 \lesssim \max_{f \in \mathcal{G}_*} \|\hat{h}(f+\xi) - f\|_n^2.$$

Picking $f \in \mathcal{G}_*$ which maximizes the error of $\hat{h}$, we have that the squared error of $\widehat{f}_n$ on f is upper-bounded by $c \cdot \mathrm{E}_{f^*}^2$ and the squared error of $\hat{h}$ on f is lower-bounded by $c_1 \cdot \mathrm{E}_{f^*}^2$, which is precisely what we want.

Remark 21. The Brouwer fixed point theorem, which we use in the first step of the proof to obtain the existence of $f^* \in \mathcal{F}$ for which $\mathbb{E}_{\overline{\xi}} \widehat{f}_n = f^*$, is a deep result, and one may ask whether it is essential to the proof. Another commonly used fixed-point theorem is that due to Banach; the Banach fixed point theorem is elementary, but requires a bound $\|F(f) - F(g)\|_n \leq c\|f - g\|_n$ for some $c < 1$ and all $f, g \in \mathcal{F}$.

One has

$$\|F(f) - F(g)\|_n \leq \mathbb{E}_{\overline{\xi}} \|P(f+\overline{\xi}) - P(g+\overline{\xi})\|_n, \tag{3.30}$$

where P denotes the projection to $\mathcal{F}$. Note that $\|P(f+\overline{\xi}) - P(g+\overline{\xi})\|_n \leq \|f - g\|_n$ because P is 1-Lipschitz. Also, it's easy to see that there exists some $\overline{\xi}$ for which $\|P(f+\overline{\xi}) - P(g+\overline{\xi})\|_n$ is strictly smaller than $\|f - g\|_n$, and continuity of P ensures that the same holds for all $\overline{\xi}'$ sufficiently close to $\overline{\xi}$, implying $\|F(f) - F(g)\|_n < \|f - g\|_n$. But this is not yet sufficient to apply the Banach fixed point theorem.

Via more delicate convex-geometric arguments, though, one can show that if $\|\xi\|_n$ is sufficiently large compared to the diameter of $\mathcal{F}$ (say, $\|\xi\|_n > C \cdot \mathrm{Diam}(\mathcal{F})$) and $\langle f - g, \overline{\xi} \rangle_n \geq \epsilon \|f - g\|_n \|\xi\|_n$ (i.e., the angle between $f - g$ and $\overline{\xi}$ is bounded away from 90 degrees) then

$$\|P(f + \overline{\xi}) - P(g + \overline{\xi})\|_n \leq (1 - \delta)\|f - g\|_n$$

for some δ depending on ϵ and C, which allows one to conclude, using (3.30), that $\|F(f) - F(g)\|_n \leq c\|f - g\|_n$ for some $c < 1$ and all $f, g \in \mathcal{F}$. Hence, the Banach fixed point theorem can be used in the proof instead of the Brouwer fixed point theorem, rendering it elementary but more technical.

3.4.4 Proof of Theorem 13

Preliminaries The main tool we use from the theory of empirical processes is Talagrand's inequality [Kol11, Theorem 2.6]:

Lemma 13. *Let $\mathcal{H}$ be a class of functions on a domain $\mathcal{Z}$ all of which are uniformly bounded by M. Let $Z_1, \ldots, Z_n \underset{i.i.d.}{\sim} \mathbb{P}$. Then, there exist universal constants $C, c > 0$ such that*

$$\Pr(|\|\mathcal{H}\|_n - \mathbb{E}\|\mathcal{H}\|_n| \geq t) \leq C \exp\left(-\frac{cnt}{M} \log\left(1 + \frac{t}{\mathbb{E}\|\mathcal{H}^2\|_n}\right)\right),$$

where $\|\mathcal{H}\|_n := \sup_{h \in \mathcal{H}} n^{-1} \sum_{i=1}^n h(Z_i)$, and $\|\mathcal{H}^2\|_n = \sup_{h \in \mathcal{H}} n^{-1} \sum_{i=1}^n h(Z_i)^2$.

We abbreviate $\mathcal{I}_L := \mathcal{I}_L(n, \mathbb{P}), \mathcal{I}_U := \mathcal{I}_U(n, \mathbb{P})$. For every fixed $\overline{X}, \overline{\xi}$, the function $\hat{\mathcal{L}} : \mathcal{F} \to \mathbb{R}$ defined by

$$\hat{\mathcal{L}}(f) := \|\overline{Y} - f\|_n^2 - \|\overline{\xi}\|_n^2 = -2G_{f - f^*} + \|f - f^*\|_n^2, \tag{3.31}$$

satisfies $\widehat{f}_n = \mathrm{argmin}_{f \in \mathcal{F}} \hat{\mathcal{L}}(f)$. (Of course, $\hat{\mathcal{L}}(f)$ is just the empirical loss of f, up to subtracting a constant.) Note that $\hat{\mathcal{L}}(f^*) = 0$, so $\hat{\mathcal{L}}(\widehat{f}_n)$ must be non-positive.

Let $\{f_1, \ldots, f_N\}$ be an ϵ_U-net of $\mathcal{F}$ with respect to $\mathbb{P}$ of cardinality $N = \mathcal{N}(\epsilon_U, \mathcal{F}, \mathbb{P})$; for each $i \in [N]$, let $B(f_i) := B(f_i, \epsilon_U)$ denote the ball of radius ϵ_U around f_i, so that the $B(f_i)$ cover

$\mathcal{F}$. For each i, let L_i denote the minimal loss on the ball $B(f_i)$:

$$L_i := \min_{f \in B(f_i)} \hat{\mathcal{L}}(f) \tag{3.32}$$

The main technical result is the following lemma:

Lemma 14. *Fix $i_* \in [N]$. For any absolute constant $A > 0$, there exist absolute constants $C_1, C, >$ 0, such that the following holds with probability of at least $1 - 2\exp(-An\epsilon_U^4 / \max\{\mathcal{I}_U, \epsilon_*^2\})$:*

$$\forall i \in [N], \ \ |(L_i - L_{i_*}) - \mathbb{E}(L_i - L_{i_*})| \leq C_1\epsilon_U^2 + \frac{1}{4}\|f_i - f_{i_*}\|^2. \tag{3.33}$$

We defer the proof of Lemma 14 to the end of the section, and show how it implies the theorem.

Proof of Theorem 13 (assuming Lemma 14). We apply Lemma 14 with $i_* = \mathrm{argmin}_{i \in [N]} \mathbb{E} L_i$. Let $\mathcal{E}$ denote the event of Lemma 14 (the constant $A > 0$ in the lemma will be chosen shortly), and let $\mathcal{E}'$ be the event that

$$\|f - g\|_n^2 \geq \frac{1}{2}\|f - g\|^2 - \mathcal{I}_L \tag{3.34}$$

for all $f, g \in \mathcal{F}$. By the definition of $\mathcal{I}_L$, $\mathcal{E}'$ holds with probability $1 - n^{-1}$; in addition, a mildly tedious computation, which we defer to Lemma 16, shows that A can be chosen such that $\Pr(\mathcal{E}) \geq 1 - n^{-1}$ as well. In the remainder of the proof, we work on $\mathcal{E} \cap \mathcal{E}'$.

Let $\widehat{f_{i^*}} = \mathrm{argmin}_{f \in B(f_{i_*})} \hat{\mathcal{L}}(f)$, so that $L_{i_*} = \hat{\mathcal{L}}(\widehat{f_{i^*}})$. Consider the function $h = \frac{\widehat{f_{i^*}} + \widehat{f_n}}{2}$, which lies in $\mathcal{F}$ as $\mathcal{F}$ is convex. We have

$$\begin{aligned} \hat{\mathcal{L}}(h) &= -2G_{h - f^*} + \|h - f^*\|_n^2 \\ &= \frac{\hat{\mathcal{L}}(\widehat{f_{i^*}}) + \hat{\mathcal{L}}(\widehat{f_n})}{2} + \left(\|h - f^*\|_n^2 - \frac{\|\widehat{f_{i^*}} - f^*\|_n^2 + \|\widehat{f_n} - f^*\|_n^2}{2} \right). \end{aligned} \tag{3.35}$$

Applying the parallelogram law

$$\|a + b\|_n^2 + \|a - b\|_n^2 = 2\|a\|_n^2 + 2\|b\|_n^2$$

with $a = \frac{\widehat{f}_{i^*} - f^*}{2}$, $b = \frac{\widehat{f}_n - f^*}{2}$ yields

$$\|h - f^*\|_n^2 - \frac{\|\widehat{f}_{i^*} - f^*\|_n^2 + \|\widehat{f}_n - f^*\|_n^2}{2} = - \left\| \frac{\widehat{f}_{i^*} - \widehat{f}_n}{2} \right\|_n^2 .$$

Combining this equation with (3.35) yields

$$\hat{\mathcal{L}}(h) \leq \frac{\hat{\mathcal{L}}(\widehat{f}_{i^*}) + \hat{\mathcal{L}}(\widehat{f}_n)}{2} - \left\| \frac{\widehat{f}_{i^*} - \widehat{f}_n}{2} \right\|_n^2 .$$

But we also know that $\hat{\mathcal{L}}(h) \geq \hat{\mathcal{L}}(\widehat{f}_n)$ by the definition of $\widehat{f}_n$, so rearranging we obtain

$$\hat{\mathcal{L}}(\widehat{f}_n) \leq \hat{\mathcal{L}}(\widehat{f}_{i^*}) - \frac{1}{2}\|\widehat{f}_{i^*} - \widehat{f}_n\|_n^2. \tag{3.36}$$

Now let $i \in N$ such that $\widehat{f}_n \in B(f_i)$ and substitute $L_i = \hat{\mathcal{L}}(\widehat{f}_n)$, giving

$$L_i \leq L_{i_*} - \frac{1}{2}\|\widehat{f}_{i^*} - \widehat{f}_n\|_n^2.$$

Since we are on $\mathcal{E}$, we may apply (3.33) and obtain

$$\mathbb{E}L_i \leq \mathbb{E}L_{i_*} + C_3\epsilon_U^2 - \frac{1}{4}\|\widehat{f}_{i^*} - \widehat{f}_n\|_n^2 \leq \mathbb{E}L_{i_*} + C_4\epsilon_U^2 + \mathcal{I}_L - \frac{1}{8}\|\widehat{f}_{i^*} - \widehat{f}_n\|^2.$$

But $\mathbb{E}L_{i_*} \leq \mathbb{E}L_i$ by our choice of i_*, which implies finally that $\|\widehat{f}_{i^*} - \widehat{f}_n\|_n^2 \lesssim \max\{\epsilon_U^2, \mathcal{I}_L\} = \epsilon_V^2$.

Recall that we are also interested in $f \in \mathcal{O}_{\delta_n}$ for $\delta_n = O(\epsilon_V^2)$. By the geometric argument in the proof of Theorem 8, for such f, we have $\|f - \hat{f}_n\|_n = O(\delta_n)$.

Applying the lower isometry property (3.34), we obtain

$$\|\widehat{f}_n - \widehat{f}_{i^*}\|, \|f - \widehat{f}_n\| \leq C\epsilon_V$$

for any $f \in \mathcal{O}_{\delta_n}$ on $\mathcal{E} \cap \mathcal{E}'$. Since $\|\widehat{f}_{i^*} - f_{i_*}\| \leq \epsilon_V$ (as $\widehat{f}_{i^*} \in B(f_{i_*})$ by definition), we also have

$$\|\widehat{f}_n - f_{i_*}\|, \|f - f_{i_*}\| \leq C\epsilon_V.$$

In sum, thus far we have shown that under $\mathcal{E} \cap \mathcal{E}'$, an event of probability at least $1 - 2n^{-1}$ any $f \in \mathcal{O}_{\delta_n}$ satisfies $\|f - f_{i_*}\| \lesssim \epsilon_V$. It remains to show that this implies that $\|f - \mathbb{E}\widehat{f}_n\| \lesssim \epsilon_V$, for which it suffices to show that $\|\mathbb{E}\widehat{f}_n - f_{i_*}\| \lesssim \epsilon_V$. But

$$\mathbb{E}\widehat{f}_n - f_{i_*} = (\mathbb{E}[\widehat{f}_n|\mathcal{E} \cap \mathcal{E}'] - f_{i_*})\Pr(\mathcal{E} \cap \mathcal{E}') + (\mathbb{E}[\widehat{f}_n|(\mathcal{E} \cap \mathcal{E}')^c] - f_{i_*})\Pr((\mathcal{E} \cap \mathcal{E}')^c).$$

By what we have shown, $\|\mathbb{E}[\widehat{f}_n|\mathcal{E} \cap \mathcal{E}'] - f_{i_*}\| \leq C\epsilon_V$, while $\|\mathbb{E}[\widehat{f}_n|(\mathcal{E} \cap \mathcal{E}')^c] - f_{i_*}\| = O(1)$ by Assumption 11 and so the norm of the second term is asymptotically bounded by $O(n^{-1}) \ll \epsilon_* \leq \epsilon_V$ because $\epsilon_* \gtrsim \sqrt{\frac{\log n}{n}}$ by Assumption 10. This concludes the proof. □

It remains to prove the deferred lemmas. We begin with the most substantial one, Lemma 14.

Proof of Lemma 14. Recall that

$$-L_i = \sup_{f \in B(f_i)} (2G_{f-f^*} - \|f - f^*\|_n^2).$$

We write $f - f^* = (f - f_i) + (f_i - f^*)$, expand, and decompose this expression into terms depending on $f - f_i$ and terms depending only on $f_i - f^*$:

$$-L_i = A_i + A_i', \tag{3.37}$$

where

$$A_i := \sup_{f \in B(f_i)} \left(2G_{f-f_i} - \|f - f_i\|_n^2 - 2\langle f - f_i, f_i - f^*\rangle_n\right),$$
$$A_i' := 2G_{f_i-f^*} - \|f_i - f^*\|_n^2.$$

We also write

$$\begin{aligned}
B_i := A'_i - A'_{i_*} &= 2G_{f_i - f^*} - \|f_i - f^*\|_n^2 - (2G_{f_{i_*} - f^*} - \|f_{i_*} - f^*\|_n^2) \\
&= 2G_{f_i - f_{i_*}} - \|f_i\|_n^2 + \|f_{i_*}\|_n^2 + 2\langle f_i - f_{i_*}, f^* \rangle \\
&= 2G_{f_i - f_{i_*}} + \|f_i - f_{i_*}\|_n^2 - 2\|f_i\|_n^2 + 2\langle f_i, f_{i^*} \rangle + 2\langle f_i - f_{i_*}, f^* \rangle \\
&= 2G_{f_i - f_{i_*}} + \|f_i - f_{i_*}\|_n^2 + 2\langle f_i - f_{i_*}, f^* - f_i \rangle
\end{aligned}$$

We claim that with probability $1 - C\exp(-cn\epsilon_U^4 / \max\{\epsilon_*^2, \mathcal{I}_U\})$ the following holds:

$$\forall i \in [N], \quad |A_i - \mathbb{E}A_i| \le C_1\epsilon_U^2, \tag{3.38}$$

$$\forall i \in [N], \quad |B_i - \mathbb{E}B_i| \le C_2\epsilon_U^2 + \frac{1}{4}\|f_i - f_{i_*}\|^2. \tag{3.39}$$

Since $L_i - L_{i_*} = A_{i_*} - A_i - B_i$, combining (3.38) and (3.39) yields the lemma.

We first prove (3.38). For each $i \in [N]$, we control fluctuations of A_i by applying Talagrand's inequality. To this end, write

$$A_i = \sup_{f \in B(f_i)} \frac{1}{n}\sum_{j=1}^n a_f(X_j, \xi_j)$$

where

$$a_f(x,\xi) = 2\xi(f(x) - f_i(x)) - 2(f(x) - f_i(x))(f_i(x) - f^*(x)) - (f(x) - f_i(x))^2.$$

To apply Talagrand's inequality, we need to bound $\mathbb{E}\sup_{f \in B(f_i)} n^{-1}\sum_{j=1}^n a_f(X_j, \xi_j)^2$.

Using the identity $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we see that

$$\begin{aligned}
&\mathbb{E}_{\overline{X},\overline{\xi}} \sup_{f \in B(f_i)} \int a_f(X_j, \xi_j)^2 d\mathbb{P}_n \\
&\le 3 \cdot \mathbb{E}_{\overline{X},\overline{\xi}} \sup_{f \in B(f_i)} \int \Big(2\xi^2(f(x) - f_i(x))^2 + 2(f(x) - f_i(x))^2(f_i(x) - f^*(x))^2 + (f(x) - f_i(x))^4\Big)\, d\mathbb{P}_n.
\end{aligned}$$

Using the assumptions $|\xi_i| \le \Gamma_1$, $\|f\|_\infty \le \Gamma_2$, one can obtain that

$$\mathbb{E}_{\overline{X},\overline{\xi}} \sup_{f \in B(f_i)} \int a_f(X_j, \xi_j)^2 d\mathbb{P}_n \le C \cdot \mathbb{E}_{\overline{X}} \sup_{f \in B(f_i)} \int (f - f_i)^2 d\mathbb{P}_n$$

for some $C = O(\max\{\Gamma_1^2, \Gamma_2^2\})$. Using the definition of the upper isometry constant $\mathcal{I}_U$ and the stationary point ϵ_U, we obtain

$$\mathbb{E}_{\overline{X},\overline{\xi}} \sup_{f \in B(f_i)} \int a_f(X_j, \xi_j)^2 d\mathbb{P}_n \lesssim \max\{\mathcal{I}_U, \epsilon_U^2\} \asymp \max\{\mathcal{I}_U, \epsilon_*^2\},$$

where the last step uses Lemma 15 below.

Thus we may apply Talagrand's inequality to A_i with $\mathbb{E}\|\mathcal{H}^2\|_n \lesssim \max\{\epsilon_*^2, \epsilon_U^2\}$, giving

$$\Pr_{\overline{X},\overline{\xi}}\{|A_i - \mathbb{E}A_i| \ge t\} \le C \exp(-cnt \log(1 + t/\max\{\epsilon_*^2, \mathcal{I}_U\})). \tag{3.40}$$

Taking a union bound over $i \in [N]$, we obtain

$$\Pr_{\overline{\xi}}\{\exists i \in [N] : |A_i - \mathbb{E}A_i| \ge t\} \le C \exp(-cnt^2/\max\{\epsilon_*^2, \mathcal{I}_U\} + \log N).$$

Choosing $t = C_1\epsilon_U^2$ for C_1 sufficiently large and recalling that $\log N = \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}) \le n\epsilon_U^4/\max(\mathcal{I}_U, \epsilon_*^2)$ by (3.11), we obtain that

$$\forall i \in [N], \ \ |A_i - \mathbb{E}A_i| \le C_1\epsilon_U^2$$

with probability at least $1 - 2\exp(-c_1 n\epsilon_U^4/\max\{\epsilon_*^2, \mathcal{I}_U\})$, which is (3.38).

Next, we handle B_i for every $i \in [N]$. As in the case of A_i, we may write $B_i = n^{-1}\sum_{j=1}^n b_i(X_i, \xi_i)$ where

$$b_i(x, \xi) = 2\xi(f_i(x) - f_{i_*}(x)) + (f_i(x) - f_{i_*}(x))^2 + 2(f_i(x) - f_{i_*}(x))(f^*(x) - f_i(x)).$$

We have $|b_i(x,\xi)| \le C|f_i(x) - f_{i_*}(x)|$, so as before,

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[b_i(X_j,\xi_j)^2] \le C\mathbb{E}[\|f_i - f_{i_*}\|_n^2] = C\|f_i - f_{i_*}\|^2,$$

and hence by Bernstein's inequality,

$$\Pr(|B_i - \mathbb{E}B_i| \ge t) \le \exp\left(-\frac{cnt^2}{C_3 t + \|f_i - f_{i_*}\|^2}\right).$$

Substituting $t = t_i := C\epsilon_U^2 + \|f_i - f_{i_*}\|^2/4$, we obtain

$$\begin{aligned}\Pr\left(|B_i - \mathbb{E}B_i| \ge C_2\epsilon_U^2 + \frac{\|f_i - f_{i_*}\|^2}{4}\right) &\le 2\exp\left(\frac{-c_1 n t_i^2}{C_3 t_i + \|f_i - f_{i_*}\|^2}\right) \\ &\le 2\exp\left(-c_2 n \max\left\{C\epsilon_U^2, \frac{\|f_i - f_{i_*}\|^2}{4}\right\}\right) \qquad (3.41) \\ &\le 2\exp(-c_3 n \cdot C\epsilon_U^2).\end{aligned}$$

By the same exact argument as in the case of A_i, we may choose $C > 0$ sufficiently large such that with probability $1 - C\exp(-cn\epsilon_U^4/\max(\mathcal{I}_U, \epsilon_*^2))$,

$$|B_i - \mathbb{E}B_i| \le C_2\epsilon_U^2 + \|f_i - f_{i_*}\|^2/4$$

for every $i \in [N]$, which is (3.39). This concludes the proof of Lemma 14. □

Lemma 15. *The following holds:*

$$\max\{\epsilon_U^2, \mathcal{I}_U\} \asymp \max\{\epsilon_*^2, \mathcal{I}_U\}. \qquad (3.42)$$

Proof. If $\mathcal{I}_U \lesssim \epsilon_*^2$, then $\epsilon_*^2 \asymp \epsilon_U^2$ by definition. If $\mathcal{I}_U \gtrsim \epsilon_*^2$, assume to the contrary $\mathcal{I}_U \ll \epsilon_U$; as we have $n\epsilon_U^4 \asymp \mathcal{I}_U \log \mathcal{N}(\epsilon_U, \mathcal{F}, \mathbb{P})$, this implies

$$\log \mathcal{N}(\epsilon_U, \mathcal{F}, \mathbb{P})/n \gg \epsilon_U^2.$$

However, this contradicts the definition of ϵ_* as the maximal choice of ϵ satisfying the last equation. $\square$

Lemma 16. *For a sufficiently large absolute constant $A > 0$, one has* $\exp(-An\epsilon_U^4/\max(\epsilon_U^2, \mathcal{I}_U)) \leq n^{-1}$.

Proof. First, we show that $N = \mathcal{N}(\mathcal{F}, \epsilon_U, \mathbb{P}) \gtrsim n^{1/4}/\log n$. Suppose to the contrary that $N \ll n^{1/4}/\log n$. Then

$$n\epsilon_U^4 \asymp \mathcal{I}_U \cdot \log N \lesssim \log n,$$

since $\mathcal{I}_U$ is at most the squared diameter of $\mathcal{F}_n$, which is $\Theta(1)$. This yields $\epsilon_U \lesssim (\log n/n)^{1/4}$. But $\mathcal{N}(\mathcal{F}, \epsilon, \mathbb{P}) \geq \frac{1}{\epsilon}$ because $\mathrm{Diam}_{\mathbb{P}}(\mathcal{F}) = \Theta(1)$ and $\mathcal{F}_n$ is convex, so we obtain $\mathcal{N}(\mathcal{F}, \epsilon_U, \mathbb{P}) \gtrsim n^{1/4}/\log n$, contradiction.

To upper-bound $\exp(-An\epsilon_U^4/\max(\epsilon_U^2, \mathcal{I}_U))$, we split into cases. If $\mathcal{I}_U \lesssim \epsilon_U^2$ then $\epsilon_U \asymp \epsilon_*$ and we have $n\epsilon_*^2 \gtrsim \log n$ by Assumption 10, so $\exp(-An\epsilon_U^4/\max\{\epsilon_U^2, \mathcal{I}_U\}) \leq \exp(-An\epsilon_*^2) \leq n^{-1}$ for sufficiently large $A > 0$.

Otherwise, if $\mathcal{I}_U \gg \epsilon_U^2$ we have $n\epsilon_U^4/\mathcal{I}_U \gtrsim \log N$ by the definition of ϵ_U, and since $N \gtrsim n^{1/5}$, we have $\log N \gg \log n$. Hence, by choosing $A > 0$ large enough we can ensure that $\exp(-An\epsilon_U^4/\mathcal{I}_U) \leq n^{-1}$ in this case as well. $\square$

3.4.5 Proof of Theorem 14

Assume for simplicity that $c_L = 1$. We say $\overline{\xi}$ has a Gaussian Isoperimetric Profile (GIP) with respect to $\|\cdot\|_n$, if for any measurable set $A \subset \mathbb{R}^n$ such that $\Pr_{\overline{\xi}}(A) \geq 1/2$, we have that

$$\Pr_{\overline{\xi}}(A_t) \geq 1 - 2\exp(-nt^2/2). \tag{3.43}$$

where $A_t = \{\overline{\xi} \in \mathbb{R}^n : \inf_{x \in A} \|x - \overline{\xi}\|_n \leq t\}$. It is not hard to verify that the GIP and LCP are equivalent (cf. [AAGM15, Thm 3.1.30]).

The main observation is the following simple and useful lemma which leverages the power of isoperimetry:

Lemma 17. *For any measurable* $A \subset \mathbb{R}^n$ *such that* $\Pr_{\overline{\xi}}(A) \geq 2\exp(-nt^2/2)$, $\Pr_{\overline{\xi}}(A_{2t}) \geq 1 - 2\exp(-nt^2/2)$.

Proof. Since $A_{2t} = (A_t)_t$, (3.43) implies that it's sufficient to show that $\Pr(A_t) \geq 1/2$, and indeed it suffices to show that $\Pr(A_{t+\epsilon}) \geq 1/2$ for any $\epsilon > 0$. Fix $\epsilon > 0$, and assume to the contrary that $\Pr(B) > 1/2$, where $B = \mathbb{R}^n \backslash A_{t+\epsilon}$. It's easy to see that $A \subset \mathbb{R}^n \backslash B_{t+\epsilon}$. Hence, using (3.43), we obtain

$$\Pr(\mathbb{R}^n \backslash A) \geq \Pr(B_{t+\epsilon}) \geq 1 - 2\exp(-n(t+\epsilon)^2/2),$$

i.e., $\Pr(A) \leq 2\exp(-n(t+\epsilon)^2/2) < 2\exp(-nt^2/2)$, contradiction. □

Denote the event of Definition 9 by $\mathcal{E}$, and recall the definition of ϵ_*^2 via $n\epsilon_*^2 \asymp \log \mathcal{N}(\epsilon_*, \mathcal{F}, \mathbb{P})$. Letting S be an ϵ_*-net of $\mathcal{F}$ of cardinality $\mathcal{N}(\epsilon_*, \mathcal{F}, \mathbb{P})$, the pigeonhole principle implies the existence of $f_c \in S$ such that

$$\Pr_{\overline{\xi}}(\underbrace{\{\widehat{f}_n \in B(f_c, \epsilon_*)\} \cap \mathcal{E}}_{A} \,|\overline{\mathrm{X}}) \geq \frac{\Pr_{\overline{\xi}}(\mathcal{E})}{\mathcal{N}(\epsilon_*, \mathcal{F}, \mathbb{P})} \geq \frac{\exp(-c_2 n\epsilon_*^2)}{\mathcal{N}(\epsilon_*, \mathcal{F}, \mathbb{P})} \geq \exp(-c_3 n\epsilon_*^2).$$

By isoperimetry, $\Pr_{\overline{\xi}}(A_{2t}) \geq 1 - 2\exp(-nt^2/2)$, where $t = M\epsilon_*/2$ and M is chosen such that $(M/2)^2 \geq 2C_3$; this fixes the value of the absolute constant M used in (3.12).

Applying (3.12) yields that if $\overline{\xi} \in A \subset \mathcal{E}$ and $\|\overline{\xi}' - \overline{\xi}\| \leq M\epsilon_* = 2t$, $\|\widehat{f}_n(\overline{\xi}) - \widehat{f}_n(\overline{\xi}')\|^2 \leq \rho_S(\overline{\mathrm{X}}, f^*)$ and so $\|\widehat{f}_n(\overline{\xi}') - f_c\| \leq \epsilon_* + \sqrt{\rho_S(\overline{\mathrm{X}}, f^*)}$. This implies

$$\Pr_{\overline{\xi}}(\{\widehat{f}_n \in B(f_c, \epsilon_* + \sqrt{\rho_S(\overline{\mathrm{X}}, f^*)})\}|\overline{\mathrm{X}}) \geq \Pr_{\overline{\xi}}(A_{2t}|\overline{\mathrm{X}}) \geq 1 - 2\exp(-nt^2/2).$$

To bound the variance of $\widehat{f}_n$, we use conditional expectation as in Theorem 13. We have

$$V(\widehat{f}_n|\overline{\mathrm{X}}) \leq \mathbb{E}\|\widehat{f}_n - f_c\|^2 \leq (1 - 2\exp(-nt^2/2)) \cdot (\epsilon_* + \sqrt{\rho_S(\overline{\mathrm{X}}, f^*)})^2 + 2C\exp(-nt^2/2),$$

where we have used the fact that $\mathrm{Diam}_{\mathbb{P}}(\mathcal{F}) = \Theta(1)$. Recalling that $\epsilon_*^2 \gtrsim \log(n)/n$, and $t = \Theta(\epsilon_*)$, we have $\exp(-nt^2) = O(\epsilon_*^2)$ and hence the RHS is bounded by $C \max\{\epsilon_*^2, \rho_S(\overline{X}, f^*)\}$, as desired.

3.4.6 Proof of Theorem 15

For simplicity, we assume that $c_X = c_L = c_I = O(1)$, We abbreviate $\rho_{\mathcal{O}} := \rho_{\mathcal{O}}(n, \mathbb{P}, f^*)$.

We shall use the joint metric on $\mathcal{X}^n \times \mathbb{R}^n$ given by

$$\Delta_n((\overline{X}_1, \overline{\xi}_1), (\overline{X}_2, \overline{\xi}_2)) := n^{-1/2} \cdot d_n(\overline{X}_1, \overline{X}_2) + \|\overline{\xi}_1 - \overline{\xi}_2\|_n.$$

As $(\overline{X}, d_n)$ and $(\overline{\xi}, \|\cdot\|_2)$ both satisfy Lipschitz concentration inequalities with parameter $\Theta(1)$, so does the product space $\mathcal{X}^n \times \mathbb{R}^n$ with the usual product metric

$$((\overline{X}_1, \overline{\xi}_1), (\overline{X}_2, \overline{\xi}_2)) \mapsto d_n(\overline{X}_1, \overline{X}_2) + \|\overline{\xi}_1 - \overline{\xi}_2\|_2,$$

and since Δ_n is obtained by scaling this metric by $n^{-1/2}$, we obtain that $(\mathcal{X}^n \times \mathbb{R}^n, \Delta)$ satisfies an LCP condition with parameter $\Theta(n)$.

Let $\mathcal{E}_1$ be the event of Assumption 14, namely, the event that the ERM is almost interpolating, and let $\mathcal{E}_2$ be the event that $\mathrm{Diam}_{\mathbb{P}}(\mathcal{O}_{M'\epsilon_*^2}) \leq \rho_{\mathcal{O}}$. Since $\Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2) > 1 + \exp(-c_I n \epsilon_*^2)$, we have $\Pr(\mathcal{E}_1 \cap \mathcal{E}_2) > \exp(-c_I n \epsilon_*^2)$.

Set $\mathcal{E}_3 = \mathcal{E}_1 \cap \mathcal{E}_2$. Since $\Pr(\mathcal{E}_3) \geq \exp(-c_I n \epsilon_*^2)$, the same pigeonhole principle argument used in the proofs of Theorems 8 and 14 shows that there exists an absolute constant $c_1 \in (0, c_I)$ and $f_c \in \mathcal{F}$ such that

$$\Pr_{\overline{X}, \overline{\xi}}(\mathcal{E}_3 \cap \{\widehat{f}_n \in B(f_c, \epsilon_*)\}) \geq \exp(-c_1 n \epsilon_*^2).$$

Denote this event by $\mathcal{E}$ (in this case, it is better to think about it as a subset of $\mathcal{X}^n \times \mathbb{R}^n$). By the same argument as in Theorem 14, $\widetilde{\mathcal{E}} := \mathcal{E}_{C_1 \epsilon_*}$ will be an event of probability $1 - \exp(-c_1 n \epsilon_*^2)$, where $A_r = \{(\overline{X}, \overline{\xi}) \in \mathcal{X}^n \times \mathbb{R}^n : \Delta_n((\overline{X}, \overline{\xi}), A) \leq r\}$ as above, and C_1 is an absolute constant depending on c_1 and the LCP parameter of Δ.

Thus, we would like to show that any $(\overline{X},\overline{\xi}) \in \widetilde{\mathcal{E}}$ is not too far from f_c. More precisely, we claim that for any $(\overline{X},\overline{\xi})$ at distance at most $C_1\epsilon_*$ from $\mathcal{E}$, the corresponding $\widehat{f}_n$ is at distance at most $C_2 \cdot \max\{\epsilon_*, \sqrt{\rho_{\mathcal{O}}}\}$ from f_c.

For $f \in \mathcal{F}$, let $f_{\overline{X}}$ denote $(f(X_1),\dots,f(X_n)) \in \mathbb{R}^n$. We claim that it suffices to prove the following: for every $(\overline{X}_1,\overline{\xi}_1) \in \mathcal{E}$ and such that $\Delta_n((\overline{X}_2,\overline{\xi}_2),(\overline{X}_1,\overline{\xi}_1)) \leq C_1\epsilon_*$, we have

$$\|\widehat{f}_n(\overline{X}_2,\overline{\xi}_2)_{\overline{X}_1} - \widehat{f}_n(\overline{X}_1,\overline{\xi}_1)_{\overline{X}_1}\|_n \lesssim \epsilon_*, \tag{3.44}$$

where $\widehat{f}_n(\overline{X}_i,\overline{\xi}_i)$ is the ERM for the input points $\overline{X}_i$ and noise $\overline{\xi}_i$,. Indeed, assuming (3.44) we have

$$\|\widehat{f}_n(\overline{X}_2,\overline{\xi}_2)_{\overline{X}_1} - \overline{Y}_1\|_n^2 \leq 2\|\widehat{f}_n(\overline{X}_1,\overline{\xi}_1)_{\overline{X}_1} - \overline{Y}_1\|_n^2 + 2\|\widehat{f}_n(\overline{X}_2,\overline{\xi}_2)_{\overline{X}_1} - \widehat{f}_n(\overline{X}_1,\overline{\xi}_1)_{\overline{X}_1}\|_n^2 \lesssim \epsilon_*^2, \tag{3.45}$$

as the first term on the RHS is bounded by $2\epsilon_*$ because $(\overline{X}_1,\overline{\xi}_1) \in \mathcal{E}$, and the second term is bounded by $4\epsilon_*^2$ by construction. We now specify the constant M' in the definition of $\rho_{\mathcal{O}}$ (Definition 10) to be any upper bound for the implicit absolute constant in (3.45). Under this definition, (3.45) implies that $\widehat{f}_n(\overline{X}_2,\overline{\xi}_2)_{\overline{X}_1} \in \mathcal{O}_{M'\epsilon_*^2}$ and hence $\|\widehat{f}_n(\overline{X}_2,\overline{\xi}_2) - \widehat{f}_n(\overline{X}_1,\overline{\xi}_1)\|^2 \leq \rho_{\mathcal{O}}$. Since $\widehat{f}_n(\overline{X}_1,\overline{\xi}_1) \in \mathcal{E}$, this implies that

$$\|\widehat{f}_n(\overline{X}_2,\overline{\xi}_2) - f_c\|^2 \leq 2(\|\widehat{f}_n(\overline{X}_2,\overline{\xi}_2) - \widehat{f}_n(\overline{X}_1,\overline{\xi}_1)\|^2 + \|\widehat{f}_n(\overline{X}_2,\overline{\xi}_2) - f_c\|^2 \lesssim \max\{\epsilon_*^2, \rho_{\mathcal{O}}\}$$

as desired.

Thus, on the high-probability event $\widetilde{\mathcal{E}}$, $\hat{f}_n \in B(f_c, C\max\{\epsilon_*^2, \rho_O\})$. As in the proof of Theorem 13, one concludes by conditional expectation that $V(\widehat{f}_n) \leq C\max\{\epsilon_*^2, \rho_O\}$.

Proof of (3.44)**:** For convenience, denote $f_{i,j} = \widehat{f}_n(\overline{X}_i,\overline{\xi}_i)_{\overline{X}_j}$, and similarly ${f^*}_j = (f^*)_{\overline{X}_j}$. As $d((\overline{X}_2,\overline{\xi}_2),(\overline{X}_1,\overline{\xi}_1)) \leq 2\epsilon_*$, we have by the Lipschitz property that $\|f_{i,1} - f_{i,2}\|_n \leq 2\epsilon_*$ and also $\|{f^*}_1 - {f^*}_2\|_n \leq 2\epsilon_*$. In addition, letting $\overline{Y}_i = {f^*}_i + \overline{\xi}_i$ be the observation vector, the Lipschitz property of f^* and the bound on $\|\overline{\xi}_1 - \overline{\xi}_2\|_n$ together imply that $\|\overline{Y}_1 - \overline{Y}_2\|_n \leq 4\epsilon_*$. The definition of $f_{i,i}$ as the ERM with data points $\overline{X}_i$ and observations $\overline{Y}_i$ implies that for $i = 1, 2$,

$$\|f_{i,i} - \overline{Y}_i\|_n \leq \|f_{\overline{X}_i} - \overline{Y}_i\|_n \tag{3.46}$$

for any $f \in \mathcal{F}$. Finally, the almost interpolating assumption (Assumption 14) yields $\|\overline{Y}_1 - f_{1,1}\|_n \leq C\epsilon_*$.

