Topics on Least Squares Estimation

Qiyang Han

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Jon A. Wellner, Chair

Richard Samworth

Mathias Drton

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Topics on Least Squares Estimation

Qiyang Han

Chair of the Supervisory Committee:
Professor Jon A. Wellner

We revisit and make progress on some old but challenging problems concerning least squares estimation. Two major problems are addressed: (i) least squares estimation with heavy-tailed errors, and (ii) least squares estimation in non-Donsker classes. For (i), we study this problem both from a worst-case perspective, and a more refined envelope perspective. For (ii), we perform two case studies in the context of (a) estimation involving sets and (b) estimation of multivariate isotonic functions. Understanding these particular aspects of least squares estimation problems requires several new tools in the empirical process theory, including a sharp multiplier inequality controlling the size of the multiplier empirical process, and matching upper and lower bounds for empirical processes indexed by non-Donsker classes.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like give my deepest gratitude to my advisor, Jon Wellner, for his tremendous support, guidance and encouragements during many difficult times. This thesis would not have been possible without all of his inputs. Jon's exceptionally high taste for research problems has been one major aspiration of mine in finding challenging and interesting research projects, from problems in shape constrained estimation and inference, Bayes nonparametrics, to the abstract empirical process theory. His deep insights into the theory of empirical processes have laid down the very fundamental part in my own theoretical understanding. I am more than honored and fortunate to be able to work with Jon in the past few years.

I would also like to particularly thank Richard Samworth for inviting to me to visit Cambridge during the summer of 2016 and the spring of 2018, where we had very productive research outputs. It is a major pleasure to work with Richard on the problem of multivariate isotonic regression, along with my co-authors Tengyao Wang and Sabyasachi Chatterjee. I learned tremendously from them.

Thanks go to Chao Gao and Johannes Schimdt-Hieber for many fruitful discussions on the frequentist theory of Bayes nonparametrics, which enabled me to write the paper [74]. I am also very grateful to Johannes for inviting me to a Bayes workshop at Leiden in the August of 2017 to present my paper.

During the last two years of my graduate studies, I benefited a lot from the fellow graduate students in the weekly reading group organized by Fang Han. I am fortunate to participate in such a nice group.

Many people gave me warm help and encouragements during the intensive job search season. I would not have been able to land in a good position without the tremendously

strong supports from Jon Wellner, Richard Samworth, Mathias Drton and Johannes Lederer, and continuous advice and encouragements from Shizhe Chen, Yen-chi Chen, Chao Gao, Fang Han, and Kean Ming Tan. Many close friends and family members accompanied me in this uncertain period of time; I am more than grateful to all of these mental supports.

Lastly, I would like to express my deep admiration for the late Evarist Giné and his work. Although I was not fortunate enough to meet him in person before his tragic and untimely passing, his influence on me has been enormous through many of his research masterpieces, and this will surely continue in the years to come.

DEDICATION

to my parents and Irene

Chapter 1

INTRODUCTION

1.1 Motivation and the questions

Suppose we observe data $\{(X_i, Y_i)\}_{i=1}^n$, where X_i are explanatory variables (covariates) and Y_i are the corresponding responses. The classical *regression model* specifies the relationship between the covariates X_i and the responses Y_i as:

$$Y_i = f_0(X_i) + \xi_i, \quad i = 1, 2, \ldots, n. \tag{1.1.1}$$

Here f_0 is considered as the 'true regression function/signal' modeling the relationship between X_i and Y_i, and the ξ_i are i.i.d. mean-zero additive measurement errors independent of the covariates X_i.

The simplest example for (1.1.1) is the *linear regression* model, where the regression function f_0 satisfies a linear contrast: $f_0(X_i) = X_i^\top \beta_0$ for some $\beta_0 \in \mathbb{R}^d$. In such a model, we often write the model in compact form as

$$Y = X\beta_0 + \xi, \tag{1.1.2}$$

where $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ is the response vector, $X = [X_1 \ldots X_n]^\top \in \mathbb{R}^{n \times d}$ is the design matrix, and $\xi = (\xi_1, \ldots, \xi_n)^\top \in \mathbb{R}^n$ is the error vector.

Although the linear regression model (1.1.2) is often the method of first choice in applications due to its easy interpretability, the overly simplified linearity assumption can often be severely violated. A first step in extending the *parametric* model (1.1.2) to cope with more complicated situations is to consider *non-parametric* models, where in general instead of specifying some exact form of f_0, smoothness or qualitative assumptions are imposed on f_0. Classical examples in this direction include Hölder smoothness conditions (and more

general Besov classes), and monotonicity or convexity qualitative constraints. Recent years (although perhaps starting as long ago as [79]) have also witnessed a tremendous interest in the *high-dimensional* models; a canonical example is given by (1.1.2) where potentially the problem dimension d can be much larger than the sample size n, so any reasonable method of estimation would assume certain low dimension structure of the model—a certain variant of sparsity.

A statistical goal for using the model (1.1.1) when observing (X_i, Y_i)'s is to recover the true regression function f_0. There are countless statistical methods to 'achieve this goal', but in this thesis, we will be interested in clarifying some mysterious aspects to the perhaps most fundamental estimator—the *least squares estimator* (LSE) defined via

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2, \tag{1.1.3}$$

where $\mathcal{F}$ is a model specified by the user that is believed to capture certain 'structure' of the true signal f_0. We assume without loss of generality throughout this thesis that the LSE is well-defined (otherwise we may take a suitable approximating least squares estimator with arbitrary precision to the global infimum).

Two central questions for the theory of the least squares estimator are to understand:

(M1) How fast does the LSE $\hat{f}_n$ converge to the true signal f_0 as the sample size grows?

(M2) Is the above convergence rate "optimal" in some sense?

(M1) has been the focus of nonparametric statistical theory for decades. In the empirical risk minimization (ERM) literature, the rate behavior of the LSE (3.1.2) has been well understood under certain canonical settings. A classical result is the following:

Theorem 1.1.1. *Suppose that:*

1. *the errors $\{\xi_i\}$ are sub-Gaussian (or at least sub-exponential);*

2. the model $\mathcal{F}$ satisfies an entropy condition with exponent $\alpha \in (0,2)$[1].

Then

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2+\alpha}}\big). \tag{1.1.4}$$

Theorem 1.1.1 is well-known in the literature, see e.g. [153, 21, 162, 160, 84, 17], to name a few.

The optimality of Theorem 1.1.1, i.e. (M2), is classically studied in a minimax framework. In particular, it is known that the rate $\mathcal{O}_P(n^{-\frac{1}{2+\alpha}})$ is the best-possible one can hope for *any* estimation procedure [169] from a worst-case perspective. The remarkable feature of Theorem 1.1.1 and this minimax optimality is that the least squares estimator is optimal in this minimax sense under the specified assumptions as an *estimation procedure*.

One may naturally wonder if the least squares estimator continues to be an optimal estimation procedure when either of the two major assumptions in Theorem 1.1.1 fails:

(1) The error distributions need to be lighted tailed (i.e. have exponentially many moments);

(2) The model cannot be too 'rich' in the entropy complexity sense.

At a more technical level, the first assumption (1) that the errors ξ_i's are sub-exponential facilitates the use of the so-called 'Bernstein-norm' [21] so that the usual sharp exponential chaining of the empirical processes is possible. When the errors ξ_i's are genuinely heavy-tailed, apparently we can no longer expect exponential chaining in a usual sense at a general level for the empirical processes. Such a difficulty is related to the following long-standing open question:

Question 1.1.2. *What is the rate performance of the LSE in a heavy-tailed regression setting (i.e. the errors do not have exponential moments)?*

---

[1] $\mathcal{F}$ satisfies an entropy condition with exponent $\alpha \in (0,2)$ if either (i) $\sup_Q \log \mathcal{N}(\varepsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)) \lesssim \varepsilon^{-\alpha}$, where the supremum is over all finitely discrete measures Q on $(\mathcal{X}, \mathcal{A})$; or (ii) $\log \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_2(P)) \lesssim \varepsilon^{-\alpha}$.

As we will review in the next section, although there are some recent breakthroughs for this question when the model satisfies very nice properties [91, 109, 113, 112], a solution at a general level, even from a worst-case perspective, is still lacking.

The second assumption (2) that the model has 'entropy complexity level' $\alpha \in (0, 2)$ is, unfortunately, *necessary* in general to guarantee that the convergence rate of the LSE is optimal in a minimax sense. When $\alpha > 2$, we have the following result due to [21]:

Theorem 1.1.3. *Suppose*

1. *the errors $\{\xi_i\}$ are sub-Gaussian (or at least sub-exponential);*

2. *the model $\mathcal{F}$ is bounded with 'entropy level' $\alpha \in (2, +\infty)$.*

Then

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_P(n^{-\frac{1}{2\alpha}}).$$

Furthermore, the rate cannot be improved under these assumptions.

The boundary case $\alpha = 2$ gives rise to the rate $\mathcal{O}_P(n^{-1/4} \log n)$ which can be recovered by both Theorems 1.1.1 and 1.1.3 up to a multiplicative logarithmic factor so we will not illustrate this case separately from now on. Comparing the rates in Theorems 1.1.1 and 1.1.3, we immediately see that the power of the convergence rate does not depend 'smoothly' on the complexity level α of the model. The serious problem here is that the minimax rate is always $\mathcal{O}_P(n^{-\frac{1}{2+\alpha}})$ (cf. [169, 20]) and hence the rate provided in Theorem 1.1.3 is strictly sub-optimal when $\alpha > 2$.

The striking phenomenon revealed here leads to the question, as to whether this sub-optimality is due to the proof or to the least squares estimation procedure. It is shown in [21] that the convergence rate derived in Theorem 1.1.3 is the best-possible one can hope for under an entropy complexity assumption on the models—a slightly 'constructed' parameter space (a subspace of the univariate Hölder functions with regularity less than $1/2$) witnesses the rate $\mathcal{O}_P(n^{-1/2\alpha})$. Therefore it is clear that from a worst-case perspective, one should not expect the LSE to converge at a minimax rate.

However, although being of great theoretical interest, a worst-case analysis can often be conservative, and sometimes perhaps even mis-leading. In fact, from a practical point of view, it is far from clear if we can give an answer to whether or not the LSE is rate-suboptimal for a given model with entropy complexity level $\alpha > 2$. The real difficult question here is:

Question 1.1.4. *What is the rate performance of the LSE for any given model with entropy complexity level $\alpha > 2$?*

Both Questions 1.1.2 and 1.1.4, posed in statistical contexts, are manifestations of major deficiencies in the current empirical process theory: it has now been well-understood that the problem of deriving convergence rates of LSEs, is essentially equivalent to understanding sharply the size of the *multiplier empirical process* ([76]; see also [36, 158] for related results):

$$\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\xi_i f(X_i)\right|. \tag{1.1.5}$$

In this thesis, we will therefore address Questions 1.1.2 and 1.1.4 by studying the size of (1.1.5) in the context of

1. the multipliers ξ_i's only have a few moments, and

2. the model is 'high-dimensional' or very 'high-dimensional'.

Question 1.1.2 then translates into an understanding for the size of the multiplier empirical process (1.1.5) when ξ_i's can be genuinely heavy-tailed, while Question 1.1.4 is asking about the size the *empirical process* (taking ξ_i's to be Rademacher random variables) when the indexing set $\mathcal{F}$ has an entropy complexity level greater than 2.