We obtain (3.44) by putting these bounds all together. Indeed, we have

$$\begin{aligned}\|f_{1,1} - f_{2,1}\|_n &\leq \|f_{1,1} - \overline{Y}_1\|_n + \|\overline{Y}_1 - \overline{Y}_2\|_n + \|\overline{Y}_2 - f_{2,2}\|_n + \|f_{2,2} - f_{2,1}\|_n \\ &\leq (C+6)\epsilon_* + \|f_{2,2} - \overline{Y}_2\|_n,\end{aligned}$$

and substituting $i = 2$, $f = f_1$ into (3.46) yields

$$\begin{aligned}\|f_{2,2} - \overline{Y}_2\|_n &\leq \|f_{1,2} - \overline{Y}_2\|_n \\ &\leq \|f_{1,2} - f_{1,1}\|_n + \|f_{1,1} - \overline{Y}_1\|_n + \|\overline{Y}_1 + \overline{Y}_2\|_n \\ &\leq (C+6)\epsilon_*,\end{aligned}$$

so we finally obtain

$$\|f_{1,1} - f_{2,1}\|_n \leq 2(C+6)\epsilon_* \lesssim \epsilon_*,$$

as desired.

3.4.7 Proof of Theorem 16

The proof strategy is identical to that of Corollary 2: use a fixed-point theorem to find a function f^* for which $f^* = \mathbb{E}_{\overline{X},\overline{\xi}}\widehat{f}_n$, for which we have $\mathbb{E}\|\widehat{f}_n - f^*\|^2 \leq \sup_{f^* \in \mathcal{F}} V(\widehat{f}_n)$. However, the infinite-dimensional random-design setting makes things a bit trickier.

For given $f^*, \overline{X}, \overline{\xi}$, let $F_{\overline{X},\overline{\xi}}(f^*)$ denote the corresponding ERM (which we have previously denoted $\widehat{f}_n$). Recall that while the ERM is uniquely defined as a vector in $\mathcal{F}_n$, its lift to $\mathcal{F}$ is in general far from unique. We will make two temporary assumptions to streamline the proof, and explain at the end of the proof how to remove them, at the cost of some additional technical complexity. First, we assume that $F_{\overline{X},\overline{\xi}}(f^*)$ is the (unique) element of $\mathcal{F}$ of minimal $L^2(\mathbb{P})$-norm mapping to the finite-dimensional ERM; second, we assume that for each $\overline{X}$, the minimal-norm lifting map, defined by

$$L_{\overline{X}}(v) = \text{argmin}\{\|f\| : f \in \mathcal{F} \mid v = (f(x_1), \dots, f(x_n))\},$$

is continuous.

The map $F_{\overline{X},\overline{\xi}}$ is the composition of the following maps:

$$\begin{tikzcd} \mathcal{F} \arrow[r, "P_n"] & \mathcal{F}_n \arrow[r, "v \mapsto P_{\mathcal{F}_n}(v+\overline{\xi})"] & \mathcal{F}_n \arrow[r, "L_{\overline{X}}"] & \mathcal{F} \end{tikzcd}$$

where $P_n(f) = (f(x_1), \dots, f(x_n))$, $P_{\mathcal{F}_n}$ is the projection from $L^2(\mathbb{P}^{(n)})$ onto the convex set $\mathcal{F}_n$, which is the LSE in fixed design, and $L_{\overline{X}}$ is the lifting map defined above. The linear map P_n is continuous by Assumption 15, and the map $v \mapsto P_{\mathcal{F}_n}(v + \overline{\xi})$ is continuous because projection onto a convex set is continuous. As we have assumed (for now) that $L_{\overline{X}}$ is continuous, this proves that for every $\overline{X}, \overline{\xi}$, $f \mapsto F_{\overline{X},\overline{\xi}}(f)$ is a continuous map of the compact set $\mathcal{F}$ to itself.

We claim that the expectation of this map, $f \mapsto \mathbb{E}_{\overline{X},\overline{\xi}}[F_{\overline{X},\overline{\xi}}(f)]$, is also continuous: indeed, if $f_k \to f$ then

$$\|F_{\overline{X},\overline{\xi}}(f_k) - F_{\overline{X},\overline{\xi}}(f)\| \to 0$$

for each $\overline{X}, \overline{\xi}$ and is bounded by the diameter $D_{\mathbb{P}}(\mathcal{F})$, so Jensen's inequality and dominated convergence imply

$$\|\mathbb{E}[F_{\overline{X},\overline{\xi}}(f_k)] - \mathbb{E}[F_{\overline{X},\overline{\xi}}(f)]\| \le \mathbb{E}[\|F_{\overline{X},\overline{\xi}}(f_k) - F_{\overline{X},\overline{\xi}}(f)\|] \to 0$$

which is continuity.

We can thus apply the Schauder fixed point theorem [AB06, Theorem 17.56]:

Theorem 1. *Let K be a nonempty compact convex subset of a Banach space, and let $f : K \to K$ be a continuous function. Then the set of fixed points of f is compact and nonempty.*

The fixed point we obtain is a function $f^* \in \mathcal{F}$ for which $f^* = \mathbb{E}[F_{\overline{X},\overline{\xi}}(f)] = \mathbb{E}\widehat{f}_n$ and hence,

$$\mathbb{E}\|\widehat{f}_n - f^*\|^2 = \mathbb{E}\|\widehat{f}_n - \mathbb{E}\widehat{f}_n\|^2 \lesssim \mathcal{V}(\widehat{f}_n, \mathcal{F}, \mathbb{P}).$$

This concludes the proof in the case that the lifting maps $L_{\overline{X}}$ are continuous.

Unfortunately, the assumption that the $L_{\overline{X}}$ are continuous turns out to be unjustified in general. Indeed, it is not difficult to construct an example of a convex set $K \subset \mathbb{R}^3$ for which the minimal-norm lift $P_{\mathbb{R}^2}(K) \to K$ is not continuous; in fact, one can construct $K \subset \mathbb{R}^3$ with no continuous section $P_{\mathbb{R}^2}(K) \to K$. So we need to explain how to proceed without this assumption.

Fortunately, each $L_{\overline{X}}$ is always continuous on the *relative interior* of $\mathcal{F}_n$ (we sketch the proof of this at the end of the section), so the following modification of $F_{\overline{X},\overline{\xi}}$ does turn out to be continuous:

$$\mathcal{F} \xrightarrow{P_n} \mathcal{F}_n \xrightarrow{v \mapsto P_{\mathcal{F}_n}(v+\overline{\xi})} \mathcal{F}_n \xrightarrow{\varphi_\delta} \mathcal{F}_n \xrightarrow{L_{\overline{X}}} \mathcal{F}, \tag{3.47}$$

where

$$\varphi_\delta(v) = (1-\delta)(v - v_0) + v_0$$

is simply a contraction of $\mathcal{F}_n$ into a $(1-\delta)$-scale copy of itself (v_0 is some arbitrarily chosen point in the interior of $\mathcal{F}_n$).

Let $\widetilde{F}_{\overline{X},\overline{\xi}}$ denote the composition of the maps in (3.47). By the argument above, $\widetilde{F}_{\overline{X},\overline{\xi}}$ is continuous and $\mathbb{E}[\widetilde{F}_{\overline{X},\overline{\xi}}]$ has a fixed point f^*.

Of course, f^* is not a fixed point of $\mathbb{E}[F_{\overline{X},\overline{\xi}}]$ as we would like. However, note that $\|\varphi_\delta(v) - v\|_n \leq 2\delta$ for any $v \in \mathcal{F}_n$ (as the diameter of $\mathcal{F}_n$ is at most 2). Hence, we have for any $v \in \mathcal{F}_n$ that

$$\|L_{\overline{X}}(\varphi_\delta(v)) - L_{\overline{X}}(v)\|^2 \leq 2\|\varphi_\delta(v) - v\|_n^2 + C\mathcal{I}_L(n, \mathbb{P}) \leq 8\delta^2 + \mathcal{I}_L(n, \mathbb{P})$$

on an event $\mathcal{E}$ of high probability; in particular this holds for $v = P_{\mathcal{F}_n}(P_n(f^*) + \overline{\xi})$, which means that on $\mathcal{E}$,

$$\|\widetilde{F}_{\overline{X},\overline{\xi}}(f^*) - F_{\overline{X},\overline{\xi}}(f^*)\| \leq 8\delta^2 + C\mathcal{I}_L(n, \mathbb{P}).$$

Choosing $\delta \lesssim \mathcal{I}_L(n, \mathbb{P})$ and applying conditional expectation (using the fact that $\mathcal{E}^c$ is negligible) and Jensen's inequality, we get that the f^* thus obtained satisfies

$$\|f^* - \mathbb{E}\hat{f}_n\| \lesssim \max\{V(\widehat{f}_n, \mathcal{F}, \mathbb{P}), \mathcal{I}_L(n, \mathbb{P})\}$$

, which shows that the ERM is admissible for this f^*.

By the same argument, we may discard the assumption that the ERM is computed by finding the element of $\mathcal{F}$ of minimal norm mapping to the finite-dimensional ERM $\hat{f}_n^{(fd)}$: indeed, under the event $\mathcal{E}$, the set of functions in $\mathcal{F}$ mapping to $\hat{f}_n^{(fd)}$ has diameter $C \cdot \mathcal{I}_L(n, \mathbb{P})$, so changing the selection rule for the ERM will shift its expectation by a perturbation of norm at most $C \cdot \mathcal{I}_L(n, \mathbb{P})$.

It remains to explain why the lifting map $L = L_{\overline{X}} : \mathcal{F}_n \to \mathcal{F}$ is continuous on the relative interior of $\mathcal{F}_n$. Replacing the ambient space with the affine hull of $\mathcal{F}_n$, we may assume $\mathcal{F}_n$ has nonempty interior.

Suppose $v_k \to v$ in $\mathcal{F}_n$ and $v \in \operatorname{int} \mathcal{F}_n$; we wish to show that $L(v_k) \to f = L(v)$. As $\mathcal{F}_n$ is compact, by passing to a subsequence we may assume $L(v_k)$ converges to some $g \in \mathcal{F}$. Since P_n is continuous, we have $v = P_n(L(v_k)) \to P_n(g)$, i.e., g is a lift of v. Hence, by definition, $\|g\| \geq \|f\|$, and we wish to show that equality holds.

Suppose not. Then $\|g\| > \|f\|$ and hence $\|L(v_k)\| \geq \|f\| + \epsilon$ for all k and some $\epsilon > 0$; that is, there exist v_k arbitrarily close to v whose minimal-norm lift has much larger norm than that of v. It suffices to show this is impossible (i.e., that $u \mapsto \|L(u)\|$ is upper semicontinuous at v). This follows from the fact that $u \mapsto \|L(u)\|$ is convex, as is easily verified, and a convex function is continuous on the interior of its domain [Sch14, Theorem 1.5.3]; for completeness, we give a direct proof.

Since $v \in \operatorname{int} \mathcal{F}_n$, there exists $r > 0$ such that $B(v, r) \subset \mathcal{F}_n$. This implies that for any $\delta > 0$, one has

$$B(v, \delta) \subset v + \frac{\delta}{r}(\mathcal{F}_n - v).$$

Let D be the diameter of $\mathcal{F}$ in $L^2(\mathbb{P})$. We have $\mathcal{F} \subset B(f, D)$ and hence $\mathcal{F}_n \subset P_n(B(f, D) \cap \mathcal{F})$. By linearity, this implies that

$$v + a(\mathcal{F}_n - v) \subset P_n(B(f, aD) \cap \mathcal{F})$$

for any $a > 0$; choosing $a = \frac{\delta}{r}$ we obtain

$$B(v,\delta) \cap \mathcal{F}_n \subset P_n\left(B\left(f, \frac{D\delta}{r}\right) \cap \mathcal{F}\right).$$

In other words, if $\|u - v\|_n < \delta$ and $u \in \mathcal{F}_n$, there exists an element of $\mathcal{F}$ in $B(f, \frac{D\delta}{r})$ mapping to u, which in particular implies that $\|L(u)\| \leq \|f\| + \frac{D\delta}{r}$. This means that $u \mapsto \|L(u)\|$ is upper semicontinuous at v, which was precisely what we needed in order to conclude that L is continuous at v.

3.4.8 Proof of Theorem 12

For the purposes of this proof, the assumption that $(\mathcal{F}, \mathbb{P}^{(n)})$ is a non-Donsker class will be formulated in the following way:

Assumption 1. *For any $n \geq 1$, $\mathbb{E}\sup_{f \in \mathcal{F}} G_f = \omega(\epsilon_*^2)$ and*

$$\frac{1}{\sqrt{n}} \int_{\epsilon_*(n)}^{1} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P}^{(n)})}\, d\epsilon = O(\epsilon_*^2).$$

In order to avoid confusion, in this model $\mathcal{F}$ is a fixed function class, and $\mathbb{P}^{(n)}$ is a sequence of fixed measures. This model is inspired from non-parametric models that appear in the shape-constraints literature such as convex regression when $d \geq 5$ or α-Hölder regression in the suitable dimension For example, in the case of convex Lipschitz regression [SS11] when $d \geq 5$ and $\mathbb{P}^{(n)}$ are n-well separated grid points on $\mathcal{X}$, Assumption 1 is valid.

Preliminaries The following is known as the maximal inequality (cf. [Ver18]):

Lemma 18. *Let $Z_1, \ldots, Z_k$ be zero mean σ-sub-Gaussian random variables with bounded variance. Then, we have that*

$$\mathbb{E} \max_{1 \leq i \leq k} Z_i \lesssim \sigma\sqrt{\log k}.$$

Next, we state the classical Dudley's lemma (cf. [Ver18]):

Lemma 19. *The following holds for all $\epsilon \in (0,1)$:*

$$\mathbb{E} \sup_{f_i \in \mathcal{N}_\epsilon} G_{f_i - f^*} \leq \frac{C_4}{\sqrt{n}} \int_{\epsilon}^{\mathrm{D}_{\mathbb{P}^{(n)}}(\mathcal{F})} \sqrt{\log \mathcal{N}(u, \mathcal{F}, \mathbb{P}^{(n)})} du, \tag{3.48}$$

where $\mathcal{N}_\epsilon$ denotes the minimal ϵ-net in $\mathcal{F}$.

Proof of Theorem 12. We start by finding a weakly admissible f^* by a more constructive method than that used in the proof of Corollary 3 (the method here is closer to the original proof of [Cha14]). Then, we prove (3.9) for this choice of f^*.

Let $S = \{f_1, \ldots, f_{\mathcal{N}}\}$ be a minimal $\epsilon_* := \epsilon_*(n)$-net of $\mathcal{F}$, and denote $B(f_i) := B_n(f_i, \epsilon_*)$, $i = 1, \ldots, \mathcal{N}$.

Our admissible $f^* \in \mathcal{F}$ is defined as

$$f^* := \operatorname{argmax}_{f_i \in S} \mathbb{E} \sup_{f \in B(f_i)} G_{f - f_i}. \tag{3.49}$$

Lemma 20. *The following event holds with probability of at least $1 - 2\exp(-cn\epsilon_*^2)$:*

$$\forall i \in [\mathcal{N}] \quad \left| \sup_{f \in B(f_i)} G_{f - f_i} - \mathbb{E} \max_{f \in B(f_i)} G_{f - f_i} \right| \leq C_1 \epsilon_*^2. \tag{3.50}$$

In addition, for a fixed $j \in 1, \ldots, \lceil \epsilon_*^{-1} \rceil$, the event

$$\sup_{f_i \in (S \cap B(f^*, j\epsilon_*))} G_{f_i - f^*} \leq C_2 j \epsilon_*^2. \tag{3.51}$$

holds with probability of at least $1 - 2\exp(-c_3 n \epsilon_*^2)$.

The proof of this lemma appears below. We denote the event of (3.50) by $\mathcal{E}$, and by $\mathcal{E}(j)$ the event of (3.51).

Following [Cha14], we define

$$\Psi_{\bar{\xi}}(t) = \sup_{f \in B_n(f^*, t)} 2G_{f - f^*} - t^2.$$

One easily verifies (see [Cha14, Proof of Theorem 1.1]) that $\Psi_{\overline{\xi}}(t)$ is strictly concave and that $\text{argmax}_t \Psi_{\overline{\xi}}(t) = \|f^* - \widehat{f}_n\|_n^2$.

This implies that if, for any particular $\overline{\xi}$, we identify t_1, t_2 such that $\Psi_{\overline{\xi}}(t_1) > \Psi_{\overline{\xi}}(t_2)$, the unique maximum of $\Psi_{\overline{\xi}}$ occurs for some t smaller than t_2, i.e., $\|f^* - \widehat{f}_n\| \leq t_2$. We will take $t_1 = \epsilon_*$ and $t_2 = D\epsilon_*$ for a sufficiently large constant D and show that $\Psi_{\overline{\xi}}(t_1) > \Psi_{\overline{\xi}}(t_2)$ on $\mathcal{E} \cap \mathcal{E}(D)$. This implies that $\|f^* - \widehat{f}_n\| \leq D\epsilon_*$ on $\mathcal{E} \cap \mathcal{E}(D)$, which precisely means that $\widehat{f}_n$ is admissible for f^*.

On the one hand, conditioned on $\mathcal{E}$ we have

$$\begin{aligned}\Psi_{\overline{\xi}}(\epsilon_*) &= \sup_{f \in B(f^*)} 2G_{f-f^*} - \epsilon_*^2 \\ &\geq \max_{f_i \in \mathcal{N}} \mathbb{E} \sup_{f \in B(f_i)} 2G_{f-f^*} - C_1\epsilon_*^2,\end{aligned} \tag{3.52}$$

where we used the definition of f^* and Eq. (3.50) above. On the other hand, for $D \geq 2$ we have under $\mathcal{E}(D) \cap \mathcal{E}$ that

$$\begin{aligned}\sup_{f \in B_n(f^*, D\epsilon_*)} G_{f-f^*} &\leq \max_{f_i \in \mathcal{N} \cap B(f^*, D\epsilon_*)} \sup_{f \in B(f_i)} G_{f-f_i} + \sup_{f_i \in \mathcal{N} \cap B(f^*, D\epsilon_*)} G_{f_i - f^*} \\ &\leq \sup_{f \in B(f^*, \epsilon_*)} G_{f-f^*} + C_2 D\epsilon_*^2,\end{aligned}$$

where we used the definition of f^* and (3.51). Substituting in the definition of Ψ, we obtain

$$\begin{aligned}\Psi_{\overline{\xi}}(D\epsilon_*) &= \sup_{f \in B_n(f^*, D\epsilon_*)} 2G_{f-f^*} - D^2\epsilon_*^2 \\ &\leq 2\left(\sup_{f \in B(f^*)} G_{f-f^*} + C_2 D\epsilon_*^2\right) - D^2\epsilon_*^2 \\ &\leq \Psi_{\overline{\xi}}(\epsilon_*) + (2C_2 + 1)D\epsilon_*^2 - D^2\epsilon_*^2.\end{aligned}$$

Comparing with (3.52) we see that for $D \geq 2C_2 + C_1 + 1$ (say) we have $\Psi_{\overline{\xi}}(D\epsilon_*) < \Psi_{\overline{\xi}}(\epsilon_*)$ on $\mathcal{E} \cap \mathcal{E}(D)$, which is what we wished to prove.

Now, we are ready to prove (3.9). We first claim that with probability $1 - 2\exp(-cn\epsilon_*^2)$,

$$\sup_{f \in B(f^*) - f^*} G_f = \omega(\epsilon_*^2), \tag{3.53}$$

where f^* is our admissible function. To prove this, we first apply the basic inequality [Gee00], which implies that

$$\mathcal{R}_n(\widehat{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) \leq \mathbb{E} \sup_{f \in \mathcal{F}} 2G_{f-f^*},$$

where we recall that the LHS denotes the risk of $\widehat{f}_n$, i.e. the maximal $L_2(\mathbb{P}^{(n)})$ error that $\widehat{f}_n$ can attain over $\mathcal{F}$. By assumption, $\mathcal{R}_n(\widehat{f}_n, \mathcal{F}, \mathbb{P}^{(n)}) = \omega(\epsilon_*^2)$.

Next, we have by Lemma 19 and (3.49)

$$\begin{aligned}
\omega(\epsilon_*^2) = \mathbb{E} \sup_{f \in \mathcal{F}} G_{f-f^*} &\leq \max_{f_i \in \mathcal{N} \cap B_n(f^*, \mathrm{E}_{f^*})} \mathbb{E} \sup_{f \in B(f_i)} G_{f-f_i} + \mathbb{E} \max_{f_i \in \mathcal{N}} G_{f_i - f^*} \\
&\leq \mathbb{E} \sup_{f \in B(f^*)} G_{f-f^*} + \frac{C_1}{\sqrt{n}} \int_{\epsilon_*}^{D_{\mathbb{P}^{(n)}}} \sqrt{\log \mathcal{N}(t, \mathcal{F}, \mathbb{P}^{(n)})} dt \\
&\leq \mathbb{E} \sup_{f \in B(f^*)} G_{f-f^*} + O(\epsilon_*^2),
\end{aligned}$$

where we used Assumption 1. Hence,

$$\mathbb{E} \sup_{f \in B(f^*)} G_{f-f^*} = \omega(\epsilon_*^2) - O(\epsilon_*^2) = \omega(\epsilon_*^2). \tag{3.54}$$

This gives us a lower bound for $\sup_{f \in B(f^*)} G_{f-f^*}$ in expectation, but we require a high-probability bound. The proof of this is the same as the proof of Lemma 20, so we only sketch it: $\sup_{f \in B(f^*)} G_{f-f^*}$ is convex and $\epsilon_* n^{-1/2}$-Lipschitz, which means that it deviates from its expectation by ϵ_*^2 with probability at most $2\exp(-cn\epsilon_*^2)$. Combining this with (3.54) proves (3.53).

Let $\mathcal{V}$ denote the set of noise vectors for which $\|\hat{f}_n - f^*\|_n \leq C\epsilon_*$ and $\sup_{f \in B(f^*)} G_{f-f^*} =$

$\omega(\epsilon_*^2)$; by what we have already proven, we have

$$\Pr(\overline{\xi} \in \mathcal{V}) \geq 1 - C\exp(-cn\epsilon_*^2).$$

Let $\mathcal{V}' = \mathcal{V} \cap (-\mathcal{V}) = \{\overline{\xi} : \overline{\xi}, -\overline{\xi} \in \mathcal{V}\}$. By the union bound,

$$\Pr(\overline{\xi} \in \mathcal{V}') \geq 1 - 2C\exp(-cn\epsilon_*^2).$$

Fix any $\overline{\xi} \in \mathcal{V}'$, and denote by $\hat{f}_n^-$ the ERM with the flipped noise vector $-\overline{\xi}$:

$$\widehat{f}_n^- := \mathrm{argmin}_{f \in \mathcal{F}} \left(\sum_i (-\xi_i + f^*(x_i) - f(x_i))^2 \right).$$

Since $\overline{\xi}, -\overline{\xi} \in \mathcal{V}' \subset \mathcal{V}$, we have

$$\|\hat{f}_n - \hat{f}_n^-\|_n \leq \|\hat{f}_n - f^*\|_n + \|\hat{f}_n^- - f^*\|_n \leq C\epsilon_*.$$

In other words, $\hat{f}_n^- \in B_n(\hat{f}_n, C\epsilon_*)$, so to prove (3.9), it thus suffices to show that $\hat{f}_n^-$ is not a δ-approximate minimizer for $\delta = \omega(\epsilon_*^2)$ with respect to the noise $\overline{\xi}$, i.e.,

$$\frac{1}{n}\sum_{i=1}^n (f^*(x_i) + \xi_i - \hat{f}_n^-(x_i))^2 \geq \frac{1}{n}\sum_{i=1}^n (f^*(x_i) + \xi_i - \widehat{f}_n(x_i))^2 + \omega(\epsilon_*^2).$$

Equivalently, (by subtracting $\|\overline{\xi}\|_n^2$ from both sides as in (3.31)), we wish to prove that

$$-2G_{\hat{f}_n^- - f^*} + \|\hat{f}_n^- - f^*\|_n^2 \geq -2G_{\widehat{f}_n - f^*} + \|\widehat{f}_n - f^*\|_n^2 + \omega(\epsilon_*^2).$$

Since $\|\hat{f}_n^- - f^*\|_n^2, \|\hat{f}_n^- - f^*\|_n^2 = O(\epsilon_*^2)$ as $\overline{\xi}, -\overline{\xi} \in \mathcal{V}$, this reduces to showing that

$$-2G_{\hat{f}_n^- - f^*} \geq -2G_{\widehat{f}_n - f^*} + \omega(\epsilon_*^2). \tag{3.55}$$

On the one hand, by using Eqs. (3.52) and (3.53) above it is easy to see that on $\mathcal{V}' \subset \mathcal{V}$, we have $G_{\widehat{f}_n - f^*} \geq \omega(\epsilon_*^2)$. On the other hand,

$$G_{\hat{f}_n^- - f^*} = \frac{1}{n}\langle \widehat{f}_n - f^*, \overline{\xi}\rangle = -\frac{1}{n}\langle \hat{f}_n^- - f^*, -\xi\rangle. \tag{3.56}$$

But note that $\frac{1}{n}\langle \hat{f}_n^- - f^*, -\overline{\xi}\rangle$ is the process for the noise vector $-\overline{\xi}$ evaluated at the corresponding ERM, namely $\hat{f}_n^-$, and we have $-\overline{\xi} \in \mathcal{V}'$, which implies that

$$\frac{1}{n}\langle \hat{f}_n^- - f^*, -\overline{\xi}\rangle \geq \omega(\epsilon_*^2).$$

Combining these last two inequalities we see that (3.55) indeed holds over all $\mathcal{V}'$, which implies that (3.9) holds on $\mathcal{V}'$, as desired. □

It remains to prove Lemma 20.

Proof of Lemma 20. First, define

$$F_i(\overline{\xi}) := 2n^{-1} \sup_{f \in B(f_i)} \sum_{k=1}^{n} (f - f_i)(x_k) \cdot \xi_k.$$

Since $\|f - f_i\|_n \leq \epsilon_*$ for all $f \in B(f_i)$, we see that $F_i(\overline{\xi})$ is a $2\epsilon_* n^{-1/2}$-Lipschitz function (with respect to the usual Euclidean norm on $\mathbb{R}^n$). Hence, we apply (3.10) and obtain

$$\Pr_{\overline{\xi}}\left\{\left|\sup_{f \in B(f_i)} G_{f-f_i} - \mathbb{E}\sup_{f \in B_n(f_i)} G_{f-f_i}\right| \geq t\right\} \leq 2\exp(-cnt^2/\epsilon_*^2). \tag{3.57}$$

Therefore, by taking a union bound over $1 \leq i \leq |\mathcal{N}|$

$$\Pr_{\overline{\xi}}\left\{\forall i \in [|\mathcal{N}|] : \left|\sup_{f \in B_n(f_i)} G_{f-f_i} - \mathbb{E}\sup_{f \in B_n(f_i)} G_{f-f_i}\right| \geq t\right\} \leq 2\exp(-cnt^2/\epsilon_*^2 + \log|\mathcal{N}|).$$

Now, recall that $\mathcal{N} := \mathcal{N}(\epsilon_*, \mathcal{F}, \mathbb{P}^{(n)})$ and that by the definition of the minimax rate, $\log|\mathcal{N}|/n \sim \epsilon_*^2$. This allows us to choose $t = C\epsilon_*^2$ (for large enough $C > 0$) such that with probability of at least

$1 - 2\exp(-cn\epsilon_*^2)$, the following holds:

$$\forall i \in [|\mathcal{N}|] : \left| \sup_{f \in B_n(f_i)} G_{f-f_i} - \mathbb{E} \sup_{f \in B_n(f_i)} G_{f-f_i} \right| \leq C\epsilon_* \sqrt{\log \mathcal{N}/n} \leq C_1 \epsilon_*^2, \tag{3.58}$$

which proves (3.50).

For the second part of the lemma, fix $j \geq 1$, and define

$$F_j(\overline{\xi}) = \sup_{f \in \mathcal{N} \cap B_n(f^*, j\epsilon_*)} 2G_{f_j - f^*}.$$

Again, it is easy to verify that $F_j(\cdot)$ is convex and $2j\epsilon_* n^{-1/2}$-Lipschitz. Using (3.10) once again, we obtain that

$$\Pr\left\{ \left| \sup_{f \in \mathcal{N} \cap B_n(f^*, j\epsilon_*)} G_{f-f^*} - \mathbb{E} \sup_{f \in \mathcal{N} \cap B_n(f^*, j\epsilon_*)} G_{f-f^*} \right| \geq t \right\} \leq 2\exp(-cn(t/j)^2/\epsilon_*^2).$$

Choosing $t \sim j\epsilon_*/\sqrt{n} \lesssim j\epsilon_*^2$, we obtain

$$\Pr\left\{ \left| \sup_{f \in \mathcal{N} \cap B_n(f^*, j\epsilon_*)} G_{f-f^*} - \mathbb{E} \sup_{f \in \mathcal{N} \cap B_n(f^*, j\epsilon_*)} G_{f-f^*} \right| \geq j\epsilon_*^2 \right\} \leq 2\exp(-c\epsilon_*^2). \tag{3.59}$$

Next, by applying the maximal inequality (Lemma 18) over $|\mathcal{N}|$ random variables, and the definition of the minimax rate,

$$\mathbb{E} \sup_{f \in \mathcal{N} \cap B_n(f^*, j\epsilon_*)} G_{f-f^*} \leq Cj\epsilon_* \sqrt{\log \mathcal{N}/n} \leq C_1 j\epsilon_*^2. \tag{3.60}$$

Combining (3.59) and (3.60) yields (3.51), concluding the proof. □

3.5 Loose Ends

Lemma 21. *Under Assumption 1 the following holds:* $\rho_S(\overline{X}) \lesssim \max\{\mathcal{I}_L(\overline{X}), \epsilon_*^2\}$ *for any realization* $\overline{X}$.

Proof. It is always the case that $\rho_S(\overline{X}) \leq \max\{\mathcal{I}_L(\overline{X}), \epsilon_*^2\}$ for any $\overline{X}$. Indeed, if (3.12) holds, then for any estimator $\bar{f}_n$ such that $\overline{\xi} \mapsto \bar{f}_n(\overline{\xi})$ is 1-Lipschitz, $\|\overline{\xi} - \overline{\xi}'\|_n \lesssim \epsilon_*$ implies that

$$\|\bar{f}_n(\overline{\xi}') - \bar{f}_n(\overline{\xi})\|^2 \leq 2(\|\bar{f}_n(\overline{\xi}') - \bar{f}_n(\overline{\xi})\|_n^2 + \mathcal{I}_L(\overline{X})) \leq 2(\|\overline{\xi}' - \overline{\xi}\|_n^2 + \mathcal{I}_L(\overline{X})) \lesssim \max\{\epsilon_*^2, \mathcal{I}_L(\overline{X})\}$$

deterministically, not just with non-negligible probability. □

Lemma 22. *Under Assumptions 1, 13-14, we have that* $\rho_{\overline{S}}(\overline{X}, f^*) \leq \rho_{\mathcal{O}}(n, \mathbb{P}, f) \lesssim \max\{\mathcal{I}_L(n, \mathbb{P}), \epsilon_*^2\}$.

Proof. Let $\mathcal{E} \subset \mathcal{X}^n \times \mathbb{R}^n$ be the event that $\mathrm{Diam}_{\mathbb{P}}(\mathcal{O}_{M'\epsilon_*^2}) \leq \rho_{\mathcal{O}}(n, \mathbb{P}, f^*)$ and $\|\hat{f}_n(\overline{X}, \overline{\xi}) - \overline{Y}\|_n^2 \leq C_I \epsilon_*^2$. For any $\overline{X} \in \mathcal{X}^n$, let $\mathcal{E}_{\overline{X}} = \{\overline{\xi} \in \mathbb{R}^n \mid (\overline{X}, \overline{\xi}) \in \mathcal{E}\}$.

Since $\Pr(\mathcal{E}) \geq \exp(-c_I n \epsilon_*^2)$ by definition, the set

$$\mathcal{E}' = \{\overline{X} \in \mathcal{X}^n \mid \Pr_{\overline{\xi}}(\mathcal{E}_{\overline{X}}) \leq \rho_{\mathcal{O}}(n, \mathbb{P}, f^*)) \geq \exp(-c_I n \epsilon_*^2) \in \mathcal{E}\}$$

satisfies $\Pr_{\mathcal{X}^n}(\mathcal{E}') \geq \exp(-(c_I/2) n \epsilon_*^2)$ by Fubini's theorem.

Fix $\overline{X} \in \mathcal{E}'$, $\overline{\xi} \in \mathcal{E}_{\overline{X}}$, and let $\overline{Y} = f^* + \overline{\xi}|_{\overline{X}}$ as usual. If $\|\overline{\xi}' - \overline{\xi}\| \leq \frac{\sqrt{M' - C_I}}{2} \epsilon_*$ then

$$\begin{aligned} \|\hat{f}_n(\overline{X}, \overline{\xi}') - \overline{Y}\|_n^2 &\leq 2(\|\hat{f}_n(\overline{X}, \overline{\xi}') - \hat{f}_n(\overline{X}, \overline{\xi})\|_n^2 + \|\hat{f}_n(\overline{X}, \xi) - \overline{Y}\|_n^2) \\ &\leq 2((M' - C_I)\epsilon_*^2 + C_I)\epsilon_*^2 = M'\epsilon_*^2 \end{aligned}$$

where we have used $\|\hat{f}_n(\overline{X}, \xi') - \hat{f}_n(\overline{\xi})\|_n \leq \|\overline{\xi} - \overline{\xi}'\|_n$ as $\hat{f}_n$ is 1-Lipschitz in the noise. In particular $\hat{f}_n(\overline{X}, \overline{\xi}') \in \mathcal{O}_{M'\epsilon_*^2}$ and so $\|\hat{f}_n(\overline{X}, \overline{\xi}') - \hat{f}_n(\overline{X}, \overline{\xi})\| \leq \rho_{\mathcal{O}}(n, \mathbb{P}, f^*)$, as $(\overline{X}, \overline{\xi}) \in \mathcal{E}$. Thus, if M' is chosen large enough so that $M \leq \frac{\sqrt{M' - C_I}}{2} \epsilon_*$, one obtains (3.12) is satisfied with $\delta(n) = \rho_{\mathcal{O}}(n, \mathbb{P}, f)$ and $c_2 = c_I/2$, implying that $\rho_{\overline{S}}(\overline{X}, f^*) \leq \rho_{\mathcal{O}}(n, \mathbb{P}, f)$.

To see that $\rho_{\mathcal{O}}(n,\mathbb{P},f^*) \lesssim \max\{\mathcal{I}_L(n,\mathbb{P}),\epsilon_*^2\}$ is even easier: it's easy to see that $\mathrm{Diam}_{\mathbb{P}_n}(\mathcal{O}_{M'\epsilon_*^2}) \lesssim \epsilon_*$ (see the end of the proof of Theorem 8 for details), and the definition of the lower isometry remainder implies that $\mathrm{Diam}_{\mathbb{P}}(\mathcal{O}_{M'\epsilon_*^2}) \lesssim \mathrm{Diam}_{\mathbb{P}}(\mathcal{O}_{M'\epsilon_*^2}) + \sqrt{\mathcal{I}_L(n,\mathbb{P})}$ on the high-probability event $\mathcal{I}_L(\overline{X}) \leq \mathcal{I}_L(n,\mathbb{P})$. This yields that for an appropriate choice of $C > 0$, $\delta(n) = C\max\{\mathcal{I}_L(n,\mathbb{P}),\epsilon_*^2\}$ satisfies (3.14) and hence $\rho_{\mathcal{O}}(n,\mathbb{P},f^*) \leq \max\{\mathcal{I}_L(n,\mathbb{P}),\epsilon_*^2\}$. □

3.5.1 Full Proof of Corollary 2

Continuing from §3.4.3 above, tote that when the convex set

$$\mathcal{F}_n := \{(f(X_1),\ldots,f(X_n)) : f \in \mathcal{F}\}$$

is not compact, we cannot apply any fixed point theorem. However, if we find $f^* \in \mathcal{F}$ such that

$$B^2(\widehat{f}_n) \leq \max\{3 \cdot V(\widehat{f}_n), C/n\}, \tag{3.61}$$

where $C > 0$, then we may repeat the same steps as in §3.4.3 above. Since, the fixed point theorem was only used to find a $f^* \in \mathcal{F}$ that satisfies the last equation.

First let us provide some intuition to our proof. The idea to find $f^* \in \mathcal{F}$ with low bias is inspired from the Banach fixed point theorem. If $\widehat{f}_n$ has a "high" bias on some underlying $f_0 \in \mathcal{F}$, then, $\widehat{f}_n$ should have a lower or equal bias when the underlying function is $f_1 = \mathbb{E}_0\widehat{f}_n$, where $\mathbb{E}_0$ means taking expectation when $f^* = f_0$. Now, if f_1 has a "low" bias, then we are done. Otherwise, consider the underlying $f_2 = \mathbb{E}_1\widehat{f}_n$, and repeat this process for n times. We will show that some $m \leq n$, $f^* = f_m$ will be our "admissible" function, i.e. a function that has a "low bias".

This idea is captured in the following lemma (that we will prove below):

Lemma 23. *Let $f_0 \in \mathcal{F}$ and for any $i \geq 1$ denote by $f_i = \mathbb{E}_{i-1}\widehat{f}_n$. Then, there exists $m = O(n)$, such that*

$$B_m^2(\widehat{f}_n) \leq 3 \cdot V_m(\widehat{f}_n) + C/n, \tag{3.62}$$

where $B_m^2(\widehat{f}_n)$ and $V_m(\widehat{f}_n)$ is the bias and variance error of $\widehat{f}_n$ when $f^* = f_m$.

Proof. Assume that $f^* = f_0$. If $B_0^2(\widehat{f}_n) \leq 3V_0(\widehat{f}_n) + C/n$, then we are done. Otherwise, when $B_0^2(\widehat{f}_n) \geq 3V_0(\widehat{f}_n) + C/n$, we proceed with the following strategy: First, recall that when $f = f^*$, then

$$\widehat{f}_n = \mathrm{argmax}_{g \in \mathcal{F}} 2\langle \overline{\xi}, g - f \rangle_n - \|g - f\|_n^2. \tag{3.63}$$

Define the score function of $\widehat{f}_n$ (when $f^* = f$) via

$$L_{\overline{\xi}}(f) = 2\langle \overline{\xi}, \widehat{f}_n - f \rangle_n - \|\widehat{f}_n - f\|_n^2.$$

and let

$$L(f) = \mathrm{Med} L_{\overline{\xi}}(f).$$

It is not hard to verify that when $\overline{\xi}$, then $0 \leq L(f) \leq 2$.

Next, we will show that

$$L(f_1) - L(f_0) \geq \frac{B_0^2(\widehat{f}_n)}{3}. \tag{3.64}$$

If we prove this equation, the lemma will follow. To see this, if f_1 satisfies (3.62), then we stop. Otherwise, repeat the same argument with f_1 and f_2. Then, we would obtain that

$$L(f_2) \geq L(f_1) + B_1^2(\widehat{f}_n)/3 \geq L(f_1) + Cn^{-1} \geq L(f_0) + 2Cn^{-1},$$

where we used the B.

Therefore, if we repeat this process for n/C times. We must have that f_m, for some $m \leq n/C$, will satisfy (3.62); since, $L(f) \leq 2$ for all $f \in \mathcal{F}$. It remains to prove (3.64).

Proof of (3.64) Let $\widetilde{V}_0(\widehat{f}_n) := \max\{3V_0(\widehat{f}_n), C/n\}$ and let $\widetilde{f}_n$ be the restricted LSE on

$$\mathcal{G} := B_n(f_1, \sqrt{\widetilde{V}_0(\widehat{f}_n)}) = \{f \in \mathcal{F} : \|f - f_1\|_n^2 \leq \widetilde{V}_0(\widehat{f}_n)\},$$

namely, $\widetilde{f}_n := \text{argmin}_{f \in \mathcal{G}} \|\overline{Y} - f\|_n^2$. Define

$$\widetilde{L}_{\overline{\xi}}(f) = 2\langle \overline{\xi}, \widetilde{f}_n - f \rangle_n - \|\widetilde{f}_n - f\|_n^2.$$

Using (3.63), we have that

$$L_{\overline{\xi}}(f_1) \geq \widetilde{L}_{\overline{\xi}}(f_1) = \sup_{f \in \mathcal{G}} 2\langle \overline{\xi}, f - f_1 \rangle_n - \|f - f_1\|_n^2 \geq \sup_{f \in \mathcal{G}} 2\langle \overline{\xi}, f - f_1 \rangle_n - \widetilde{V}_0^2(\widehat{f}_n),$$

where the first inequality follows from $\mathcal{G} \subseteq \mathcal{F}$, and the second equality follows from that $\widetilde{f}_n = \text{argmax}_{f \in \mathcal{G}} \widetilde{L}_{\overline{\xi}}(f_1)$, and finally that $\text{Diam}(\mathcal{G})^2 \leq \widetilde{V}_0(\widehat{f}_n)$.

Now, by Chebyshev's inequality, with probability of at least $2/3$, we have that $\widehat{f}_n \in \mathcal{G}$ (when $f^* = f_0$) denote this event by $\mathcal{E}$. Under $\mathcal{E}$, we clearly have that

$$\sup_{f \in \mathcal{G}} \langle \overline{\xi}, f \rangle_n \geq \langle \overline{\xi}, \widehat{f}_n \rangle_n.$$

Hence, using the last two equations, the following holds on the event $\mathcal{E}$:

$$L_{\overline{\xi}}(f_1) - L_{\overline{\xi}}(f_0) \geq \widetilde{L}_{\overline{\xi}}(f_1) - L_{\overline{\xi}}(f_0) \geq \|\widehat{f}_n - f_0\|_n^2 - \langle \overline{\xi}, f_1 - f_0 \rangle_n - \widetilde{V}_0(\widehat{f}_n).$$

Now, note that $\langle \overline{\xi}, f_1 - f_0 \rangle_n \sim N(0, \|f_1 - f_0\|_n^2) = N(0, B_0^2(\widehat{f}_n))$. Hence, there exists an event $\mathcal{E}_1 \subset \mathcal{E}$ that holds with probability of at least 0.5 such that

$$L_{\overline{\xi}}(f_1) - L_{\overline{\xi}}(f_0) \geq \|\widehat{f}_n - f_0\|_n^2 - \widetilde{V}_0(\widehat{f}_n) - C \cdot \sqrt{B_0^2(\widehat{f}_n)/n}. \tag{3.65}$$

Next, using the fact that $\|\widehat{f}_n - f_0\|_n$ is $1/\sqrt{n}$ Lipschitz and the LCP inequality (3.10), we know that $\|\widehat{f}_n - f_0\|_n - \mathbb{E}\|\widehat{f}_n - f_0\|_n$ is $1/\sqrt{n}$-subgaussian. Hence, by using standard argument of integration of tails (see for example (3.29) above), it is easy to see that

$$\mathbb{E}_0\|\widehat{f}_n - f_0\|_n^2 = \text{Med}_0\|\widehat{f}_n - f_0\|_n^2 + O(\max\{\mathbb{E}_0\|\widehat{f}_n - f_0\|_n/\sqrt{n}, 1/n\}).$$

Finally, we take a median on (3.65), and use the last equation and obtain:

$$\begin{aligned}L(f_1) - L(f_0) &\geq \mathbb{E}_0\|\widehat{f}_n - f_0\|_n^2 - 3V_0^2(\widehat{f}_n) - O(\max\{\sqrt{B_0^2(\widehat{f}_n)/n}, \mathbb{E}_0\|\widehat{f}_n - f_0\|_n/\sqrt{n}, 1/n\})\\ &= B_0^2(\widehat{f}_n) + V_0^2(\widehat{f}_n) - 3V_0^2(\widehat{f}_n) - O(\max\{\mathbb{E}_0\|\widehat{f}_n - f_0\|_n/\sqrt{n}, 1/n\})\\ &= B_0^2(\widehat{f}_n) - 2V_0^2(\widehat{f}_n) - O(\max\{\mathbb{E}_0\|\widehat{f}_n - f_0\|_n/\sqrt{n}, 1/n\})\\ &\geq B^2(\widehat{f}_n)/3,\end{aligned}$$

where we used that our assumption of $B_0^2(\widehat{f}_n) \geq C/n$ for C that is large enough; and the claim follows. □

3.5.2 Addendum to §3.2.5

Example 1. Let $\mathcal{X} = \mathbb{S}^n$, the Euclidean unit sphere of dimension n contained in $\mathbb{R}^{n+1}$, let $\mathbb{P}$ be the uniform measure on $\mathcal{X}$, and for each hyperplane H passing through the origin, let $\mathbb{P}^H$ be the uniform measure on $\mathcal{X} \cap H \cong \mathbb{S}^{n-1}$. The set of such hyperplanes is the real $(n+1, n)$ Grassmanian, denoted $\mathrm{Gr_{n+1,n}}$.

For each H, let 1_H denote the characteristic function of H, and let $\mathcal{F} = \mathrm{conv}\{1_H, 1 - 1_H : H \in \mathrm{Gr_{n+1,n}}\}$. Note that in $L^2(\mathbb{P})$, $\mathcal{F}$ reduces to the class of constant functions between 0 and 1 because for any H, $1 - 1_H \equiv 0$ and $1 - 1_H \equiv 1$ almost everywhere on $\mathcal{X}$. Similarly, in $L^2(\mathbb{P}^H)$, $1_H \equiv 1$ and $1 - 1_H \equiv 0$, while for any $H' \neq H$, $1'_H \equiv 1$ and $1 - 1_H \equiv 0$ because $H \cap H'$ has n-dimensional Hausdorff measure 0. In particular, regressing $\mathcal{F}$ on $L^2(\mathbb{P})$ or $L^2(\mathbb{P}^H)$ reduces to estimating an element of $[0, 1]$ given n noisy observations, for which the minimax rate is $\frac{1}{n}$.

On the other hand, let $\mathcal{P} = \{\mathbb{P}\} \cup \{\mathbb{P}^H : H \in \mathrm{Gr_{n+1,n}}\}$, and consider the distribution-unaware minimax risk $\mathcal{M}_n^{(du)}(\mathcal{F}, \mathcal{P})$. We claim that $\mathcal{M}_n^{(du)}(\mathcal{F}, \mathcal{P}) = \omega(1)$ *even when there is no noise*. The intuition behind this is that when the estimator $\bar{f}_n$ observes $(X_1, Y_1), \ldots, (X_n, Y_n)$, it does not "know" whether the distribution is $\mathbb{P}$ or $\mathbb{P}^H$ where $H = \mathrm{Span}(X_1, \ldots, X_n)$, and thus it does not know whether to generalize the observations in a way which is consistent with the $L^2(\mathbb{P})$ norm or the $L^2(\mathbb{P}^H)$ norm.

To see this formally, let $\bar{f}_n : \mathcal{D} \to \mathcal{F}$ be some estimator. For any $\overline{X} \in \mathcal{X}^n$, let $\mathsf{Span}(\overline{X}) := \mathsf{Span}(X_1, \ldots, X_n)$, which is an element of $\mathrm{Gr_{n+1,n}}$ with probability 1. In the noiseless setting, any sample $\overline{X}, \overline{Y}$ such that $\mathsf{Span}(\overline{X}) \in \mathrm{Gr_{n+1,n}}$ will have $\overline{Y}$ a multiple of $(1, \ldots, 1)$ with probability 1, so we may as well consider $\overline{Y}$ to be a constant function. To get a lower bound on the minimax rate, it is also sufficient to consider only the extreme points of $\mathcal{F}$, namely the functions in $\mathcal{F}' = \{1_H, 1 - 1_H : H \in \mathrm{Gr_{n+1,n}}\}$, which are $\{0, 1\}$-functions, so we think of our estimators as functions $\bar{f}_n : \mathcal{X}^n \times \{0, 1\} \to \mathcal{F}$. We also assume for simplicity that $\bar{f}_n$ always returns a function in $\mathcal{F}'$; if $\bar{f}_n$ is allowed to take values in the full convex hull $\mathcal{F}$, one obtains the same lower bound on the risk of $\bar{f}_n$ by a more complicated version of the argument below.

Suppose, for example, that the estimator is given the sample $(\overline{X}, 1)$. In the noiseless setting, there is no point in returning a function inconsistent with the observations, so $\bar{f}_n$ must return either 1_H for $H = \mathsf{Span}(\overline{X})$ or $1 - 1_{H'}$ for some $H' \neq H$ (not containing any of the X_i). Similarly, on the sample $(\overline{X}, 0)$ an optimal estimator $\bar{f}_n$ will return either $1 - 1_H$ or $1_{H'}$ for some $H' \neq H$. Let

$$\begin{aligned} p_{H,0} &= \mathrm{Pr}_{X_i \sim \mathbb{P}^H}\{\bar{f}_n(\overline{X}, 0) = 1 - 1_H\} \\ p_{H,1} &= \mathrm{Pr}_{X_i \sim \mathbb{P}^H}\{\bar{f}_n(\overline{X}, 1) = 1_H\} \\ p_0 &= \mathrm{Pr}_{X_i \sim \mathbb{P}}\{\bar{f}_n(\overline{X}, 1) = 1 - 1_{\mathsf{Span}(\overline{X})}\} \\ p_1 &= \mathrm{Pr}_{X_i \sim \mathbb{P}}\{\bar{f}_n(\overline{X}, 1) = 1_{\mathsf{Span}(\overline{X})}\}. \end{aligned}$$

The Grassmannian $\mathrm{Gr_{n+1,n}}$, which is itself isomorphic to $\mathbb{S}^n$, has a uniform (i.e., rotationally invariant) probability measure, and one has $p_i = \mathbb{E}_{H \sim U(\mathrm{Gr_{n+1,n}})}[p_{H,i}]$: choosing n points from the unit sphere in $\mathbb{R}^{n+1}$ is the same as choosing a uniform hyperplane and then choosing n points uniformly from that hyperplane, by rotational invariance.

One computes that $p_{H,i}$ and p_i determine the error of $\bar{f}_n$ on $\mathcal{F}'$ as follows:

$$\varepsilon^2_{f^*, \mathbb{P}^H} := \mathbb{E} \int (\bar{f}_n - f^*)^2 \, d\mathbb{P}^H = \begin{cases} p_{H,0} & f^* \in \{1 - 1_H\} \cup \{1'_H\}_{H' \neq H} \\ p_{H,1} & f^* \in \{1_H\} \cup \{1 - 1'_H\}_{H' \neq H} \end{cases},$$

$$\varepsilon^2_{f^*,\mathbb{P}} := \mathbb{E}\int (\bar{f}_n - f^*)^2\, d\mathbb{P} = \begin{cases} 1 - p_0 & f^* \in \{1_H\}_{H \in \mathrm{Gr_{n+1,n}}} \\ 1 - p_1 & f^* \in \{1 - 1_H\}_{H \in \mathrm{Gr_{n+1,n}}} \end{cases}.$$

To lower-bound the distribution-unaware minimax risk, we consider the expected error of $\bar{f}_n$ under two different scenarios: when we choose a hyperplane H uniformly at random and measure the error of $\bar{f}_n$ on the function 1_H when the input distribution is $\mathbb{P}^H$, and when we fix $f^* = 1 - 1_{H_0}$ and measure the expected error of $\bar{f}_n$ when the distribution is $\mathbb{P}$. By the above, the error in the first scenario is $\mathbb{E}_{H \sim U(\mathrm{Gr_{n+1,n}})}[p_{H,1}] = p_1$, while the error in the second scenario is $1 - p_1$. As $\max\{1 - p_1, p_1\} \geq \frac{1}{2}$, this shows that $\mathcal{M}_n^{(du)}(\mathcal{F}, \mathcal{P}) = \omega(1)$, as desired.

Remark 22. Note that the $L^2(\mathbb{P})$-diameter of the set of functions which generalize the observation vector $(\overline{X}, 1)$ (say) is huge, as this set includes both 1_H and $1 - 1_{H'}$, where $H = \mathrm{Span}(\overline{X})$ and H' is any other hyperplane.

Remark 23. Example 1 may seem unnatural, as the measures $\mathbb{P}$ and $\mathbb{P}^H$ are all mutually singular, which leads to the "collapse" of the function class in different ways in $L^2(\mathbb{P})$ and each $L^2(\mathbb{P}^H)$. To exclude such pathology, one might wish to consider only families of distributions all of which are absolutely continuous with respect to some reference measure $\mathbb{P}^{(0)}$. It is not difficult, though, to modify Example 1 in such a way which avoids any measurability issues: for given n, let $F = \mathbb{F}_q$ be the finite field of cardinality q for some $q > n$, let $\mathcal{X} = F^{n+1} \backslash \{0\}$, let $\mathbb{P}$ be the uniform probability measure on $\mathcal{X}$, and for each H in the set $\mathrm{Gr_{n+1,n}}(F)$ of n-dimensional linear subspaces of F^{n+1}, let $\mathbb{P}^H$ be the uniform probability measure on $H \backslash \{0\} \subset \mathcal{X}$, let $\mathcal{F} = \mathrm{conv}\{1_H, 1 - 1_H : H \in \mathrm{Gr_{n+1,n}}(F)\}$, and let $\mathcal{P} = \{\mathbb{P}\} \cup \{\mathbb{P}^H\}_{H \in \mathrm{Gr_{n+1,n}}(F)}$ as above. (Note that all measures in $\mathcal{P}$ are absolutely continuous with respect to $\mathbb{P}$.) As $q > n$, each hyperplane H has $\mathbb{P}$-measure $O(n^{-1})$ in $\mathcal{X}$ and for any $H' \neq H$, $H \cap H'$ has $\mathbb{P}^H$-measure $O(n^{-1})$, so one easily verifies all the computations in the example are still valid, up to errors of order $O(n^{-1})$.

3.5.3 Proof of Theorem 10

Recall the following classical characterization for the minimax rate of [YB99] (that holds for Gaussian noise and for any class whose diameter is at most 1):

$$\mathcal{M}(\mathcal{H}_*, \mathbb{P}^{(n)}) \sim \min_{\epsilon} \left[\frac{\log \mathcal{N}(\epsilon, \mathcal{H}_*, \mathbb{P}^{(n)})}{n} + \epsilon^2 \right]. \tag{3.66}$$

Case I: $V(\widehat{f}_n) \leq Sn^{-1}$**.** It is well known (see, e.g., [Wai19, Example 15.4]) that the minimax rate is at least $\Theta(V(\widehat{f}_n))$ the restricted class $\mathcal{H}_*$, when the diameter of $\mathcal{H}_*$ is less than $O(1/\sqrt{n})$. Hence, the claim follows.

Case II: $V(\widehat{f}_n) \geq Sn^{-1}$**.** We will prove that the covering number of $\mathcal{H}_*$ by balls of radius $\sqrt{V(\widehat{f}_n)}/2$ is not too small:

$$\log \mathcal{N}(\sqrt{V(\widehat{f}_n)}/2, \mathcal{H}_*, \mathbb{P}^{(n)}) \geq c_1 n \cdot V(\widehat{f}_n), \tag{3.67}$$

So it suffices to prove that (3.67) holds for an appropriate $c_1 > 0$, as the minimax of $\mathcal{H}_*$ cannot be larger than its diameter which is $2\sqrt{V(\widehat{f}_n)}$. The proof strategy is very similar to that of Theorem 8. Suppose for the sake of contradiction that $\log \mathcal{N}(\sqrt{V(\widehat{f}_n)}/2, \mathcal{H}_*, \mathbb{P}^{(n)}) < c_1 n \cdot V(\widehat{f}_n)$. We consider the distribution of $\widehat{f}_n$ when the true function is f^*. First, note that as $\mathbb{E}\|\widehat{f}_n - \mathbb{E}\widehat{f}_n\|^2 = V(\widehat{f}_n)$, we have that

$$\Pr(\widehat{f}_n \in \mathcal{H}_*) = \Pr(\widehat{f}_n \in B_n(\mathbb{E}_{\mathcal{D}}\widehat{f}_n, 2\sqrt{V(\widehat{f}_n)})) = 1 - \Pr(\|\widehat{f}_n - \mathbb{E}_{\mathcal{D}}\widehat{f}_n\|_n^2 \geq 4V(\widehat{f}_n)) \geq 3/4$$

by Chebyshev's inequality. Let $A = \{f_1, \ldots, f_{\mathcal{N}}\}$ be a minimal $\sqrt{V(\widehat{f}_n)/2}$ - net in $\mathcal{H}_*$; by the pigeonhole principle, there exists at least one element $g \in A$ such that

$$\Pr(\|\widehat{f}_n - g\|_n \leq \sqrt{V(\widehat{f}_n)}/2) \geq \frac{1}{2\mathcal{N}} \geq 3\exp(-c_1 n V(\widehat{f}_n))/4. \tag{3.68}$$

Next, we apply (3.18) with $f = g$ and $t = \sqrt{V(\widehat{f}_n)}/6$, to obtain

$$\Pr_{\overline{\xi}}\left(|\|\widehat{f}_n - g\|_n - \mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - g\|_n| \le \sqrt{V(\widehat{f}_n)}/6\right) \ge 1 - 2\exp(-nV(\widehat{f}_n)/18). \tag{3.69}$$

Recalling that we are in the case $V(\widehat{f}_n) \ge Sn^{-1}$, by choosing $c_1 > 0$ small enough and $S > 0$ large enough we can ensure that $\exp(nV(\widehat{f}_n)(1/18 - c_1)) > 8/3$, or equivalently

$$\frac{3}{4}\exp(-c_1 nV(\widehat{f}_n)) - 2\exp(-nV(\widehat{f}_n)/18) > 0. \tag{3.70}$$

Combining (3.68), (3.69), and (3.70) yields

$$\Pr(\|\widehat{f}_n - g\|_n \le \sqrt{V(\widehat{f}_n)}/2) + \Pr\left(\left|\|\widehat{f}_n - g\|_n - \mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - g\|_n\right| < \sqrt{V(\widehat{f}_n)}/6\right) > 1,$$

so the two events

$$\left\{\left|\|\widehat{f}_n - g\|_n - \mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - g\|_n\right| < \sqrt{V(\widehat{f}_n)}/6\right\}, \{\|\widehat{f}_n - g\|_n \le \sqrt{V(\widehat{f}_n)}/2\}$$

have nonempty intersection, which implies that $\mathbb{E}_{\overline{\xi}}\|\widehat{f}_n - g\|_n < 2\sqrt{V(\widehat{f}_n)}/3$.

Let $h(\xi) = \|\widehat{f}_n - g\|_n$. We have $\mathbb{E}h^2 = (\mathbb{E}h)^2 + \mathbb{E}(h - \mathbb{E}h)^2 < 4V(\widehat{f}_n)/9$. As h is $\frac{1}{\sqrt{n}}$-Lipschitz, the LCP implies that h is $\frac{1}{\sqrt{n}}$-subgaussian. Thus $h - \mathbb{E}h$ is a centered $\frac{1}{\sqrt{n}}$-subgaussian random variable, so $\mathbb{E}(h - \mathbb{E}h)^2 \le \frac{2}{n}$ [Ver18, Proposition 2.5.2], and hence

$$\mathbb{E}_{\mathcal{D}}\|\widehat{f}_n - g\|_n^2 < \frac{4}{9}V(\widehat{f}_n) + \frac{2}{n}.$$

Again recalling that $V(\widehat{f}_n) > Sn^{-1}$, by taking S large enough we can ensure that $\mathbb{E}_{\mathcal{D}}\|\widehat{f}_n - \mathbb{E}\widehat{f}_n\|_n^2 < V(\widehat{f}_n)$, which contradicts the definition of $V(\widehat{f}_n)$.

Chapter 4

Efficient Estimators For Multivariate Convex Regression

4.1 Main Results

This chapter is dedicated to prove the results that appeared in §1.3 above.

Theorem 4. *Let $d \geq 5$ and $n \geq d+1$. Then, under Assumption 6, for the task of L-Lipschitz convex regression on a convex polytope $\Omega \subset B_d$, there exists an efficient estimator, $\widehat{f}_{L,n}$, with runtime of at most $n^{O(d)}$ such that*

$$\mathcal{R}(\widehat{f}_{L,n}, \mathcal{F}_L(\Omega), \mathbb{P}) \leq (\sigma + L)^2 n^{-\frac{4}{d+4}} \log(n)^{h(d)} + C(\Omega) n^{-\frac{d}{d+4}} \log(n)^{2 \cdot h(d)}, \tag{1.16}$$

where $h(d) \leq 3d$ and $C(\Omega)$ is a constant that only depends on the polytope Ω.[1]

Theorem 5. *Let $d \geq 5$ and $n \geq d+1$. Then, under Assumption 6, for the task of Γ-bounded convex regression on the polytope $\Omega \subset B_d$, there exists an efficient estimator, $\widehat{f}^{\Gamma,n}$, with runtime of at most $O_d(n^{O(d)})$ such that*

$$\mathcal{R}(\widehat{f}^{\Gamma,n}, \mathcal{F}^{\Gamma}(\Omega), \mathbb{P}) \leq C_1(\Omega)(\sigma + \Gamma)^2 n^{-\frac{4}{d+4}} \log(n)^{h(d)},$$

[1] Specifically, it depends on the flags number of polytope Ω (cf. [RSW19]).

where $h(d) \leq 3d$ and $C_1(\Omega)$ is a constant that only depends on Ω.

4.1.1 Notations and Preliminaries

Throughout this text, $C, C_1, C_2 \in (1,\infty)$ and $c, c_1, c_2, \ldots \in (0,1)$ are positive absolute constants that may change from line to line. Similarly, $C(d), C_1(d), C_2(d), \ldots \in (1,\infty)$ and $c(d), c_1(d), c_2(d), \ldots \in (0,1)$ are positive constants that only depend on d that may change from line to line. We also often use expressions such as $g(n) \leq O_d(f(n))$ to mean that there exists $C_d \geq 0$ such that $g(n) \leq C_d f(n)$ for all n.

For any probability measure $\mathbb{Q}$ and $m \geq 0$, we introduce the notation $\mathbb{Q}_m$ for the random empirical measure of $Z_1, \ldots, Z_m \underset{i.i.d.}{\sim} \mathbb{Q}$, i.e. $\mathbb{Q}_m = m^{-1} \sum_{i=1}^m \delta_{Z_i}$. Also, given a subset $A \subset \Omega$ of positive measure, we let $\mathbb{P}_A$ denote the conditional probability measure on A. For a positive integer k, $[k]$ denotes $\{1, \ldots, k\}$.

Definition 11. *A simplex in $\mathbb{R}^d$ is the convex hull of $d+1$ points $v_1, \ldots, v_{d+1} \in \mathbb{R}^d$ which do not all lie in any hyperplane.*

Definition 12. *A convex function $f : \Omega \to \mathbb{R}$ is defined to be k-simplicial if there exists $\triangle_1, \ldots, \triangle_k \subset \mathbb{R}^{\dim(\Omega)}$ simplices such that $\Omega = \bigcup_{i=1}^k \triangle_i$ and for each $1 \leq i \leq k$, we have that $f : \triangle_i \to \mathbb{R}$ is affine.*

Note that the definition is more restrictive than the usual definition of a k-max affine function (see Remark 3), since the affine pieces of a k-max affine function are not constrained to be simplices.

The following result from empirical process theory is a corollary of the peeling device [Gee00, Ch. 5], [Bou02] and Bronshtein's entropy bound [Bro76].

Lemma 24. *Let $d \geq 5$, $m \geq C^d$ and $\mathbb{Q}$ be a probability measure on $\Omega' \subset B_d$. Suppose $Z_1, \ldots, Z_m$ are drawn independently from $\mathbb{Q}$; then with probability at least $1 - C_1(d)\exp(-c_1(d)\sqrt{m})$, the following holds uniformly for all $f, g \in \mathcal{F}_L(\Omega')$:*

$$2^{-1} \int_{\Omega'} (f-g)^2 d\mathbb{Q} - CL_2 m^{-\frac{4}{d}} \leq \int (f-g)^2 d\mathbb{Q}_m \leq 2 \int_{\Omega'} (f-g)^2 d\mathbb{Q} + CL_2 m^{-\frac{4}{d}}. \quad (4.1)$$

Next, we introduce a statistical estimator with the high-probability guarantees we shall need, based on a recent result that is presented in [MVZ21, Prop. 1]. Its statistical aspects are proven in the seminal works [Tsy03b; LM19], and guarantees on its runtime are given in [Hop18; DL19; HLZ20].

Lemma 25. *Let $m \geq d+1$, $d \geq 1$, $\delta \in (0,1)$ and $\mathbb{Q}$ be a probability measure that is supported on $\Omega' \subset B_d$ with* a known *covariance matrix Σ. Consider the regression model $W = f^*(Z) + \xi$, where $\|f^*\|_\infty \leq L$, and let $Z_1, \dots, Z_m \underset{i.i.d.}{\sim} \mathbb{Q}$. Then, there exists an estimator $\hat{f}_{R,\delta}$ that has an input of $(\Sigma, \{(Z_i, W_i)\}_{i=1}^m)$ and runtime of $\widetilde{O}_d(m)$ and outputs an L-Lipschitz affine function that satisfies with probability at least $1-\delta$*

$$\int (\hat{f}_{R,\delta}(x) - w^*(x))^2 d\mathbb{Q}(x) \leq \frac{C(\sigma+L)^2(d+\log(1/\delta))}{m},$$

where $w^* = \operatorname{argmin}_{w \text{ affine}} \int (w - f^*)^2 d\mathbb{Q}$.

4.1.2 The Proposed Estimator of Theorem 4

For simplicity assume that $L = 1$ and that $\sigma = \Theta(1)$. Without loss of generality, we may assume that $\|f^*\|_\infty \leq L = 1$, since our function is 1-Lipschitz and $\Omega \subset B_d$. Also, without loss of generality, we further assume that $\mathbb{P} = \mathrm{U}(\Omega)$, where $\mathrm{U}(\Omega)$ denotes the uniform measure over Ω. To see this, note that can always simulate $\Theta_d(n)$ uniform samples using the method of rejection sampling given samples from any distribution $\mathbb{P}$ which satisfies Assumption 6 (cf. [Dev86]). *To conclude we assume that $L = 1, \sigma = \Theta(1)$, $\mathbb{P} = \mathrm{U}(\Omega)$, and recall that we assume that $d \geq 5$.*

In order on prove the correctness of our estimator, we develop the following approximation theorem for convex functions (its proof appears below).

Theorem 17. *Let $\Omega \subset B_d$ be a convex polytope, $f \in \mathcal{F}_L(\Omega)$, and k an integer greater than $(Cd)^{d/2}$, for some large enough $C \geq 0$. Then, there exists a convex set $\Omega_k \subset \Omega$ and a k-simplicial convex function $f_k : \Omega_k \to \mathbb{R}$ such that*

$$\mathbb{P}(\Omega \setminus \Omega_k) \leq C(\Omega) k^{-\frac{d+2}{d}} \log(k)^{d-1}. \tag{4.2}$$

and

$$\int_{\Omega_k} (f_k - f)^2 d\mathbb{P} \leq L^2 \cdot O_d(k^{-\frac{4}{d}} + C(\Omega) k^{-1} \log(k)^{d-1}). \tag{4.3}$$

Note that both Ω_k, as well as f_k, depend on f^*. The bound of Eq. (4.3) is in fact tight, up to a constant that only depends on d, cf. [LSW06]. Also note that f_k is not necessarily an L-Lipschitz function, i.e., it may be an "improper" approximation to f. We remark that the constant $C(\Omega)$ depends on the flag number of the polytope Ω; for more details see [RSW19]. As we mentioned earlier, we assume that the number of vertices or facets of Ω is bounded by C_d, so the definition of the flag number and the upper bound theorem of McMullen [McM70] implies that we can assume $C(\Omega) \leq C_1(d)$.

Fix $n \geq (Cd)^{d/2}$, $f^* \in \mathcal{F}_1(\Omega)$, and set $k(n) := n^{\frac{d}{d+4}}$. Let $f_{k(n)} : \Omega_{k(n)} \to \mathbb{R}$ be the convex function whose existence is guaranteed by Theorem 17 for $f = f^*$. We have

$$\int (f^* - f_{k(n)})^2 d\mathbb{P} \leq O_d(n^{-\frac{4}{d+4}}), \tag{4.4}$$

and there exist $\triangle_1, \dots, \triangle_{k(n)} \subset \Omega$ simplices such that $f_{k(n)}\Big|_{\triangle_i}$ is affine on each i.

If we were given the decomposition of Ω into pieces on which f^* is near-affine, it would be relatively simple to estimate f^*, as we show in §4.4 below. We recommend reading it, to get some intuition for our approach, before attempting the description and correctness proof for our "full" estimator below.

To overcome the fact that we do not know the simplices $\triangle_i$ on which f_k is affine, we need another lemma, which says that if we randomly sample a set of n points $\{X_{n+1}, \dots, X_{2n}\}$ from Ω, there exists a collection of at most $\widetilde{O}_d(k(n))$ simplices covering "most" of Ω_k, such that the vertices of each simplex belong to $\{X_{n+1}, \dots, X_{2n}\}$ and f_k is affine on each simplex in the collection:

Lemma 26. *Let $n \geq d+1$, and $\triangle_1, \dots, \triangle_{k(n)}$ that are defined above, and let $X_{n+1}, \dots, X_{2n} \underset{i.i.d.}{\sim} \mathbb{P}$. Then, with probability at least $1 - n^{-1}$, there exist $k(n)$ disjoint sets $\mathcal{S}_X^1, \dots \mathcal{S}_X^{k(n)}$ of simplices with disjoint interiors such that*

1. *The vertices of each simplex in $\bigcup_{i=1}^{k(n)} \mathcal{S}_X^i$ lie in $\{X_{n+1}, \dots, X_{2n}\}$. Moreover, for each $1 \leq i \leq k(n)$, we have that $|\mathcal{S}_X^i| \leq O_d(\log(n\mathbb{P}(\triangle_i))^{d-1})$.*

2. *For each* $1 \leq i \leq k(n)$*, we have that* $\bigcup \mathcal{S}_X^i \subset \triangle_i$*, and*

$$\mathbb{P}(\bigcup \mathcal{S}_X^i) \geq \mathbb{P}(\triangle_i) - \min\left\{ O_d\left(\frac{\log(n)\log(n\mathbb{P}(\triangle_i))^{d-1}}{n}\right), \mathbb{P}(\triangle_i)\right\}.$$

The proof of this lemma appears in subsection 4.2.2. Essentially, this lemma states that we can triangulate "most" of each simplex $\triangle_i$ with "few" simplices whose vertices lie among the points $X_{n+1}, \ldots, X_{2n}$ which fall in $\triangle_i$, so long as $\triangle_i$ is large enough. From now on, we condition on the high-probability event of Lemma 26.

Note that if we were given the set of simplices $\mathcal{S}_X := \bigcup_{i=1}^{k(n)} \mathcal{S}_X^i$, we could use the same strategy as in §4.4 to obtain a minimax optimal estimator for this task as well. Unfortunately, we do not know how to identify the simplices of $\mathcal{S}_X$, but we do know that they belong to the collection of all simplices with vertices in $\{X_{n+1}, \ldots, X_{2n}\}$,

$$\mathcal{S} := \{\text{conv}\{X_{n+i} : i \in S\} : S \subset [n], |S| = d+1\}. \tag{4.5}$$

Note that $|\mathcal{S}| = O_d(n^{d+1})$, which is polynomial in n.

Instead of trying to identify the simplices $\mathcal{S}_X \subset \mathcal{S}$ on which f is close to being linear, our algorithm finds a function $\hat{f}$ which, on *every* simplex $\triangle \in \mathcal{S}$, is "not much farther" from the best linear approximation to f on $\triangle$ then f is. Since f itself is close to its best linear approximation on each simplex in $\mathcal{S}_X$, $\hat{f}$ will be close to f on $\bigcup \mathcal{S}_X$, which is most of Ω.

We restate this a bit more precisely: if $\hat{f} : \Omega \to \mathbb{R}$ is a convex Lipschitz function such that

$$\forall \triangle \in \mathcal{S} : \int (\hat{f} - w_\triangle^*)^2 d\mathbb{P}_\triangle \leq \widetilde{O}_d\left(\int (f^* - w_\triangle^*)^2 d\mathbb{P}_\triangle + (\mathbb{P}(\triangle)n)^{-1}\right) \tag{4.6}$$

where $w_\triangle^* = \inf_{w \text{ affine}} \int (f^* - w)^2 d\mathbb{P}_\triangle$, then $\hat{f}$ satisfies $\int_\Omega (\hat{f} - f^*)^2 d\mathbb{P} \leq \widetilde{O}_d(n^{-\frac{4}{d+4}})$, and the RHS is the minimax optimal rate. (The idea of the proof is to use the fact that f^* is close to $w_\triangle^*$ for each $\triangle \in \mathcal{S}_X$, along with the triangle inequality, and then sum over all simplices in $\mathcal{S}_X$; the full justification is given at (4.15)-(4.18) below.) In the remainder of this section, we will describe an efficient algorithm which constructs a function $\hat{f}$ which comes "close enough" to satisfying (4.6) that it manages to attain the minimax optimal rate.