1.2 Summary of the results and our contribution

In this thesis, we attempt to make progress on both Questions 1.1.2 and 1.1.4 mentioned above.

1.2.1 Least squares estimation with heavy-tailed errors I: a worst-case analysis

In Chapter 2, we show that if the errors have finite $L_{p,1}(p \geq 1)$ moments, and the underlying function class is uniformly bounded with an entropy of order $\varepsilon^{-\alpha}$ for some $\alpha \in (0,2)$, then the rate of convergence for the LSE is no worse than

$$\mathcal{O}_P\left(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\right). \tag{1.2.1}$$

Furthermore, this rate is unimprovable under entropy conditions alone. To the best knowledge of the author, this rate offers the first quantitative answer to Question 1.1.2 for a general regression function class, and shows a clear tradeoff between the complexity of the function class and the heaviness of the tail of the errors.

As indicated in the previous section, the problem of deriving convergence rates of the LSE is essentially equivalent to that of obtaining sharp estimates for the associated multiplier empirical processes. The key technical tool in this regard, as we will show, is a sharp multiplier inequality characterizing the size of the multiplier empirical processes. Very informally,

$$\begin{aligned}&\mathbb{E}\|\sum_{i=1}^{n}\xi_i f(X_i)\|_{\mathcal{F}}\\ &\qquad\approx \max\left\{\text{`growth rate' of } k \mapsto \mathbb{E}\|\sum_{i=1}^{k}\varepsilon_i f(X_i)\|_{\mathcal{F}}, \mathbb{E}\max_{1\leq i\leq n}|\xi_i|\right\}.\end{aligned}$$

In particular, the size of the multiplier empirical process is determined not only by the size of the corresponding empirical process, but also by the moments of the multipliers.

Aside from the classical non-parametric regression models, we also consider in Chapter 2 high-dimensional models exemplified by the sparse linear regression model, where both the design matrix and errors can be heavy-tailed. We suggest a new proof method that combines our new multiplier inequality with Gaussian approximation techniques to quantify the growth of the empirical process when both the design matrix and the errors can be heavy-tailed. The new technique allows us to prove rate-optimality of the Lasso estimator under a wide range of distributional assumptions on the design and the errors.

1.2.2 Least squares estimation with heavy-tailed errors II: an envelope perspective

In Chapter 2, we investigate Question 1.1.2 from a worst-case perspective. As is often the case, worst-case analysis can be conservative in specific situations. In Chapter 3, we consider further a more difficult regression model where the errors only have a second (or $L_{2,1}$) moment, in which the worst case rate (1.2.1) derived in Chapter 2 is not very informative in that it is always $\mathcal{O}_P(n^{-1/4})$. On the other hand, there are many natural and important examples for which the LSE is known to converge at a much faster rate than the worst-case $\mathcal{O}_P(n^{-1/4})$ rate.

To this end, we will show in Chapter 3 that, the size of the 'localized envelope' of $\mathcal{F}$ gives a sharp interpolation between the parametric rate $\mathcal{O}_P(n^{-1/2})$ and the worst-case rate $\mathcal{O}_P(n^{-1/4})$, and thereby determining the actual convergence rate of the L_2 loss of the LSE when the errors only admit an $L_{2,1}$ moment. More specifically, let $F_0(\delta)$ be the envelope for $\mathcal{F}_0(\delta) \equiv \{f \in \mathcal{F}_0 : Pf^2 \leq \delta^2\}$ where $\mathcal{F}_0 \equiv \mathcal{F} - f_0$. We show that under certain uniform entropy condition on the function class, if for some $0 \leq \gamma \leq 1$, the localized envelope has the growth rate

$$\|F_0(\delta)\|_{L_2(P)} \sim \delta^\gamma : \tag{1.2.2}$$

then the convergence rate of the LSE in the L_2 loss is no worse than

$$\mathcal{O}_P\big(n^{-\frac{1}{2(2-\gamma)}}\big). \tag{1.2.3}$$

Furthermore, the rate (1.2.3) cannot be improved under the condition (1.2.2). It is easily seen from (1.2.3) that, as the size of the localized envelope increases, the rate of the L_2 loss of the LSE deteriorates from the parametric rate $\mathcal{O}_P(n^{-1/2})$ to the worst-case rate $\mathcal{O}_P(n^{-1/4})$.

The envelope approach also gives a systematic approach to proving oracle inequalities in the random design regression setting for these LSEs under an $L_{2,1}$ moment condition on the errors. More specifically, we first prove the following oracle inequality that holds for the canonical isotonic and convex LSEs in the simple regression models: Suppose that $\|f_0\|_\infty < \infty$ and the errors $\{\xi_i\}$ are i.i.d. mean-zero with $\|\xi_1\|_{2,1} < \infty$. Then for any

$\delta \in (0,1)$, there exists some constant $c > 0$ such that with probability $1 - \delta$,

$$\|\hat{f}_n - f_0^*\|_{L_2(P)}^2 \leq c \inf_{m \in \mathbb{N}} \left(\inf_{f_m \in \mathcal{G}_m} \|f_m - f_0^*\|_{L_2(P)}^2 + \frac{m}{n} \cdot \log^2 n \right), \tag{1.2.4}$$

where f_0^* is the $L_2(P)$-projection of f_0 onto the space of square integrable monotonic non-decreasing (resp. convex) functions, and $\mathcal{G}_m$ is the class of piecewise constant non-decreasing (resp. linear convex) functions on $[0,1]$ with at most m pieces in the isotonic (resp. convex) model. The oracle inequality (1.2.4) is further verified for the shape-restricted LSEs in the additive model, where now f_0 is the marginal L_2 projection of the true regression function. One striking message of the oracle inequality for the shape-restricted LSEs in the additive model is the following: both the adaptation and $L_{2,1}$-robustness properties of the LSE can be preserved, up to error distributions with an $L_{2,1}$ moment, for estimating the shape-constrained proxy of the marginal L_2 projection of the true regression function, *essentially regardless of whether or not the additive structure is correctly specified*.

1.2.3 Least squares estimation in non-Donsker models I: estimation involving sets

In Chapter 4, we attempt to give a solution to Question 1.1.4 in various problems involving the class of indicators over measurable sets. Somewhat surprisingly, we show that global ERM procedures are *rate-optimal* for the edge estimation problem in both additive and multiplicative regression models (cf. [87, 86]), and the binary classification problem in the learning theory (cf. [42, 100, 150, 108, 84, 89]) even if the indexing sets are non-Donsker. More specifically, the convergence rate is shown to achieve the minimax optimal rate $\mathcal{O}_P(n^{-\frac{1}{2+\alpha}})$ rather than the (expected) sub-optimal rate $\mathcal{O}_P(n^{-\frac{1}{2\alpha}})$ by using general empirical processes tools.

The key technical ingredient is an understanding for the size of the associated empirical process: if the L_2-size of the indexing set $\mathscr{C}$ is not too small ($\sigma^2 \gtrsim n^{-1/(\alpha+1)}$), then

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} |\mathbb{G}_n(C)| \asymp \max\{\sigma^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\}, \tag{1.2.5}$$

where $\mathscr{C}(\sigma) \equiv \{C \in \mathscr{C} : P(C) \leq \sigma^2\}$, and α is an entropy measurement of the complexity of $\mathscr{C}$ in the sense that an (appropriate) entropy of $\mathscr{C}$ scales as $\mathcal{O}(\varepsilon^{-\alpha})$. Here $\mathbb{G}_n(C) \equiv$

$\sqrt{n}(\mathbb{P}_n - P)(C)$ is the empirical process. The unified perspective (1.2.5) helps us to identify an important phase transition phenomenon: the size of the empirical process indexed by a Donsker class of sets is determined by its L_2-size along with its entropy complexity, while for a non-Donsker class of sets only the complexity of the class $\mathscr{C}$ matters.

Using the sharp bounds (1.2.5), we further investigate the behavior of various ratio-type empirical processes, complementing the results of [5, 60, 59] beyond the Donsker classes of measurable sets. In particular, we obtain the exact order of the normalizing factor for these ratio-type empirical processes, instead of only an upper bound as in [5, 60, 59]. The associated notion of local asymptotic moduli (originated in [5]) is also examined.

1.2.4 Least squares estimation in non-Donsker models II: multivariate isotonic model

In Chapter 5, we attempt to shed light on Question 1.1.4 by another case study, where the model is the class containing multivariate isotonic (block non-decreasing) functions on $[0,1]^d$:

$$\mathcal{F}_d := \big\{ f : [0,1]^d \to \mathbb{R},\ f(x_1, \ldots, x_d) \le f(x_1', \ldots, x_d') \\ \text{when } x_j \le x_j' \text{ for } j = 1, \ldots, d \big\}.$$

Under the assumptions that (i) $X_1, \ldots, X_n$ follow a uniform random design on $[0,1]^d$ and (ii) the errors $\xi_1, \ldots, \xi_n$ are Gaussian, we prove the following results: Let $d \ge 2$. There exists $C_d, \gamma_d > 0$ such that

$$\sup_{f_0 \in \mathcal{F}_d \cap L_\infty(1)} \mathbb{E} \|\hat{f}_n - f_0\|^2_{L_2(\mathbb{P}_n)} \le C_d n^{-1/d} \log^{\gamma_d} n, \tag{1.2.6}$$

and that

$$\mathbb{E} \|\hat{f}_n - f_0\|^2_{L_2(\mathbb{P}_n)} \le \inf_{k \in \mathbb{N}} \left\{ \inf_{f \in \mathcal{F}_d^{(k)}} \|f - f_0\|^2_{L_2(P)} + C_d \left(\frac{k}{n}\right)^{2/d} \log^{2\gamma_d} \left(\frac{n}{k}\right) \right\}, \tag{1.2.7}$$

where $\mathcal{F}_d^{(k)}$ denotes the class of all multivariate isotonic functions on $[0,1]^d$ that take constant values on at most k rectangular regions. Inequalities of type (1.2.6) and (1.2.7) are usually referred to as *worst-case* and *adaptive* risk bounds in the literature, cf. [172, 33, 18].

The rate in (1.2.6) matches the corresponding minimax lower bound, and hence the class $\mathcal{F}_d$ serves as another example for which the global ERM procedures can be rate-optimal (up to multiplicative logarithmic factors) in non-Donsker problems. The rate in (1.2.7) also reveals some surprising features: Sharp adaptive behavior for shape-constrained estimators has previously only been shown when the adaptive rate is nearly parametric (cf. [172, 33, 18]); our results here show that the least squares estimator in the d-dimensional isotonic regression problem necessarily adapts at a strictly nonparametric rate. Clearly, the minimax optimal rate for constant functions is parametric. Hence, the least squares estimator in this problem adapts at a strictly suboptimal rate while at the same time being nearly rate optimal from a worst-case perspective.

Similar inequalities to (1.2.6)-(1.2.7) are derived for the $L_2(P)$ loss in the random design case, and for the $L_2(\mathbb{P}_n)$ loss in the fixed lattice design case.

1.3 Notation

For a real-valued random variable ξ and $1 \leq p < \infty$, let $\|\xi\|_p := \left(\mathbb{E}|\xi|^p\right)^{1/p}$ denote the ordinary p-norm. The $L_{p,1}$ norm for a random variable ξ is defined by

$$\|\xi\|_{p,1} := \int_0^\infty \mathbb{P}(|\xi| > t)^{1/p} \, \mathrm{d}t.$$

It is well known that $L_{p+\varepsilon} \subset L_{p,1} \subset L_p$ holds for any underlying probability measure, and hence a finite $L_{p,1}$ condition requires slightly more than a p-th moment, but no more than any $p+\varepsilon$ moment, see Chapter 10 of [93]. In this paper, we will primarily be concerned with the case $p = 2$.

For a real-valued measurable function f defined on $(\mathcal{X}, \mathcal{A}, P)$, $\|f\|_{L_p(P)} \equiv \left(P|f|^p\right)^{1/p}$ denotes the usual L_p-norm under P, and $\|f\|_\infty \equiv \|f\|_{L_\infty} \equiv \sup_{x \in \mathcal{X}} |f(x)|$. f is said to be P-centered if $Pf = 0$. $L_p(g, B)$ denotes the $L_p(P)$-ball centered at g with radius B. For simplicity we write $L_p(B) \equiv L_p(0, B)$.

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of the normed space of real functions $f : \mathcal{X} \to \mathbb{R}$. Let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-covering number, and let $\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-bracketing number; see page 83 of

[162] for more details. To avoid unnecessary measurability digressions, we assume that $\mathcal{F}$ is countable throughout the thesis. As usual, for any $\phi : \mathcal{F} \to \mathbb{R}$, we write $\|\phi(f)\|_{\mathcal{F}}$ for $\sup_{f\in\mathcal{F}}|\phi(f)|$.

Throughout the thesis $\varepsilon_1, \ldots, \varepsilon_n$ will be i.i.d. Rademacher random variables independent of all other random variables. C_x will denote a generic constant that depends only on x, whose numeric value may change from line to line unless otherwise specified. $a \lesssim_x b$ and $a \gtrsim_x b$ mean $a \leq C_x b$ and $a \geq C_x b$ respectively, and $a \asymp_x b$ means $a \lesssim_x b$ and $a \gtrsim_x b$ [$a \lesssim b$ means $a \leq Cb$ for some absolute constant C]. For two real numbers a, b, $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. We slightly abuse notation by defining $\log(x) \equiv \log(x \vee e)$.

1.4 Organization

The thesis is organized as follows. In Chapter 2, which largely follows [76], we develop a general sharp multiplier inequality to study the size of the multiplier empirical process, and perform a worst-case analysis for the rate behavior of the least squares estimator in a heavy-tailed regression setting. In Chapter 3, which large follows [77], we work in a more restrictive heavy-tailed regression setup where the errors only admit a second moment, but take a further step beyond a worst-case analysis, by characterizing the difficulty of the least squares estimator in terms of the 'localized envelopes' of the model. In Chapter 4, we examine the size of the empirical processes indexed by non-Donsker classes by a special case—the class of indicators over measurable sets, and the associated least squares estimators in various related learning problems. In Chapter 5, which is based on joint work with Sabyasachi Chatterjee, Richard Samworth and Tengyao Wang [75], we study the behavior of the least squares estimator in the multivariate isotonic model, and derive worst-case and adaptive risk bounds for both fixed and random designs.

Chapter 2

LEAST SQUARES ESTIMATION WITH HEAVY-TAILED ERRORS I: A WORST-CASE ANALYSIS

2.1 Introduction

2.1.1 Motivation and problems

Consider the classical setting of nonparametric regression: suppose that

$$Y_i = f_0(X_i) + \xi_i \quad \text{for} \quad i = 1, \ldots, n \tag{2.1.1}$$

where $f_0 \in \mathcal{F}$, a class of possible regression functions f where $f : \mathcal{X} \to \mathbb{R}$, $X_1, \ldots, X_n$ are i.i.d. P on $(\mathcal{X}, \mathcal{A})$, and $\xi_1, \ldots, \xi_n$ are i.i.d "errors" independent of $X_1, \ldots, X_n$. We observe the pairs $\{(X_i, Y_i) : \ 1 \leq i \leq n\}$ and want to estimate f_0.

While there are many approaches to this problem, the most classical approach has been to study the Least Squares Estimator (or LSE) $\hat{f}_n$ defined by

$$\hat{f}_n = \text{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2. \tag{2.1.2}$$

The LSE is well-known to have nice properties (e.g. rate-optimality) when:

(E) the errors $\{\xi_i\}$ are sub-Gaussian or at least sub-exponential;

(F) the class $\mathcal{F}$ of regression functions satisfies a condition slightly stronger than a *Donsker* condition: namely, either a uniform entropy condition or a bracketing entropy condition with exponent $\alpha \in (0, 2)$:

$$\sup_Q \log \mathcal{N}(\varepsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)) \lesssim \varepsilon^{-\alpha},$$

where the supremum is over all finitely discrete measures Q on $(\mathcal{X}, \mathcal{A})$, or

$$\log \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_2(P)) \lesssim \varepsilon^{-\alpha};$$

See for example [21] and [160], chapter 9, and Section 1.3 for notation. In spite of a very large literature, there remains a lack of clear understanding of the properties of $\hat{f}_n$ in terms of assumptions concerning the heaviness of the tails of the errors and the massiveness or "size" of the class $\mathcal{F}$.

Our interest here is in developing further tools and methods to study properties of $\hat{f}_n$, especially its convergence rate when the error condition (E) is replaced by:

(E′) the errors $\{\xi_i\}$ have only a p-moment for some $1 \leq p < \infty$.

This leads to our first question:

Question 2.1.1. *What determines the convergence rate b_n of $\hat{f}_n$ with respect to some risk or loss functions? When is this rate b_n determined by p (and hence the tail behavior of the ξ_i's), and when is it determined by α (and hence the size of $\mathcal{F}$)?*

There are a variety of measures of loss and risk in this setting. Two of the most common are:

(a) Empirical L_2 loss: $\|\hat{f}_n - f_0\|_{L_2(\mathbb{P}_n)}$[1], and the corresponding risk $\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(\mathbb{P}_n)}$.

(b) Population (or prediction) L_2 loss $\|\hat{f}_n - f_0\|_{L_2(P)}$, and the corresponding risk $\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P)}$.

Here we will mainly focus on measuring loss or risk in the sense of the prediction loss (b) since it corresponds to the usual choice in the language of Empirical Risk Minimization; see e.g. [16, 17, 21, 84, 85, 108, 153, 160, 162]. Thus we will (usually) measure loss or risk in $L_2(P)$ and hence study rates of convergence of

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \left[\int_{\mathcal{X}} |\hat{f}_n(x; (X_1, Y_1), \ldots, (X_n, Y_n)) - f_0(x)|^2 \, \mathrm{d}P(x) \right]^{1/2},$$

[1] We write $\mathbb{P}_n$ for the empirical measure of the (X_i, Y_i) pairs: $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{(X_i, Y_i)}$.

or, in somewhat more compact notation,

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P)} = \mathbb{E}\bigg[\int_{\mathcal{X}} |\hat{f}_n(x) - f_0(x)|^2 \, \mathrm{d}P(x)\bigg]^{1/2}.$$

As we will see in Section 3, the rate of convergence of the LSE $\hat{f}_n$ under conditions (E′) and (F) is

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big). \tag{2.1.3}$$

So, the dividing line between p and α in determining the rate of convergence of the LSE is given by

$$p = 1 + 2/\alpha$$

in the following sense:

(R_α) If $p \geq 1 + 2/\alpha$, then for any function class with entropy exponent α, the rate of convergence of the LSE is $\mathcal{O}_{\mathbf{P}}(n^{-1/(2+\alpha)})$.

(R_p) If $p < 1 + 2/\alpha$, then there exist model classes $\mathcal{F}$ with entropy exponent α such that the rate of convergence of the LSE is $\mathcal{O}_{\mathbf{P}}(n^{-1/2+1/(2p)})$.

These rates in R_α and R_p indicate both some positive and negative aspects of the LSE in a heavy-tailed regression setting:

- If $p \geq 1 + 2/\alpha$, then the heaviness of the tails of the errors (E′) does not play a role in the rate of convergence of the LSE, since the rate in R_α coincides with the usual rate under the light-tailed error assumption (E) and the entropy condition (F).

- If $p < 1+2/\alpha$, there exist (many) hard models at any entropy level α for which the LSE converges only at a slower rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2+1/(2p)})$ compared with the faster (optimal) rate $\mathcal{O}_{\mathbf{P}}(n^{-1/(2+\alpha)})$—a rate that can be achieved by other robust estimation procedures. See Section 3 for examples and more details.

It should be noted that the assumption of independence of the errors ξ_i's and the X_i's in the regression model (2.1.1) is crucial for the above results to hold. In fact, when the errors ξ_i's can be dependent on the X_i's, there is no longer any universal moment condition on the ξ_i's alone that guarantees the rate-optimality of the LSE, as opposed to (R_α) (cf. Proposition 2.3.12).

To briefly introduce the main new tool we develop in Section 2 below, we first recall the classical methods used to prove consistency and rates of convergence of the LSE (and many other contrast-type estimators). These methods are based on a "basic inequality" which lead naturally to a multiplier empirical process. This is well-known to experts in the area, but we will briefly review the basic facts here. Since $\hat{f}_n$ minimizes the functional $f \mapsto \mathbb{P}_n(Y - f(X))^2 = n^{-1}\sum_{i=1}^n (Y_i - f(X_i))^2$, it follows that

$$\mathbb{P}_n(Y - \hat{f}_n(X))^2 \leq \mathbb{P}_n(Y - f_0(X))^2.$$

Adding and subtracting f_0 on the left side, some algebra yields

$$\mathbb{P}_n(Y - f_0(X))^2 + 2\mathbb{P}_n(Y - f_0)(f_0 - \hat{f}_n) + \mathbb{P}_n(f_0 - \hat{f}_n)^2 \leq \mathbb{P}_n(Y - f_0(X))^2.$$

Since $\xi_i = Y_i - f_0(X_i)$ under the model given by (2.1.1) we conclude that

$$\mathbb{P}_n(\hat{f}_n(X) - f_0(X))^2 \leq 2\mathbb{P}_n\left(\xi(\hat{f}_n(X) - f_0(X))\right) \leq 2\sup_{f\in\mathcal{F}} \mathbb{P}_n\left(\xi(f(X) - f_0(X))\right) \qquad (2.1.4)$$

where the process

$$f \mapsto n(\mathbb{P}_n - P)(\xi f(X)) = n\mathbb{P}_n(\xi f(X)) = \sum_{i=1}^n \xi_i f(X_i) \qquad (2.1.5)$$

is a *multiplier empirical process*. This is exactly as in Section 4.3 of [160]. When the ξ_i's are Gaussian, the process in (2.1.5) is even a Gaussian process conditionally on the X_i's, and is relatively easy to analyze. If the $\{\xi_i\}$'s are integrable and $\mathcal{F}$ is a *Glivenko-Cantelli class* of functions, then the inequality (2.1.4) leads easily to consistency of the LSE in the sense of the loss and risk measures (a); see e.g. [160].