We now begin the description of our algorithm. In the notation of (4.6), for each simplex $\triangle \in \mathcal{S}$, we estimate $w^*_\triangle$ by applying Lemma 25, with the data points of $\mathcal{D}^1 = \{(X_i, Y_i)\}_{i=1}^{n/2}$ that lie in $\triangle$ as input; denote the regressor we obtain by $\hat{w}_\triangle$.

Next, we shall need to estimate (with high probability) the squared error of the regressor on each simplex in $\mathcal{S}$, i.e. $\ell^2_\triangle := \|f^* - \hat{w}_\triangle\|^2_{L_2(\mathrm{U}(\triangle))}$, up to a polylogarithmic multiplicative factor, using the data points in $\mathcal{D}^2 = \{(X_i, Y_i)\}_{i=n/2+1}^{n}$ that lie in $\triangle$. Letting $w^*_\triangle$ to be defined as above, we have

$$\ell^2_\triangle = \|w^*_\triangle - \hat{w}_\triangle\|^2_{L_2(\mathrm{U}(\triangle))} + \|f^* - w^*_\triangle\|^2_{L_2(\mathrm{U}(\triangle))};$$

the first term is called the (squared) estimation error and the second is called the (squared) approximation error. By Lemma 25, with probability at least $1 - n^{-2d}$ the estimation error will be at most $Cd\log(n)/(\mathbb{P}(\triangle)n)$, which is no more than a $O(d\log(n))$ factor times the expected estimation error. However, f^* may not be affine on $\triangle$, and the squared approximation error may be significantly larger than the squared estimation error. When this occurs, the estimation of $\ell^2_\triangle$ by noisy samples is challenging, even in the (unrealistic) setting of sub-Gaussian noise with known variance σ^2. Indeed, it would be natural to estimate the approximation error by the (centered) empirical mean of the squared loss, namely

$$\frac{1}{\mathbb{P}(\triangle)n} \sum_{(X,Y)\in\mathcal{D}_2, X\in\triangle} (Y - \hat{w}_\triangle(X))^2 - \sigma^2.$$

However, the additive deviation of this estimate is of order $\Omega_d(n_\triangle^{-\frac{1}{2}})$, where $n_\triangle \approx \mathbb{P}(\triangle)n$ is the number of data points falling in $\triangle$, and therefore when $\ell^2_\triangle$ is in the range $[O_d(n_\triangle^{-1}), \Omega_d(n_\triangle^{-\frac{1}{2}})]$ we will not be able to estimate $\ell^2_\triangle$ even *up to a multiplicative constant*, which is what our algorithm requires in order to succeed.

Overcoming this problem necessitates constructing a new efficient procedure to estimate $\ell^2_\triangle$, up to a multiplicative constant with an *additive* deviation of $\widetilde{O}_d((\mathbb{P}(\triangle)n)^{-1})$, using the data points of $\mathcal{D}_2$ that fall in $\triangle$. Using the convexity of $f^* - \hat{w}_\triangle$ and techniques from potential theory and stochastic geometry, we show that such a procedure indeed exists. Specifically, we prove the following:

Lemma 27. *Let $\triangle \subset \Omega$ and let $g : \triangle \to [-2L, 2L]$ be a convex function such $\int_{\triangle} g^2 d\mathbb{P} \geq CdL_2 \log(n)^2/n$. Then, there exists an estimator $\hat{f}_{E,\delta(n)}$ (that uses that points of $\mathcal{D}_2$) with a runtime $O_d(n^{O(1)})$ and with probability at least $1 - n^{-2d}$ satisfies*

$$\|g\|^2_{L_2(\mathrm{U}(\triangle))} \leq \hat{f}_{E,\delta(n)} \leq C(d) \log(n)^{2d-1} \left(\|g\|^2_{L_2(\mathrm{U}(\triangle))} + (L+\sigma)^2 \frac{\log(2/\delta)}{\mathbb{P}(\triangle) n} \right),$$

where $C(d)$ is a constant that only depends on d.

We remark that our estimator requires an upper bound on L and σ (up to multiplicative constants that only depend on d). Both can be found using standard methods when $L = \Theta(\sigma) = \Theta(1)$. We provide additional information about this estimator in §4.1.3 below.

We denote the output of the estimator of Lemma 27 for $g = f^* - \hat{w}_{\triangle}$ by $\hat{\ell}^2_{\triangle}$ for any $\triangle \in \mathcal{S}$. Given our regressors $\hat{w}_{\triangle}$ and squared error estimates $\hat{\ell}^2_{\triangle}$, we proceed to solve the quadratic program which encodes the conditions $\|\widetilde{f} - \hat{w}_{\triangle}\|^2_{L_2(\mathrm{U}(\triangle))} \leq \hat{\ell}^2_{\triangle}$ for all simplices with large enough volume. (We rely on the fact that the L_2-norm on each simplex can be approximated by the empirical L_2-norm, again using Lemma 24.) This program is feasible, since f^* itself is a solution. $\widetilde{f}$ is close to f^* on every simplex in our collection and in particular on the simplices restricted to which f^* is near-affine (which we don't know how to identify), which allows us to conclude that $\int_{\widetilde{\Omega}_{k(n)}} (\widetilde{f} - f^*)^2 d\mathbb{P} \leq \widetilde{O}_d(n^{-\frac{4}{d+4}})$ with high probability, where $\widetilde{\Omega}_{k(n)}$ is the union of the simplices in Lemma 26.

So we have constructed a function $\widetilde{f}$ which closely approximates f^* on $\widetilde{\Omega}_{k(n)}$. $\widetilde{\Omega}_{k(n)}$ is not known to us, but as we shall see, $\Omega \backslash \widetilde{\Omega}_{k(n)}$ has asymptotically negligible volume, so the function $\min\{\widetilde{f}, L\}$ turns out to be a minimax optimal improper estimator (up to logarithmic factors) of f^* on all of Ω. In order to transform this improper estimator to a proper estimator, i.e., one whose output is a convex L-Lipschitz function, we use a standard procedure (denoted by MP), as described in §4.5 below. This concludes the sketch of our algorithm.

Pseudocode for the algorithm is given in Algorithm 1 below. In its formulation, note that the procedure $\hat{f}_{R,\delta(n)}$ is described in Lemma 25, $\hat{f}_{E,\delta(n)}$ is described in Lemma 27, and MP is described in § 4.5.

```
Require: $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$
Ensure: A random $\widehat{f}_L \in \mathcal{F}_L(\Omega)$ s.t. w.h.p. $\|\widehat{f}_L - f^*\|_{\mathbb{P}}^2 \leq \widetilde{O}_d((L+\sigma)^2 n^{-\frac{4}{d+4}})$.
  Draw $X_{n+1}, \ldots, X_{2n} \underset{i.i.d.}{\sim} \mathbb{P}$
  $\mathcal{S} \leftarrow \{\mathrm{conv}\{X_{n+i} : i \in S\} : S \subset [n], |S| = d+1\}$
  Part I:
  $\mathcal{D}^1 \leftarrow \{(X_i, Y_i)\}_{i=1}^{n/2}$
  $\mathcal{D}^2 \leftarrow \{(X_i, Y_i)\}_{i=n/2+1}^{n}$
  for $\triangle_1, \ldots, \triangle_i, \ldots \in \mathcal{S}$ do
    $\hat{w}_i \leftarrow \hat{f}_{R,\delta(n)}(\{(X,Y) \in \mathcal{D}^1 : X \in \triangle_i\})$.
    $\hat{\ell}_i \leftarrow \min(4, \hat{f}_{E,\delta(n)}(\{(X, Y - \hat{w}_i(X)) : (X,Y) \in \mathcal{D}^2, X \in \triangle_i\}))$
  end
  Part II:
  for $i \in 1, \ldots |\mathcal{S}|$ do
    Draw $Z_{i,1}, \ldots, Z_{i,n^2} \sim \mathbb{P}_{\triangle_i}$
    Define an inequality constraint
    $I_j := \frac{1}{n^2}\sum_{j=1}^{n^2}(f(Z_{i,j}) - \hat{w}_i^\top(Z_{i,j}))^2 \leq \hat{\ell}_i^2 + CL^2\frac{\sqrt{d\log(n)}}{n}$.
    Construct $\widetilde{f} \in \mathcal{F}_L(\Omega)$ satisfying the constraints $I_1, I_2 \ldots, I_{|S|}$ (cf. Eqs. (4.11)-(4.13))
  end
  return $MP(\min\{\widetilde{f}, L\})$
```

Algorithm 1: A Minimax Optimal For L-Lipschitz Multivariate Convex Regression

We now turn to the proof that Algorithm 1 succeeds with high probability. In the analysis, we assume for simplicity that $L = \sigma = 1$. Let $\mathcal{S}$ be as defined in Algorithm 1, and let $\mathcal{S}^T := \{\triangle : \triangle \in \mathcal{S}, \int_{\triangle} g^2 d\mathbb{P} \geq Cd\log(n)^2/n\}$, for some sufficiently large C. In particular, we have $\mathbb{P}(S) \geq C_1(C)d\log(n)/n$ for all $S \in \mathcal{S}^T$. We first note that our samples may be assumed to be close to uniformly distributed on the simplices in $\mathcal{S}^T$. Indeed, by standard concentration bounds, we have

$$\forall \triangle \in \mathcal{S}^T, j \in \{1,2,3\} : \quad \frac{1}{2} \leq \frac{\mathbb{P}_n^{(j)}(\triangle)}{\mathbb{P}(\triangle)} \leq 2, \tag{4.7}$$

with probability $1 - 3n^{-3d}$, where $\mathbb{P}_n^{(1)} = \frac{2}{n}\sum_{i=1}^{n/2}\delta_{X_i}$, $\mathbb{P}_n^{(2)} = \frac{2}{n}\sum_{i=n/2+1}^{n}\delta_{X_i}$ and $\mathbb{P}_n^{(3)} = \frac{1}{n}\sum_{i=n+1}^{2n}\delta_{X_i}$ (see Lemma 29 in sub-Section 4.2.2). From now on, we condition on the intersection of the events of (4.7) and Lemma 26.

The first step in the algorithm is to apply the estimator of Lemma 25 for each $\triangle_i \in \mathcal{S}^T$ with

$\mathbb{Q} := \mathbb{P}_{\triangle_i}$, and $\delta = n^{-(d+2)}$, using those points among of $\mathcal{D}^1$ that fall in $\triangle_i$. (By the preceding paragraph, under our conditioning, we may assume $\mathbb{P}(\triangle_i)n$ of the points in $X_1, \ldots, X_{\frac{n}{2}}$ fall in each $\triangle_i$, up to absolute constants. We will silently use the same argument several more times below.) By the lemma and a union bound, we know that the following event has probability at least $1 - n^{-1}$:

$$\forall 1 \leq i \leq |\mathcal{S}^T| \;\; \int_{\triangle_i} (\hat{w}_i(x) - f^*(x))^2 d\mathbb{P}_{\triangle_i} \leq \frac{2Cd\log(n)}{\mathbb{P}(\triangle_i)n} + \int_{\triangle_i} (w_i^*(x) - f^*(x))^2 d\mathbb{P}_{\triangle_i}, \tag{4.8}$$

where $w_i^* = \operatorname{argmin}_{w \text{ affine}} \int_{\triangle_i} (w(x) - f^*)^2 d\mathbb{P}_{\triangle_i}$. We condition also on the event of (4.8). Next, we apply Lemma 27 (with $\delta = n^{-(d+2)}$) on each $\triangle_i$, with $g = f^* - \hat{w}_i$, and using those points among of $\mathcal{D}^2$ that fall in $\triangle_i$, and obtain that

$$\forall 1 \leq i \leq |\mathcal{S}| : \;\; \int_{\triangle_i} (\hat{w}_i - f^*)^2 d\mathbb{P}_{\triangle_i} \leq \hat{\ell}_i^2, \tag{4.9}$$

with $\hat{\ell}_i^2$ as defined in Algorithm 1. Note that for $\triangle \in \mathcal{S} \setminus \mathcal{S}^T$, taking $\hat{w}_i = 0$ suffices, since f^* is bounded by 1, the loss is bounded by 4. Finally, we further condition on the event of the last equation.

We proceed to explain and analyze Part II of Algorithm 1. We first claim that conditioned on (4.9), the function f^* satisfies the constraints $I_1, I_2, \ldots$ defined in the algorithm with probability at least $1 - n^{-1}$. Indeed, for each $1 \leq i \leq |\mathcal{S}|$, $\|(f^* - \hat{w}_i)^2\|_{L^\infty(\triangle_i)} \leq 4$, so by Hoeffding's inequality and (4.9) we know that with probability at least $1 - n^{-(d+2)}$, we have that

$$\begin{aligned} \frac{1}{n^2} \sum_{j=1}^{n^2} (f^*(Z_{i,j}) - \hat{w}_i(Z_{i,j}))^2 &\leq \int_{\triangle_i} (f^* - \hat{w}_i)^2 d\mathbb{P}_{\triangle_i} + \frac{\sqrt{Cd\log(n)}}{n} \\ &\leq \hat{\ell}_i^2 + \frac{\sqrt{Cd\log(n)}}{n}. \end{aligned} \tag{4.10}$$

Taking a union over i, we know that (4.10) holds for all i with probability at least $1 - n^{-1}$.

We also note (for later use) that applying Lemma 24 to the measures $\mathbb{P}_{\triangle_i}$ and using a union bound, it holds with probability at least $1 - Cn^d e^{-c\sqrt{n}}$ that for all i, the empirical measure $\mathbb{P}_{\triangle_i, n} = \frac{1}{n} \sum_{j=1}^{n} \delta_{Z_{i,j}}$ on $\triangle_i$ approximates $\mathbb{P}_{\triangle_i}$ in the sense of (4.1). We condition on the intersection of these two events as well.

We now explain how to algorithmically construct $\widetilde{f} \in \mathcal{F}_L(\Omega)$ satisfying all the constraints I_j. The idea is to mimic the computation of the convex LSE [SS11], by considering the values of the unknown function $y_{i,j} = \widetilde{f}(Z_{i,j})$ and the subgradients $\xi_{i,j} \in \partial f(Z_{i,j})$ at each $Z_{i,j}$ as variables. More precisely, we search for $y_{i,j} \in \mathbb{R}$ and $\xi_{i,j} \in \mathbb{R}^d$ satisfying the following set of constraints (here $L = 1$):

$$\forall i \leq |\mathcal{S}| : \qquad \frac{1}{n^2}\sum_{j=1}^{n^2}(y_{i,j} - \hat{w}_i(Z_{i,j}))^2 \leq \hat{\ell}_i^2 + \frac{\sqrt{Cd\log(n)}}{n} \tag{4.11}$$

$$\forall (i,j) \in [|\mathcal{S}|] \times [n] : \qquad \|\xi_{i,j}\|^2 \leq L^2 \tag{4.12}$$

$$\forall (i_1,j_1),(i_2,j_2) \in [|\mathcal{S}|] \times [n] : \qquad y_{i_2,j_2} \geq \langle \xi_{i_1,j_1}, Z_{i_2,j_2} - Z_{i_1,j_1} + y_{i_1,j_1} \rangle. \tag{4.13}$$

For any feasible solution $(y_{i,j}, \xi_{i,j})_{i,j}$ of (4.11)-(4.13), define the affine functions $a_{i,j}(x) = y_{i,j} + \langle \xi_{i,j}, x - Z_{i,j} \rangle$. We claim that the function $\widetilde{f} = \max_{i,j} a_{i,j}$ is a 1-Lipschitz convex function which satisfies the constraints I_j. Indeed, (4.13) guarantees that $\widetilde{f}(Z_{i,j}) = y_{i,j}$ for each i, so the I_j are satisfied due to (4.11); moreover, the $a_{i,j}$ are convex and 1-Lipschitz (the latter because of (4.12)), so $\widetilde{f}$ is convex as a maximum of convex functions and 1-Lipschitz as a maximum of 1-Lipschitz functions.

Conditioned on (4.10) there exists a feasible solution to the problem (4.11)-(4.13), namely that obtained by taking $y_{i,j} = f^*(Z_{i,j})$, $\xi_{i,j} \in \partial f^*(Z_{i,j})$ (where $\partial f(x)$ denotes the subgradient set of a convex function f at the point x). Moreover, the constraints in (4.11)-(4.13) are either linear or convex and quadratic in $f(Z_{i,j}), u_{i,j}$, and hence the problem can be solved efficiently. For instance, it can be expressed as a second-order cone program (SOCP) with $O_d(n^{2d+2})$ variables and constraints, which can be solved in time $O_d(n^{O(d)})$ (see, e.g., [BTN01]).

Next, recall that under our conditions on the $Z_{i,j}$, we have for each i that

$$\int_{\triangle_i} (\widetilde{f}(x) - f^*(x))^2 d\mathbb{P}_{\triangle_i} \leq \frac{1}{n^2}\sum_{j=1}^{n^2}(\widetilde{f}(Z_{i,j}) - f^*(Z_{i,j}))^2 + Cn^{-\frac{4}{d}}, \tag{4.14}$$

since both f^* and $\widetilde{f}$ lie in $\mathcal{F}_1(\Omega)$. Recall also that under our conditioning, for each i, the constraint

$$\frac{1}{n^2}\sum_{j=1}^{n^2}(y_{i,j}-\hat{w}_i(Z_{i,j}))^2 \leq \hat{\ell}_i^2+\frac{\sqrt{Cd\log(n)}}{n}$$

holds whether we take $y_{i,j}=\widetilde{f}(Z_{i,j})$ or $y_{i,j}=f^*(Z_{i,j})$. Using this bound along with the inequality $(f^*-\widetilde{f})^2 \leq 2(\widetilde{f}-\hat{w}_i)^2+2(\hat{w}_i-f^*)^2$ in (4.14), we obtain

$$\int_{\triangle_i}(\widetilde{f}(x)-f^*(x))^2 d\mathbb{P}_{\triangle_i} \leq 2\hat{\ell}_i^2+\frac{\sqrt{Cd\log(n)}}{n}+Cn^{-\frac{4}{d}} \leq 2\hat{\ell}_i^2+C'n^{-\frac{4}{d+4}}, \tag{4.15}$$

where we used our assumption of $d \geq 5$. Now, recalling that ℓ_i^2 denotes the LSE error $\|f^*-\hat{w}_i\|_{L_2(\triangle_i)}^2$, which is bounded by (4.8), we have

$$\hat{\ell}_i^2 \leq C_d\log(n)^{3d}\ell_i^2 \leq C_d\log(n)^{2d-1}\left(\|f^*-w_i^*\|_{L_2(\triangle_i)}^2+\frac{C(d)\log n}{\mathbb{P}(\triangle_i)n}\right).$$

Substituting in (4.15), we obtain for any $\triangle_i$ that

$$\int_{\triangle_i}(\widetilde{f}(x)-f^*(x))^2 d\mathbb{P}_{\triangle_i} \leq C(d)\left(\log(n)^{2d-1}\|f^*-w_i^*\|_{L_2(\triangle_i)}^2+\frac{\log(n)^{2d}}{\mathbb{P}(\triangle_i)n}+n^{-\frac{4}{d+4}}\right). \tag{4.16}$$

Now recall that $f_{k(n)}$ is our $k(n)$-simplicial approximation to f, and that $f_{k(n)}\Big|_{S_{i,j}}$ is affine for each i and $S_{i,j} \in \mathcal{S}_X^i$, where the sets $\mathcal{S}_X^i$ are defined in Lemma 26. Define $\widetilde{\Omega}_{k(n)} := \bigcup\bigcup_{i=1}^{k(n)}\mathcal{S}_X^i$. Recall that by definition, $\|f^*-w_m^*\|^2 = \inf_{w \text{ affine}}\|f^*-w\|_{L_2(S_m)}^2$ for any $S_m \in \mathcal{S}$, in particular for those S_m which belong to one of the $\mathcal{S}_X^i$. Hence, multiplying (4.16) by $\mathbb{P}(\triangle_i)$ and summing over all the $\triangle_i$ belonging to any of the $\mathcal{S}_X^i$, we obtain

$$\begin{aligned}\int_{\widetilde{\Omega}_{k(n)}} (\widetilde{f}-f^*)^2 d\mathbb{P} &\leq C_1(d) \sum_{\widetilde{\triangle}\in\bigcup_{i=1}^{k(n)} \mathcal{S}_X^i} \left(\log(n)^{2d-1} \inf_{w \text{ affine}} \int_{\widetilde{\triangle}} (f^*-w)^2 d\mathbb{P} + \frac{\log(n)^{2d}}{n} + \mathbb{P}(\widetilde{\triangle}) n^{-\frac{4}{d}}\right) \\ &\leq C_d \left(\log(n)^{2d-1} \int_{\widetilde{\Omega}_{k(n)}} (f_{k(n)}-f^*)^2 d\mathbb{P} + \log(n)^{3d} n^{-\frac{4}{d+4}}\right) \\ &\leq O_d(n^{-\frac{4}{d+4}} \log(n)^{3d}),\end{aligned} \tag{4.17}$$

where we used the fact that the cardinality of $\bigcup_{i=1}^{k(n)} \mathcal{S}_X^i$ is bounded by $O_d(n^{\frac{d}{d+4}} \log(n)^d)$, the disjointness of all the simplices comprising $\widetilde{\Omega}_{k(n)}$, and finally the definition of $f_{k(n)}$. Next, it is not hard to show that $\|\widetilde{f}\|_\infty \leq C$ (simply because $\widetilde{f}$ is 1-Lipschitz, Ω is contained in the unit ball, and $\int_{\Omega_{k(n)}} (\widetilde{f}-f^*)^2 \leq 4$). Thus, we obtain that

$$\begin{aligned}\int_{\Omega_{k(n)} \setminus \widetilde{\Omega}_{k(n)}} (\widetilde{f}-f^*)^2 d\mathbb{P} \leq \|\widetilde{f}\|_\infty \mathbb{P}(\Omega_{k(n)} \setminus \widetilde{\Omega}_{k(n)}) &\leq C_1 \sum_{i=1}^{k(n)} \mathbb{P}(\triangle_i \setminus \bigcup \mathcal{S}_X^i) \\ &\leq O_d(n^{-\frac{4}{d+4}} \log(n)^d).\end{aligned}$$

where we used part (2) of Lemma 26. Combining the last two equations, we obtain that

$$\int_{\Omega_{k(n)}} (\widetilde{f}-f^*)^2 d\mathbb{P} \leq O_d(n^{-\frac{4}{d+4}} \log(n)^{3d}). \tag{4.18}$$

Finally, since $\mathbb{P}(\Omega \setminus \Omega_{k(n)}) \leq C(d) n^{-\frac{d+2}{d+4}} \log(n)^{d-1}$, we can estimate f^* simply by 1 on $\Omega \setminus \Omega_{k(n)}$ and the error of doing so will be asymptotically negligible ($n^{-\frac{d+2}{d+4}} \ll n^{-\frac{4}{d+4}}$), so $\min\{\widetilde{f}, 1\}$ is a minimax optimal estimator on all of Ω. (Recall that this is an improper estimator, and we can apply the procedure *MP* described in §4.5 to obtain a proper estimator.) It is not hard to see that the runtime of the above algorithm is $O_d(n^{O(d)})$. The proof of Theorem 4 is complete.

4.1.3 On The Estimator of Lemma 27

The construction of this estimator has a middle step that may be independent of interest. We develop a new estimator for the $L_1(\mathrm{U}(K))$ norm of any convex function $g : K \to \mathbb{R}$:

Lemma 28. *Let K be any convex body in $\mathbb{R}^d$. Let $\delta \in (0,1)$ and $f : K \to \mathbb{R}$ be a convex function, and suppose that $m \geq d+1$ i.i.d. samples are drawn from the regression model $Y = f(Z) + \xi$, where $Z \sim \mathrm{U}(K)$. There exists an estimator $\bar{f}_m^\delta$ taking these samples as input which satisfies, with probability at least $1 - 3\max\{\delta, e^{-cm}\}$,*

$$\|f\|_{L_1(\mathrm{U}(K))} \leq \bar{f}_m^\delta \leq C(d,K)\|f\|_{L_1(\mathrm{U}(K))} + C_1(d,K)\sqrt{\frac{\log(2/\delta)}{m}} \cdot (\|f\|_{L_2(\mathrm{U}(K))} + \sigma),$$

where $C(d,K), C_1(d,K)$ are constants that only depend on the convex set K and the dimension d.

The estimator of Lemma 28 is invariant with respect to affine transformations of the domain. Thus, the constants $C(d,K), C_1(d,K)$ are the same for all K in the class of affine images of a fixed convex body in $\mathbb{R}^d$, such as the class of simplices. Furthermore, it has the optimal error rate, with respect to the number of samples m, for the $L_1(\mathrm{U}(K))$-norm of any convex function g (with no restriction on its uniform norm or Lipschitz constant). However, we cannot extend this estimate for the $L_2(\mathrm{U}(K)$ for any convex K and any convex $g : K \to \mathbb{R}$. Unfortunately, it is essential for both the statistical guarantees and the computational aspects of the proposed estimator of Lemma 27 to assume that the domain of g is a simplex and that $\|g\|_\infty \leq L$.

4.2 All Remaining Proofs

Basic notions regarding polytopes A quick but thorough treatment of the basic theory is given in, e.g. [Sch14]. A set $P \subset \mathbb{R}^d$ is called a polyhedral set if it is the intersection of a finite set of half-spaces, i.e., sets of the form $\{x \in \mathbb{R}^d : x \cdot a \leq c\}$ for some $a \in \mathbb{R}^d$, $c \in \mathbb{R}$. A polyhedral set P is called a polytope if it is bounded and has nonempty interior; equivalently, a set P is a polytope if it is the convex hull of a finite set of points and has nonempty interior.

The affine hull of a set $S \subset \mathbb{R}^d$ is defined as

$$\operatorname{affHull} S = \bigcup_{k=1}^{\infty} \{ \sum_{i=1}^{k} a_i x_i : x_i \in K, a_i \in \mathbb{R} \mid \sum_{i=1}^{k} a_i = 1 \},$$

which is the minimal affine subspace of $\mathbb{R}^d$ containing S. For a convex set K, we define its dimension to be the linear dimension of its affine hull.

For any unit vector u and any convex set K, the support set $F(K, u)$ is defined as

$$F(K, u) = \{ x \in K : x \cdot u = \max_{y \in K} y \cdot u \}.$$

(If $\max_{y \in K} y \cdot u = \infty$ then $F(K, u)$ is defined to be the empty set.)

Suppose P is a polyhedral set. For any $u \in \mathbb{S}^{m-1}$, $F(P, u)$ is a polyhedral set of smaller dimension than K. Any such $F(P, u)$ is called a face of P, and if $F(P, u)$ has dimension $m - 1$, it is called a facet of P. A polyhedral set P which is neither empty nor the whole space $\mathbb{R}^d$ has a finite and nonempty set of facets, and every face of P is the intersection of some subset of the set of facets of P. If P is a polytope, all of its faces, and in particular all of its facets, are bounded. A polytope is called simplicial if all of its facets are $(m - 1)$-dimensional simplices, which is to say, each facet F of P is the convex hull of precisely m points in $\operatorname{affHull} F$.

4.2.1 Proof of Theorem 17

Since the squared L_2-error scales quadratically with the function to be estimated, it suffices to prove the theorem for the class of 1-Lipschitz functions. Since the range of a 1-Lipschitz function on a domain of diameter at most 1 is contained in an interval of length 1, it is no loss to assume that the range of f^* is contained in $[0, 1]$.

The construction of a k-affine approximation to any convex 1-Lipschitz function $f^* : \Omega \to [0, 1]$, uses a combination of two tools: the theory of random polytopes in convex sets, and empirical processes.

Fix a convex body $K \subset \mathbb{R}^d$ and $n \geq d + 1$. The random polytope K_n is defined to be the convex hull of n random points $X_1, \ldots, X_n \sim \mathrm{U}(K)$. It is well-known and easy to justify that K_n is a simplicial polytope with probability 1: Indeed, if $X_1, \dots, X_n$ form a facet of K_n then in particular they lie in the same affine hyperplane, and if $k \geq d+1$, the probability that X_k lies in the affine hull H of $X_1, \dots, X_n$ is 0, since $K \cap H$ has volume 0. For future use we note that with probability 1, the projection of every facet of K_n on the first $d-1$ coordinates is a $(d-1)$-dimensional simplex, by similar reasoning.

For $s \in \{0, 1, \dots, d-1\}$ and P a polytope, we let $f_s(P)$ denote the number of s-dimensional faces of P. The first result regarding random polytopes that we need appears in [Bá89, Corollary 3]:

Theorem 18. *Let $d \geq 1$, $1 \leq s \leq d-1$ and a convex body $K \subset \mathbb{R}^d$. Then, there exists $C(d,s) \leq C_1(d)$ such that*

$$\mathbb{E}[f_s(K_n)] \leq C(d,s) n^{\frac{d-1}{d+1}}.$$

We will also use the following result that was derived in [Dwy88]:

Theorem 19. *Let $P \subset B_d$ be a polytope, and let $Y_1, \dots, Y_m \underset{i.i.d.}{\sim} \mathbb{P}_P$. Then, $P_m = \operatorname{conv}(Y_1, \dots, Y_m)$ is a simplicial polytope with probability* 1*, and the following holds:*

$$\mathbb{E}\mathbb{P}_P(P \setminus P_m) = O_d(C(P) m^{-1} \log(m)^{d-1}),$$

The other result that we need from empirical processes appears as Lemma 24 in the main text.

We now describe our construction. Given a 1-Lipschitz function $f^* : \Omega \to [0,1]$, define the convex body

$$K = \{(x,y) : x \in \Omega, y \in [0,2] \,|\, f^*(x) \leq y\}.$$

In other words, K is the epigraph of the function f^*, intersected with the slab $\mathbb{R}^d \times [0,2]$. Note that $\mathrm{Vol}_{d-1}(\Omega) \leq \mathrm{Vol}_d(K) \leq 2\mathrm{Vol}_{d-1}(\Omega)$, since $\operatorname{Im} f^* \subset [0,1]$.

Let $n = \lfloor k^{\frac{d+2}{d}} \rfloor$, and consider the random polytope $K_n \subset K$. Let Ω_k be the projection of K_n to $\mathbb{R}^d$, and define the function $f_k : \Omega_k \to [0,2]$ by

$$f_k(x) = \min\{y \in \mathbb{R} : (x,y) \in K_n\},$$

i.e., f_k is the lower envelope of K_n. In particular, since $K_n \subset K$, f_k lies above the graph of f^*. We would like to show that with positive probability, f_k satisfies the properties in the statement of the theorem. We treat each property in turn.

f_k is k-simplicial with probability at least $9/10$: Using Theorem 18, and Markov's inequality K_n has at most $10C(d)n^{\frac{d}{d+2}} = C'(d)k$ facets (recall that all facets of K_n are simplices with probability 1). Letting $\triangle_1, \ldots, \triangle_F$ be the bottom facets of K_n, and letting $\pi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ be the projection onto the first factor, $\pi(\triangle_1), \ldots, \pi(\triangle_F)$ is a triangulation of Ω_k and for each $i = 1, \ldots, F$, $f_k|_{\triangle_i}$ is affine, as its graph is simply $\triangle_i$.

Bounding $\mathbb{P}(\Omega \backslash \Omega_k)$ with probability at least $9/10$: Since Ω_k is the projection of K_n to $\mathbb{R}^d$, it is equivalently defined as $\operatorname{conv}(\pi(X_1), \ldots, \pi(X_n))$ where $X_1, \ldots, X_n$ are independently chosen from the uniform distribution on K, and π is the projection onto the first d coordinates as above. $\pi(X_i)$ is not uniformly distributed on Ω, so we cannot apply Theorem 19 (and Markov's inequality directly). Instead, we re-express $\pi(X_i)$ as a mixture of a uniform distribution and another distribution, and apply Theorem 19 to the points which come from the uniform distribution.

In more detail, note that we may write $K = K_1 \cup K_2$ where $K_1 = (\Omega \times [0,1]) \cap \operatorname{epi} f$ and $K_2 = \Omega \times [1,2]$, since $f \leq 1$. Let $p = \frac{\operatorname{Vol}(K_2)}{\operatorname{Vol}(\Omega)} \geq \frac{1}{2}$. The uniform distribution from K can be sampled from as follows: with probability p, sample uniformly from K_2, and with probability $1-p$ sample uniformly from K_1. Clearly, if X is uniformly distributed from K_2 then $\pi(X)$ is uniformly distributed on Ω. Hence, Ω_k can be constructed as follows: draw M from the binomial distribution $B(n,p)$ with n trials and success probability p, then sample M points $X_1, \ldots, X_M$ uniformly from Ω and sample $k - M$ points $X'_1, \ldots, X'_{k-m}$ from some other distribution on Ω, which doesn't interest us; then set $\Omega_k = \operatorname{conv}(X_1, \ldots, X_M, X'_1, \ldots, X'_{k-M})$. In particular, $\mathbb{P}(\Omega \backslash \Omega_k) \geq \mathbb{P}(\Omega \backslash \Omega_M)$, so it is sufficient to bound the RHS with high probability.

By the usual tail bounds on the binomial distribution, $M \geq \frac{np}{2} \geq \frac{n}{4}$ with probability $1 - e^{-\Omega(n)}$. Hence, by Theorem 19 we obtain

$$\begin{aligned}\mathbb{E}\mathbb{P}(\Omega \backslash \Omega_M) &\leq \mathbb{E}\mathbb{P}(\Omega \backslash \operatorname{conv}\{X_1, \ldots, X_{n/4}\}) + C(d)e^{-c(d)n} \leq O_d(C(\Omega)n^{-1}\log(n)^{d-1}) \\ &\leq O_d(C(\Omega)k^{-\frac{d+2}{d}}\log(k)^{d-1}),\end{aligned}$$

and we obtain $\mathbb{P}(\Omega\backslash\Omega_M) \leq 10C(\Omega)k^{-\frac{d+2}{d}}\log(k)^{d-1}$ with probability at least $\frac{9}{10}$ by Markov's inequality.

Bounding $\int(f-f_k)^2 d\mathbb{P}$ with probability at least $9/10$: Finally, we wish to bound the $\mathbb{L}_2(\mathbb{P})$-norm of $f^* - f_k$. To do this, we use the same strategy, arguing that on average, k of the points of K_n can be thought of as drawn from the uniform distribution on a thin shell of width $k^{-\frac{2}{d}}$ lying above the graph of f^*, which automatically bounds the empirical L_2-norm $\int(f^*-f_k)^2\, d\mathbb{P}_n$ and hence the L_2-norm by Lemma 24.

Now for the details. Set $\epsilon = k^{-\frac{2}{d}}$, and define

$$K_\epsilon = \{(x,y) : x \in \mathbb{R}^d, y \in [0,2] \,|\, f^*(x) \leq y \leq f^*(x)+\epsilon\},$$

i.e., $K_\epsilon \subset K$ is just the strip of width ϵ lying above the graph of f^*. By Fubini, K_ϵ has volume $\epsilon\text{Vol}(\Omega) \geq \frac{\epsilon}{2\text{Vol}(K)}$, and if X is uniformly distributed on K_ϵ, $\pi(X)$ is uniformly distributed on Ω. Hence, we can argue precisely as in the preceding: with probability $1-e^{-\Omega(n)}$,

$$L := |\{X_i : X_i \in K_\epsilon\}| \geq \frac{\epsilon n}{4} = \frac{k}{4}.$$

Conditioning on L for some $L \geq \frac{k}{4}$ and letting $X_1,\ldots,X_L$ be the points drawn from K which lie in K_ϵ we have that $\pi(X_1),\ldots,\pi(X_L)$ are uniformly distributed on Ω. Moreover, for any $i \in \{1,\ldots,n\}$, $X_i \in K_n$ and so it lies above the graph of f_k, but also $X_i \in K_\epsilon$ and so it lies below the graph of $f^*+\epsilon$. Combining these two facts yields

$$\forall 1 \leq i \leq L: \quad f_k(\pi(X_i)) \leq (X_i)_{d+1} \leq f^*(\pi(X_i))+\epsilon,$$

where $(\cdot)_{d+1}$ denotes the $d+1$ coordinate. Hence,

$$\forall 1 \leq i \leq L: \quad f^*(\pi(X_i)) \leq f_k(\pi(X_i)) \leq f^*(\pi(X_i)) + Ck^{-2/d}. \tag{4.19}$$

Thus, letting $\mathbb{P}_L = \frac{1}{L}\sum_{i=1}^{L}\delta_{\pi(X_i)}$ denote the empirical measure on $\pi(X_1),\ldots,\pi(X_L)$, we obtain

$$\int_{\Omega}(f^* - f_k)^2\,d\mathbb{P}_L \le \frac{1}{L}\sum_{i=1}^{L}\epsilon^2 = \epsilon^2 = k^{-\frac{4}{d}}.$$

Since the $\pi(X_i)$ are drawn uniformly from Ω, if we knew that f_k were 1-Lipschitz it would follow from Lemma 24 that

$$\int_{\Omega}(f^* - f_k)^2\,d\mathbb{P} \le k^{-\frac{4}{d}} + CL^{-\frac{4}{d}} = C'k^{-\frac{4}{d}},$$

with high probability.

We do not know, however, that f_k is 1-Lipschitz. To get around this, define the function $\hat{f}_k$ as the function on Ω_k whose graph is $\operatorname{conv}\{(\Pi(X_i), f^*(\Pi(X_i)))\}_{i=1}^{L}$. Unlike f_k, $\hat{f}_k$ is necessarily 1-Lipschitz since f^* is (see, e.g., the argument in the paragraph below equations (4.11)-(4.13)), so by Lemma 24, it follows that

$$\int_{\Omega}(f^* - \hat{f}_k)^2\,d\mathbb{P} \le C_1 k^{-\frac{4}{d}}$$

with probability at least $1 - C(d)\exp(-c(d)k)$. Also, by (4.19),

$$\forall 1 \le i \le L: \quad \hat{f}_k(\pi(X_i)) \le f_k(\pi(X_i)) \le \hat{f}_k(\pi(X_i)) + C_1 k^{-2/d}.$$

It easily follows by the definitions of f_k and $\hat{f}_k$ as convex hulls that on the domain $\Omega_{\Pi(X)} := \operatorname{conv}\{(\Pi(X_i))\}_{i=1}^{L}$, we have

$$f^* \le \hat{f}_k \le f_k \le \hat{f}_k + Ck^{-2/d}.$$

Hence, we conclude that

$$\begin{aligned}\int_{\Omega_{\Pi(X)}}(f_k - f^*)^2\,d\mathbb{P} &\le 2\int_{\Omega_{\Pi(X)}}(\hat{f}_k - f^*)^2\,d\mathbb{P} + 2\int_{\Omega_{\Pi(X)}}(f_k - \hat{f}_k)^2\,d\mathbb{P} \\ &\le 2\int_{\Omega_{\Pi(X)}}(\hat{f}_k - f^*)^2\,d\mathbb{P} + 2\|f_k - \hat{f}_k\|^2_{L^\infty(\Pi(X))} \le C_2 k^{-4/d}\end{aligned}$$

with high probability.

Now, using Theorem 19 and Markov's inequality, we also know that

$$\mathbb{P}(\Omega \setminus \Omega_{\Pi(X)}) \leq 20C(\Omega)k^{-1}\log(k)^{d-1}.$$

with probability at least $\frac{19}{20}$. Conditioned on this event, and using the fact that f_k is uniformly bounded by 1, we obtain

$$\int_{\Omega \setminus \Omega_{\Pi(X)}} (f_k - f^*)^2 d\mathrm{U}(x) \leq C(\Omega)k^{-1}\log(k)^{d-1}.$$

On the intersection of the two events defined above, which has probability at least $\frac{9}{10}$, we have $\int_\Omega (f_k - f^*)^2 \leq C_3 k^{-\frac{4}{d}} + C(\Omega)k^{-1}\log(k)^{d-1}$.

Deriving the theorem Since we have three events each of which hold with probability at least $9/10$, then the intersection of these events is not empty. Therefore, an f_k satisfying all the desired properties exists, and the theorem follows.

4.2.2 Proof of Lemma 26

We start with the following easy lemma:

Lemma 29. *The following event holds with probability at least* $1 - n^{-3d}$*:*

$$\forall 1 \leq i \leq k(n) \text{ s.t. } \mathbb{P}(\triangle_i) \geq C_3 d\log(n)/n: \; 2^{-1}\mathbb{P}(\triangle_i) \leq \mathbb{P}_n(\triangle_i) \leq 2\mathbb{P}(\triangle_i). \tag{4.20}$$

Proof. The lemma follows for the fact that $n \cdot \mathbb{P}_n(S) \sim Bin(n, \mathbb{P}(S))$, along with the concentration inequality (cf. [BLM13]) for binomial random variables: for all $\epsilon \in (0,1)$,

$$\Pr\left(\left|\frac{\mathbb{P}_n(S)}{\mathbb{P}(S)} - 1\right| \leq \epsilon\right) \leq 2\exp(-c\min\{\mathbb{P}(S), 1 - \mathbb{P}(S)\}n\epsilon^2).$$

By taking $\epsilon = 1/2$, and choosing C to be large enough, we conclude that for any particular $\triangle_i$,

$$\mathbb{P}(\triangle_i) \geq C_3 d\log(n)/n: \; 2^{-1}\mathbb{P}(\triangle_i) \leq \mathbb{P}_n(\triangle_i) \leq 2\mathbb{P}(\triangle_i)$$

with probability at least $1 - n^{-(3d+1)}$. Taking the union bound over all $k(n)$ simplices, the claim follows. □

The main step is the following lemma, which shows that for any given simplex $\triangle_i$, if we draw $Cd \log n$ points from the uniform distribution on $\triangle_i$ for sufficiently large C, then there exists some subset S of these points whose convex hull P covers almost all of the simplex and can also be triangulated by a polylogarithmic number of simplices whose vertices lie in S.

Lemma 30. *Let $S \subset \mathbb{R}^d$ be a simplex, and $m \geq C_3 d \log(n)$, for some large enough $C_3 \geq 0$. Let $Y_1, \ldots, Y_m \sim \mathbb{P}_S$. Then, with probability at least $1 - n^{-3d}$ there exists a set $\mathcal{A}$ of simplices contained in S with disjoint interiors of cardinality $|\mathcal{A}| \leq C_d \log(m)^{d-1}$ such that*

$$\mathbb{P}_S(S \setminus \bigcup \mathcal{A}) = O_d(m^{-1} \log(n) \log(m)^{d-1}).$$

Proof. For each $s \in \{0, 1, \ldots, d-1\}$ and P a polytope, we let $f_s(P)$ denote the number of s-dimensional faces of P. We need the following result, which was first proven in [Dwy88]; for more details see the recent paper [RSW19].

Theorem 20. *[[Dwy88]] Let $S \subset \mathbb{R}^d$ be a simplex, and let $Y_1, \ldots, Y_m \sim \mathbb{P}_S$. Then, $S_m = \mathrm{conv}(Y_1, \ldots, Y_m)$ is a simplicial polytope with probability 1, and the following holds:*

$$\mathbb{E}\mathbb{P}_S(S \setminus S_m) = O_d(m^{-1} \log(m)^{d-1}),$$

and

$$\mathbb{E} f_{d-1}(S_m) = O_d(\log(m)^{d-1}).$$

This theorem does not give us what we need directly, since it treats only expectation while we require high-probability bounds. (To the best of our knowledge, sub-Gaussian concentration bounds are not known for the random variables $f_{d-1}(P_m), \mathbb{P}(S \setminus S_m)$ when S is a simplex, cf. [Vu05].) This necessitates using a partitioning strategy. We divide our $Y_1, \ldots, Y_m$ into $B := C_1 d \log(n)$ blocks, for C_1 to be chosen later, each with $m(n) := \frac{m}{C_1 d \log(n)}$ samples drawn uniformly from

$\triangle$. Let $P_1, \ldots, P_B$ be the convex hulls of the points in each block, each of which are independent realizations of the random polytope $S_{m(n)}$. For each P_i, Markov's inequality and a union bound yield that with probability at least $\frac{1}{3}$,

$$\begin{aligned}\mathbb{P}_S(S \setminus P_i) &\leq 3 \cdot \mathbb{E}\mathbb{P}_S(S \setminus P_i) \leq C_1(d) m(n)^{-1} \log(m(n))^{d-1} \\ &= \frac{C_2(d) \log(m)^{d-1} \log(n)}{n},\end{aligned} \tag{4.21}$$

and

$$f_{d-1}(P_i) \leq 3\mathbb{E} f_{d-1}(S_{m(n)}) \leq O_d(\log(m(n))^{d-1}) \leq O_d(\log(m)^{d-1}). \tag{4.22}$$

Since there are $C_1 d \log(n)$ independent P_i, at least one of them will satisfy these conditions with probability $1 - \left(\frac{2}{3}\right)^{C_1 d \log n}$, and we may choose C_1 so that this is at least $1 - n^{-3d}$.

Conditioned on the existence of P_i satisfying (4.21) and (4.22), we take one such P_i and triangulate it by picking any point among the original $Y_1, \ldots, Y_m$ lying in the interior of P_i and connecting it to each of the $(d-1)$-simplices making up the boundary of P_i. The set $\mathcal{A}$ is simply the set of d-simplices in this triangulation. □

Now, to obtain Lemma 26, we condition on the event of Lemma 29 and apply Lemma 30 to each $\triangle_i$ such that $\mathbb{P}(\triangle_i) \geq Cd \log(n)/n$, with the $Y_1, \ldots, Y_m$ taken to be the points of $X_{n+1}, \ldots, X_{2n}$ drawn from $\mathbb{P}$ which fall inside of $\triangle_i$. Using the fact that $\mathbb{P}_n(\triangle_i) \geq 0.5\mathbb{P}(\triangle_i)$, we see that $m \geq Cd \log n$ for each $\triangle_i$, so Lemma 30 is in fact applicable. In addition, the bounds on the cardinality of $\mathcal{S}_X^i$ and on the volume of $\triangle_i$ left uncovered by the simplices in $\mathcal{S}_X^i$ follow immediately by substituting $c\mathbb{P}(\triangle_i)$ for m in the conclusions of Lemma 30. For i such that $\mathbb{P}(\triangle_i) \leq Cd \log(n)/n$, we take $\mathcal{S}_X^i$ to be the empty set.

4.2.3 Proofs of Lemmas and 28 and 27

In several places, we will use a high-probability estimator for the mean of a random variable presented in [Dev+16]:

Lemma 31. *Let $\delta \in (0,1)$ and let $Z_1, \ldots, Z_k$ be i.i.d. samples from a distribution on $\mathbb{R}$ with finite variance σ_Z^2. There exists an estimator $\hat{f}_\delta : \mathbb{R}^k \to \mathbb{R}$ with a runtime of $O(k)$, such that with probability at least $1-\delta$,*

$$(\hat{f}_\delta(Z_1, \ldots, Z_k) - \mathbb{E}Z)^2 \leq \frac{8\sigma_Z^2 \cdot \log(2/\delta)}{k}.$$

In the first sub-subsection, we construct the estimator for the L_1 norm of a convex function g defined on a convex body K, which is the content of Lemma 28. In the second sub-subsection, we show how to "upgrade" this estimator to an estimator of the L_2-norm in the special case that K is a simplex.

The final step will be to estimate the L_2 norm of g under the assumptions of Lemma 27, by using Lemma 28 and the claim of Lemma 27 will follow.

Proof of Lemma 28

We only prove this this for K a simplex, for a general K can be done in similar fashion, by placing K in John's position (besides the computational aspects).

Let S be the regular simplex inscribed in the unit ball B_d, and for each $t \in [0,1]$, denote by $S^t := (1-t)S$. We will use the following geometric facts, which can be extracted from the statements and proofs of [GW17, Lemmas 2.6-2.7].

Lemma 32. *Let $g : S \to \mathbb{R}$ be a convex function, and let $\|g\|_1 := \int_S |g| d\mathrm{U}$. Then,*

- *$g(x) \geq -C_d\|g\|_1$ on every $x \in S$.*
- *For each $\delta \in (0,1)$, $g|_{S^\delta}$ is is $C_d\delta^{-(d+1)}\|g\|_1$-Lipschitz and satisfies $g|_{S^\delta} \leq C_d\delta^{-(d+1)}\|g\|_1$.*

We immediately obtain the following corollary:

Corollary 1. *For any $a \in (0,1)$, there exists a constant δ such that g restricted to S^δ is uniformly bounded by $C_d s^{-(d+1)}\|g\|_1$; moreover, letting $g_- = \min(g, 0)$, we have*

$$\int_{S \setminus S^\delta} |g_-| \, d\mathrm{U}(x) \leq a\|g\|_1.$$

Define a probability density p_S on S by the formula

$$p_S(x) := \int_S \frac{1_{B_y}(x)}{\mathrm{U}(B_y)} d\mathrm{U}(y), \tag{4.23}$$

where we define B_y to be the largest ball centered on y which is contained in S. For any simplex $\triangle$, we define the density $p_\triangle$ as the pushforward of p_S under the affine transformation T sending S to T.

Lemma 33. *Let M, M' be positive constants, let $\triangle$ be a simplex contained in the unit ball B_d, and let $g : \triangle \to [-M'\|g\|_{L_1(\mathrm{U}(\triangle))}, M'\|g\|_{L_1(\mathrm{U}(\triangle))}]$ be a convex $M\|g\|_{L_1(\mathrm{U}(\triangle))}$-Lipschitz function which is orthogonal to affine functions, i.e.,*

$$\int_\triangle gw \, d\mathrm{U} = 0$$

for any affine function w. Then there exist positive constants $c_1 = c_1(M, M', d)$, $C_1 = C_1(M, M', d)$ such that:

$$c_1\|g\|_{L_1(\mathrm{U}(\triangle))} \le \int g(x)p_\triangle(x)\, dx \le C_1\|g\|_{L_1(\mathrm{U}(\triangle))}, \tag{4.24}$$

In addition, for every affine function w,

$$\int wp_\triangle \, dx = \int_\triangle w \, d\mathrm{U}(x).$$

Moreover,

$$\max_{x\in\triangle} p_\triangle(x) \le \alpha_d := 2^d \frac{d+1}{d-1}\frac{v_{d-1}}{v_d}$$

where v_d is the volume of the unit ball in dimension d, and there exists an efficient algorithm to compute $p_\triangle(x)$ for any $x \in \triangle$.

The idea behind the proof of this lemma is that for any point $x \in \triangle$ and a ball $B_x \subset \triangle$, the average $\bar{g}(x)$ of g over B_x is at least $g(x)$, with equality iff g is affine on B_x. If g is nonzero and orthogonal to affine functions, the averaged function $\bar{g}$ must have positive integral, and a compactness argument then yields a lower bound on $\frac{1}{\|g\|_1}\int \bar{g}$. The full proof is given at the end of the sub-subsection.

Using the above results we can estimate the L_1 norm $g : \triangle \to [-1, 1]$.

Letting T be the unique affine transformation such that $T\triangle = S$, we define the shrunken simplex $\triangle_\delta$ by the $\triangle_\delta := T^{-1}(T\triangle)^{\delta(l,d)}$, where $a = 1/10$ and $\delta(a, d)$ is defined in Corollary 1. The proof involves analysis of several cases.

Case 1: $\|g\mathbb{1}_{\triangle\setminus\triangle_\delta}\|_1 \geq 2^{-1}\|g\|_1$, i.e. most of the L_1-norm of g comes from the shell $\triangle \setminus \triangle_\delta$. Using Corollary 1, we know that

$$\|g\|_1 \geq \int_{\triangle\setminus\triangle_\delta} g d\mathrm{U}(x) = \int_{\triangle\setminus\triangle_\delta} g^+ d\mathrm{U}(x) + \int_{\triangle\setminus\triangle_\delta} g^- d\mathrm{U}(x) \geq (3/20) \cdot \|g\|_1.$$

Therefore it is enough to estimate the mean (scaled by $\mathrm{U}(\triangle \setminus \triangle_\delta)$) of the r.v. $g(X)$ where $X \sim \mathrm{U}(\triangle \setminus \triangle_\delta)$, which can be done using the samples that fall in $\triangle \setminus \triangle_\delta$. By Lemma 31 above, we conclude that using these samples we have an estimator $\hat{f}_{(1)}$ such that with probability at least $1 - \delta$,

$$\begin{aligned}\left|\hat{f}_{(1)} - \int_{\triangle\setminus\triangle_\delta} g\, d\mathrm{U}\right|^2 &\leq \mathrm{U}(\triangle \setminus \triangle_\delta)^2 \frac{Cd(\sigma^2 + \|g\mathbb{1}_{\triangle\setminus\triangle_\delta}\|_2^2)\log(2/\delta)}{\mathrm{U}(\triangle \setminus \triangle_\delta)m} \\ &\leq C_d \cdot \frac{(\sigma^2 + \|g\|_2^2)\log(2/\delta)}{m}, \qquad (4.25)\end{aligned}$$

where we used that fact that $\mathrm{U}(\triangle \setminus \triangle_\delta) \geq c_d$.

Case 2: If we are not in Case 1, we must have $\|g\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{2}\|g\|_1$, i.e., most of the L^1-norm of g comes from the inner simplex $\triangle_\delta$. Decompose $g = w_g + (g - w_g)$, where $w_g = \mathrm{argmin}_{w \text{ affine}} \|g - w\|_{L_2(\mathrm{U}(\triangle_\delta))}$ is the $L_2(\mathrm{U}(\triangle_\delta))$-projection of g onto the space of affine functions. Note that by orthogonality, we have

$$\max(\|w_g\|_{L_2(\mathrm{U}(\triangle_\delta))}, \|g - w_g\|_{L_2(\mathrm{U}(\triangle_\delta))}) \leq \|g\|_{L_2(\mathrm{U}(\triangle_\delta))} \qquad (4.26)$$

by orthogonality, while by Lemma 32, we have

$$\|g\|_{L_1(\mathrm{U}(\triangle_\delta))} \leq \|g\|_{L_2(\mathrm{U}(\triangle_\delta))}\| \leq C_d \|g\|_{L_1(\mathrm{U}(\triangle_\delta))}. \tag{4.27}$$

By the triangle inequality, we must have either $\|w_g \mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$ or $\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$; we analyze each case below.

Case 2a: First, suppose $\|w_g \mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$. Using half of the samples that fall into $\triangle_\delta$, we may apply Lemma 25, giving us an affine $\hat{w}_g$ such that with probability at least $1 - \delta$,

$$\|\hat{w}_g - w_g\|_2^2 \leq C_d(\sigma^2 + \|g\|_2^2)\frac{\log(2/\delta)}{m}.$$

Writing $\bar{f}_{(2a)} := \|\hat{w}_g\|_1$, we conclude that

$$\frac{1}{4}\|g\|_1 - C_d(\sigma + \|g\|_2)\sqrt{\frac{\log(2/\delta)}{m}} \leq \bar{f}_{(2a)} \leq \|g\|_1 + C_d(\sigma + \|g\|_2)\sqrt{\frac{\log(2/\delta)}{m}}. \tag{4.28}$$

Note that the right-hand inequality does not require the assumption $\|w_g \mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$.

Case 2b: Now suppose $\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$. To estimate $\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1$, we will use our Lemma 33. Note that by the definition of $\triangle_\delta$, we may assume by Lemma 1 that $\max\{M', M\} \leq C(d)$ (in Lemma 33). Therefore, we conclude that

$$c_1(d)\|g - w_g\|_1 \leq \int_{\triangle_\delta} (g - w_g) \cdot p_{\triangle_\delta} \leq C_1(d)\|g - w_g\|_1.$$

By our assumption $\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$; we also have

$$\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1 \leq \|g\|_1 + \|\hat{w}_g\|_1 \leq \|g\|_1 + \|\hat{w}_g - w$$

so it follows that

$$c_2(d)\|g\|_1 \leq \int_{\triangle_\delta} (g - \hat{w}_g) \cdot p_{\triangle_\delta} \leq C_2(d)\|g\|_1.$$

Therefore, it is enough to estimate $\int (g - \hat{w}_g) \cdot p_{\triangle_\delta}$, using the the second half of the samples that fall into $\triangle_\delta$.

To do this, we simulate sampling from $p_{\triangle_\delta}$ given samples from $\mathrm{U}(\triangle_\delta)$ and their corresponding noisy samples of $g - \hat{w}_g$. The idea is simply to use rejection sampling [Dev86]: given a single sample $X \sim \mathrm{U}(\triangle_\delta)$, we keep it with probability $p_{\triangle_\delta}(x) \cdot \frac{1}{\alpha_d}$. Conditioned on keeping the sample, X is distributed according to $p_{\triangle_\delta}$. If we are given $m/2$ i.i.d. samples from $\mathrm{U}(\triangle_\delta)$, then with probability $1 - e^{-c_d' m}$ the random number of samples N we obtain from $p_\triangle$ by this method is at least $c(\alpha_d) \cdot m \geq c_1(d)m$, and conditioned on N, these samples are i.i.d. $p_{\triangle_\delta}$. We condition on this event going forward. Now, using these N samples, Lemma 31 gives an estimator $\bar{f}_{(2b)}$ such that

$$\left| \bar{f}_{(2b)} - \int (g - \hat{w}_g) \cdot p_{\triangle_\delta} \right| \leq C_d(\sigma + \|g\|_2) \sqrt{\frac{\log(2/\delta)}{m}}. \tag{4.29}$$

with probability at least $1 - \max\{\delta, e^{-cm}\}$.

Note that each of $\bar{f}_{(1)}, \bar{f}_{(2a)}, \bar{f}_{(2b)}$ is bounded from above by $C_d\|g\|_1 + C_d(\sigma + \|g\|_2)\sqrt{\frac{\log(2/\delta)}{m}}$, irrespective of whether we are in the case for which the estimator was designed; this follows from (4.25), (4.28), (4.29), respectively.

Finally, by using $\bar{f}_{(1)}, \bar{f}_{(2a)}, \bar{f}_{(2b)}$, we conclude that with probability $1 - 3\max\{\delta, e^{-cm}\}$ at least

$$c_1(d)\|g\|_1 - C_d(\sigma + \|g\|_2)\sqrt{\frac{\log(2/\delta)}{m}} \leq \bar{f}_{(1)} + \bar{f}_{(2a)} + \bar{f}_{(2b)} \leq C_1(d)\|g\|_1 + C_d|\sigma + \|g\|_2|\sqrt{\frac{\log(2/\delta)}{m}},$$

and the claim follows.

Proof of Lemma 33. Recall that it suffices to show that

$$c_1(M, M') \leq \int_S g(x) p_S(x) dx \leq C_1(M, M').$$

Note that the function g is convex and in particular subharmonic, i.e., for any ball B_x with center x contained in S we have

$$\frac{1}{\mathrm{U}(B_x)} \int_{B_x} g d\mathrm{U}(x) \geq g(x),$$

where U denotes the uniform measure on the regular simplex S. g is non-affine and hence strictly subharmonic (as convex harmonic functions are affine), so there exists some x such that for any ball $B_x \subset S$ centered on x, the above inequality is strict, since subharmonicity is a local property. As g is convex and in particular continuous, the inequality is strict on some open set of positive measure. We obtain that for a *non*-affine convex function that

$$\begin{aligned}\int_S g(x)p_S(x)\, d\mathrm{U}(x) &= \int_S \int_S g(x)1_{B_y}(x)\, d\mathrm{U}(y)\, d\mathrm{U}(x) \\ &= \int_S \left(\frac{1}{\mathrm{U}(B_y)} \int_{B_y} g(x)\, dx\right) d\mathrm{U}(y) > \int_S g(y) d\mathrm{U}(y) = 0,\end{aligned} \tag{4.30}$$

i.e. we showed that for a non-harmonic g that $\int_S g(x)p_S(x)\, dx > 0$.

Now, we show why Eq. (4.30) actually implies the lower bound of Eq. (4.24), which is certainly not obvious a priori. However, it follows from a standard compactness argument. The set $\mathcal{C}$ of convex M-bounded, M'-Lipschitz functions with norm 1 that is orthogonal to the affine functions is closed in $L^\infty(S)$, and also equicontinuous due to the Lipschitz condition. Hence, by the Arzela-Ascoli theorem it is compact in $L^\infty(S)$, and we conclude that

$$A = \left\{ \int_S g(x)p_S(x)dx : g \in \mathcal{C} \right\}$$

is compact; but (4.30) implies that $A \subset (0, \infty)$, which finally implies the existence of $c(M, M', d) > 0$ such that $S \subset [c(M, M', d), \infty)$. As for the upper bound in (4.24), it follows immediately from the boundedness of p_S, which we prove below.

We claim that in this case (4.23) can be evaluated analytically as a function of x, though the formulas are sufficiently complicated that this is best left to a computer algebra system. Indeed, we note that $y \in S$ contributes to the integral at x if and only if x is closer to y than y is to the boundary of S. The regular simplex can be divided into $d+1$ congruent cells $C_1, \ldots, C_{d+1}$ such that the points in C_i are closer to the i-th facet of the simplex than to any other facet (in fact, C_i is simply the convex hull of the barycenter of S and the ith facet); for any $y \in C_i$, $x \in B_y$ if and only if x is closer to y than y is to the hyperplane H_i containing C_i. But the locus of points equidistant from a fixed point x and a hyperplane is the higher-dimensional analog of an elliptic paraboloid, for which it's easy to write down an explicit equation. Letting $P_{i,x}$ be the set of points on x's side of the paraboloid (namely, those closer to x than to H_i), we obtain

$$p_S(x) = \sum_{i=1}^{d+1} \int_{C_i \cap P_{i,x}} \frac{dy}{v_d \cdot d(y, H_i)^d}.$$

Each region of integration $C_i \cap P_{i,x}$ is defined by several linear inequalities and a single quadratic inequality, and the integrand can be written simply as $\frac{1}{y_i^d}$ in an appropriate coordinate system. It is thus clear that the integral can be evaluated analytically, as claimed.

Finally, we need to show that $p_S(x)$ is bounded above by α_d. By symmetry,

$$p_S(x) \leq \sum_{i=1}^{d+1} \frac{dy}{\Omega_d \cdot d(y, H_i)^d} \leq \frac{d+1}{\Omega_d} \cdot \sup_{x \in \mathbb{R}^n} \int_{P_{1,x}} \frac{dy}{d(y, H_1)^d}. \tag{4.31}$$

Fix x, and choose coordinates such that $H_1 = \{x_1 = 0\}$ and $x = (x_0, 0, \ldots, 0)$ with $x_0 > 0$. Then for any $y = (t, z)$ with $t \in \mathbb{R}, z \in \mathbb{R}^{d-1}$, y lies in $P_{1,x}$ if $t^2 \geq (x_0 - w)^2 + |z|^2$, or $2tx_0 - x_0^2 \geq |z|^2$. Hence,

$$\begin{aligned}
\int_{P_{1,x}} \frac{dy}{d(y, H_1)^d} &= \int_{\frac{x_0}{2}}^{\infty} \frac{dt}{t^d} \int_{\mathbb{R}^{d-1}} 1_{|z|^2 \leq 2tx_0 - x_0^2}(z)\, dz \\
&\leq \int_{\frac{x_0}{2}}^{\infty} \frac{dt}{t^d} \cdot \Omega_{d-1} (2tx_0 - x_0^2)^{\frac{d-1}{2}} \leq \Omega_{d-1} \int_{\frac{x_0}{2}}^{\infty} \frac{dt}{t^d} (2tx_0)^{\frac{d-1}{2}} \\
&= \Omega_{d-1} 2^{\frac{d-1}{2}} x_0^{\frac{d-1}{2}} \int_{\frac{x_0}{2}}^{\infty} \frac{dt}{t^{\frac{d+1}{2}}} \\
&= \Omega_{d-1} 2^{\frac{d-1}{2}} x_0^{\frac{d-1}{2}} \cdot \left(\frac{d-1}{2} \right)^{-1} \left(\frac{x_0}{2} \right)^{-\frac{d-1}{2}} = \Omega_{d-1} \cdot \frac{2^d}{d-1},
\end{aligned}$$

and substituting in (4.31) gives the desired bound. □

Proof of Lemma 27

Recall that we are given a convex L-Lipschitz function g satisfying $\|g\|_\infty \leq L$; by homogeneity, we may assume $L = 1$. Our goal in this subsection is to estimate $\|g\|_2$ up to polylogarithmic factors given an estimate of $\|g\|_1$, where $g : \triangle \to [-1, 1]$. This part requires the additional assumption that $\|g\|_2 \geq \frac{Cd^{1/2}\log n}{n^{1/2}}$.

For this section, we will need the following classical result about the floating body of a simplex [BL88; SW90].

Lemma 34. *For a simplex S and $\epsilon \in (0, 1)$. let S_ϵ be its ϵ-convex floating body, defined as*

$$S_\epsilon := \bigcap\{K : K \subset S \text{ convex }, \mathrm{Vol}(S \setminus K) \leq \epsilon \mathrm{Vol}(S)\},$$

and let $S(\epsilon) = S \backslash S_\epsilon$ be the so-called wet part of S. Then $\mathrm{Vol}(S(\epsilon)) \leq C_d \epsilon \log(\epsilon^{-1})^{d-1}\mathrm{Vol}(S)$.

We also note that for any particular ϵ and $x \in S$ one can check in polynomial time whether $x \in S(\epsilon)$: indeed, letting

$$H^+_{x,u} = \{y \in \mathbb{R}^n : \langle y, u\rangle \geq \langle x, u\rangle\}$$
$$H_{x,u} = \partial H^+_{x,u} = \{y \in \mathbb{R}^n : \langle y, u\rangle = \langle x, u\rangle\},$$

the function $u \mapsto \mathrm{U}(S \cap H^+_{x,u})$ is smooth on S^{d-1} outside of the closed, lower-dimensional subset A where $H_{x,u}$ is not in general position with respect to some face of u, and, moreover, is given by an analytic expression in each of the connected components C_i of $S^{d-1}\backslash A$. It can thus be determined algorithmically whether $\min_i \inf_{x \in C_i} \mathrm{U}(S \cap H^+_{x,u}) \leq \epsilon$, i.e., whether $x \in S(\epsilon)$.

Let $v = \mathbb{P}(S)$, and let $i_{min} = \min(\lfloor \log_2(C_d \log(n)^{d+1} v^{-1})\rfloor, 0)$; note that since $\|g\|_1 \geq v \geq \frac{\log(n)^d}{n}$, $|i_{min}| \leq C \log n$. Set $V = g^{-1}((-\infty, 2^{i_{min}}])$, and for $i = i_{min}, i_{min} + 1, \ldots, 0$, set $U_i = g^{-1}((2^i, 1])$. Note that V is convex, while each U_i, $i \geq 0$, is the complement of a convex subset of S.

We will use the following lemma:

Lemma 35. *For g and V, U_i as defined above, at least one of the following alternatives holds:*

1. $c_d \log(n)^{-d+1/2}\|g\|_2 \leq v^{-\frac{1}{2}}\|g\|_1 \leq \|g\|_2$.

2. *There exists* $i_0 \in [i_{min}, 0]$ *such that* $2^{-i_0} \geq C_d(\log n)^{d-1}\frac{\|g\|_1}{\mathbb{P}(S)}$ *and*

$$c\log(n)^{-1/2}\|g\|_2 \leq \mathbb{P}(U_{i_0})^{-1/2}\int_{U_{i_0}} g\, d\mathbb{P}. \tag{4.32}$$

The proof of this lemma appears at the end of this subsection.

If alternative (1) of the lemma holds, the L_1-norm of g is only a polylogarithmic factor away from the L_2-norm (up to normalizing by the measure of S, which is known to us). Therefore, we may use the L_1-estimator of the previous subsection and estimate the L_2 norm of g, up to a larger polylogarithmic factor, as we will see below.

We must therefore consider what happens when alternative (2) of Lemma 35 holds. If we could estimate the integral of g over U_{i_0}, we'd be done, but neither the index i_0 nor the set U_{i_0} are given to us. So we make use of the fact that each such U_{i_0}, being the complement of a convex subset of $\triangle$, is contained in the wet part $S(\mathbb{P}(U_i))$, which has volume at most $C_d\log(n)^{d-1}\mathbb{P}(U_i)$ by Lemma 34. We will show in the next lemma that this replacement costs us a $C_d\log(n)^{d/2}$ factor in the worst case.

More precisely, let $\epsilon_j = 2^{-2j}$, and let $S(\epsilon_j)$ be the corresponding wet part of S, as defined in Lemma 34. Then we have the following:

Lemma 36. *With* g *as above, we have*

$$\max_{j\in[i_{min},0]} \mathbb{P}(S(\epsilon_j))^{-1/2}\int_{S(\epsilon_j)} gd\mathbb{P} \leq \|g\|_2,$$

and moreover, if alternative (2) of Lemma 34 holds, then there exists j *such that* $\mathbb{P}(S(\epsilon_j)) \geq \frac{Cd\log n}{n}$ *and*

$$c_d\log(n)^{-d/2}\|g\|_2 \leq \mathbb{P}(S(\epsilon_j))^{-1/2}\int_{S(\epsilon_j)} gd\mathbb{P}.$$

Using the last lemma, we can construct an estimator for $\|g\|_2$ that is at most a polylogarithmic factor away from the true value, whether we are in case (1) or case (2) of Lemma 35.

Indeed, note that we if alternative (2) holds, we have $\mathbb{P}(S_{\epsilon_j}) \geq \frac{Cd\log n}{n}$ and hence, as in Section 2, we can assume by a union bound that the number of sample points falling in $S(\epsilon_j)$ is proportional to $\mathbb{P}(S(\epsilon_j))$. Hence, for each j the estimation of $\int_{S(\epsilon_j)} g \, d\mathbb{P}_{S(\epsilon_j)} = \mathbb{P}(S(\epsilon_j))^{-1} \int_{S(\epsilon_j)} g \, d\mathbb{P}$ can be done in a similar fashion as in §4.2.3, with an additive deviation that is proportional to $\sqrt{\frac{\log(2/\delta)}{n\mathbb{P}(S(\epsilon_j))}}$. However, since we need to estimate $\mathbb{P}(S(\epsilon_j))^{-\frac{1}{2}} \int_{S(\epsilon_j)} g \, d\mathbb{P}$, we can multiply it by $\sqrt{\mathbb{P}(S(\epsilon_j))}$, and obtain the correct deviation of $O(\sqrt{\log(2/\delta)/(\mathbb{P}(S)n)})$ (as usual, we work on the event $\mathbb{P}(S(\epsilon_j))/2 \le \mathbb{P}_n(S(\epsilon_j))$, which holds with probability at least $1 - n^{-2d}$).

We conclude that the maximum over $j \in [-c\log(n) \le i_{min}, 0]$, estimators of the means of the random variables $\mathbb{P}(S(\epsilon_j))^{\frac{1}{2}} \int_{S(\epsilon_j)} g \, d\mathbb{P}_{S(\epsilon\triangle_i)}$ and the L_1 estimator of the above sub-sub section give the claim.

Proof of Lemma 35. Let $g_- = \min(g, 0)$, $g_+ = \max(g, 0)$, so that $\|g\|_2^2 = \|g_-\|^2 + \|g_+\|^2$.

We claim that alternative (1) holds if $\|g_-\|_2^2 \ge \frac{1}{2}\|g\|_2^2$. Indeed, by Lemma 32, $g_- \ge -C_d v^{-1}\|g\|_1$, which immediately yields

$$\int_S g_-^2 \, d\mathbb{P} \le -C_d v^{-1}\|g\|_1 \cdot \int g_- \, d\mathbb{P} \le -C_d v^{-1}\|g\|_1^2$$

i.e.,

$$v^{-1/2}\|g\|_1 \ge c_d\|g_-\|_2 \ge c_d'\|g\|_2.$$

Note that by Jensen's inequality we have that $v^{-1/2}\|g\|_1 \le \|g\|_2$. Otherwise, we have $\|g_+\|^2 \ge \frac{1}{2}\|g\|_2^2$. Let $T_i = U_i \backslash U_{i+1} = g^{-1}((2^i, 2^{i+1}])$. We have

$$\frac{1}{2}\|g\|_2^2 \le \|g_+\|_2^2 \le \sum_{i=-\infty}^{0} 2^{2(i+2))}\mathbb{P}(T_i).$$

By our assumption $\|g\|_2^2 \ge C\frac{\log n}{n}$ and the fact that $v \le 1$, the terms in the sum with $i \le i_{min} = \log\left(C\frac{\log n}{n}\right) + 2$ cannot contribute more than half of the sum, so we have .

$$\frac{1}{4}\|g\|_2^2 \le \|g_+\|_2^2 \le \sum_{i=i_{min}}^{0} 2^{2(i+2))}\mathbb{P}(T_i).$$

Hence there exists $i_0 \in [i_{min}, 0]$ such that

$$\frac{\|g\|_2^2}{4\log n} \leq 2^{2(i_0+2))}\mathbb{P}(T_{i_0}) \leq 4\min_{x\in U_i} g(x)^2 \cdot \mathbb{P}(U_i) \leq 4\mathbb{P}(U_{i_0})^{-1}\left(\int_{U_i} g\, d\mathbb{P}\right)^2,$$

or

$$c\log(n)^{-\frac{1}{2}}\|g\|_2 \leq \mathbb{P}(U_{i_0})^{-\frac{1}{2}}\int_{U_{i_0}} g\, d\mathbb{P}.$$

We consider two cases: either $2^{-i} \geq C(\log n)^{d-1}v^{-1}\|g\|_1$, or $2^{-i} \leq C(\log n)^{d-1}v^{-1}\|g\|_1$. The first case leads immediately to alternative (2), while in the second case we have

$$c\log(n)^{-\frac{1}{2}}\|g\|_2 \leq \mathbb{P}(U_{i_0})^{-\frac{1}{2}}\int_{U_{i_0}} g\, d\mathbb{P} \leq 4C(\log n)^{d-1}v^{-1}\|g\|_1 \cdot \mathbb{P}(U_{i_0})^{\frac{1}{2}} \leq C(\log n)^{d-1}\|g\|_1 v^{-\frac{1}{2}},$$

which is another instance of alternative (1).