To obtain rates of convergence we need to consider localized versions of the processes in (2.1.4), much as in Section 3.4.3 of [162]. As in Section 3.4.3 of [162], (but replacing their $\theta \in \Theta$ and ε by our $f \in \mathcal{F}$ and ξ) we consider

$$\mathbb{M}_n(f) = 2\mathbb{P}_n\xi(f - f_0) - \mathbb{P}_n(f - f_0)^2,$$

and note that $\hat{f}_n$ maximizes $\mathbb{M}_n(f)$ over $\mathcal{F}$. Since the errors have zero mean and are independent of the X_i's, this process has mean $M(f) \equiv -P(f - f_0)^2$. Since $\mathbb{M}_n(f_0) = 0 = M(f_0)$, centering then yields the process

$$\begin{aligned} f \mapsto \mathbb{Z}_n(f) &\equiv \mathbb{M}_n(f) - \mathbb{M}_n(f_0) - (M(f) - M(f_0)) \\ &= 2\mathbb{P}_n\xi(f - f_0) - (\mathbb{P}_n - P)(f - f_0)^2. \end{aligned}$$

Establishing rates of convergence for $\hat{f}_n$ then boils down to bounding

$$\mathbb{E} \sup_{f \in \mathcal{F}: P(f-f_0)^2 \le \delta^2} \mathbb{Z}_n(f)$$

as a function of n and δ; see e.g. [162] Theorem 3.4.1, pages 322-323. It is clear at least for $\mathcal{F} \subset L_\infty$ that this can be accomplished if we have good bounds for the multiplier empirical process (2.1.5) in terms of the empirical process itself

$$f \mapsto n(\mathbb{P}_n - P)(f(X)\big) = \sum_{i=1}^{n} \big(f(X_i) - Pf\big), \tag{2.1.6}$$

or, in view of standard symmetrization inequalities (as in Section 2.3 of [162]), its symmetrized equivalent,

$$f \mapsto \sum_{i=1}^{n} \varepsilon_i f(X_i), \tag{2.1.7}$$

where the ε_i are i.i.d. Rademacher random variables $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$ independent of the X_i's. This leads naturally to:

Question 2.1.2. *Under what moment conditions on the ξ_i's can we assert that the multiplier empirical process (2.1.5) has (roughly) the same "size" as the empirical process (2.1.6) (or equivalently the symmetrized empirical process (2.1.7) for (nearly) all function classes $\mathcal{F}$ in a non-asymptotic manner?*

In Section 2.2 below we provide simple moment conditions on the ξ_i's which yield a positive answer to Question 2, when the ξ_i's are independent from the X_i's. We then give some comparisons to the existing multiplier inequalities which illustrate the improvement possible via the new bounds in non-asymptotic settings, and show that our bounds also yield the asymptotic equivalence required for multiplier CLT's (cf. Section 2.9 of [162]). Further impossibility results are demonstrated, showing that there is no positive solution to Question 2 when the ξ_i's and the X_i's can be dependent.

In Section 2.3 we address Question 1 by applying the new multiplier inequality to derive the convergence rate of the LSE (2.1.3) in the context of the nonparametric regression model (2.1.1), and indicate in greater detail both the positive and negative aspects of the LSE due to this rate. We further show that no solution to Question 1 exists when the errors ξ_i's and the covariates X_i's can be dependent.

Not surprisingly, the new bounds for the multiplier empirical process have applications to many settings in which the Least Squares criterion plays a role, for example the Lasso in the sparse linear regression model. In Section 2.4 we give an application of the new bounds in a Lasso setting with both heavy-tailed errors and heavy-tailed covariates. Most detailed proofs are given in Sections 2.5-2.8.

2.2 The multiplier inequality

Multiplier inequalities have a long history in the theory of empirical processes. Our new multiplier inequality in this section is closest in spirit to the classical multiplier inequality, cf. Section 2.9 of [162] or [62], but strictly improves the classical one in a non-asymptotic setting (see Section 2.2.3).

Our work here is also related to [110], who derived bounds for the multiplier empirical process, assuming: (i) ξ_i's have a $2+\varepsilon$ moment, and (ii) $\{(\xi_i, X_i)\}$ are i.i.d. (i.e. ξ_i need not be independent from X_i). The bounds in [110] use techniques from generic chaining [146], and work particularly well for 'sub-Gaussian classes' (defined in [110]). Our setting here will be different: we assume that: (i) ξ_i's have a $L_{p,1} (p \geq 1)$ moment and (ii) ξ_i's are independent

from X_i's, but the ξ_i's need not be independent from each other.

We make this choice in view of a negative result of Alexander [2], stating that there is no universal moment condition on ξ_i's for a multiplier CLT to hold when ξ_i's need not be independent from X_i's, while a $L_{2,1}$ moment condition is known to be universal in the independent case [62, 93, 162]. The complication here makes it more hopeful to work in the independent case for a precise understanding of the multiplier empirical process. In fact:

- In the independent case we are able to quantify the exact *structural interplay* between the moment of the multipliers and the complexity of the indexing function class in the size of the multiplier empirical process (cf. Theorems 2.2.1-2.2.6), thereby giving a satisfactory answer to Question 2;

- Such an interplay fails when the X_i's may not be independent from the ξ_i's. Moreover, no simple moment condition on the ξ_i's alone can lead to a solution to Question 2 in the dependent case (cf. Proposition 2.2.10).

2.2.1 Upper bound

We first state the assumptions.

Assumption A. Suppose that $\xi_1, \ldots, \xi_n$ are independent of the random variables $X_1, \ldots, X_n$, and *either* of the following conditions holds:

(A1) $X_1, \ldots, X_n$ are i.i.d. with law P on $(\mathcal{X}, \mathcal{A})$, and $\mathcal{F}$ is P-centered.

(A2) $X_1, \ldots, X_n$ are permutation invariant, and $\xi_1, \ldots, \xi_n$ are independent mean-zero random variables.

Theorem 2.2.1. *Suppose Assumption A holds. Let $\{\mathcal{F}_k\}_{k=1}^n$ be a sequence of function classes such that $\mathcal{F}_k \supset \mathcal{F}_n$ for any $1 \leq k \leq n$. Assume further that there exist non-decreasing concave*

functions $\{\psi_n\} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *with* $\psi_n(0) = 0$ *such that*

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_k} \leq \psi_n(k) \tag{2.2.1}$$

holds for all $1 \leq k \leq n$. *Then*

$$\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} \leq 4 \int_0^\infty \psi_n\left(\sum_{i=1}^{n} \mathbb{P}(|\xi_i| > t) \right) \mathrm{d}t. \tag{2.2.2}$$

The primary application of Theorem 2.2.1 to non-parametric regression problems in Section 2.3 involves a non-increasing sequence of function classes $\mathcal{F}_1 \supset \ldots \supset \mathcal{F}_n$. It is also possible to use Theorem 2.2.1 for the case $\mathcal{F}_1 = \cdots = \mathcal{F}_n$; see Section 2.4 for an application to the sparse linear regression model.

The following corollary provides a canonical concrete application of Theorem 2.2.1.

Corollary 2.2.2. *Consider the same assumptions as in Theorem 2.2.1. Assume for simplicity that* ξ_i*'s have the same marginal distributions. Suppose that for some* $\gamma \geq 1$, *and some constant* $\kappa_0 > 0$,

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_k} \leq \kappa_0 \cdot k^{1/\gamma} \tag{2.2.3}$$

holds for all $1 \leq k \leq n$. *Then for any* $p \geq 1$ *such that* $\|\xi_1\|_{p,1} < \infty$,

$$\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} \leq 4\kappa_0 \cdot n^{\max\{1/\gamma, 1/p\}} \|\xi_1\|_{\min\{\gamma,p\},1}.$$

Proof. First consider $\gamma \leq p$. In this case, letting $\psi_n(t) \equiv \kappa_0 t^{1/\gamma}$ in Theorem 2.2.1, we see that $\mathbb{E}\|\sum_{i=1}^n \xi_i f(X_i)\|_{\mathcal{F}_n} \leq 4\kappa_0 \cdot n^{1/\gamma} \|\xi_1\|_{\gamma,1}$. On the other hand, if $\gamma > p$, we can take $\psi_n(t) \equiv \kappa_0 t^{1/p}$ to conclude that $\mathbb{E}\|\sum_{i=1}^n \xi_i f(X_i)\|_{\mathcal{F}_n} \leq 4\kappa_0 \cdot n^{1/p} \|\xi_1\|_{p,1}$. Note that $\gamma \geq 1$ ensures the concavity of ψ_n. □

Corollary 2.2.2 says that the upper bound for the multiplier empirical process has two components: one part comes from the growth rate of the empirical process; another part comes from the moment barrier of the multipliers ξ_i's.

Remark 2.2.3. One particular case for application of Theorem 2.2.1 and Corollary 2.2.2 is the following. Let $\delta_1 \geq \ldots \geq \delta_n \geq 0$ be a sequence of non-increasing non-negative real numbers, and $\mathcal{F}$ be an arbitrary function class. Let $\mathcal{F}_k \equiv \mathcal{F}(\delta_k) \equiv \{f \in \mathcal{F} : Pf^2 < \delta_k^2\}$ be the 'local' set of $\mathcal{F}$ with L_2-radius at most δ_k. There exists a large literature on controlling such localized empirical processes; a classical device suited for applications in nonparametric problems is to use local maximal inequalities under either the uniform or bracketing entropy conditions (cf. Proposition 2.5.1).

An important choice in statistical applications for δ_k is given by

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}(\delta_k)} \lesssim k\delta_k^2. \tag{2.2.4}$$

As will be seen in Section 2.3, the above choice $\{\delta_k\}$ corresponds to the rate of convergence of the LSE in the nonparametric regression model (2.1.1).

In this case Theorem 2.2.1 and Corollary 2.2.2 yield that

$$\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}(\delta_n)} \lesssim n\delta_n^2 \tag{2.2.5}$$

given sufficient moments of the ξ_i's.

Remark 2.2.4. Choosing $\gamma \geq 2$ in Corollary 2.2.2 corresponds to the bounded Donsker regime[2] for the empirical process. In this case we only need $\|\xi_1\|_{2,1} < \infty$ to ensure the multiplier empirical process to also be bounded Donsker. This moment condition is generally unimprovable in view of [92]. On the other hand, such a choice of γ can fail due to: (i) failure of integrability of the envelope functions of the classes $\{\mathcal{F}_k\}$, or (ii) failure of the classes $\{\mathcal{F}_k\}$ to be bounded Donsker. (i) is related to the classical Marcinkiewicz-Zygmund strong laws of large numbers and the generalizations of those to empirical measures, see [7, 103, 105]. For (ii), some examples in this regard can be found in [141], Chapter 11 of [48], see also Proposition 17.3.7 of [137].

---

[2] $\mathcal{F}$ is said to be *bounded Donsker* if $\sup_{n\in\mathbb{N}} \mathbb{E} \sup_{f\in\mathcal{F}} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i)| < \infty$.

Theorem 2.2.1 and Corollary 2.2.2 only concern the first moment of the suprema of the multiplier empirical process. For higher moments, we may use the following Hoffmann-Jørgensen/Talagrand type inequality relating the q-th moment estimate with the first moment estimate.

Lemma 2.2.5 (Proposition 3.1 of [61]). *Let $q \geq 1$. Suppose $X_1, \ldots, X_n$ are i.i.d. with law P and $\xi_1, \ldots, \xi_n$ are i.i.d. with $\|\xi_1\|_{2\vee q} < \infty$. Let $\mathcal{F}$ be a class of functions with $\sup_{f\in\mathcal{F}} Pf^2 \leq \sigma^2$ such that either $\mathcal{F}$ is P-centered, or ξ_1 is centered. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \xi_i f(X_i)\right|^q \leq K^q\Bigg[\left(\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \xi_i f(X_i)\right|\right)^q + q^{q/2}(\sqrt{n}\|\xi_1\|_2\sigma)^q + q^q\mathbb{E}\max_{1\leq i\leq n}|\xi_i|^q \sup_{f\in\mathcal{F}}|f(X_i)|^q\Bigg].$$

Here $K > 0$ is a universal constant.

2.2.2 Lower bound

Theorem 2.2.1 and Corollary 2.2.2 do not require any structural assumptions on the function class $\mathcal{F}$. [110] showed that for a 'sub-Gaussian' class, a $2+\varepsilon$ moment on i.i.d. ξ_i's suffices to conclude that the multiplier empirical process behaves like the canonical Gaussian process. One may therefore wonder if the moment barrier for the multipliers in Corollary 2.2.2 is due to an artifact of the proof. Below in Theorem 2.2.6 we show that this barrier is intrinsic for general classes $\mathcal{F}$.

Theorem 2.2.6. *Let $\mathcal{X} = [0,1]$ and P be a probability measure on $\mathcal{X}$ with Lebesgue density bounded away from 0 and ∞. Let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables such that $\mathbb{E}\max_{1\leq i\leq n}|\xi_i| \geq \kappa_0 n^{1/p}$ for some $p > 1$ and some constant κ_0 independent of ξ_1. Then for any $\gamma > 2$, there exists a sequence of function classes $\{\mathcal{F}_k\}_{k=1}^n$ defined on $\mathcal{X}$ with $\mathcal{F}_k \supset \mathcal{F}_n$ for any $1 \leq k \leq n$ such that for n sufficiently large,*

$$\mathbb{E}\left\|\sum_{i=1}^k \varepsilon_i f(X_i)\right\|_{\mathcal{F}_k} \leq \kappa_1 \cdot k^{1/\gamma},$$

holds for all $1 \le k \le n$*, and that*

$$\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} \ge \kappa_1^{-1} n^{\max\{1/\gamma, 1/p\}}.$$

Here κ_1 *is a constant depending on* κ_0, γ *and* P*.*

Remark 2.2.7. The condition on the ξ_i's will be satisfied, for example if the ξ_i's are i.i.d. with the tail condition $\mathbb{P}(|\xi_i| > t) \ge \kappa_0'/(1+t^p)$ for $t > 0$.

Combined with Corollary 2.2.2, it is seen that the growth rate $n^{\max\{1/\gamma, 1/p\}}$ of the multiplier empirical process cannot be improved in general. This suggests an interesting phase transition phenomenon from a worst-case perspective: if the complexity of the function class dominates the effect of the tail of the multipliers, then the multiplier empirical process essentially behaves as the empirical process counterpart; otherwise the tail of the multipliers governs the growth of the multiplier empirical process.

Remark 2.2.8. The function class we constructed that witnesses the moment barrier rate $n^{1/p}$ in Theorem 2.2.6 can be simply taken to be the class of indicators over closed intervals on $[0, 1]$. Although being the 'simplest' function class in the theory of empirical processes, this class serves as an important running example that achieves the bad rate $n^{1/p}$.

2.2.3 Comparison of Theorem 2.2.1 with the multiplier inequality in [162]

In this section we compare the classical multiplier inequality in Theorem 2.2.1 with the one in Section 2.9 of [162], which originates from [64, 65, 92]; see also [62]: for i.i.d. mean-zero ξ_i's and i.i.d. X_i's, and for any $1 \le n_0 \le n$,

$$\mathbb{E}\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}} \lesssim (n_0 - 1)\mathbb{E}\|f(X_1)\|_{\mathcal{F}} \frac{\mathbb{E}\max_{1\le i\le n}|\xi_i|}{\sqrt{n}} + \|\xi_1\|_{2,1} \max_{n_0 \le k \le n} \mathbb{E}\left\| \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}}. \tag{2.2.6}$$

Non-asymptotic setting

The major drawback of (2.2.6) is that it is not sharp in a non-asymptotic setting. For an illustration, let $\xi_1, \dots, \xi_n$ be i.i.d. multipliers with $\|\xi_1\|_{p,1} < \infty$ $(p \geq 2)$, X_i's be i.i.d. uniformly distributed on $[0,1]$, and $\mathcal{F}$ be a uniformly bounded function class on $[0,1]$ satisfying the entropy condition (F) with $\alpha \in (0,2)$. We apply (2.2.6) with $\mathcal{F}(n^{-1/(2+\alpha)})$ (note that $n^{-1/(2+\alpha)}$ is the usual local radius for $1/\alpha$-smooth problems) and local maximal inequalities for the empirical process (Proposition 2.5.1 in Section 2.5 below) to see that

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}(n^{-1/(2+\alpha)})} &\lesssim \inf_{1\leq n_0\leq n} n_0 \cdot n^{-1/2+1/p} + n_0^{-\frac{(2-\alpha)}{2(2+\alpha)}} \\ &\asymp n^{-\frac{2-\alpha}{6+\alpha}\left(\frac{1}{2}-\frac{1}{p}\right)} \equiv n^{-\delta_1(\alpha,p)}.\end{aligned} \tag{2.2.7}$$

On the other hand, Corollary 2.2.2 gives the rate:

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}(n^{-1/(2+\alpha)})} \lesssim n^{-\min\{\frac{2-\alpha}{2(2+\alpha)}, 1/2-1/p\}} \equiv n^{-\delta_2(\alpha,p)}. \tag{2.2.8}$$

In the above inequalities we used the following bound for the symmetrized empirical process (for illustration we only consider bracketing entropy):

$$\begin{aligned}&\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}(n^{-1/(2+\alpha)})} \\ &\lesssim J_{[]}(n^{-1/(2+\alpha)}, \mathcal{F}, L_2(P))\left(1 + \frac{J_{[]}(n^{-1/(2+\alpha)}, \mathcal{F}, L_2(P))}{\sqrt{n}\cdot n^{-2/(2+\alpha)}}\right) \lesssim n^{\frac{2-\alpha}{2(2+\alpha)}},\end{aligned}$$

where in the last line of the above display we used

$$J_{[]}(n^{-1/(2+\alpha)}, \mathcal{F}, L_2(P)) = \int_0^{n^{-1/(2+\alpha)}} \sqrt{1 + \log \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_2(P))}\, \mathrm{d}\varepsilon \lesssim n^{\frac{2-\alpha}{2(2+\alpha)}}.$$

It is easily seen that the bound (2.2.7) calculated from (2.2.6) is worse than (2.2.8) because $\delta_1(\alpha,p) < \delta_2(\alpha,p)$ for all $\alpha \in (0,2)$ and $p \geq 2$. Moreover, if $p \geq 1 + 2/\alpha$, the bound (2.2.8) becomes $n^{-\frac{2-\alpha}{2(2+\alpha)}}$, which matches the rate for the symmetrized empirical process $\mathbb{E}\|n^{-1/2}\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}(n^{-1/(2+\alpha)})}$.

Asymptotic setting

The primary application of (2.2.6) rests in studying asymptotic equicontinuity of the multiplier empirical process in the following sense. Suppose that $\mathcal{F}$ is Donsker. Then by the integrability of the empirical process (see Lemma 2.3.11 of [162])[3], $\mathbb{E}\|n^{-1/2}\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}_\delta} \to 0$ as $n \to \infty$ followed by $\delta \to 0$. Now apply (2.2.6) via $n \to \infty$, $n_0 \to \infty$ followed by $\delta \to 0$ we see that $\mathbb{E}\|n^{-1/2}\sum_{i=1}^n \xi_i f(X_i)\|_{\mathcal{F}_\delta} \to 0$ as $n \to \infty$ followed by $\delta \to 0$ if $\|\xi_1\|_{2,1} < \infty$. This shows that $\left(n^{-1/2}\sum_{i=1}^n \xi_i f(X_i)\right)_{f\in\mathcal{F}}$ satisfies a CLT in $\ell^\infty(\mathcal{F})$ if $\mathcal{F}$ is Donsker and the ξ_i's are i.i.d. with $\|\xi_1\|_{2,1} < \infty$.

Our new multiplier inequality, Theorem 2.2.1, can also be used to study asymptotic equicontinuity of the multiplier empirical process with the help of the following lemma.

Lemma 2.2.9. *Fix $0 \le \gamma \le 1$ and a concave function $\varphi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ such that $\varphi(x) \to \infty$ as $x \to \infty$. Let $\{a_n\} \subset \mathbb{R}_{\ge 0}$ be such that $a_n \to 0$ as $n \to \infty$, and $\psi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be the least concave majorant of $\{(n, a_n\varphi(n))\}_{n=0}^\infty$. Then $\psi(t)/\varphi(t) \to 0$ as $t \to \infty$.*

The proof of this lemma can be found in Section 3.5. Take any sequence $\delta_n \to 0$ and let $a_n \equiv \mathbb{E}\|n^{-1/2}\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}_{\delta_n}}$. By Lemma 2.2.9, the least concave majorant function $\psi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ of the map $n \mapsto a_n n^{1/2} (n \ge 0)$ satisfies $\psi(t)/t^{1/2} \to 0$ as $t \to \infty$. Now an application of Theorem 2.2.1 and the dominated convergence theorem shows that

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}_{\delta_n}} \le 4\int_0^\infty \frac{\psi(n\mathbb{P}(|\xi_1| > t))}{\sqrt{n\mathbb{P}(|\xi_1| > t)}} \cdot \sqrt{\mathbb{P}(|\xi_1| > t)}\, \mathrm{d}t \to 0$$

as $n \to \infty$.

We note that the moment conditions of Theorem 2.2.1 and 2.2.6 have a small gap: in essence we require an $L_{p,1}$ moment in Theorem 1, while an L_p moment is required in Theorem 2. In the context of multiplier CLTs discussed above, [92] showed that the $L_{2,1}$ moment condition is sharp—there exists a construction of a Banach space of X on which a multiplier CLT fails for ξX if $\|\xi_1\|_{2,1} = \infty$. It remains open in our setting if $L_{p,1}$ (or L_p) is the exact moment requirement.

---

[3]Here $\mathcal{F}_\delta \equiv \{f - g : f, g \in \mathcal{F}, \|f - g\|_{L_2(P)} \le \delta\}$.

2.2.4 An impossibility result

In this section we formally prove an impossibility result, showing that the independence assumption between the X_i's and the ξ_i's is crucial for Theorem 2.2.1 and Corollary 2.2.2 to hold.

Proposition 2.2.10. *Let $\mathcal{X} \equiv \mathbb{R}$. For every triple (δ, γ, p) such that $\delta \in (0, 1/2)$, $2 < \gamma < 1 + 1/(2\delta)$ and $2 \le p < \min\{4/\delta, 2\gamma/(1+\gamma\delta)\}$, there exist X_i's and ξ_i's satisfying: (i) $\{(X_i, \xi_i)\}$'s are i.i.d.; (ii) ξ_i is not independent from X_i but $\mathbb{E}[\xi_1|X_1] = 0$, $\|\xi_1\|_{p,1} < \infty$, and a sequence of function classes $\{\mathcal{F}_k\}_{k=1}^n$ defined on $\mathcal{X}$ with $\mathcal{F}_k \supset \mathcal{F}_n$ for any $1 \le k \le n$, such that*

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_k} \lesssim k^{1/\gamma},$$

holds for all $1 \le k \le n$, and that

$$\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} \gtrsim_p \omega(n),$$

where $\omega(n) \ge n^{\beta} \cdot n^{\max\{1/\gamma, 1/p\}}$ for some $\beta = \beta(\delta, \gamma, p) > 0$. In other words, $\omega(n)$ grows faster than $n^{\max\{1/\gamma, 1/p\}}$ (= the upper bound in Theorem 2.2.1 and Corollary 2.2.2) by a positive power of n.