As for the right-hand inequality in alternative (1), this is simply Cauchy-Schwarz: $\left(\int |g|\, d\mathbb{P}\right)^2 \leq \int g^2\, d\mathbb{P} \cdot v$. □

Proof of Lemma 36. The first inequality is again Cauchy-Schwarz: for any subset A of S, we have

$$\int_A g\, d\mathbb{P} \leq \left(\int_A g^2\, d\mathbb{P}\right)^{\frac{1}{2}}\left(\int_A 1\, d\mathbb{P}\right)^{\frac{1}{2}} \leq \|g\|_2 \cdot \mathbb{P}(A)^{\frac{1}{2}}.$$

As for the second statement, first note that since $\|g\|_\infty \leq 1$ and we have

$$c\log(n)^{-1/2}\|g\|_2 \leq \mathbb{P}(U_{i_0})^{-1/2}\int_{U_{i_0}} g\, d\mathbb{P} \leq \mathbb{P}(U_{i_0})^{\frac{1}{2}}.$$

Let $j = \lceil \log \mathbb{P}(U_{i_0}) \rceil \geq i_{min}$, $\epsilon_j = 2^j$, so that $U_{i_0} \subset S(\epsilon_j)$ and

$$\mathbb{P}(S_{\epsilon_j}) \leq C\mathbb{P}(U_{i_0})\log(\mathbb{P}(U_{i_0})^{-1})^{d-1} \leq C\mathbb{P}(U_{i_0})(\log n)^{d-1}.$$

Recalling again that by (32), $g \geq -C\|g\|_1 \mathbb{P}(S)^{-1}$, we have

$$\begin{aligned}
\int_{S(\epsilon_j)} g\, d\mathbb{P} - 2^{-1} \int_{U_{i_0}} g\, d\mathbb{P} &\geq 2^{-1} \int_{U_{i_0}} g\, d\mathbb{P} + \int_{S(\epsilon_j)\setminus U_{i_0}} g\, d\mathbb{P} \\
&\geq \mathbb{P}(U_{i_0}) 2^{i_0} - \mathbb{P}(S(\epsilon_j)) \cdot C\|g\|_1 \mathbb{P}(S)^{-1} \\
&\geq \mathbb{P}(U_{i_0}) \left(2^{i_0} - C_d (\log n)^{d-1} \|g\|_1 \mathbb{P}(S)^{-1}\right) \\
&\geq c \cdot \mathbb{P}(U_{i_0}) \cdot 2^{i_0} > 0,
\end{aligned}$$

where we used our assumption on i_0 in the last line. Therefore, by the last two inequalities

$$\mathbb{P}(S(\epsilon_j))^{-\frac{1}{2}} \int_{S(\epsilon_j)} g\, d\mathbb{P} \geq c(\log n)^{-(d-1)/2} \mathbb{P}(U_{i_0})^{-\frac{1}{2}} \int_{U_{i_0}} g\, d\mathbb{P} \geq c(\log n)^{-d/2} \|g\|_2,$$

as claimed. Finally, note that by the assumptions of $\|g\|_2^2 \geq \frac{Cd(\log n)^2}{n}$ and $\|g\|_\infty \leq 1$, we obtain that

$$\mathbb{P}(S(\epsilon_j)) \geq \mathbb{P}(U_{i_0}) \geq (c \log(n)^{-\frac{1}{2}} \|g\|_2 / \sqrt{\|g\|_\infty})^2 \geq C_1 d \frac{\log n}{n}.$$

□

4.3 Sketch of the Proof of Theorem 5

The modifications of Algorithm 1 to work in this setting are minimal: we simply need to replace L by Γ, and replace (4.12) with

$$\forall (i,j) \in [|\mathcal{S}|] \times [n]: \qquad |y_{i,j}| \leq \Gamma \tag{4.33}$$

$$\begin{aligned}
&\forall\, 1 \leq i \leq |\mathcal{S}| \;\; \frac{1}{n} \sum_{j=1}^{n} (f(Z_{i,j}) - \hat{w}_i^\top (Z_{i,j}, 1))^2 \leq \hat{l}_i^2 + \Gamma \sqrt{\frac{Cd \log(n)}{n}} \\
&\forall (i,j) \in [|\mathcal{S}|] \times [n] \quad |f(Z_{i,j})| \leq \Gamma \\
&\forall (i_1, j_1), (i_2, j_2) \in [|\mathcal{S}|] \times [n] \quad f(Z_{i_2,j_2}) \geq \nabla f(Z_{i_1,j_1})^\top (Z_{i_2,j_2} - Z_{i_1,j_1}).
\end{aligned}$$

For the correctness proof, we need some additional modifications. First, we replace Lemma 24 with a similar bound in the Γ-bounded setting. The following lemma is based on the L_4 entropy bound of [GW17, Thm 1.1] and the peeling device [Gee00, Ch. 5]; it appears explicitly in [HW16]):

Lemma 37. *Let $d \geq 5$, $m \geq C^d$ and $\mathbb{Q}$ be a uniform measure on a convex polytope $P' \subset B_d$ and $Z_1, \dots, Z_m \sim \mathbb{Q}$. Then, the following holds uniformly for all $f, g \in \mathcal{F}^\Gamma(P')$*

$$2^{-1}\int_{P'}(f-g)^2 d\mathbb{Q} - C(P')\Gamma^2 m^{-\frac{4}{d}} \leq \int (f-g)^2 d\mathbb{Q}_m \leq 2\int_{P'}(f-g)^2 d\mathbb{Q} + C(P')\Gamma^2 m^{-\frac{4}{d}},$$

with probability at least $1 - C_1(P')\exp(-c_1(P')\sqrt{m})$.

Note that differently from Lemma 24, the constant before $m^{-4/d}$ depends on the domain P', and this dependence cannot be removed.

Since $\mathcal{F}^\Gamma(P')$ has finite L_2-entropy for every ϵ, it is in particular compact in $L_2(P')$, which means that the proof of Lemma 27 in sub-Section 4.2.3 works for this class of functions as well.

The proof of Theorem 17 also goes through for this case, by replacing Lemma 24 by Lemma 37. The precise statement we obtain is the following:

Theorem 21. *Let $P \subset B_d$ be a convex polytope, $f \in \mathcal{F}^\Gamma(P)$, and some integer $k \geq (Cd)^{d/2}$, for some large enough $C \geq 0$, there exists a convex set $P_k \subset P$ and a k-simplicial convex function $f_k : P_k \to \mathbb{R}$ such that*

$$\mathbb{P}(P \setminus P_k) \leq C(P)k^{-\frac{d+2}{d}}\log(k)^{d-1}.$$

and

$$\int_{P_k}(f_k - f)^2 d\mathbb{P} \leq \Gamma^2 \cdot C(P)k^{-\frac{4}{d}}$$

The remaining lemmas and arguments in the proof of Theorem 4, can easily be seen to apply in the setting of Γ-bounded regression under polytopal support P.

4.4 Simplified version of our estimator

Like the estimator for our original problem, the simplified version of our estimator is based on the existence of a simplicial approximation $\hat{f}_{k(n)} : \Omega_{k(n)} \to [0,1]$ to the unknown convex function f^* (Theorem 4). Here we demonstrate how to recover f^* to within the desired accuracy if we are given the simplicial structure of $f_{k(n)}$, i.e., the set $\Omega_{k(n)}$ and the decomposition $\bigcup_{i=1}^{k(n)} \triangle_i$ of $\Omega_{k(n)}$ into simplices such that $f_{k(n)}|_{\triangle_i}$ is affine for each i. In this case the performance of our algorithm is rather better: it runs in time $O_d(n^{O(1)})$ rather than $O_d(n^{O(d)})$, and is minimax optimal up to a constant that depends on d. We can also slightly weaken the assumptions: it is no longer required that the variance σ^2 of the noise be given.

We will use the following classical estimator [Gyö+02, Thm 11.3]; it is quoted here with an improved bound which is proven in [MVZ21, Theorem A]:

Lemma 38. *Let $m \geq d+1$, $d \geq 1$ and $\mathbb{Q}$ be a probability measure that is supported on some $\Omega' \subset \mathbb{R}^d$. Consider the regression model $W = f^*(Z) + \xi$, where f^* is L-Lipschitz and $\|f^*\|_\infty \leq L$, and $Z_1, \dots, Z_m \underset{i.i.d.}{\sim} \mathbb{Q}$. Then, the exists an estimator $\hat{f}_R$ that has an input of $\{(Z_i, W_i)\}_{i=1}^m$ and runtime of $O_d(n)$ and outputs a function such that*

$$\mathbb{E} \int (\hat{f}_R(x) - f^*(x))^2 d\mathbb{Q}(x) \leq \frac{Cd(\sigma + L)^2}{m} + \inf_{w \in \mathbb{R}^{d+1}} \int (w^\top (x,1) - f^*(x))^2 d\mathbb{Q}(x).$$

Note that this estimator is *distribution-free*: it works irrespective of the structure of $\mathbb{Q}$, nor does it require that $\mathbb{Q}$ be known.

The first step of the simplified algorithm is estimating $f^*|_{\triangle_i}$ on each $\triangle_i \subset \Omega_{k(n)}$ ($1 \leq i \leq k(n)$) with the estimator $\hat{f}_R$ defined in Lemma 38 (with respect to the probability measure $\mathbb{P}(\cdot|\triangle_i)$) with the input of the data points in $\mathcal{D}$ that lie in $\triangle_i$. We obtain independent regressors $\hat{f}_1, \dots, \hat{f}_{k(n)}$ such that

$$\mathbb{E} \int_{\triangle_i} (\hat{f}_i(x) - f^*(x))^2 \frac{d\mathbb{P}}{\mathbb{P}(\triangle_i)} \leq \inf_{w \in \mathbb{R}^{d+1}} \int_{\triangle_i} (w^\top (x,1) - f^*(x))^2 \frac{d\mathbb{P}}{\mathbb{P}(\triangle_i)} + \mathbb{E} \min\{\frac{Cd}{\mathbb{P}_n(\triangle_i) n}, 1\}, \tag{4.34}$$

where the $\min\{\cdot, 1\}$ part follows from the fact that when we have less than Cd points, we can always set $\hat{f}_i$ to be the zero function.

Now, we define the function $f'(x) := \sum_{i=1}^{k(n)} \hat{f}_i(x)\mathbb{1}_{x\in\triangle_i}$, and by multiplying the last equation by $\mathbb{P}(\triangle_i)$ for each $1 \leq i \leq k(n)$ and taking a sum over i, we obtain that

$$\begin{aligned}\mathbb{E}\int_{\Omega_{k(n)}}(f'-f^*)^2 d\mathbb{P} &\leq \sum_{i=1}^{k(n)} \inf_{w_i\in\mathbb{R}^{d+1}} \int_{\triangle_i}(w_i^\top(x,1)-f^*)^2 d\mathbb{P} + \mathbb{E}\sum_{i=1}^{k(n)} \min\{\frac{Cd\cdot\mathbb{P}(\triangle_i)}{n\cdot\mathbb{P}_n(\triangle_i)}, \mathbb{P}(\triangle_i)\} \\ &\leq \int_{\Omega_{k(n)}}(f_{k(n)}-f^*)^2 d\mathbb{P} + C_1 dk(n)\cdot n^{-1} = O_d(n^{-\frac{4}{d+4}}),\end{aligned} \tag{4.35}$$

where in the first equation, we used the the fact that $n\cdot\mathbb{P}_n(\triangle_i) \sim Bin(n, \mathbb{P}(\triangle_i))$ (for completeness, see Lemma 29), and in the last inequality we used Eq. (4.4). Next, recall that Theorem 17 implies that

$$\mathbb{P}(\Omega\setminus\Omega_{k(n)}) \leq C(d)k(n)^{-\frac{d+2}{d}} \leq O_d(n^{-\frac{d+2}{d+4}}).$$

Therefore, if we consider the (not necessarily convex) function $\widetilde{f} = f'\mathbb{1}_{\Omega_{k(n)}} + \mathbb{1}_{\Omega\setminus\Omega_{k(n)}}$, we obtain that

$$\begin{aligned}\mathbb{E}\int_{\Omega}(f'-f^*)^2 d\mathbb{P} = \mathbb{E}\int_{\Omega\setminus\Omega_{k(n)}}(f'-f^*)^2 d\mathbb{P} + \mathbb{E}\int_{\Omega_{k(n)}}(f'-f^*)^2 d\mathbb{P} &\leq O_d(n^{-\frac{d+2}{d+4}} + n^{-\frac{4}{d+4}}) \\ \leq O_d(n^{-\frac{4}{d+4}}).\end{aligned}$$

Thus, $\widetilde{f}$ is a minimax optimal *improper* estimator. To obtain a proper estimator, we simply need to replace $\widetilde{f}$ by $MP(\widetilde{f})$, where MP is the procedure defined in §4.5.

It remains only to point out that the runtime of this estimator is of order $O_d(n^{O(1)})$. Indeed, the procedure MP is essentially a convex LSE on n points, which can be formulated as a quadratic programming problem with $O(n^2)$ constraints, and hence can be computed in $O_d(n^{O(1)})$ time [SS11]. In addition, the runtime of the other estimator we use, namely the estimator of Lemma 38, is linear in the number of inputs.

4.5 From an Improper to a Proper Estimator

The following procedure, which we named MP, is classical and we give its description and prove its correctness here for completeness. However, note that we only give a proof for optimality in expectation; high-probability bounds can be obtained using standard concentration inequalities.

The procedure MP is defined as follows: given an improper estimator $\widetilde{f}$, draw $X_1', \dots, X_{k(n)}' \underset{i.i.d.}{\sim} \mathbb{P}$, and apply the convex LSE with the input $\{(X_i', \widetilde{f}(X_i'))\}_{i=1}^{k(n)}$, yielding a function $\hat{f}_1$. We remark that the convex LSE is only unique on the convex hull of the data-points $X_1', \dots, X_{k(n)}'$, and not on the entire domain Ω [SS11], so we will show that any solution $\hat{f}_1$ of the convex LSE is optimal.

First off, we have

$$\mathbb{E}\int_\Omega (\widetilde{f} - f^*)^2 d\mathbb{P}_{k(n)}' = \mathbb{E}\int_\Omega (\widetilde{f} - f^*)^2 d\mathbb{P} \tag{4.36}$$

Also recall the classical observation that for $\hat{f}_1$ that is defined above, we know that $(\hat{f}_1(X_1'), \dots, \hat{f}_1(X_{k(n)}'))$ is precisely the projection of $(\widetilde{f}(X_1'), \dots, \widetilde{f}(X_{k(n)}'))$ on the *convex set*

$$\mathcal{F}_{k(n)} := \{(f(X_1'), \dots, f(X_{k(n)}')) : f \in \mathcal{F}_1(\Omega)\} \subset \mathbb{R}^{k(n)},$$

cf. [Cha14]. Now, the function $\Pi_{\mathcal{F}_{k(n)}}$ sending a point to its projection onto $\mathcal{F}_{k(n)}$, like any projection to a convex set, is a 1-Lipschitz function, i.e.,

$$\|\Pi_{\mathcal{F}_{k(n)}}(x) - \Pi_{\mathcal{F}_{k(n)}}(y)\| \leq \|x - y\| \quad \forall x, y \in \mathbb{R}^{k(n)}.$$

We also know that $(\hat{f}_1(X_i'))_{i=1}^{k(n)} = \Pi_{\mathcal{F}_{k(n)}}(\widetilde{f})$ and $\Pi_{\mathcal{F}_{k(n)}}((f^*(X_i'))_{i=1}^{k(n)}) = (f^*(X_i'))_{i=1}^{k(n)}$; substituting in the preceding equation, we therefore obtain

$$\mathbb{E}\int_\Omega (\hat{f}_1 - f^*)^2 d\mathbb{P}_{k(n)}' \leq \mathbb{E}\int_\Omega (\widetilde{f} - f^*)^2 d\mathbb{P}_{k(n)}' = \mathbb{E}\int_\Omega (\widetilde{f} - f^*)^2 d\mathbb{P},$$

since $\int (\cdot)^2 d\mathbb{P}_{k(n)}'$ is just $\|\cdot\|^2/k(n)$. In order to conclude the minimax optimality of $\hat{f}_1$, we know by Lemma 24 that for any function in

$$\mathcal{O} := \left\{ f \in \mathcal{F}_1 : \int_\Omega (f - \widetilde{f}')^2 d\mathbb{P}'_{k(n)} = 0 \right\},$$

it holds that

$$\mathbb{E} \int (f - f^*)^2 d\mathbb{P} \leq 2\mathbb{E} \int (\widetilde{f} - f^*)^2 d\mathbb{P}'_{k(n)} + Ck(n)^{-\frac{4}{d}} \leq 2\mathbb{E} \int (\widetilde{f} - f^*)^2 d\mathbb{P} + C_1 n^{-\frac{4}{d+4}},$$

where we used Eq. (4.36) and the fact that $k(n) = n^{\frac{d}{d+4}}$. Since we showed that $\hat{f}_1$ must lie in $\mathcal{O}$, the minimax optimality of this proper estimator follows.

Chapter 5

On the Minimal Error of Empirical Risk Minimization

Recall the definitions of the risk and minimax optimality from Chapter 1. These definitions measure the "worst case scenario" of a given estimator, and may hide the true statistical performance of the ERM in real-life applications (cf. [Bel17]). For example, as mentioned in the introduction, if f^* is known to belong to a smaller class $\mathcal{H} \subset \mathcal{F}$, the relevant quantity is

$$\mathcal{R}_{\mathcal{H}}(\widehat{f}_n, \mathcal{F}, \mathbb{Q}) := \sup_{f^* \in \mathcal{H}} \mathbb{E} \int (\widehat{f}_n - f^*)^2 d\mathbb{Q},$$

where $\mathbb{Q} \in \{\mathbb{P}^{(n)}, \mathbb{P}\}$ that may be significantly smaller than the minimax risk. We remark that the LSE is still defined over $\mathcal{F}$, due to computational or other considerations.

As an example, consider linear regression in $\mathbb{R}^d$ when the true coefficient vector is sparse, i.e. supported on $k \ll d$ coordinates. Then, due to computational considerations, it is standard to replace the original problem of minimizing square loss over sparse vectors in $\mathbb{R}^d$ by minimization over a larger ℓ_1 ball in $\mathbb{R}^d$ (the Lasso procedure).

The second example was already briefly mentioned in the introduction, and we expand on it here. Let $\mathcal{F}_d$ be the family of convex 1-Lipschitz functions on $\mathcal{X} = [0,1]^d$, and let $\mathbb{P} = \text{Unif}(\mathcal{X})$. The subset $\mathcal{H}_d$ (of 'simple functions') is the set of 1-Lipschitz k-affine piece-wise linear functions with

$k = \Theta(1)$. It is well known that ERM over $\mathcal{H}_d$ is NP-hard since the problem is highly non-convex; moreover, even estimating the number of pieces is computationally hard (cf. the recent paper [Gho+21] for more details). In contrast, ERM over $\mathcal{F}_d$ can be efficiently computed [Gho+21]. While the minimax rate for $(\mathcal{F}_d, \mathbb{P}_d)$ is $\Theta(n^{-\frac{4}{d+4}})$ [Dud99; Bro76], it was proved in another paper of author [Kur+20], that the risk of LSE is $\Theta_d(n^{-\min\{\frac{2}{d},\frac{4}{d+4}\}})$, which is minimax-suboptimal when $d \geq 5$. Furthermore, it was shown in [HW16; Fen+18] that

$$\mathcal{R}_{\mathcal{H}_d}(\widehat{f}_n, \mathcal{F}_d, \mathbb{P}_d) = \widetilde{O}_d(n^{-\min\{4/d,1\}}), \tag{5.1}$$

which is *significantly* smaller than both the risk of ERM and the minimax rate. When the ERM (or MLE) satisfies such improved bounds, we say that it exhibits *adaptation* (cf. [Fen+18; KGS18; Sam18; Han+19; Kur+20]).

Here, we answer the two following questions: Does there exist a *uniform* lower bound on the *minimal error* of ERM, i.e.

$$\inf_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{Q},$$

where $\mathbb{Q} \in \{\mathbb{P}, \mathbb{P}^{(n)}\}$. Does the richness of the entire class $\mathcal{F}$ affect the minimal error, or is there a more refined notion of complexity that governs its behavior?

5.1 Main Results

5.1.1 Fixed Design

Notation: For n points $\overline{x} := \{x_1, \ldots, x_n\}$ in $\mathcal{X}$ and $\mathcal{G} \subseteq \mathcal{F}$, we remind the definition of Gaussian complexity of $\mathcal{G}$ as

$$\mathcal{W}_{\mathbf{x}}(\mathcal{G}) := \mathbb{E}_{\xi} \sup_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(x_i).$$

Let $\mathbb{P}^{(n)} := n^{-1} \sum_{i=1}^{n} \delta_{x_i}$ and for any $f : \mathcal{X} \to \mathbb{R}$ we denote by $\|f\|_n$ the $L_2(\mathbb{P}^{(n)})$ norm of f. Next, for any $f \in \mathcal{F}$ and $r \geq 0$ we denote by $B_n(f, r) := \{g \in \mathcal{F} : \|g - f\|_n \leq r\}$, i.e the intersection of the $L_2(\mathbb{P}^{(n)})$ ball around f and the class $\mathcal{F}$.

We now state our sharp lower bound for the fixed design error, for simplicity of exposition under the assumption of uniform boundedness of $\mathcal{F}$ (the general statement is given below in Lemma 39).

Corollary 5. *Under Assumptions 1,2,3, the minimal error of $\widehat{f}_n$ satisfies*

$$\inf_{f^* \in \mathcal{F}} \mathbb{E}_{\xi} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}^{(n)} \gtrsim \mathcal{W}_{\mathbf{x}}(\mathcal{F})^2. \tag{5.2}$$

When $\mathcal{F}$ is uniformly bounded (say, by 1), a classical result in non-parametric statistics [Gee00] and our theorem imply that

$$\mathcal{W}_{\mathbf{x}}(\mathcal{F})^2 \lesssim \inf_{f^* \in \mathcal{F}} \mathbb{E}_{\xi} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}^{(n)} \leq \underbrace{\sup_{f^* \in \mathcal{F}} \mathbb{E}_{\xi} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}^{(n)}}_{=\mathcal{R}(\widehat{f}_n, \mathcal{F}, \mathbb{P}^{(n)})} \leq 2\mathcal{W}_{\mathbf{x}}(\mathcal{F}).$$

Moreover, both of these bounds are tight, in the sense that they can be attained on certain families of functions, up to constants (cf. [BM98; Han+19; KDR19]). Therefore, we conclude that in the *fixed design* case, both the minimax risk and the minimal error of the ERM depend on the entire Gaussian complexity of $\mathcal{F}$ (when it is convex and uniformly bounded). In particular, for the case of convex regression, Corollary 5 recovers the rate in (5.1) (up to logarithmic factors) for the fixed design case, since with high probability the global complexity $\mathcal{W}_{\mathbf{x}}(\mathcal{F})$ is of the order $\max\{n^{-2/d}, n^{-1/2}\}$.

5.1.2 Random Design

We begin and remind some notation:

Notation: We denote by $\mathbb{P}_n$ to be the *random* empirical measure of $X_1, \ldots, X_n \underset{i.i.d.}{\sim} \mathbb{P}$. Next, we denote the averaged Gaussian complexity over $X_1, \ldots, X_n$ in

$$\mathcal{W}(\mathcal{G}) := \mathbb{E}\mathcal{W}_{\mathbf{x}}(\mathcal{G}).$$

With some ambiguity in the notation, for every $f : \mathcal{X} \to \mathbb{R}$ we denote by $\|f\|_n$ the $L_2(\mathbb{P}_n)$ norm of f. Next, for any $f \in \mathcal{F}$ and $r \geq 0$ we denote by $B_n(f,r) := \{g \in \mathcal{F} : \|g - f\|_n \leq r\}$, i.e the intersection of the $L_2(\mathbb{P}_n)$ ball around f and the class $\mathcal{F}$. Next, for every $f : \mathcal{X} \to \mathbb{R}$ we denote by $\|f\|$ the $L_2(\mathbb{P})$ norm of f. Next, for any $f \in \mathcal{F}$ and $r \geq 0$ we denote by $B(f,r) := \{g \in \mathcal{F} : \|g - f\| \leq r\}$, i.e the intersection of the $L_2(\mathbb{P})$ ball around f and the class $\mathcal{F}$.

We now can turn to the random design setting, which is significantly more subtle. Before stating the result, we describe a direct proof strategy that fails. This approach would attempt to pass from the fixed design lower bound to the random design lower bound by relating the population and empirical norms $\|\cdot\|$ and $\|\cdot\|_n$, uniformly over the class. A statement of this type (which may be called "upper isometry," in contrast with "lower isometry" studied, for instance, in [Men14]) could be derived under additional assumptions on the geometry of $(\mathcal{F}, \mathbb{P})$, such as a small-ball condition [Men14], Kolchinskii-Pollard entropy [RST17], or an ϵ-covering with respect to the sup-norm [Gee00]. To the best of our knowledge, such upper-isometry statements can at best read

$$\|f - g\|^2 \geq \frac{1}{2}\|f - g\|_n^2 - C \cdot \mathcal{W}(\mathcal{F})^2 \quad \forall f, g \in \mathcal{F},$$

where $C \geq 1$. Since $\mathcal{W}(\mathcal{F})^2$ is larger than the lower bound on the fixed design error, this technique does not appear to work.

Moreover, a uniform lower bound of order $\mathcal{W}(\mathcal{F})^2$ in random design cannot be true in general. For instance, it was shown in a string of recent works [LRZ20; BRT19; Bar+20] that it is possible to completely interpolate $Y_1, \ldots, Y_n$ (i.e. achieve zero empirical error) and still have a small generalization error (of order n^{-c}, for some $c \in (0,1)$), and even be minimax optimal (with an appropriate function Ψ in the definition of LSE (see (1.2) above). In these examples, because of the ability to interpolate any data, we know that $\mathcal{W}(B_n(f^*, 1)) = \Theta(1)$; therefore, the lower bound in the fixed design case cannot be always true in random design.

The last paragraph motivates the need to consider additional properties of the model $\mathcal{F}$ and the underlying distribution $\mathbb{P}$. With the interpolation examples in mind, we might hope that the relation between the global complexity of the class and complexity of local neighborhoods around the regression function f^* may play a role in determining rates of convergence of ERM. To this end, for every n and $f^* \in \mathcal{F}$, we define the following notion of complexity:

$$t_{n,\mathbb{P}}(f^*, \mathcal{F}) := \max\{t \in \mathbb{R}^+ : \mathcal{W}(B(f^*, t)) \leq l_\xi \mathcal{W}(B_n(f^*, 1))\}, \tag{5.3}$$

where $l_\xi \in (0, 1)$ is a small absolute constant that will be chosen in the proofs. We remark that under the additional assumption of $\mathcal{F}$ being uniformly bounded by 1, we have that $\mathcal{W}(\mathcal{F}) \geq \mathcal{W}(B_n(f^*, 1)) \geq \frac{1}{2}\mathcal{W}(B_n(f^*, 2)) = \frac{1}{2}\mathcal{W}(\mathcal{F})$, and up to a different l_ξ, we may replace the term on the right-hand side of (5.3) with global Gaussian averages $\mathcal{W}(\mathcal{F})$.

The quantity $t_{n,\mathbb{P}}(f^*, \mathcal{F})$ is the maximal radius of the population ball around f^* that has Gaussian complexity comparable to that of the entire class (in the uniformly bounded case), up to some absolute constant, or to a ball of constant radius within the class. As we show next, this local richness is necessary in order to avoid the rate being dominated by the global complexity of $\mathcal{F}$. In the aforementioned interpolation examples we have both $t_{n,\mathbb{P}}(f^*, \mathcal{F}) = O(n^{-c})$ and $\mathcal{W}(B_n(f^*, 1)) = \Theta(1)$. The last two relations must be true for any $f^* \in \mathcal{F}$ for which ERM attains perfect fit to data, and yet a small generalization error of order n^{-c}.

We now state the main result of this paper for the random design setting, under the additional assumption of $\mathcal{F}$ being uniformly bounded. Remarkably, $t_{n,\mathbb{P}}(f^*, \mathcal{F})$ is the only additional quantity that we need to consider for a general uniform lower bound on a general family $\mathcal{F}$. Specifically, we prove the following:

Theorem 22. *Let $\mathcal{F}$ be a convex class of functions*[1] *uniformly bounded by one. Then, for large enough n, the following holds for any $f^* \in \mathcal{F}$*

$$\mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{P} \gtrsim \cdot \min\{\mathcal{W}(\mathcal{F})^2, t_{n,\mathbb{P}}(f^*, \mathcal{F})^2\},$$

where $c \in (0, 1)$.

Using the last theorem and the classical basic inequality (cf . [Gee00, Chp. 5]), we derive the

[1] We assume that $\mathcal{F}$ is non-degenerate and contains at least two functions such that $\|f_1 - f_2\| \geq 1/2$.

analogue bound to 5.2, in the case of a uniformly bounded by one convex $\mathcal{F}$:

$$\forall f^* \in \mathcal{F} \quad \min\{\mathcal{W}(\mathcal{F})^2, t_{n,\mathbb{P}}(f^*, \mathcal{F})^2\} \lesssim \mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{P} \leq 8 \cdot \mathcal{W}(\mathcal{F}) \tag{5.4}$$

and as in the fixed design class, there exist functions classes that attain both of these bounds.

Remark 24. Notably, Theorem 22 holds under only convexity and uniform boundedness assumptions on the class $\mathcal{F}$. Furthermore, one can easily design a convex uniformly bounded family and an $f^* \in \mathcal{F}$ such that the ERM attains an error of order $t_{n,\mathbb{P}}(f^*, \mathcal{F})^2 \ll \mathcal{W}(\mathcal{F})^2$ for all n that are large enough (for completeness see Section 5.5.1). Therefore, under no additional assumption on $\mathcal{F}$, the above lower bound is sharp up to absolute constants.

An almost immediate corollary of this theorem is the following key insight on the behavior of the ERM procedure in the random design setting:

Corollary 6. *Let $\mathcal{F}$ be convex and uniformly bounded by* 1. *For any $f^* \in \mathcal{F}$ such that*

$$\underbrace{\mathbb{E}_{\mathcal{D}} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}}_{:=E^2(f^*)} \ll \mathcal{W}(\mathcal{F})^2,$$

there must *exists some constant $t(f^*) \leq c_1 \cdot E(f^*)$ such that*

$$\mathcal{W}(B(f^*, t(f^*))) = \Theta(\mathcal{W}(\mathcal{F})),$$

where $c_1 \in (0, 1)$ is some absolute constant.

Informally speaking, if ERM learns some $f^* \in \mathcal{F}$ at a rate faster than $\mathcal{W}(\mathcal{F})^2$, then the local complexity of a population ball centered at f^* with a very small radius must be as rich as the entire complexity of $\mathcal{F}$. A more prescriptive recipe for guaranteeing such fast rates is an interesting direction of further work.

5.1.3 Donsker and non-Donsker Classes

Our next result shows that without further assumptions we cannot learn *any* function in a convex uniformly bounded $\mathbb{P}$-Donsker class faster than a parametric rate.

Corollary 7. *Let $\mathcal{F}$ be a $\mathbb{P}$-Donsker class (see Definition 4 above). Then,*

$$n^{-1} \lesssim \inf_{f^* \in \mathcal{F}} \mathbb{E} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}_n \asymp \inf_{f^* \in \mathcal{F}} \mathbb{E} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}$$

This lower bound is sharp, namely there are classical $\mathbb{P}$-Donsker classes, such as the convex regression example mentioned in the introduction and Section 5.1, where ERM can attain a parametric rate (up to logarithmic factors) when optimizing over all convex Lipschitz functions, but only for $d \leq 4$ which puts us in the Donsker regime.

For non-Donsker classes, i.e when $\alpha > 2$, the ERM procedure may not be optimal. One can show that

$$n^{-\frac{2}{2+\alpha}} \lesssim \mathcal{R}(\widehat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim n^{-\frac{1}{\alpha}}$$

and both of these bounds can be tight, up to logarithmic factors. Furthermore, one can show that

$$n^{-\frac{2}{2+\alpha}} \lesssim \mathcal{W}(\mathcal{F}) \lesssim n^{-\frac{1}{\alpha}}$$

and, again, both of these can be tight. Our next corollary shows that in this regime, the fixed-design error is at least of the order $\mathcal{W}(\mathcal{F})^2$, i.e. it is impossible to learn at a parametric rate in the non-Donsker regime (in terms of the $L_2(\mathbb{P}_n)$).

Corollary 8. *Let $\mathcal{F}$ be a convex uniformly bounded non-$\mathbb{P}$-Donsker class, and let $X_1, \ldots, X_n \sim \mathbb{P}$. Then the following holds:*

$$n^{-\frac{4}{2+\alpha}} \lesssim \mathcal{W}(\mathcal{F})^2 \lesssim \inf_{f^* \in \mathcal{F}} \mathbb{E} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}_n$$

The proof of these two corollaries appears in the appendix.

Remark 25. Due to the geometry of general non-Donsker classes, in random design case the same lower bound may not hold. However, in all the examples in the literature [HW16; Fen+18; KGS18; Han+19; Kur+20] that study the adaptivity of ERM in non-Donsker families (such as convex functions when $d \geq 5$, isotonic functions when $d \geq 3$), the term of $t_{n,\mathbb{P}}(f^*, \mathcal{F})$ of Theorem 22 is significantly larger than $\mathcal{W}(\mathcal{F})$. As a consequence, one may use Theorem 22 to show that the bound in (5.1) is tight up to logarithmic factors.

5.1.4 General Lower Bound for Fixed Design

In this section, we state the general lower bound for fixed design. In comparison to its consequence, Corollary 5, the version below captures complexity of local neighborhoods around regression functions that are close to f^*. Note that this lemma holds for any convex family (and not necessarily bounded).

Lemma 39. *Let $\mathcal{F}$ be a convex family of functions and and let $x_1, \dots, x_n \in \mathcal{X}$ be some n points, and let $\mathbb{P}^{(n)} := n^{-1} \sum_{i=1}^{n} \delta_{x_i}$. For all $f^* \in \mathcal{F}$ define*

$$r(f^*) := \operatorname{argmax}_{r \geq 0} \mathcal{W}_{\mathbf{x}}(B_n(f^*, r)) - \frac{r^2}{2} \tag{5.5}$$

and

$$L_x(f^*) := \max_{g \in B_n(f^*,1), t \geq 0} \frac{\mathcal{W}_{\mathbf{x}}(B_n(g,t)) - \mathcal{W}_{\mathbf{x}}(B_n(f^*, r(f^*)) - Cn^{-1}}{\|g - f^*\|_n + t}$$

where $C \in (1, \infty)$ is some absolute constant. Then the following lower bound holds:

$$\mathbb{E}_{\xi} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}^{(n)} \geq \max\{\frac{(\mathcal{W}_{\mathbf{x}}(B_n(f^*,1)) - Cn^{-1})^2}{4}, L_x(f^*)^2\}.$$

Note that Corollary 5 follows almost immediately from the last lemma. To see this, the convexity of $\mathcal{F}$, and the uniform bounded by 1 assumption imply that

$$\widehat{\mathcal{W}}(\mathcal{F}) = \mathcal{W}_{\mathbf{x}}(B_n(f^*,2)) \leq 2\mathcal{W}_{\mathbf{x}}(B_n(f^*,1)).$$

Remark 26. The second term in our lower bound may be significantly larger than $\mathcal{W}_{\mathbf{x}}(\mathcal{F})^2$. For example, the second term may be equal to $\mathcal{W}_{\mathbf{x}}(\mathcal{F})$ in several non-Donsker families that appear in [BM93; Kur+20; Bir06]. We also remark that constant $\frac{1}{4}$ is tight (up to $o_n(1)$).

The rest of this paper is devoted to proofs. While the fixed design lower bound follows a rather simple argument, the corresponding lower bound in the random design case is more subtle. In particular, we employ a particular version of Talagrand's inequality that, in our particular regime, provides control on certain empirical processes, while the more commonly used versions (including Bousquet's inequality) result in vacuous estimates.

5.2 Proof of Lemma 39

Notation Throughout this section, $c, c_1, c_2 \in (0,1)$ and $C, C_1, C_2 \in (1,\infty)$ are some absolute constants that may change from to line to line. Also S_1, s_1, S_2, s_2 are absolute constants, but we use this notation to emphasize that we have some freedom to control their size. We also use the notation $C(c_1, C_2)$ to mean that the constant depends on c_1, C_2.