Proposition 2.2.10 is a negative result for the multiplier empirical processes in the similar vein as in [2], but more quantitatively: there is no universal moment condition for the multipliers that yield a positive solution to Question 2 when the X_i's and the ξ_i's are allowed to be dependent.

Remark 2.2.11. The basic trouble for removing the independence assumption between the X_i's and the ξ_i's can be seen by the following example. Let X_i's be i.i.d. mean-zero random variables with a finite second moment. Then clearly $\sum_{i=1}^n X_i$ grows at a rate $\mathcal{O}_{\mathbf{P}}(n^{1/2})$ by the CLT. On the other hand, let $\xi_i = \varepsilon_i X_i$ where ε_i's are independent Rademacher random variables. Then the multiplier sum $\sum_{i=1}^n \xi_i X_i = \sum_{i=1}^n \varepsilon_i X_i^2$ may grow at a rate as fast as

$\mathcal{O}_{\mathbf{P}}(n^{1-\delta})$, if $\varepsilon_1 X_1^2$ is in the domain of attraction of a symmetric stable law with index close to 1.

2.3 Nonparametric regression: least squares estimation

In this section, we apply our new multiplier inequalities in Section 2.2 to study the least squares estimator (LSE) (3.1.2) in the nonparametric regression model (2.1.1) when the errors ξ_i's are heavy-tailed (E′), independent of the X_i's (but need not be independent of each other), and the model satisfies the entropy condition (F).

Our results here are related to the recent ground-breaking work of Mendelson and his coauthors [91, 109, 112, 113]. These papers proved rate-optimality of ERM procedures under a $2+\varepsilon$ moment condition on the errors, in a general structured learning framework that contains models satisfying sub-Gaussian/small-ball conditions. Their framework also allows arbitrary dependence between the errors ξ_i's and the X_i's. See [111] for some recent development. Here the reasons for our focus on the different structure—models with entropy conditions, are twofold:

- Entropy is a standard and well-understood notion for the complexity of a large class of models, see examples in [62, 162].

- The moment condition on the errors needed to guarantee rate-optimality of the LSE in our setting is no longer a $2+\varepsilon$ moment. In fact, as we will show, $p \geq 1+2/\alpha$ (cf. Theorems 2.3.1-2.3.7) moments are needed for such a guarantee.

The reason that we work with independent errors is more fundamental: when the errors ξ_i's are allowed to be dependent on the X_i's, there is no universal moment condition on the ξ_i's alone that guarantees the rate-optimality of the LSE (cf. Proposition 2.3.12). In fact, even in the family of one-dimensional linear regression models with heteroscedastic errors of any finite p-th moment, the convergence rate of the LSE can be as slow as specified (cf. Remark 2.3.13).

2.3.1 Upper bound for the convergence rates of the LSE

Theorem 2.3.1. *Suppose that $\xi_1, \dots, \xi_n$ are mean-zero errors independent of $X_1, \dots, X_n$ with the same marginal distributions, and $\|\xi_1\|_{p,1} < \infty$ for some $p \geq 1$. Further suppose that $\mathcal{F}$ is a P-centered function class (if the ξ_i's are i.i.d. $\mathcal{F}$ need not be P-centered) such that $\mathcal{F} - f_0 \subset L_\infty(1)$ satisfies the entropy condition (F) with some $\alpha \in (0,2)$. Then the LSE $\hat{f}_n$ in (3.1.2) satisfies*

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big). \tag{2.3.1}$$

Furthermore, if ξ_i's are i.i.d. and $p \geq 2$, then (2.3.1) holds in expectation:

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}\big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big). \tag{2.3.2}$$

One interesting consequence of Theorem 2.3.1 is a convergence rate of the LSE when the errors only have a $L_{p,1}$ moment ($1 < p \leq 2$).

Corollary 2.3.2. *Suppose the assumptions in Theorem 2.3.1 hold with $p \in (1,2]$. Then*

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2}+\frac{1}{2p}}\big) = \mathfrak{o}_{\mathbf{P}}(1).$$

Consistency of the LSE has been a classical topic, see e.g. [152, 159] for sufficient and necessary conditions in this regard under a second moment assumption on the errors. Here Theorem 2.3.1 provides a quantitative rate of convergence of the LSE when the errors may not even have a second moment (under stronger conditions on $\mathcal{F}$).

The connection between the proof of Theorem 2.3.1 and the new multiplier inequality in Section 2.2 is the following reduction scheme.

Proposition 2.3.3. *Suppose that $\xi_1, \dots, \xi_n$ are mean-zero random variables independent of $X_1, \dots, X_n$, and $\mathcal{F} - f_0 \subset L_\infty(1)$. Further assume that*

$$\mathbb{E} \sup_{f \in \mathcal{F}: \|f - f_0\|_{L_2(P)} \leq \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i (f - f_0)(X_i) \right| \lesssim \phi_n(\delta), \tag{2.3.3}$$

and

$$\mathbb{E} \sup_{f \in \mathcal{F}: \|f - f_0\|_{L_2(P)} \leq \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i (f - f_0)(X_i) \right| \lesssim \phi_n(\delta). \tag{2.3.4}$$

hold for some ϕ_n *such that* $\delta \mapsto \phi_n(\delta)/\delta$ *is non-increasing. Then* $\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}(\delta_n)$ *holds for any* δ_n *such that* $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$. *Furthermore, if* $\xi_1, \ldots, \xi_n$ *are i.i.d. mean-zero with* $\|\xi_1\|_p < \infty$ *for some* $p \geq 2$, *then* $\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}(\delta_n)$ *for any* $\delta_n \geq n^{-\frac{1}{2}+\frac{1}{2p}}$ *such that* $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$.

The remaining task in the proof of Theorem 2.3.1 is a calculation of the modulus of continuity of the (multiplier) empirical process involved in (2.3.3) and (2.3.4) using Theorem 2.2.1 and local maximal inequalities for the empirical process (see Proposition 2.5.1).

Remark 2.3.4. Some remarks on the assumptions on $\mathcal{F}$.

1. The entropy condition (F) is standard in nonparametric statistics literature. The condition $\alpha \in (0, 2)$ additionally requires $\mathcal{F}$ to be a *Donsker* class. Although the proof applies to non-Donsker function classes with $\alpha \geq 2$, the first term in (2.3.1) becomes *sub-optimal* in general, see [21].

2. $\mathcal{F}$ is assumed to be P-centered when the errors ξ_i's have an arbitrary dependence structure. It is known from [168] (see Theorem 1, page 638) that for a centered function class, the minimax risk of estimating a regression function under arbitrary errors with second moments uniformly bounded, is no worse than that for i.i.d. Gaussian errors. If the errors are i.i.d., then $\mathcal{F}$ need not be P-centered (as stated in the theorem).

3. The uniform boundedness assumption on $\mathcal{F}$, including many classical examples (cf. Section 9.3 of [160]), should be primarily viewed as *a method of proof*: all that we need is $\|\hat{f}_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$. In subsequent work of the authors [77], this method is applied to shape-restricted regression problems in a heavy-tailed regression setting.

Remark 2.3.5. Here in Theorem 2.3.1 we focus on the regression model (2.1.1) with errors ξ_i's independent from X_i's. This is crucial: we show below in Proposition 2.3.12 that the independence assumption between the X_i's and ξ_i's cannot be relaxed for the rate in Theorem 2.3.1 to hold.

On the other hand, our Theorem 2.3.1 is useful in handling centered models with arbitrarily dependent errors in the regression model. This complements Mendelson's work [91, 109, 112, 113, 114] that allows arbitrary dependence between ξ_i and X_i's with independent observations in a learning framework.

Remark 2.3.6. In Theorem 2.3.1 the results are 'in probability' and 'in expectation' statements. It is easy to see from the proof that a tail estimate can be obtained for $\|\hat{f}_n - f_0\|_{L_2(P)}$: if $\|\xi_1\|_{p,1} < \infty$ for some $p \geq 2$, then

$$\mathbb{P}\big(\delta_n^{-1}\|\hat{f}_n - f_0\|_{L_2(P)} > t\big) \leq Ct^{-p},$$

where $\delta_n \equiv n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}$. Constructing estimators other than the LSE that give rise to *exponential tail bound* under a heavy-tailed regression setting is also of significant interest. We refer the readers to, e.g. [43, 94, 96] and references therein for this line of research.

2.3.2 Lower bound for the convergence rates of the LSE

At this point, (2.3.1) only serves as an *upper bound* for the convergence rates of the LSE. Since the rate $n^{-\frac{1}{2+\alpha}}$ corresponds to the optimal rate in the Gaussian regression case [169], it is natural to conjecture that this rate cannot be improved. On the other hand, the 'noise' rate $n^{-\frac{1}{2}+\frac{1}{2p}}$ is due to the reduction scheme in Proposition 2.3.3, which relates the convergence rate of the LSE to the size of the multiplier empirical process involved. It is natural to wonder if this 'noise rate' is a proof artifact due to some possible deficiency in Proposition 2.3.3.

Theorem 2.3.7. *Let $\mathcal{X} = [0,1]$ and P be a probability measure on $\mathcal{X}$ with Lebesgue density bounded away from 0 and ∞, and ξ_i's are i.i.d. mean-zero errors independent of X_i's. Then*

for each $\alpha \in (0,2)$ *and* $2 \vee \sqrt{\log n} \le p \le (\log n)^{1-\delta}$ *with some* $\delta \in (0,1/2)$*, there exists a function class* $\mathcal{F} \equiv \mathcal{F}_n$*, and some* $f_0 \in \mathcal{F}$ *with* $\mathcal{F} - f_0$ *satisfying the entropy condition (F), such that the following holds: there exists some law for the error* ξ_1 *with* $\|\xi_1\|_{p,1} \lesssim \log n$*, such that for* n *sufficiently large, there exists some least squares estimator* f_n^* *over* $\mathcal{F}_n$ *satisfying*

$$\mathbb{E}\|f_n^* - f_0\|_{L_2(P)} \ge \rho \cdot \big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big)(\log n)^{-2}.$$

Here $\rho > 0$ *is a (small) constant independent of* n*.*

Theorem 2.3.7 has two claims. The first claim justifies the heuristic conjecture that the convergence rate for the LSE with heavy-tailed errors under entropy conditions, should be no better than the optimal rate in the Gaussian regression setting. Although here we give an existence statement, the proof is constructive: in fact we use (essentially) a Hölder class. Other function classes are also possible if we can handle the Poisson (small-sample) domain of the empirical process indexed by these classes.

The second claim asserts that for any entropy level $\alpha \in (0,2)$, there exist 'hard models' for which the noise level dominates the risk for the least squares estimator. Here are some examples for these hard models:

Example 2.3.8. A benchmark model witnessing the worst case rate $\mathcal{O}(n^{-\frac{1}{2}+\frac{1}{2p}})$ (up to logarithmic factors) is (almost) the one we used in Theorem 2.2.6, i.e. the class of indicators [4] over closed intervals in $[0,1]$.