To recap, we assume that $\mathcal{F}$ is a convex family of functions, $Y_i = f^*(x_i) + \xi_i$, where $\xi_i \sim N(0,1)$ i.i.d., $x_1, \dots, x_n \in \mathcal{X}$, and $f^* \in \mathcal{F}$. We write $\langle f, g\rangle_n = \int fgd\mathbb{P}^{(n)}$. With slight abuse of notation, we write $\langle \boldsymbol{\xi}, f\rangle_n = \frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)$ for $\overline{\xi} := (\xi_1, \dots, \xi_n)$.

Recall the definition of $r(f^*)$ in (5.5). The following lemma that was proven in [Cha14]:

Lemma 40. *[[Cha14, Thm 1.1]] The following holds under the above assumptions:*

$$\Pr\Big(|\|\widehat{f}_n - f^*\|_n - r(f^*)| \geq t\Big) \leq \begin{cases} 3\exp(-\frac{nt^2}{64}) & t \geq r(f^*) \\ 3\exp(-\frac{nt^4}{64r(f^*)^2}) & 0 \leq t \leq r(f^*) \end{cases} \tag{5.6}$$

Moreover, for each $t \geq 0$ the following holds

$$\Pr\Big(|\langle \widehat{f}_n - f^*, \boldsymbol{\xi}\rangle_n - \mathcal{W}_{\mathbf{x}}(B_n(f^*, r(f^*)))| \geq t\cdot r(f^*)\Big) \leq \begin{cases} 3\exp(-\frac{nt^2}{64}) & t \geq r(f^*) \\ 3\exp(-\frac{nt^4}{64r(f^*)^2}) & 0 \leq t \leq r(f^*) \end{cases} \tag{5.7}$$

Also, we state a simple corollary that follows from this lemma (cf. [BLM13],[Cha14, Thm 1.2])

Corollary 2. *The following two bounds hold*

$$\mathbb{E}\left|\langle \widehat{f}_n - f^*, \boldsymbol{\xi}\rangle_n - \mathcal{W}_{\mathbf{x}}(B_n(f^*, r(f^*)))\right| \leq C_1 \max\{r(f^*)^{3/2} n^{-1/4}, n^{-1}\}.$$

and

$$\mathbb{E}\left|\|\widehat{f}_n - f^*\|_n^2 - r(f^*)^2\right| \leq C_2 \max\{r(f^*)^{3/2} n^{-1/4}, n^{-1}\}.$$

Proof of Lemma 39.

For brevity, denote $\widehat{r} := r(f^*)$, where $r(f^*)$ is defined in Lemma 40. Define

$$g_\xi := \operatorname{argmax}_{h \in B_n(g,t)} \langle h - g, \overline{\xi}\rangle_n.$$

Optimality of $\widehat{f}_n$ and convexity of $\mathcal{F}$ imply that $\langle \nabla_f \|f - \boldsymbol{y}\|_n^2|_{f=\widehat{f}_n}, g - \widehat{f}_n\rangle_n \geq 0$ for any $g \in \mathcal{F}$. In particular, for $g = g_\xi$ this implies

$$0 \geq \mathbb{E}\langle \boldsymbol{\xi} + f^* - \widehat{f}_n, g_\xi - \widehat{f}_n\rangle_n,$$

where the expectation is over $\boldsymbol{\xi}$, conditionally on $x_1, \ldots, x_n$. For any $g \in \mathcal{F}$, we may write the right-hand side as

$$\begin{aligned}&\mathbb{E}\langle \boldsymbol{\xi} + f^* - \widehat{f}_n, g_\xi - g + g - f^* + f^* - \widehat{f}_n\rangle_n \\ &= \mathcal{W}_{\mathbf{x}}(B_n(g,t)) - \mathbb{E}\left[\langle \boldsymbol{\xi}, \widehat{f}_n - f^*\rangle_n + \langle \widehat{f}_n - f^*, g_\xi - g\rangle_n + \langle \widehat{f}_n - f^*, g - f^*\rangle_n\right] + \mathbb{E}\|\hat{f}_n - f^*\|_n^2\end{aligned}$$

where we used the definition of g_ξ and the fact that $\mathbb{E}\langle \boldsymbol{\xi}, g - f^*\rangle_n = 0$. Using Corollary 2, we obtain a further lower bound of

$$\begin{aligned}
&\mathcal{W}_{\mathbf{x}}(B_n(g,t)) - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - \mathbb{E}\left[\langle \widehat{f}_n - f^*, g_\xi - g\rangle_n + \langle \widehat{f}_n - f^*, g - f^*\rangle_n\right] \\
&\quad + \widehat{r}^2 - C\widehat{r}^{3/2}n^{-1/4} - Cn^{-1} \\
&\geq \mathcal{W}_{\mathbf{x}}(B_n(g,t)) - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - \mathbb{E}\left[\langle \widehat{f}_n - f^*, g_\xi - g\rangle_n + \langle \widehat{f}_n - f^*, g - f^*\rangle_n\right] \qquad (5.8) \\
&\quad + \widehat{r}^2/2 - C_1 n^{-1}.
\end{aligned}$$

To verify the last inequality, observe that $\widehat{r}^2/2 \geq C_1\widehat{r}^{3/2}n^{-1/4}$ when $\widehat{r} \geq C_2 n^{-1/2}$ for C_2 that is large enough; on the other hand, if $\widehat{r} \leq C_2 n^{-1/2}$, the Cn^{-1} term is dominant for C large enough. Since $\langle \widehat{f}_n - f^*, g - f^*\rangle_n \leq \|\widehat{f}_n - f^*\|_n \|g - f^*\|_n$ and $\langle \widehat{f}_n - f^*, g_\xi - g\rangle_n \leq t \cdot \|\widehat{f}_n - f^*\|_n$, we conclude that

$$0 \geq \mathcal{W}_{\mathbf{x}}(B_n(g,t)) - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - \mathbb{E}[\|\widehat{f}_n - f^*\|_n](\|g - f^*\|_n + t) - Cn^{-1}. \qquad (5.9)$$

By re-arranging the terms and using Jensen's inequality, we have

$$\mathbb{E}\|\hat{f}_n - f^*\|_n^2 \geq \left(\mathbb{E}\|\hat{f}_n - f^*\|_n\right)^2 \geq \left(\frac{\mathcal{W}_{\mathbf{x}}(B_n(g,t)) - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - Cn^{-1}}{t + \|f^* - g\|_n}\right)_+^2$$

where $(a)_+^2 = \max\{a, 0\}^2$. Since $L_x(f^*)$ in the statement of the Lemma is non-negative, the lower bound of $L_x(f^*)^2$ follows.

Now, for the first part of the lower bound, we have to consider two cases. The first one is when $\mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) \geq 2^{-1}\mathcal{W}_{\mathbf{x}}(B_n(f^*,1)) + \widehat{r}^2/2$, and we have

$$\begin{aligned}
\mathbb{E}\|\widehat{f}_n - f^*\|_n &\geq \mathbb{E}\|\overline{\xi}\|_n \cdot \mathbb{E}\|\widehat{f}_n - f^*\|_n \geq \mathbb{E}\langle \overline{\xi}, \widehat{f}_n - f^*\rangle_n \\
&\geq 2^{-1}\mathcal{W}_{\mathbf{x}}(B_n(f^*,1)) + \widehat{r}^2/2 - C\widehat{r}^{3/2}n^{-1/4} \geq 2^{-1}\mathcal{W}_{\mathbf{x}}(B_n(f^*,1)) - C_1 n^{-1},
\end{aligned}$$

where we used Cauchy-Schwartz inequality and Corollary 2. In the other case, we use (5.9) with

$g = f^*$ and $t = 1$:

$$\mathbb{E}\|\widehat{f}_n - f^*\|_n \geq \mathcal{W}_{\mathbf{x}}(B_n(f^*,1)) - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - Cn^{-1} \geq 2^{-1}\mathcal{W}_{\mathbf{x}}(B_n(f^*,1)) - Cn^{-1},$$

concluding the proof. □

5.3 Proof of Theorem 22

Throughout the proof of Theorem 22, $\mathbb{P}_n$ denotes the random empirical measure of $X_1,\dots,X_n$. Denote by $\widehat{r} := \operatorname{argmax} \mathcal{W}_{\mathbf{x}}(B_n(f^*,r)) - \frac{r^2}{2}$, with the hat emphasizing the dependence on $\mathbf{x}_n = (X_1,\dots,X_n)$. We adopt the notation $\|\cdot\|_n, \langle\cdot,\cdot\rangle_n$, B_n in the previous section for the norm and the inner product with respect to $\mathbb{P}_n$, and the $L_2(\mathbb{P}_n)$ ball. Recall that we assumed that $\mathcal{F}$ is not degenerate: $\mathcal{W}(\mathcal{F}) \geq c/\sqrt{n}$, for some $c \in (0,1)$.[2]

Proof of Theorem 22. Denote

$$t_* := \min\{t_{n,\mathbb{P}}(f^*,\mathcal{F}), s_1\sqrt{\mathcal{W}(\mathcal{F})}, s_2\mathcal{W}(\mathcal{F})\} \tag{5.10}$$

where $s_1, s_2 \in (0,1)$ are small enough absolute constants that will be defined in the proof, and $t_{n,\mathbb{P}}(f^*,\mathcal{F})$ is defined in (5.3).

Denote by $\mathcal{M}$ the maximal separated set with respect to $L_2(\mathbb{P})$ at scale $6\sqrt{\mathcal{W}(\mathcal{F})}$, and let

$$M = \mathcal{M}(6\sqrt{\mathcal{W}(\mathcal{F})}, \mathcal{F}, \mathbb{P}) \tag{5.11}$$

denote its size.

For a constant $K_1 \in (1,\infty)$, let $\mathcal{E}_1$ denote the high-probability event that is defined by the

[2] See the proof of Lemma 43 for further details

intersection of the events of Lemma 44 and Lemma 42:

$$\begin{aligned}\mathcal{E}_1 :=\Big\{ \mathbf{x}_n : &\sup_{f,g\in\mathcal{F}} \left| \|f-g\|_n^2 - \|f-g\|^2 \right| \leq 10\mathcal{W}(\mathcal{F}), \\ &\sup_{h\in B(f^*,t_*),g\in\mathcal{M}} |\langle (g-f^*),(h-f^*)\rangle_n - \mathbb{E}[(g-f^*)(h-f^*)]| \leq (8K_1)^{-1}\mathcal{W}(\mathcal{F}) \Big\}.\end{aligned} \tag{5.12}$$

Further, define the events

$$\begin{aligned}\mathcal{E}_2 &= \left\{ \mathbf{x}_n : K_1^{-1}\mathcal{W}(\mathcal{F}) \leq \mathcal{W}_{\mathbf{x}}(\mathcal{F}) \leq K_1\mathcal{W}(\mathcal{F}) + Cn^{-1/2} \right\}, \\ \mathcal{E}_3 &= \left\{ \mathbf{x}_n : \mathcal{W}_{\mathbf{x}}(B(f^*,t_*)) \leq K_1\mathcal{W}(B(f^*,t_*)) + C_1\mathcal{W}(\mathcal{F})^{1/2}n^{-1/2} \right\}, \\ \mathcal{E} &= \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3.\end{aligned} \tag{5.13}$$

Lemma 43, proved in the appendix, shows that the event $\mathcal{E}$ holds with probability of at least 0.9. Note that under the event $\mathcal{E}_1$, $\mathcal{M}$ is also a $2\sqrt{\mathcal{W}(\mathcal{F})}$ separated set with respect to the random empirical measure $\mathbb{P}_n$. Hence, we may apply Sudakov's minoration (Lemma 45) with $\epsilon = 2\sqrt{\mathcal{W}(\mathcal{F})}$ and empirical measure $\mathbb{P}_n$ defined on any $\mathbf{x}_n \in \mathcal{E}$:

$$c_1\sqrt{4\mathcal{W}(\mathcal{F})\cdot\frac{\log M}{n}} \leq \mathcal{W}_{\mathbf{x}}(\mathcal{F}) \leq K_1\mathcal{W}(\mathcal{F}) + Cn^{-1/2} \leq C_1K_1\mathcal{W}(\mathcal{F}), \tag{5.14}$$

where in the last inequality we used the assumption that $\mathcal{W}(\mathcal{F}) \geq c\cdot n^{-1/2}$, and $C_1 \geq 0$ is defined to be large enough to satisfy the last inequality. Hence, the last equation implies that

$$M \leq \exp(C_2K_1^2 n\mathcal{W}(\mathcal{F})). \tag{5.15}$$

First, recall the definition of $t_{n,\mathbb{P}}(f^*,\mathcal{F})$ where in Lemma 42 (that appears in the supplementary) we set $l_{\xi} = (256K_1)^{-3}$. Recall (5.10), where in Lemma 42 we set $s_1 = c(K,K_1,C_2)$, and the three constants K,K_1,C_2 follow from Sudakov's minoration lemma, Talagrand's inequality, and Adamzcak's bound. We define $s_2 := 16^{-1}(K_1)^{-1}$.

Define the event

$$\mathcal{A} = \left\{ (\boldsymbol{\xi}, \mathbf{x}_n) : \widehat{f}_n \in B(f^*, t_*) \right\} \tag{5.16}$$

and, for any $\mathbf{x}_n$, define the conditional event

$$\mathcal{A}(\mathbf{x}_n) = \left\{ \boldsymbol{\xi} : \widehat{f}_n \in B(f^*, t_*) \right\}. \tag{5.17}$$

Assume by the way of contradiction that $\Pr_{x,\xi}(\mathcal{A}) > 0.5$. Then, using the average principle (Fubini) and the fact $\Pr(\mathcal{E}) \geq 0.9$, we can find an event $\mathcal{E}_4 \subseteq \mathcal{E}$ that has a probability of at least 0.4 (when n is large enough) such that

$$\forall \mathbf{x}_n \in \mathcal{E}_4 \quad \Pr_{\xi}(\mathcal{A}(\mathbf{x}_n)) \geq 0.5.$$

Our first step is to prove that, for all $\mathbf{x}_n \in \mathcal{E}_4$,

$$K_1 \mathcal{W}(B(f^*, t_*)) + l_{\xi} \mathcal{W}(\mathcal{F}) \geq \mathcal{W}_{\mathbf{x}}(B_n(f^*, \widehat{r})). \tag{5.18}$$

First, recall that $t_* \leq s_1 \sqrt{\mathcal{W}(\mathcal{F})}$ and therefore under the event $\mathcal{E}_1$, we have

$$\sup_{h \in B(f^*, t_*)} \|h - f^*\|_n^2 \leq 11 \cdot \mathcal{W}(\mathcal{F}). \tag{5.19}$$

Now, for each $\mathbf{x}_n \in \mathcal{E}_4 \subseteq \mathcal{E}$, the map $\boldsymbol{\xi} \mapsto \sup_{h \in B(f^*, t_*)} \langle \overline{\xi}, h - f^* \rangle_n$ is Lipschitz with constant at most

$$\sup_{h \in B(f^*, t_*)} n^{-1/2} \|h - f^*\|_n \leq C_3 \sqrt{\mathcal{W}(\mathcal{F}) n^{-1}}$$

by (5.19), and thus by Lipschitz concentration (Lemma 48), conditionally on $\mathbf{x}_n$,

$$\Pr_{\xi} \left(| \sup_{h \in B(f^*, t_*)} \langle \overline{\xi}, h - f^* \rangle_n - \mathcal{W}_{\mathbf{x}}(B(f^*, t_*)) | \geq \epsilon \right) \leq 2 \exp(-C n \mathcal{W}(\mathcal{F})^{-1} \epsilon^2)$$

for some absolute constant C. By setting $\epsilon = C_5 (n^{-1} \mathcal{W}(\mathcal{F}))^{1/2}$ in the last equation, we may define the event

$$\mathcal{A}_1(\mathbf{x}_n) = \left\{ \boldsymbol{\xi} : |\sup_{h \in B(f^*,t_*)} \langle \overline{\xi}, h - f^* \rangle_n - \mathcal{W}_{\mathbf{x}}(B(f^*,t_*))| \leq C_5 \sqrt{\mathcal{W}(\mathcal{F}) n^{-1}} \right\} \cap \mathcal{A}(\mathbf{x}_n).$$

that holds with probability of at least 0.25 (over ξ) for any $\mathbf{x}_n \in \mathcal{E}_4$.

Before defining the next event, observe that

$$\widehat{r} = \mathrm{argmax}_{r \geq 0} \mathcal{W}_{\mathbf{x}}(B_n(f^*,r)) - r^2/2 \leq 2\sqrt{\mathcal{W}_{\mathbf{x}}(\mathcal{F})},$$

according to Lemma 40 and the fact that for $r_n := 2\sqrt{\mathcal{W}_{\mathbf{x}}(\mathcal{F})}$ we have $\mathcal{W}_{\mathbf{x}}(B_n(f^*,r_n)) - r_n^2/2 \leq 0$. As we already argued in (5.14), for any $\mathbf{x}_n \in \mathcal{E}_2$ we have that $\mathcal{W}_{\mathbf{x}}(\mathcal{F}) \leq C_1 K_1 \mathcal{W}(\mathcal{F})$ for some absolute constant C_1, and thus

$$\forall \mathbf{x}_n \in \mathcal{E}_2, \quad \widehat{r} \leq C\sqrt{K_1 \mathcal{W}(\mathcal{F})}. \tag{5.20}$$

Now, from (5.7) in Lemma 40, for C_8 large enough, the event

$$\left\{ \boldsymbol{\xi} : |\langle \overline{\xi}, \widehat{f}_n - f^* \rangle_n - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r}))| \leq C_8(n^{-1/4}\widehat{r}^{3/2} + n^{-1}) \right\}$$

holds with probability of at least 0.5, and thus, in view of (5.20), for all $\mathbf{x}_n \in \mathcal{E}_4$, the event

$$\mathcal{A}_2(\mathbf{x}_n) = \left\{ \boldsymbol{\xi} : |\langle \overline{\xi}, \widehat{f}_n - f^* \rangle_n - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r}))| \leq C_6 K_1 \mathcal{W}(\mathcal{F})^{3/4} n^{-1/4} \right\} \cap \mathcal{A}_1(\mathbf{x}_n)$$

that holds with probability of at least 0.1 over ξ.

We are now ready to prove (5.18), using the fact that $\mathcal{A}_2(\mathbf{x}_n)$ is not empty for each $\mathbf{x}_n \in \mathcal{E}_4$. To this end, fix $\mathbf{x}_n \in \mathcal{E}_4$ and $\overline{\xi} \in \mathcal{A}_2(\mathbf{x}_n)$. First, by definition of $\mathcal{E}_3$, we have

$$K_1 \mathcal{W}(B(f^*,t_*)) \geq \mathcal{W}_{\mathbf{x}}(B(f^*,t_*)) - C\sqrt{\mathcal{W}(\mathcal{F}) n^{-1}}$$

which can be further lower bounded, by definition of $\mathcal{A}_1(\mathbf{x}_n)$, by

$$\sup_{h \in B(f^*,t_*)} \langle \boldsymbol{\xi}, h - f^* \rangle_n - C\sqrt{\mathcal{W}(\mathcal{F})n^{-1}}.$$

Since $\overline{\xi} \in \mathcal{A}_2(\mathbf{x}_n) \subseteq \mathcal{A}(\mathbf{x}_n)$, the above expression is further lower bounded by

$$\langle \boldsymbol{\xi}, \widehat{f}_n - f^* \rangle_n - C\sqrt{\mathcal{W}(\mathcal{F})n^{-1}}$$

which, under the assumption of $\overline{\xi} \in \mathcal{A}_2(\mathbf{x}_n)$, is lower bounded by

$$\mathcal{W}_{\mathbf{x}}(B_n(f^*, \widehat{r})) - C_2\sqrt{\mathcal{W}(\mathcal{F})n^{-1}} - C_6 K_1 \mathcal{W}(\mathcal{F})^{3/4} n^{-1/4}.$$

When n is large enough, the above estimate is lower bounded by

$$\mathcal{W}_{\mathbf{x}}(B_n(f^*, \widehat{r})) - l_\xi \mathcal{W}(\mathcal{F}).$$

To see this, observe that under the assumption of $\mathcal{W}(\mathcal{F}) \geq c/\sqrt{n}$, both $\sqrt{\mathcal{W}(\mathcal{F})n^{-1}} = o_n(\mathcal{W}(\mathcal{F}))$ and $\mathcal{W}(\mathcal{F})^{3/4}n^{-1/4} = o_n(\mathcal{W}(\mathcal{F}))$. Therefore, we proved (5.18) holds, namely that

$$K_1 \mathcal{W}(B(f^*, t_*)) + l_\xi \mathcal{W}(\mathcal{F}) \geq \mathcal{W}_{\mathbf{x}}(B_n(f^*, \widehat{r}))$$

for all $\mathbf{x}_n \in \mathcal{E}_4$. Using the definition of $l_\xi = (256K_1)^{-3}$, we have

$$\mathcal{W}(B(f^*, t_*)) \leq l_\xi \mathcal{W}(\mathcal{F}) \leq 128^{-3} K_1^{-3} \mathcal{W}(\mathcal{F}),$$

and thus for any $\mathbf{x}_n \in \mathcal{E}_4$,

$$8^{-1} K_1^{-1} \mathcal{W}(\mathcal{F}) \geq \mathcal{W}_{\mathbf{x}}(B_n(f^*, \widehat{r})). \tag{5.21}$$

By Lemma 41 and (5.21), for any $\mathbf{x}_n \in \mathcal{E}_4$,

$$0 \geq (2^{-1}K_1^{-1} - 8^{-1}K_1^{-1})\mathcal{W}(\mathcal{F}) - \mathbb{E}_\xi \max_{g\in\mathcal{M}} \langle g - f^*, \hat{f}_n - f^* \rangle_n - 16\sqrt{\mathcal{W}(\mathcal{F})\widehat{r}},$$

and since $\mathbf{x}_n \in \mathcal{E}_4 \subseteq \mathcal{E}_1$, we also have

$$\begin{aligned} 0 \geq &(2^{-1}K_1^{-1} - 8^{-1}K_1^{-1} - 8^{-1}K_1^{-1})\mathcal{W}(\mathcal{F}) - \sup_{h\in B(f^*,t_*), g\in\mathcal{M}} \int (g - f^*)(h - f^*) d\mathbb{P} - 16\sqrt{\mathcal{W}(\mathcal{F})\widehat{r}} \\ &\geq (4K_1)^{-1}\mathcal{W}(\mathcal{F}) - \max_{g\in\mathcal{M}} \|g - f^*\| t_* - 16\sqrt{\mathcal{W}(\mathcal{F})\widehat{r}} \\ &\geq (4K_1)^{-1}\mathcal{W}(\mathcal{F}) - 2t_* - 16\sqrt{\mathcal{W}(\mathcal{F})\widehat{r}} \end{aligned} \tag{5.22}$$

where we used the Cauchy-Schwartz inequality, the fact that $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, and the definition of $l_\xi = (256K_1)^{-3}$.

If $16\sqrt{\mathcal{W}(\mathcal{F})\widehat{r}} < (8K_1)^{-1}\mathcal{W}(\mathcal{F})$, then the last equation implies that

$$s_2\mathcal{W}(\mathcal{F}) = (16K_1)^{-1}\mathcal{W}(\mathcal{F}) < t_*.$$

However, this inequality contradicts the definition of t_*, and thus cannot hold for any $\mathbf{x}_n \in \mathcal{E}_4$. In the other case, we assume that $16\sqrt{\mathcal{W}(\mathcal{F})\widehat{r}} \geq (8K_1)^{-1}\mathcal{W}(\mathcal{F})$, or equivalently, $\widehat{r} \geq (128K_1)^{-1}\sqrt{\mathcal{W}(\mathcal{F})}$. Now, from Lemma 40 one can see that the maximizing value $\widehat{r}$ ensures

$$\mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - 2^{-1}\widehat{r}^2 > 0$$

and hence

$$\mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) > 2^{-1}(128K_1)^{-2}\mathcal{W}(\mathcal{F}).$$

Therefore, under the event $\mathcal{E}_4$ and by (5.18)

$$2K_1 l_\xi \mathcal{W}(\mathcal{F}) \geq K_1\mathcal{W}(B(f^*,t_*)) + l_\xi\mathcal{W}(\mathcal{F}) \geq \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) > 2^{-1}(128K_1)^{-2}\mathcal{W}(\mathcal{F}).$$

Once again, we have a contradiction for any $\mathbf{x}_n \in \mathcal{E}_4$, since we assumed that $l_\xi = (256K_1)^{-3}$.

Therefore, we showed that (5.22), cannot hold under the event $\mathcal{E}_4$, i.e. the set $\mathcal{E}_4$ is empty. This contradicts our earlier conclusion that $\Pr(\mathcal{E}_4) \geq 0.4$, which was made under the assumption that event $\mathcal{A}$ has probability at least 0.5. Hence, we conclude that $\Pr(\mathcal{A}) \leq 0.5$, or, equivalently, with probability at least 0.5, $\hat{f}_n \notin B(f^*, t_*)$. Therefore, we must have that

$$\begin{aligned}\mathbb{E}\int(\widehat{f}_n - f^*)^2 d\mathbb{P} \geq \frac{t_*^2}{2} &= \frac{1}{2}\min\{t_{n,\mathbb{P}}(f^*,\mathcal{F})^2, s_1^2\mathcal{W}(\mathcal{F}), s_2^2\mathcal{W}(\mathcal{F})^2\} \\ &\geq c_1 \cdot \min\{t_{n,\mathbb{P}}(f^*,\mathcal{F})^2, \mathcal{W}(\mathcal{F})^2\}.\end{aligned}$$

where in the last inequality, we used the fact that

$$\mathcal{W}(\mathcal{F}) \leq \mathcal{W}([-1,1]^{\mathcal{X}}) \leq \mathbb{E}|\xi| \leq \sqrt{\mathbb{E}\xi^2} = 1.$$

The theorem follows. □

5.4 Lemmas

Lemma 41. *Under the event $\mathcal{E}$ in* (5.13)*, and for n that is large enough, the following holds:*

$$0 \geq 2^{-1}K_1^{-1}\mathcal{W}(\mathcal{F}) - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - \mathbb{E}_\xi \max_{g\in\mathcal{M}}\langle g - f^*, \hat{f}_n - f^*\rangle_n - 16\sqrt{\mathcal{W}(\mathcal{F})\widehat{r}}, \qquad (5.23)$$

where K_1 is defined in (5.13)*, and set $\mathcal{M}$ is defined in* (5.11).

Lemma 42. *Let $X_1, \ldots, X_n \underset{i.i.d}{\sim} \mathbb{P}$, then the following holds with probability of at least $1 - K\exp(-n\mathcal{W}(\mathcal{F}))$*

$$\sup_{g\in\mathcal{M}, h\in B(f^*,t_*)} |\langle h - f^*, g - f^*\rangle_n - \int_{\mathcal{X}}(h - f^*)(g - f^*)d\mathbb{P}| \leq (8K_1)^{-1}\mathcal{W}(\mathcal{F}).$$

where $\mathcal{M}$ is defined in (5.11)*, t_* is defined in* (5.10)*, and K_1, K are defined in* (5.13)*, Lemma 46.*

Lemma 43. *The event $\mathcal{E}$ defined in* (5.13) *holds with probability of at least* 0.9.

5.4.1 Auxiliary Lemmas

Lemma 44. *[[Kol11, pgs. 25-26]] Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ be family of functions. Then with probability of at least $1 - 2\exp(-c_1 n \mathcal{W}(\mathcal{F}))$,*

$$\forall f, g \in \mathcal{F} \quad \left| \|f-g\|_n^2 - \|f-g\|^2 \right| \leq 10\mathcal{W}(\mathcal{F}),$$

and

$$\|\widehat{f}_n - f^*\|_n^2 \leq 10\mathcal{W}(\mathcal{F}).$$

Lemma 45 (Sudakov's minoration lemma)**.** *Let $\mathcal{H} \subset [-1,1]^{\mathcal{X}}$. There exists a constant c_1 such that for any $\mathbb{P}_n$,*

$$c_1 \sup_{\epsilon \geq 0} \epsilon \sqrt{\frac{\log \mathcal{M}(\epsilon, \mathcal{H}, \mathbb{P}_n)}{n}} \leq \mathcal{W}_{\mathbf{x}}(\mathcal{H}).$$

where $\mathcal{M}(\epsilon, \mathcal{H}, \mathbb{P}_n)$ denotes the size of the largest ϵ-separated set in $\mathcal{H}$ with respect to $L_2(\mathbb{P}_n)$.

The next two lemmas appear in [Kol11, pgs. 24-25], [Ada08].

Lemma 46 (Talagrand's inequality)**.** *Let $X_1, \ldots, X_n \underset{i.i.d.}{\sim} \mathbb{P}$, and $\mathcal{H} \subseteq [-U,U]^{\mathcal{X}}$ be a family of functions. Let $Z = \sup_{f \in \mathcal{H}} |n^{-1} \sum_{i=1}^n f(X_i) - \mathbb{E}[f]|$. Then there exists an absolute constant $K \geq 0$ such that for any $s \geq 0$*

$$\Pr\left(|Z - \mathbb{E}Z| \geq s\right) \leq K \exp\left(-K^{-1} U^{-1} \log(1 + \frac{sU}{V^2}) ns\right),$$

where $V^2 = \sup_{f \in \mathcal{H}} \int f^2 d\mathbb{P}$.

Lemma 47 (Adamczak's inequality)**.** *Let $\mathcal{G}$ be a centred family of functions supported on $\mathcal{D}$, and $\mathbb{Q}$ be some distribution on $\mathcal{D}$. Let $Z = \sup_{g \in \mathcal{G}} |n^{-1} \sum_{i=1}^n g(X_i)|$. Assume that there exists an envelope function G such that $|g(x)| \leq G(x)$ for all $g \in \mathcal{G}, x \in \mathcal{D}$. Then, the following holds for all $t \geq 0$*

$$K_2^{-1} \mathbb{E}Z - V\sqrt{\frac{t}{n}} - \frac{\|\max_{1 \leq i \leq n} f'(X_i)\|_{\psi_1} t}{n} \leq Z \leq K_2 \mathbb{E}Z + V\sqrt{\frac{t}{n}} + \frac{\|\max_{1 \leq i \leq n} f(X_i)\|_{\psi_1} t}{n},$$

where $V^2 := \sup_{g \in \mathcal{G}} \int g^2 d\mathbb{Q}$, and $K_2 \in (1, \infty)$ is some universal constant, and ψ_1 is the Orlicz norm.

Lemma 48 (Lipschitz Concentration)**.** *Let $\xi_1, \ldots, \xi_n \underset{i.i.d}{\sim} N(0,1)$, and $f : \mathbb{R}^n \to \mathbb{R}$ be a L-Lipschitz function with respect to $\| \cdot \|_2$. Then, for all $\epsilon > 0$,*

$$\Pr(|f - \mathbb{E}[f]| \geq \epsilon) \leq \exp(-c\epsilon^2 L^{-2}).$$

5.5 Proofs

Proof of Lemma 41.

We invoke the lower bound of (5.8) with $g = f^*$ and $t = 2$, implying

$$\begin{aligned} 0 &\geq \mathcal{W}_{\mathbf{x}}(B_n(f^*, 2)) - \mathcal{W}_{\mathbf{x}}(B_n(f^*, \widehat{r})) - \mathbb{E}\langle \widehat{f}_n - f^*, g_\xi - f^* \rangle_n - Cn^{-1} \\ &= \mathcal{W}_{\mathbf{x}}(\mathcal{F}) - \mathcal{W}_{\mathbf{x}}(B_n(f^*, \widehat{r})) - \mathbb{E}\langle \widehat{f}_n - f^*, g_\xi - \Pi(g_\xi) + \Pi(g_\xi) - f^* \rangle_n - Cn^{-1}, \end{aligned} \tag{5.24}$$

where $\Pi(g_\xi) := \operatorname{argmin}_{g \in \mathcal{M}} \|g_\xi - g\|_n$, and the equality follows for the fact that for $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ we have $B_n(f^*, 2) = \mathcal{F}$.

Now, recall that $\mathcal{M}$ is a maximal $6\sqrt{\mathcal{W}(\mathcal{F})}$-separated set with respect to $L_2(\mathbb{P})$, and therefore also a $12\sqrt{\mathcal{W}(\mathcal{F})}$-net with respect to $L_2(\mathbb{P})$. Therefore, under the event $\mathcal{E}$ it is also a $16\sqrt{\mathcal{W}(\mathcal{F})}$-net with respect to $L_2(\mathbb{P}_n)$, and, in particular, $\|\Pi(g_\xi) - g_\xi\|_n \leq 16\sqrt{\mathcal{W}(\mathcal{F})}$. Hence, we can rewrite (5.24) as

$$\begin{aligned} &\mathbb{E}_\xi \max_{g \in \mathcal{M}} \langle \widehat{f}_n - f^*, g - f^* \rangle_n \\ &\geq \mathcal{W}_{\mathbf{x}}(\mathcal{F}) - \mathcal{W}_{\mathbf{x}}(B_n(f^*, \widehat{r})) - \mathbb{E}_\xi \langle \widehat{f}_n - f^*, g_\xi - \Pi(g_\xi) \rangle_n - Cn^{-1} \\ &\geq K_1^{-1}\mathcal{W}(\mathcal{F}) - \mathcal{W}_{\mathbf{x}}(B_n(f^*, \widehat{r})) - 16\sqrt{\mathcal{W}(\mathcal{F})}\mathbb{E}_\xi \|f - \widehat{f}_n\|_n - Cn^{-1} \end{aligned}$$

Now, we proceed by using the first part of Corollary 2 and the assumption of lying in $\mathcal{E}$. The last expression is lower-bounded by

$$K_1^{-1}\mathcal{W}(\mathcal{F}) - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - 16\sqrt{\mathcal{W}(\mathcal{F})}(\widehat{r} + C\widehat{r}^{1/2}n^{-1/4}). \tag{5.25}$$

According to (5.20), under the event $\mathcal{E}$, we have

$$\widehat{r} \le C_3\sqrt{K_1\mathcal{W}(\mathcal{F})}$$

for some constant C_3. Thus the expression in (5.25) is further lower-bounded by

$$\begin{aligned} &K_1^{-1}\mathcal{W}(\mathcal{F}) - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - 16\sqrt{\mathcal{W}(\mathcal{F})}\widehat{r} - C_4\sqrt{K_1}\mathcal{W}(\mathcal{F})^{3/4}n^{-1/4} - Cn^{-1} \\ &\ge (2K_1)^{-1}\mathcal{W}(\mathcal{F}) - \mathcal{W}_{\mathbf{x}}(B_n(f^*,\widehat{r})) - 16\sqrt{\mathcal{W}(\mathcal{F})}\widehat{r} \end{aligned}$$

where the last inequality holds when n is large enough. To see this, recall that $\mathcal{W}(\mathcal{F}) \ge c/\sqrt{n}$ and under this assumption both $n^{-1} = o_n(\mathcal{W}(\mathcal{F}))$ and $\mathcal{W}(\mathcal{F})^{3/4}n^{-1/4} = o_n(\mathcal{W}(\mathcal{F}))$ hold. Therefore, the lemma follows. □

Proof of Lemma 42. First, denote by $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{H}} := \sup_{h\in\mathcal{H}} |n^{-1}\sum_{i=1}^n h(X_i) - \mathbb{E}[h]|$, and for each $g_i \in \mathcal{M}$, define $\mathcal{G}_i = \{(h-f^*)(g_i - f^*) : h \in B(f^*,t_*)\}$. By Talagrand's inequality (Lemma 46), the following holds for and $u \ge 0$

$$\Pr\left(|\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i}| \ge u\right) \le K\exp\left(-nK^{-1}\log(1+\frac{4^{-1}us_1^{-2}}{\mathcal{W}(\mathcal{F})})u\right)$$

where we used the fact that $V^2 \le \sup_{h\in\mathcal{F}}\|g-f^*\|_\infty^2 t_*^2 \le 4s_1^2\mathcal{W}(\mathcal{F})$. Now, we set $u = (16K_1)^{-1}\mathcal{W}(\mathcal{F})$ in the last equation

$$\begin{aligned} &\Pr\left(|\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i}| \ge (16K_1)^{-1}\mathcal{W}(\mathcal{F})\right) \\ &\le K\exp\left(-n(16K\cdot K_1)^{-1}\mathcal{W}(\mathcal{F})\log(1+4^{-1}(16K_1)^{-1}s_1^{-2})\right). \end{aligned}$$

Next, we aim to take a union bound over $\mathcal{M}$, and recall that $\log M \leq C_2 K_1^2 n\mathcal{W}(\mathcal{F}) \leq C(K_1)n\mathcal{W}(\mathcal{F})$, for some absolute constant that does not depend on s_1. Therefore, we may choose

$$s_1 := c(K, K_1, C_2) \tag{5.26}$$

where $c(K, K_1, C_2)$ is a constant that satisfies the following:

$$\Pr\Big(| \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} | \geq (16K_1)^{-1}\mathcal{W}(\mathcal{F}) \Big) \leq K\exp(-2C_2K_1 n\mathcal{W}(\mathcal{F})).$$

Therefore, we have

$$\begin{aligned} \Pr\Big(\exists 1 \leq i \leq M : | \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} | \geq (16K_1)^{-1}\mathcal{W}(\mathcal{F}) \Big) &\leq MK\exp(-2C_2K_1 n\mathcal{W}(\mathcal{F}))) \\ &\leq K\exp(-C_2K_1 n\mathcal{W}(\mathcal{F})) \\ &\leq K\exp(-n\mathcal{W}(\mathcal{F})). \end{aligned}$$

We conclude that with probability of at least $1 - K\exp(-n\mathcal{W}(\mathcal{F}))$ the following holds for $\mathcal{G} := \{(h - f^*)(g - f^*) : g \in \mathcal{M}, h \in B(f^*, t_*)\}$:

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}} \leq \max_{1 \leq i \leq M} \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} + (16K_1)^{-1}\mathcal{W}(\mathcal{F}). \tag{5.27}$$

The lemma will follow as soon as we show that

$$\max_{1 \leq i \leq M} \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} \leq (16K_1)^{-1}\mathcal{W}(\mathcal{F}).$$

In order to prove the last inequality, we first apply the symmetrization lemma (cf. [Kol11, p. 20]) and majorize the resulting Rademacher averages by a constant multiple of the Gaussian averages

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} \leq 8\mathcal{W}(\mathcal{G}_i) \tag{5.28}$$

where we used the fact that $0 \in \mathcal{G}_i$.

Next, since $\|g_i - f^*\|_\infty \le 2$, a standard argument (e.g. [GN16, Theorem 3.1.17]) gives

$$\mathbb{E}_\xi \sup_{h \in B(f^*,t_*)} n^{-1} \sum_{k=1}^{n} (h - f^*)(g_i - f^*)(X_k)\xi_k \le 2\mathbb{E}_\xi \sup_{h \in B(f^*,t_*)} n^{-1} \sum_{k=1}^{n} (h - f^*)(X_i)\xi_k.$$

Then, by taking expectation over $X_1, \dots, X_n$ over the last equation and by (5.28), we conclude

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} \le 16\mathcal{W}(B(f^*, t_*)) \le 16 l_\xi \mathcal{W}(\mathcal{F}) \le (16K_1)^{-1}\mathcal{W}(\mathcal{F}),$$

where we set $l_\xi = (256K_1)^{-3}$. Then, by (5.27) and the last equation, the claim follows. □

Proof of Lemma 43. It is enough to show that $\mathcal{E}_2$, $\mathcal{E}_3$ hold with probability of at least 0.99 for n large enough. First, we prove this claim for $\mathcal{E}_2$.

We aim to apply Adamczak bound for concentration of the suprema of unbounded empirical processes (Lemma 47). For this purpose, define the family of functions $\mathcal{G} := \{yf(x), y \in \mathbb{R}, f \in \mathcal{F} - f^*\}$, and the distribution $\mathbb{Q} = \mathbb{P} \otimes N(0,1)$. Note that $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ and, ξ is Gaussian. Therefore, by Pisier's inequality (cf. [Pis83],[Ada08, p. 13]), we have

$$\| \max_{1 \le i \le n} |\xi_i f(X_i)| \|_{\psi_1} \le C \log(n) \max_{1 \le i \le n} \| |\xi_i f(X_i)| \|_{\psi_1} \le C_2 \log(n).$$

By Adamczak's bound (Lemma 47),

$$\begin{aligned} &K_2^{-1}\mathbb{E}_{\mathcal{D}} \sup_{f \in \mathcal{F}-f^*} |\frac{1}{n}\sum_{i=1}^{n} f(X_i)\xi_i| - \frac{10}{\sqrt{n}} - \frac{C\log(n)}{n} \\ &\le \sup_{f \in \mathcal{F}-f^*} |\frac{1}{n}\sum_{i=1}^{n} f(X_i)\xi_i| \le K_2\mathbb{E}_{\mathcal{D}} \sup_{f \in \mathcal{F}-f^*} |\frac{1}{n}\sum_{i=1}^{n} f(X_i)\xi_i| + \frac{10}{\sqrt{n}} + \frac{C\log(n)}{n}, \end{aligned} \tag{5.29}$$

with probability of at least 0.99 both $X_1, \dots, X_n$ and $\overline{\xi}$.

Now, using the average principle, for n large enough, we can find an event $\mathcal{E}_7$ (that depends only on $X_1, \dots, X_n$) that holds with probability 0.98, such that for any fixed $\mathbf{x}_n \in \mathcal{E}_7$, there exists an event $\mathcal{A}_3(\mathbf{x}_n)$ of probability at least 0.98 (over $\overline{\xi}$) such that (5.29) holds. For each $\mathbf{x}_n \in \mathcal{E}_7$, Lemma 48 (with Lipschitz constant $\sup_{f \in \mathcal{F}} \|f - f^*\|_n \le 2$) implies that the middle term in (5.29) is, with high probability, within $Cn^{-1/2}$ from its expectation (with respect to $\overline{\xi}$). Therefore, we have for all $\mathbf{x}_n \in \mathcal{E}_7$:

$$\begin{aligned}
&K_2^{-1}\mathbb{E}_{\mathcal{D}} \sup_{f\in\mathcal{F}-f^*} |n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i| - \frac{C}{\sqrt{n}} \\
&\le \mathbb{E}_{\xi} \sup_{f\in\mathcal{F}-f^*} |n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i| \le K_2\mathbb{E}_{\mathcal{D}} \sup_{f\in\mathcal{F}-f^*} |n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i| + \frac{C}{\sqrt{n}}.
\end{aligned}$$

Finally, since $0 \in \mathcal{F} - f^*$, we have

$$\mathbb{E}_{\xi} \sup_{f\in\mathcal{F}-f^*} n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i \le \mathbb{E}_{\xi} \sup_{f\in\mathcal{F}-f^*} |n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i| \le 2\mathbb{E}_{\xi} \sup_{f\in\mathcal{F}-f^*} n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i.$$

Hence, the last two equations imply that when $\mathcal{W}(\mathcal{F}) \ge C_1 n^{-1/2}$, for C_1 that is large enough, the claim follows for $\mathcal{E}_2$. To handle the remaining case of $\mathcal{W}(\mathcal{F}) \le C_1 n^{-1/2}$, recall that we assumed that our class is not degenerate (i.e it has two functions that are $\|f_1 - f_2\| \ge 0.5$. Then, it is easy to see that with probability of 0.99 it holds that

$$\mathcal{W}_{\mathbf{x}}(\mathcal{F} - f^*) \ge \mathcal{W}(\{0, f_2 - f^*, f_1 - f^*\}) \ge \mathbb{E}\max\{n^{-0.5}g, 0\} \ge c \cdot n^{-1/2} \ge c \cdot C_1^{-1}\mathcal{W}(\mathcal{F}),$$

where $g \sim N(0, 1/4)$. Therefore, for some $K_1^{-1} = c(K_2, c)$, the claim follows for $\mathcal{E}_2$.

Next, we handle $\mathcal{E}_3$. By using the definition of $B(f^*, t_*)$, and similar considerations that led to (5.29), we have

$$\sup_{f\in B(f^*,t_*)-f^*} |n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i| \le K_2\mathbb{E}_{\mathcal{D}} \sup_{B(f^*,t_*)-f^*} |n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i| + \frac{10t_*}{\sqrt{n}} + \frac{C\log(n)}{n}, \tag{5.30}$$

with probability of at least 0.99 over both $X_1, \ldots, X_n$ and $\overline{\xi}$.

As above, for n large enough, we can find an event $\mathcal{E}_8 \subseteq \mathcal{E}_1$ (where $\mathcal{E}_1$ is defined in (5.12)) of probability at least 0.98 (over $X_1, \ldots, X_n$), such that for any $\mathbf{x}_n \in \mathcal{E}_8$, there exists an event $\mathcal{A}_4(\mathbf{x}_n)$ of probability at least 0.98 (over $\overline{\xi}$) such that (5.30) holds. Then, similarly to the case of $\mathcal{E}_2$, we will employ Lipschitz concentration for the middle term in (5.30), for each $\mathbf{x}_n \in \mathcal{E}_8$. To estimate the

Lipschitz constant, recall that under $\mathcal{E}_1$ (more precisely, under the event of Lemma 44), we also have that

$$\|f - f^*\|_n^2 \leq s_1^2 \mathcal{W}(\mathcal{F}) + 10\mathcal{W}(\mathcal{F}) \leq 11\mathcal{W}(\mathcal{F})$$

for all $f \in B(f^*, t_*)$, under the choice t_* in (5.10). Then, using the fact that $\mathcal{A}_4(\mathbf{x}_n)$ holds with probability of at least 0.98, and Lemma 48 with Lipschitz constant $\sup_{f \in B(f^*,t_*)} \|f - f^*\|_n \leq \sqrt{11\mathcal{W}(\mathcal{F})}$, imply that for each $\mathbf{x}_n \in \mathcal{E}_8$, the middle term in (5.30) is within an additive factor of $C_1\sqrt{\mathcal{W}(\mathcal{F})}n^{-1/2}$ from its expectation over $\overline{\xi}$. Namely, we have for all $\mathbf{x}_n \in \mathcal{E}_8$:

$$\mathbb{E}_{\xi} \sup_{f \in \mathcal{F} - f^*} |n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i| \leq K_2 \mathbb{E}_{\mathcal{D}} \sup_{f \in \mathcal{F} - f^*} |n^{-1}\sum_{i=1}^{n} f(X_i)\xi_i| + \frac{C\sqrt{\mathcal{W}(\mathcal{F})}}{\sqrt{n}}.$$

where we used the fact that $t_* \leq s_1\sqrt{\mathcal{W}(\mathcal{F})}$. The claim for $\mathcal{E}_3$ follows by similar considerations that we used earlier. □

Proof of Corollary 7. For any $\mathbb{P}$-Donsker class we have with probability at least 0.9 [Gee00, Chap. 5]

$$\mathcal{W}_{\mathbf{x}}(\mathcal{F}) \asymp \mathcal{W}(\mathcal{F}) \asymp n^{-1/2}.$$

Then, by Corollary 5, we have that

$$\mathbb{E}\int (\widehat{f}_n - f^*)^2 d\mathbb{P}_n \gtrsim n^{-1}.$$

In order to prove the second part of the bound, we apply Theorem 22,

$$\mathbb{E}\int (\widehat{f}_n - f^*)^2 d\mathbb{P} \gtrsim \max\{n^{-1}, t_{n,\mathbb{P}}(f^*, \mathcal{F})^2\}.$$

The corollary will follow if we show that for any $f^* \in \mathcal{F}$, we have that $t_{n,\mathbb{P}} \gtrsim 1$. To see this, we use [Gee00, Thm 5.11] that shows that for all $t \geq 0$, we have

$$\mathcal{W}(B(f^*, t)) \lesssim n^{-1/2}\int_0^t u^{-\alpha/2} du \lesssim t^{\frac{-\alpha+2}{2}} n^{-1/2}.$$

Since $\alpha \in (0,2)$, the right hand side is decreasing in t, therefore we know that if

$$\mathcal{W}(B(f^*, t_*)) \gtrsim \mathcal{W}(\mathcal{F}) \gtrsim n^{-1/2}$$

then we have $t_* \gtrsim 1$. Hence, $t_{n,\mathbb{P}}(f^*, \mathcal{F}) \gtrsim 1$, and the claim follows. □

Proof of Corollary 8. For any non $\mathbb{P}$-Donsker class we have with probability of at least 0.9 [Gee00, Chap. 5]

$$n^{-\frac{2}{2+\alpha}} \lesssim \mathcal{W}_{\mathbf{x}}(\mathcal{F}) \asymp \mathcal{W}(\mathcal{F}) \lesssim n^{-\frac{1}{\alpha}}.$$

Then, by Corollary 5, we have that

$$\mathbb{E} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}_n \gtrsim n^{-\frac{4}{2+\alpha}},$$

and the claim follows. □

5.5.1 An example to the tightness of Theorem 22 (a sketch)

Let $\mathbb{P}$ be the uniform density of $[0,1]$, and denote by $I(x_i, l_i)$ to be an interval with center x_i and length l_i. For each $m \geq 0$ we define

$$\mathcal{F}_m := \{m^{-1/6} \sum_{i=1}^{m} \epsilon_i 1_{I(x_i, m^{-5/4})} : \forall x_1, \ldots, x_m \text{ s.t. } 1 \leq j \neq k \leq m \;\; I(x_k, m^{-5/4}) \cap I(x_j, m^{-5/4}) = \emptyset,$$
$$\forall (\epsilon_1, \ldots, \epsilon_m) \in \{-1, 1\}^m\}.$$

Now, we define $\mathcal{F} := \text{conv}\{0, \bigcup_{m=1}^{\infty}\{\mathcal{F}_m\}\}$. Clearly, this family is uniformly bounded by one. Also, we assume that $f^* = 0$.

Using a classical fact, we have that with probability of at least $1 - n^2$,

$$\max_{1 \leq i \neq j \leq n} |X_j - X_i| \geq c \cdot (n \log n)^{-1},$$

and denote this event by A.

Clearly, for each $\mathbf{x}_n \in A$, we can find a function $f_\xi \in \mathcal{F}_n$ (that depends on $\mathbf{x}_n$ as well) such that

$$\langle f_\xi, \overline{\xi} \rangle_n = n^{-1/6} \cdot n^{-1} \sum_{i=1}^{n} |\xi_i|. \tag{5.31}$$

To see this, under the event A, we can place $X_1, \dots, X_n$ in n disjoint intervals of length $n^{-5/4}$, therefore f_ξ is the function that is supported on these intervals, with the suitable signs.

Clearly, under the event A, $\widehat{f}_n \notin \bigcup_{m=n+1}^{\infty} \{\mathcal{F}_m\}$, since there is no "advantage" to more than n intervals. Therefore, one can easily show that

$$\mathcal{W}_{\mathbf{x}}(B_n(f^*, n^{-1/6})) \asymp n^{-1/6}.$$

Now, denote by $C(n) := C_1(n \log(n))^{4/5}$ for C_1 that is large enough. Note that any $\mathcal{F}_m$ such that $C(n) \leq m \leq n-1$, we can only place m intervals with length of at most $(c/2) \cdot (n \log(n))^{-1}$. Therefore, under the event A, each of these intervals has at most one point. Hence, we have that

$$\max_{f_m \in \mathcal{F}_m} \langle f_m, \overline{\xi} \rangle_n = m^{-1/6} n^{-1} \max_{S \in \binom{n}{m}, |S| = m} \sum_{i \in S} |\xi_i|.$$

Now, for any fixed $C(n) \leq m \leq n-1$, by standard concentration inequalities,

$$\mathbb{E} \max_{f_m \in \mathcal{F}_m} \langle f_m, \overline{\xi} \rangle_n \lesssim m^{-1/6}(m/n) + m^{-1/6}(m/n)\sqrt{\log(n/m)} \lesssim m^{-1/6}(m/n)\sqrt{\log(n/m)}. \tag{5.32}$$

In the remaining case of $m \leq C(n)$, using standard arguments, it can be shown that with probability of at least $1 - n^2$ (over $X_1, \dots, X_n$) the following holds:

$$\mathbb{E}_\xi \sup_{f_m \in \mathcal{F}_m} \langle f_m, \overline{\xi} \rangle_n \asymp \mathbb{E}_\mathcal{D} \sup_{f_m \in \mathcal{F}_m} \langle f_m, \overline{\xi} \rangle_n \ll n^{-1/6}. \tag{5.33}$$

By using Eqs. (5.31),(5.32),(5.33), one can show that with high probability $\widehat{f}_n \in B_n(f^*, Cn^{-1/6})$, for some $C \geq 0$, and also

$$\mathcal{W}(\mathcal{F}) \asymp n^{-1/6}.$$

Therefore, one can conclude that

$$\mathbb{E} \int (\widehat{f}_n - f^*)^2 d\mathbb{P}_n \asymp n^{-\frac{1}{3}} \asymp \mathcal{W}(\mathcal{F})^2,$$

and

$$\mathbb{E} \int (\widehat{f}_n - f^*)^2 d\mathbb{P} \asymp n^{-(\frac{1}{4}+\frac{1}{3})} \ll \mathcal{W}(\mathcal{F})^2 \asymp n^{-\frac{1}{3}}.$$

Finally, it is easy to see that $t_{n,\mathbb{P}}(f^*, \mathcal{F}) \gtrsim n^{-(\frac{1}{8}+\frac{1}{6})}$, and therefore, by using the last equation

$$\mathbb{E} \int (\widehat{f}_n - f^*)^2 d\mathbb{P} \asymp t_{n,\mathbb{P}}(f^*, \mathcal{F})^2,$$

and the claim follows.

Bibliography

[AAGM15] S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman. *Asymptotic geometric analysis, Part I*. Vol. 202. Mathematical Surveys and Monographs; v. 202. American Mathematical Society, 2015.

[AB06] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Berlin; London: Springer, 2006. ISBN: 9783540326960 3540326960.

[Ada08] R. Adamczak. "A tail inequality for suprema of unbounded empirical processes with applications to Markov chains". In: *Electronic Journal of Probability* 13 (2008), pp. 1000–1034.

[AS16] N. Alon and J. H. Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

[Bah58] R. Bahadur. "Examples of inconsistency of maximum likelihood estimates". In: *Sankhyā: The Indian Journal of Statistics* (1958), pp. 207–210.

[Bal16] G. Balázs. "Convex regression: theory, practice, and applications". In: (2016).

[Bal22] G. Balázs. "Adaptively Partitioning Max-Affine Estimators for Convex Regression". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 860–874.

[Bar+20] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070.

[BBM05] P. L. Bartlett, O. Bousquet, and S. Mendelson. "Local rademacher complexities". In: *The Annals of Statistics* 33.4 (2005), pp. 1497–1537.

[BBV04] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Bel17] P. C. Bellec. “Optimistic lower bounds for convex regularized least-squares”. In: *arXiv preprint arXiv:1703.01332* (2017).

[BGS15] G. Balázs, A. György, and C. Szepesvári. “Near-optimal max-affine estimators for convex regression”. In: *Artificial Intelligence and Statistics*. PMLR. 2015, pp. 56–64.

[BHM18] M. Belkin, D. Hsu, and P. Mitra. “Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate”. In: *arXiv preprint arXiv:1806.05161* (2018).

[Bir06] L. Birgé. “Model selection via testing: an alternative to (penalized) maximum likelihood estimators”. In: *Annales de l’IHP Probabilités et statistiques*. Vol. 42. 3. 2006, pp. 273–325.

[BL88] I. Bárány and D. G. Larman. “Convex bodies, economic cap coverings, random polytopes”. In: *Mathematika* 35.2 (1988), 274–291.

[Bla+19] J. Blanchet, P. W. Glynn, J. Yan, and Z. Zhou. “Multivariate distributionally robust convex regression under absolute error loss”. In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 11817–11826.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[BM21] D. Bertsimas and N. Mundru. “Sparse convex regression”. In: *INFORMS Journal on Computing* 33.1 (2021), pp. 262–279.

[BM93] L. Birgé and P. Massart. “Rates of convergence for minimum contrast estimators”. In: *Probability Theory and Related Fields* 97.1-2 (1993), pp. 113–150.

[BM98] L. Birgé and P. Massart. “Minimum contrast estimators on sieves: exponential bounds and rates of convergence”. In: *Bernoulli* 4.3 (1998), pp. 329–375.

[Bou02] O. Bousquet. “Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms”. In: (2002).

[Bro76] E. Bronshtein. “ε-entropy of convex sets and functions”. In: *Siberian Mathematical Journal* 17.3 (1976), pp. 393–398.

[BRT19] M. Belkin, A. Rakhlin, and A. B. Tsybakov. “Does data interpolation contradict statistical optimality?” In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1611–1619.

[Bru13] V.-E. Brunel. “Adaptive estimation of convex polytopes and convex sets from noisy data”. In: *Electronic Journal of Statistics* 7 (2013), pp. 1301–1327.

[Bru16] V.-E. Brunel. “Adaptive estimation of convex and polytopal density support”. In: *Probability Theory and Related Fields* 164.1-2 (2016), pp. 1–16.

[BS23] S. Bubeck and M. Sellke. “A universal law of robustness via isoperimetry”. In: *Journal of the ACM* 70.2 (2023), pp. 1–18.

[BTN01] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[BW03] K. Böröczky and G. Wintsche. “Covering the sphere by equal spherical balls”. In: *Discrete and computational geometry*. Springer, 2003, pp. 235–251.

[Bá89] I. Bárány. “Intrinsic volumes and f-vectors of random polytopes.” In: *Mathematische Annalen* 285.4 (1989), pp. 671–699.

[Car+18] T. Carpenter, I. Diakonikolas, A. Sidiropoulos, and A. Stewart. “Near-Optimal Sample Complexity Bounds for Maximum Likelihood Estimation of Multivariate Log-concave Densities”. In: *Conference On Learning Theory*. 2018, pp. 1234–1262.

[CGZ17] X. Chen, A. Guntuboyina, and Y. Zhang. “A note on the approximate admissibility of regularized estimators in the Gaussian sequence model”. In: (2017).

[Cha14] S. Chatterjee. “A new perspective on least squares under convex constraint”. In: *The Annals of Statistics* 42.6 (2014), pp. 2340–2381.

[CM20] W. Chen and R. Mazumder. “Multivariate convex regression at scale”. In: *arXiv preprint arXiv:2005.11588* (2020).

[CMS21] W. Chen, R. Mazumder, and R. J. Samworth. “A new computational framework for log-concave density estimation”. In: *arXiv preprint arXiv:2105.11387* (2021).

[CR06] A. Caponnetto and A. Rakhlin. “Stability Properties of Empirical Risk Minimization over Donsker Classes.” In: *Journal of Machine Learning Research* 7.12 (2006).

[CS20] E. J. Candès and P. Sur. “The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression”. In: *The Annals of Statistics* 48.1 (2020), pp. 27–42.

[CSS10] M. Cule, R. Samworth, and M. Stewart. “Maximum likelihood estimation of a multidimensional log-concave density”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.5 (2010), pp. 545–607.

[Dev+16] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. “Sub-Gaussian mean estimators”. In: *The Annals of Statistics* 44.6 (2016), pp. 2695–2725.

[Dev86] L. Devroye. *Non-uniform random variate generation*. Springer-Verlag, 1986.

[DKS16] I. Diakonikolas, D. M. Kane, and A. Stewart. “Efficient robust proper learning of log-concave distributions”. In: *arXiv preprint arXiv:1606.03077* (2016).

[DL12] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[DL19] J. Depersin and G. Lecué. “Robust subgaussian estimation of a mean vector in nearly linear time”. In: *arXiv preprint arXiv:1906.03058* (2019).

[DSS18] I. Diakonikolas, A. Sidiropoulos, and A. Stewart. “A Polynomial Time Algorithm for Maximum Likelihood Estimation of Multivariate Log-concave Densities”. In: *arXiv preprint arXiv:1812.05524* (2018).

[Dud99] R. M. Dudley. *Uniform central limit theorems*. 63. Cambridge university press, 1999.

[Dwy88] R. A. Dwyer. "On the convex hull of random points in a polytope". In: *Journal of Applied Probability* 25.4 (1988), pp. 688–699.

[Fen+18] O. Y. Feng, A. Guntuboyina, A. K. Kim, and R. J. Samworth. "Adaptation in multivariate log-concave density estimation". In: *arXiv preprint arXiv:1812.11634* (2018).

[Fer82] T. S. Ferguson. "An Inconsistent Maximum Likelihood Estimate". In: *Journal of the American Statistical Association* 77.380 (1982), pp. 831–834. ISSN: 01621459.

[Fis+97] N. I. Fisher, P. Hall, B. A. Turlach, and G. Watson. "On the estimation of a convex set from noisy data on its support function". In: *Journal of the American Statistical Association* 92.437 (1997), pp. 84–91.

[Gar95] R. J. Gardner. *Geometric tomography*. Vol. 6. Cambridge University Press Cambridge, 1995.

[Gee00] S. A. van de Geer. *Empirical Processes in M-estimation*. Vol. 6. Cambridge university press, 2000.

[Gho+21] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran. "Max-affine regression: Parameter estimation for Gaussian designs". In: *IEEE Transactions on Information Theory* (2021).

[GKM06] R. J. Gardner, M. Kiderlen, and P. Milanfar. "Convergence of algorithms for reconstructing convex bodies and directional measures". In: *The Annals of Statistics* 34.3 (2006), pp. 1331–1374.

[GN16] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. 40. Cambridge University Press, 2016.

[GS13] A. Guntuboyina and B. Sen. "Covering numbers for convex functions". In: *IEEE Transactions on Information Theory* 59.4 (2013), pp. 1957–1965.

[Gun12] A. Guntuboyina. "Optimal rates of convergence for convex set estimation from support functions". In: *The Annals of Statistics* 40.1 (2012), pp. 385–411.

[GW17] F. Gao and J. A. Wellner. “Entropy of Convex Functions on $\mathbb{R}^d$”. In: *Constructive approximation* 46.3 (2017), pp. 565–592.

[Gyö+02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Vol. 1. Springer, 2002.

[Han19] Q. Han. “Global empirical risk minimizers with “shape constraints" are rate optimal in general dimensions”. In: *arXiv preprint arXiv:1905.12823* (2019).

[Han+19] Q. Han, T. Wang, S. Chatterjee, and R. J. Samworth. “Isotonic regression in general dimensions”. In: *The Annals of Statistics* 47.5 (2019), pp. 2440–2471.

[Has+22] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. “Surprises in high-dimensional ridgeless least squares interpolation”. In: *The Annals of Statistics* 50.2 (2022), pp. 949–986.

[HD12] L. Hannah and D. Dunson. “Ensemble methods for convex regression with applications to geometric programming based circuit design”. In: *arXiv preprint arXiv:1206.4645* (2012).

[HD13] L. A. Hannah and D. B. Dunson. “Multivariate convex regression with adaptive partitioning”. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 3261–3294.

[HLZ20] S. B. Hopkins, J. Li, and F. Zhang. “Robust and heavy-tailed mean estimation made simple, via regret minimization”. In: *arXiv preprint arXiv:2007.15839* (2020).

[Hop18] S. B. Hopkins. “Sub-gaussian mean estimation in polynomial time”. In: *arXiv preprint arXiv:1809.07425* (2018), p. 120.

[HW16] Q. Han and J. A. Wellner. “Multivariate convex regression: global risk bounds and adaptation”. In: *arXiv preprint arXiv:1601.06844* (2016).

[JM18] A. Javanmard and A. Montanari. “Debiasing the lasso: Optimal sample size for gaussian designs”. In: *The Annals of Statistics* 46.6A (2018), pp. 2593–2622.

[KDR19] G. Kur, Y. Dagan, and A. Rakhlin. “Optimality of Maximum Likelihood for Log-Concave Density Estimation and Bounded Convex Regression”. In: *arXiv preprint arXiv:1903.05315* (2019).

[KGS18] A. K. Kim, A. Guntuboyina, and R. J. Samworth. “Adaptation in log-concave density estimation”. In: *The Annals of Statistics* 46.5 (2018), pp. 2279–2306.

[Kid+08] M. Kiderlen et al. “A new algorithm for 3D reconstruction from support functions”. In: *IEEE transactions on pattern analysis and machine intelligence* 31.3 (2008), pp. 556–562.

[KM13] V. Koltchinskii and S. Mendelson. “Bounding the smallest singular value of a random matrix without concentration”. In: *arXiv preprint arXiv:1312.3580* (2013).

[Kol11] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d’Eté de Probabilités de Saint-Flour XXXVIII-2008*. Vol. 2033. Springer Science & Business Media, 2011.

[KP22] G. Kur and E. Putterman. “An Efficient Minimax Optimal Estimator For Multivariate Convex Regression”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 1510–1546.

[KPR23] G. Kur, E. Putterman, and A. Rakhlin. “On the Variance, Admissibility, and Stability of Empirical Risk Minimization”. In: *arXiv preprint arXiv:2305.18508* (2023).

[KR21] G. Kur and A. Rakhlin. “On the Minimal Error of Empirical Risk Minimization”. In: *Conference on Learning Theory*. PMLR. 2021, pp. 2849–2852.

[KRG20] G. Kur, A. Rakhlin, and A. Guntuboyina. “On suboptimality of least squares with application to estimation of convex bodies”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 2406–2424.

[KS16] A. K. Kim and R. J. Samworth. “Global rates of convergence in log-concave density estimation”. In: *The Annals of Statistics* 44.6 (2016), pp. 2756–2779.

[Kur+20] G. Kur, F. Gao, A. Guntuboyina, and B. Sen. "Convex regression in multidimensions: Suboptimality of least squares estimators". In: *arXiv preprint arXiv:2006.02044* (2020).

[LeC73] L. LeCam. "Convergence of estimates under dimensionality restrictions". In: *The Annals of Statistics* (1973), pp. 38–53.

[Led01] M. Ledoux. *The concentration of measure phenomenon*. 89. American Mathematical Soc., 2001.

[LKW92] A. S. Lele, S. R. Kulkarni, and A. S. Willsky. "Convex-polygon estimation from support-line measurements and applications to target reconstruction from laser-radar data". In: *JOSA A* 9.10 (1992), pp. 1693–1714.

[LM13] G. Lecué and S. Mendelson. "Learning subgaussian classes: Upper and minimax bounds". In: *arXiv preprint arXiv:1305.4825* (2013).

[LM19] G. Lugosi and S. Mendelson. "Sub-Gaussian estimators of the mean of a random vector". In: *The Annals of Statistics* 47.2 (2019), pp. 783–794.

[LR20] T. Liang and A. Rakhlin. "Just interpolate: Kernel "ridgeless" regression can generalize". In: *The Annals of Statistics* 48.3 (2020), pp. 1329–1347.

[LRZ20] T. Liang, A. Rakhlin, and X. Zhai. "On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels". In: *Conference on Learning Theory*. PMLR. 2020, pp. 2683–2711.

[LST20] M. Lin, D. Sun, and K.-C. Toh. "Efficient algorithms for multivariate shape-constrained convex regression problems". In: *arXiv preprint arXiv:2002.11410* (2020).

[LSW06] M. Ludwig, C. Schütt, and E. Werner. *Approximation of the Euclidean ball by polytopes*. Citeseer, 2006.

[LV07] L. Lovász and S. Vempala. "The geometry of logconcave functions and sampling algorithms". In: *Random Structures & Algorithms* 30.3 (2007), pp. 307–358.

[Maz+19] R. Mazumder, A. Choudhury, G. Iyengar, and B. Sen. “A computational framework for multivariate convex regression and its variants”. In: *Journal of the American Statistical Association* 114.525 (2019), pp. 318–331.

[McM70] P. McMullen. “The maximum numbers of faces of a convex polytope”. In: *Mathematika* 17.2 (1970).

[Men14] S. Mendelson. “Learning without concentration”. In: *Conference on Learning Theory*. 2014, pp. 25–39.

[Men17] S. Mendelson. “On aggregation for heavy-tailed classes”. In: *Probability Theory and Related Fields* 168.3-4 (2017), pp. 641–674.

[Mil09] E. Milman. “On the role of convexity in isoperimetry, spectral gap and concentration”. In: *Inventiones mathematicae* 177.1 (2009), pp. 1–43.

[MVZ21] J. Mourtada, T. Vaškevičius, and N. Zhivotovskiy. “Distribution-Free Robust Linear Regression”. In: *arXiv preprint arXiv:2102.12919* (2021).

[NS48] J. Neyman and E. L. Scott. “Consistent Estimates Based on Partially Consistent Observations”. In: *Econometrica* 16.1 (1948), pp. 1–32. ISSN: 00129682, 14680262.

[OC21] E. O’Reilly and V. Chandrasekaran. “Spectrahedral Regression”. In: *arXiv preprint arXiv:2110.14779* (2021).

[Pis83] G. Pisier. “Some applications of the metric entropy condition to harmonic analysis”. In: *Banach Spaces, Harmonic Analysis, and Probability Theory*. Springer, 1983, pp. 123–154.

[Pol02] D. Pollard. “Maximal inequalities via bracketing with adaptive truncation”. In: *Annales de l’Institut Henri Poincare (B) Probability and Statistics*. Vol. 38. 6. Elsevier. 2002, pp. 1039–1052.

[PS22] A. Pananjady and R. J. Samworth. “Isotonic regression with unknown permutations: Statistics, computation and adaptation”. In: *The Annals of Statistics* 50.1 (2022), pp. 324–350.

[PT03] W. B. Powell and H. Topaloglu. “Stochastic programming in transportation and logistics”. In: *Handbooks in operations research and management science* 10 (2003), pp. 555–635.

[PW90] J. L. Prince and A. S. Willsky. “Reconstructing convex sets from support line measurements”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.4 (1990), pp. 377–389.

[Roc70] R. T. Rockafellar. *Convex analysis*. 28. Princeton university press, 1970.

[RST17] A. Rakhlin, K. Sridharan, and A. B. Tsybakov. “Empirical entropy, minimax regret and minimax risk”. In: *Bernoulli* 23.2 (2017), pp. 789–824.

[RSW19] M. Reitzner, C. Schuett, and E. Werner. “The convex hull of random points on the boundary of a simple polytope”. In: *arXiv preprint arXiv:1911.05917* (2019).

[Sam18] R. J. Samworth. “Recent progress in log-concave density estimation”. In: *Statistical Science* 33.4 (2018), pp. 493–509.

[SC19a] Y. S. Soh and V. Chandrasekaran. “Fitting tractable convex sets to support function evaluations”. In: *arXiv preprint arXiv:1903.04194* (2019).

[SC19b] P. Sur and E. J. Candès. “A modern maximum-likelihood theory for high-dimensional logistic regression”. In: *Proceedings of the National Academy of Sciences* 116.29 (2019), pp. 14516–14525.

[SC21] Y. S. Soh and V. Chandrasekaran. “Fitting tractable convex sets to support function evaluations”. In: *Discrete & Computational Geometry* (2021), pp. 1–42.

[Sch14] R. Schneider. *Convex bodies: the Brunn–Minkowski theory*. 151. Cambridge university press, 2014.

[Sia+21] A. Siahkamari, D. A. E. Acar, C. Liao, K. Geyer, V. Saligrama, and B. Kulis. “Faster Convex Lipschitz Regression via 2-block ADMM”. In: *arXiv preprint arXiv:2111.01348* (2021).

[Sim+18] M. Simchowitz, K. Jamieson, J. W. Suchow, and T. L. Griffiths. “Adaptive sampling for convex regression”. In: *arXiv preprint arXiv:1808.04523* (2018).

[SS11] E. Seijo and B. Sen. “Nonparametric least squares estimation of a multivariate convex regression function”. In: *The Annals of Statistics* 39.3 (2011), pp. 1633–1657.

[SS18] R. J. Samworth and B. Sen. “Special Issue on" Nonparametric Inference Under Shape Constraints"”. In: (2018).

[SW08] R. Schneider and W. Weil. *Stochastic and integral geometry*. Springer Science & Business Media, 2008.

[SW90] C. Schütt and E. Werner. “The convex floating body”. In: *Mathematica Scandinavica* (1990), pp. 275–290.

[Tal14] M. Talagrand. *Upper and lower bounds for stochastic processes*. Vol. 60. Springer, 2014.

[Tsy03a] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2003.

[Tsy03b] A. B. Tsybakov. “Optimal rates of aggregation”. In: *Learning theory and kernel machines*. Springer, 2003, pp. 303–313.

[Var82] H. R. Varian. “The nonparametric approach to demand analysis”. In: *Econometrica: Journal of the Econometric Society* (1982), pp. 945–973.

[Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[Vu05] V. H. Vu. “Sharp concentration of random polytopes”. In: *Geometric & Functional Analysis* 15.6 (2005), pp. 1284–1318.

[VW96] A. W. van der Vaart and J. A. Wellner. “Weak convergence”. In: *Weak convergence and empirical processes*. Springer, 1996, pp. 16–28.

[Wai19] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[Wal49] A. Wald. “Note on the consistency of the maximum likelihood estimate”. In: *The Annals of Mathematical Statistics* 20.4 (1949), pp. 595–601.

[Wyn+17] A. J. Wyner, M. Olson, J. Bleich, and D. Mease. “Explaining the success of adaboost and random forests as interpolating classifiers”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 1558–1590.

[Yan04] Y. Yang. “Aggregating regression procedures to improve performance”. In: *Bernoulli* 10.1 (2004), pp. 25–47.

[YB99] Y. Yang and A. Barron. “Information-theoretic determination of minimax rates of convergence”. In: *Annals of Statistics* (1999), pp. 1564–1599.