Example 2.3.9. Consider more general classes [4]

$$\mathcal{F}_k \equiv \Big\{ \sum_{i=1}^{k} c_i \mathbf{1}_{[x_{i-1},x_i]} : |c_i| \le 1, \\ 0 \le x_0 < x_1 < \ldots < x_{k-1} < x_k \le 1 \Big\}, k \ge 1.$$

The classes $\mathcal{F}_k$ also witness the worst case rate $\mathcal{O}(n^{-\frac{1}{2}+\frac{1}{2p}})$ (up to logarithmic factors) since they contain all indicators over closed intervals on $[0,1]$, and are closely related to problems

[4] excluding the indicators indexed by intervals that are too short.

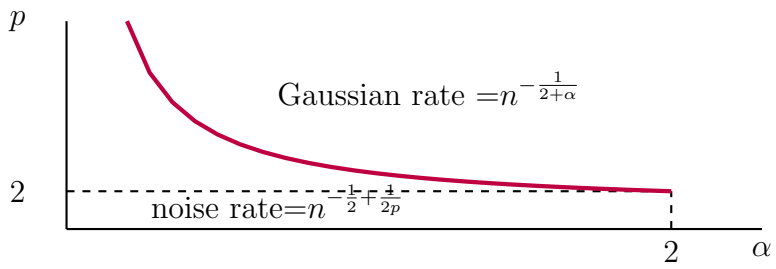

Figure 2.1: Tradeoff between the complexity of the function class and the noise level of the errors in the convergence rates for the LSE. The critical curve (purple): $p = 1 + 2/\alpha$.

in the change-point estimation/detection literature. For instance, the case $k = 1$ is of particular importance in epidemic and signal processing applications; see [8, 170] from a testing perspective of the problem. From an estimation viewpoint, [24] proposed an ℓ_0-type penalized LSE for estimating regression functions in $\mathcal{F}_k$, where a (nearly) parametric rate is obtained under a sub-Gaussian condition on the errors. Our results here suggest that such least-squares type estimators may not work well for estimating step functions with multiple change-points if the errors are heavy-tailed.

Example 2.3.10. Yet another class is given by the regression problem involving image restoration (or edge estimation), see e.g. [86, 87] or Example 9.3.7 of [160] (but we consider a random design). In particular, the class $\mathscr{C} \equiv \{\mathbf{1}_C : C \subset [0,1]^d \text{ is convex}\}$[5] also witnesses the lower bound $\mathcal{O}(n^{-\frac{1}{2}+\frac{1}{2p}})$ (up to logarithmic factors) since it contains all indicators over hypercubes on $[0,1]^d$.

2.3.3 Some positive and negative implications for the LSE

Combining Theorems 2.3.1 and 2.3.7, we see that the tradeoff in the size of the multiplier empirical process between the complexity of the function class and the heaviness of the tail of the errors (multipliers) tranlates into the convergence rate of the LSE (cf. Figure 2.1). In particular, Theorems 2.3.1 and 2.3.7 indicate both some positive and negative aspects of the

[5] excluding the indicators indexed by sets with too small volume.

LSE in a heavy-tailed regression setting:

(Positive implications for the LSE):

If $p \geq 1 + 2/\alpha$, then $\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2+\alpha}}\big)$. In this case, the noise level is 'small' compared with the complexity of the function class so that the LSE achieves the optimal rate as in the case for i.i.d. Gaussian errors (see [169]).

(Negative implications for the LSE):

If $p < 1 + 2/\alpha$, then $\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2}+\frac{1}{2p}}\big)$. In this case, the noise is so heavy-tailed that the *worst-case* rate of convergence of the LSE is governed by this noise rate (see above for examples). The negative aspect of the LSE is that this noise rate reflects a genuine deficiency of the LSE as an estimation procedure, rather than the difficulty due to the 'hard model' in such a heavy-tailed regression setting. In fact, we can design simple robust procedures to outperform the LSE in terms of the rate of convergence.

To see this, consider the least-absolute-deviation(LAD) estimator $\tilde{f}_n$ (see e.g. [58, 124, 126], or page 336 of [162]) defined by $\tilde{f}_n = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|$. It follows from a minor modification of the proof [6] in page 336 of [162] that as long as the errors $\xi_i \equiv M\eta_i$'s for some η_i admitting a smooth enough density, median zero and a first moment, and $M > 0$ not too small, then under the same conditions as in Theorem 2.3.1, the LAD estimator $\tilde{f}_n$ satisfies

$$\sup_{f_0\in\mathcal{F}} \mathbb{E}_{f_0}\|\tilde{f}_n - f_0\|_{L_2(P)} \leq \mathcal{O}\big(n^{-\frac{1}{2+\alpha}}\big),$$

where clearly the noise rate $\mathcal{O}(n^{-\frac{1}{2}+\frac{1}{2p}})$ induced by the moment of the errors does not occur. For statistically optimal procedures that do not even require a first moment on the errors, we refer the reader to [14].

---

[6] More specifically, we can proceed by replacing the empirical measure $\mathbb{P}_n$ by P, slightly restricting the suprema of the empirical process to $1/n \lesssim P(f - f_0)^2 < \delta^2$ in the third display on page 336 of [162], and noting that Theorem 3.4.1 of [162] can be strengthened to an expectation since the empirical processes involved are bounded.

It is worthwhile to note that the shortcomings of the LSE quantified here also rigorously justify the motivation of developing other robust procedures (cf. [9, 25, 28, 31, 32, 43, 80, 81, 95, 94, 96, 117]).

Remark 2.3.11. Our Theorems 2.3.1 and 2.3.7 show that the moment condition

$$p \geq 1 + 2/\alpha$$

that guarantees the LSE to converge at the optimal rate (as in the case for Gaussian errors), is the best one can hope *under entropy conditions alone*. On the other hand, this condition may be further improved if additional structure is available. For instance, in the isotonic regression case ($\alpha = 1$), our theory requires $p \geq 3$ to guarantee an optimal $n^{-1/3}$ rate for the isotonic LSE, while it is known (cf. [172]) that a second moment assumption on the errors ($p = 2$) suffices. The benefits of this extra structure due to shape constraints are investigated in further work by the authors [77].

2.3.4 An impossiblility result

In this section, dual to the impossibility result in Proposition 2.2.10 for the multiplier empirical process, we formally prove that the independence assumption between the X_i's and the ξ_i's is necessary for the rate in Theorems 2.3.1 and 2.3.7 to hold.

Proposition 2.3.12. *Consider the regression model (2.1.1) without assuming independence between the X_i's and the ξ_i's. Let $\mathcal{X} \equiv \mathbb{R}$. For every triple (δ, α, p) such that $\delta \in (0, 1/2)$, $4\delta < \alpha < 2$ and $2 \leq p < \min\{4/\delta, (2 + 4/\alpha)/(1 + (1 + 2/\alpha)\delta)\}$, there exist*

- *X_i's and ξ_i's satisfying: (i) $\{(X_i, \xi_i)\}$'s are i.i.d.; (ii) ξ_i is not independent from X_i but $\mathbb{E}[\xi_1|X_1] = 0$, $\|\xi_1\|_{p,1} < \infty$;*

- *a function class $\mathcal{F} \equiv \mathcal{F}_n$, and some $f_0 \in \mathcal{F}$ with $\mathcal{F} - f_0$ satisfying the entropy condition (F),*

such that the following holds: for n sufficiently large, there exists some least squares estimator f_n^* over $\mathcal{F}_n$ satisfying

$$\mathbb{E}\|f_n^* - f_0\|_{L_2(P)} \geq \delta_n$$

where $\delta_n \geq n^\beta \cdot (n^{-1/(2+\alpha)} \vee n^{-1/2+1/(2p)})$ for some $\beta = \beta(\delta, \alpha, p) > 0$. In other words, δ_n shrinks to 0 slower than $n^{-1/(2+\alpha)} \vee n^{-1/2+1/(2p)}$ (= the rate of the LSE in Theorems 2.3.1 and 2.3.7) by a positive power of n.

Proposition 2.3.12 is a negative result on the LSE: there is no universal moment condition on ξ_i's that guarantees the rate-optimality of the LSE when the errors ξ_i's can be dependent on the X_i's.

Remark 2.3.13. One basic model underlying the construction of Proposition 2.3.12 is the following: consider the (one-dimensional) linear regression model with heteroscedastic errors

$$Y_i = \alpha_0 X_i + \xi_i, \quad i = 1, \ldots, n$$

where $\xi_i = \varepsilon_i X_i$ for some independent Rademacher random variables ε_i's. Clearly $\mathbb{E}[\xi_i | X_i] = 0$, but ξ_i is (highly) dependent on X_i. The least squares estimator $\hat{\alpha}_n \equiv \arg\min_{\alpha \in \mathbb{R}} n^{-1} \sum_{i=1}^n (Y_i - \alpha X_i)^2$ has a closed form:

$$\hat{\alpha}_n \equiv \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \alpha_0 + \frac{\sum_{i=1}^n \varepsilon_i X_i^2}{\sum_{i=1}^n X_i^2}.$$

Suppose X_i's have a finite second moment, then by the SLLN, $\hat{\alpha}_n \to \alpha_0$ a.s., but the convergence rate of $\|\hat{f}_n - f_0\|_{L_2(P)} = |\hat{\alpha}_n - \alpha_0| \|X_1\|_2$ can be as slow as any $n^{-\delta}$: note that $\sum_{i=1}^n X_i^2 = \mathcal{O}(n)$ under the assumed second moment condition on X_i's, while the sum of the centered random variables $\sum_{i=1}^n \varepsilon_i X_i^2$ may have a growth rate $\mathcal{O}(n^{1-\delta})$ if $\varepsilon_1 X_1^2$ is in the domain of attraction of a symmetric stable law with index close to 1 (recall Remark 2.2.11).

A simple modification of the construction along the lines of the proof of Proposition 2.2.10 allows the situation where ξ_i's have a finite p-th moment ($p \geq 2$), while the convergence rate of the LSE can be as slow as $n^{-\delta}$.

So in order to derive the rate-optimality of the LSE under any universal moment condition on the errors ξ_i's, in a framework that allows arbitrary dependence between the ξ_i's and the X_i's, it is necessary to impose conditions on the model $\mathcal{F}$ to exclude the counter-examples (as in [91, 109, 112, 113, 114]).

2.4 Sparse linear regression: Lasso revisited

In this section we consider the sparse linear regression model:

$$Y = X\theta_0 + \xi \tag{2.4.1}$$

where $X \in \mathbb{R}^{n\times d}$ is a (random) design matrix and $\xi = (\xi_1, \ldots, \xi_n)$ is a mean-zero noise vector independent of X. When the true signal $\theta_0 \in \mathbb{R}^d$ is sparse, one popular estimator is the Lasso [147]:

$$\hat{\theta}(\lambda) \equiv \arg\min_{\theta\in\mathbb{R}^d} \left(\frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1 \right). \tag{2.4.2}$$

The lasso estimator has been thoroughly studied in an already vast literature; we refer readers to the monograph [29] for a comprehensive overview.

Our main interest here concerns the following question: *under what moment conditions on the distributions of X and ξ can the lasso estimator enjoy the optimal rate of convergence?* In particular, neither X nor ξ need be light tailed apriori (i.e. not sub-Gaussian), and the components $\xi_1, \ldots, \xi_n$ of the vector ξ need not be independent.

Previous work guaranteeing rate-optimality of the Lasso estimator typically assumes that both X and ξ are sub-Gaussian, see [29, 121, 156]. Relaxing the sub-Gaussian conditions in the Lasso problem is challenging: [91] showed how to remove the sub-Gaussian assumption on ξ in the case X is sub-Gaussian. The problem is even more challenging if we relax the sub-Gaussian assumption on the design matrix X. Our goal in this section is to demonstrate how the new multiplier inequality in Theorem 2.2.1, combined with (essentially) existing techniques, can be used to give a systematic treatment to the above question, in a rather straightforward fashion.

Before stating the result, we need some notion of the *compatibility condition*: For any $L > 0$ and $S \subset \{1, \ldots, d\}$, define

$$\phi(L, S) = \sqrt{|S|} \min \left\{ \frac{1}{\sqrt{n}} \|X\theta_S - X\theta_{S^c}\|_2 : \|\theta_S\|_1 = 1, \|\theta_{S^c}\|_1 \leq L \right\}.$$

Here for any $\theta = (\theta_i) \in \mathbb{R}^d$, $\theta_S \equiv (\theta_i \mathbf{1}_{i \in S})$ and $\theta_{S^c} \equiv (\theta_i \mathbf{1}_{i \notin S})$. Let $B_0(s)$ be the set of s-sparse vectors in $\mathbb{R}^d$, i.e. $\theta \in B_0(s)$ if and only if $|\{i : \theta_i \neq 0\}| \leq s$. Further let $\Sigma = \mathbb{E}\hat{\Sigma}$ where $\hat{\Sigma} = X^\top X/n$ is the sample covariance matrix, and $\underline{\sigma}_d = \sigma_{\min}(\Sigma)$ and $\bar{\sigma}_d = \sigma_{\max}(\Sigma)$ be the smallest and largest singular value of the population covariance matrix, respectively. Here $d = d_n$ and $s = s_n$ can either stay bounded or blow up to infinity in asymptotic statements.

Theorem 2.4.1. *Let X be a design matrix with i.i.d. mean-zero rows, and $0 < \liminf \underline{\sigma}_d \leq \limsup \bar{\sigma}_d < \infty$. Suppose that*

$$\min_{|S| \leq s} \phi(3, S) \geq c_0 \tag{2.4.3}$$

holds for some $c_0 > 0$ with probability tending to 1 as $n \to \infty$, and that for some $1/4 \leq \alpha \leq 1/2$,

$$\limsup_{n \to \infty} \frac{\log d \cdot \big(M_4(X) \vee \log^2 d\big)}{n^{2-4\alpha}} < \infty, \tag{2.4.4}$$

where $M_4(X) \equiv \mathbb{E} \max_{1 \leq j \leq d} |X_{1j}|^4$. Then for $\hat{\theta}^L \equiv \hat{\theta}(2L\|\boldsymbol{\xi}_n\|_{1/\alpha,1}\sqrt{\log d/n})$,

$$\lim_{L \to \infty} \limsup_{n \to \infty} \sup_{\theta_0 \in B_0(s)} \mathbb{P}_{\theta_0} \left(\frac{1}{n} \|X(\hat{\theta}^L - \theta_0)\|_2^2 > \frac{16L^2 \|\boldsymbol{\xi}_n\|_{1/\alpha,1}^2}{c_0^2} \cdot \frac{s \log d}{n} \right) = 0. \tag{2.4.5}$$

Here $\|\boldsymbol{\xi}_n\|_{1/\alpha,1} \equiv \int_0^\infty \big(\frac{1}{n}\sum_{i=1}^n \mathbb{P}(|\xi_i| > t)\big)^\alpha \, \mathrm{d}t$.

The rate $\sqrt{s \log d/n}$ in the above theorem is well-known to be (nearly) minimax optimal for prediction in the sparse linear regression model (e.g. [128]). The quantity $\|\boldsymbol{\xi}_n\|_{1/\alpha,1}$ should be thought as the 'noise level' of the regression problem. For instance, if the ξ_i's are i.i.d, and $\alpha = 1/2$, then $\|\boldsymbol{\xi}_n\|_{1/\alpha,1} = \|\xi_1\|_{2,1}$.

Although in Theorem 2.4.1 we only consider prediction error, the estimation error $\|\hat{\theta}^L - \theta_0\|_1$ can be obtained using completely similar arguments by noting that Lemma 2.4.5 below also holds for estimation error.

Remark 2.4.2. Two technical remarks.

1. As in Theorem 2.3.1, we assume in Theorem 2.4.1 that the rows of X have zero-mean as vectors in $\mathbb{R}^d$ so that arbitrary dependence structure among ξ_i's can be allowed. For i.i.d. errors, the zero-mean assumption is not needed.

2. (2.4.5) is of an asymptotic nature mainly due to the weak asymptotic assumptions made in (2.4.3) and (2.4.4). It is clear from the proof that concrete probability estimates can be obtained if a probability estimate for (2.4.3) is available.

As an illustration of the scope of Theorem 2.4.1, we consider several different scaling regimes for the parameter space (d, n, s). For simplicity of discussion we assume that the errors $\xi_1, \ldots, \xi_n$ have the same marginal distributions and the design matrix X has i.i.d. entries such that X_{11} has a Lebesgue density bounded away from ∞ and $\mathbb{E}X_{11}^2 = 1$.

Example 2.4.3. Consider the scaling regime $d/n \to \lambda \in (0, 1)$. We claim that $\mathbb{E}|X_{11}|^{4+\varepsilon} \vee \|\xi\|_{4,1} < \infty$ for some $\varepsilon > 0$ guarantees the validity of (2.4.5). First, (2.4.3) holds under the finite fourth moment condition, see [12]. Second, (2.4.4) holds under the assumed moment conditions. Note that a fourth moment condition on X_{11} is necessary: if $\mathbb{E}X_{11}^4 = \infty$, then $\limsup \bar{\sigma}_d = \infty$ a.s., see [11]. This corollary of Theorem 2.4.1 appears to be a new result; [94] considered a different 'tournament' Lasso estimator with best tradeoff between confidence statement and convergence rate under heavy-tailed designs and errors.

Example 2.4.4. If $\|X_{11}\|_p \lesssim p^\beta$ for some $\beta \geq 1/2$ and all $p \lesssim \log n$, then Theorem E of [90] showed that the compatibility condition (2.4.3) holds under $n \gtrsim s \log d \vee (\log d)^{(4\beta-1)}$. Condition (2.4.4) is satisfied if $\|\xi\|_{2+\varepsilon} < \infty$ and $\log d \lesssim \log n$.

The condition $\log d \lesssim \log n$ requires polynomial growth of d with n; this can be improved if X_{11} is light tailed. In particular, if $\mathbb{E}\exp(\mu|X_{11}|^\gamma) < \infty$ for some $\mu, \gamma > 0$, then we can take $\beta = 1/\gamma$ so that (2.4.3) holds under $n \gtrsim s \log d \vee (\log d)^{(4/\gamma)-1}$, while (2.4.4) is satisfied if $\|\xi\|_{2+\varepsilon} < \infty$ and $d \leq \exp(n^{c_{\varepsilon,\gamma}})$ for some constant $c_{\varepsilon,\gamma} > 0$. Different choices of γ lead to:

- If the entries of X have sub-exponential tails, then we may take $\gamma = 1$. In this case, (2.4.5) is valid under $\|\xi\|_{2+\varepsilon} < \infty$ subject to $n \gtrsim s \log d \vee \log^3 d$ and $d \leq \exp(n^{c_{\varepsilon,1}})$ for some constant $c_{\varepsilon,1} > 0$. This seems to be a new result; the recent result of [138] considered the similar tail condition on X along with a sub-exponential tail for the errors ξ_i's, while their rates come with additional logarithmic factors.

- If the entries of X have sub-Gaussian tails, then we may take $\gamma = 2$. In this case, (2.4.5) is valid under $\|\xi\|_{2+\varepsilon} < \infty$ subject to $n \gtrsim s \log d$ and $d \leq \exp(n^{c_{\varepsilon,2}})$ for some constant $c_{\varepsilon,2} > 0$. This recovers a recent result of [91] in the case where X and ξ are independent (up to the mild dimension constraint on d).

Now we prove Theorem 2.4.1. The following reduction (basic inequality) is well-known, cf. Theorem 6.1 of [29].

Lemma 2.4.5. *On the event* $\mathcal{E}_L \equiv \{\max_{1\leq j\leq d}|\frac{2}{n}\sum_{i=1}^n \xi_i X_{ij}| \leq L\sqrt{\log d/n}\}$, *with tuning parameter* $\lambda \equiv 2L\sqrt{\log d/n}$, *it holds that* $n^{-1}\|X(\hat{\theta}^L - \theta_0)\|_2^2 \leq 16L^2\phi^{-2}(3, S_0) \cdot s_0 \log d/n$ *where* $S_0 = \{i : (\theta_0)_i \neq 0\}$ *and* $s_0 = |S_0|$.

The difficulty involved here is that *both* X and ξ can be heavy tailed. By Theorem 2.2.1, to account for the effect of the ξ_i's, we only need to track the size of $\mathbb{E}\max_{1\leq j\leq d}|\sum_{i=1}^k \varepsilon_i X_{ij}|$ at each scale $k \leq n$. This is the content of the following Gaussian approximation lemma.

Lemma 2.4.6. *Let* $X_1, \ldots, X_n$ *be i.i.d. random vectors in* $\mathbb{R}^d$ *with covariance matrix* Σ. *If* $\sup_d \sigma_{\max}(\Sigma) < \infty$, *then for all* $k, d \in \mathbb{N}$,

$$\mathbb{E}\max_{1\leq j\leq d}\left|\sum_{i=1}^k \varepsilon_i X_{ij}\right| \lesssim \left(k\log^3 d \cdot \left(M_4(X) \vee \log^2 d\right)\right)^{1/4} + (k\log d)^{1/2}.$$

The proof of the lemma is inspired by the recent work [39] who considered Gaussian approximation of the maxima of high-dimensional random vectors by exploiting *second* moment information for the X_i's. We modify their method by taking into account the *third* moment information of X_i's induced by the symmetric Rademacher ε_i's; such a modification

proves useful in identifying certain sharp moment conditions considered in the examples (in particular Example 2.4.3). See Section 2.7.2 for a detailed proof.

Proof of Theorem 2.4.1. By Lemma 2.4.5 and the assumption on the compatibility condition (2.4.3), we see that with the choice for tuning parameter $\lambda \equiv 2L\|\boldsymbol{\xi}_n\|_{1/\alpha,1}\sqrt{\log d/n}$, the left side of (2.4.5) can be bounded by

$$\begin{aligned}\mathbb{P}_{\theta_0}\left(\frac{1}{n}\|X(\hat{\theta}^L-\theta_0)\|_2^2 > \frac{16L^2\|\boldsymbol{\xi}_n\|_{1/\alpha,1}^2}{\phi^2(3,S_0)}\cdot\frac{s\log d}{n}\right)+\mathfrak{o}(1) \qquad (2.4.6)\\ \leq \mathbb{P}\left(\max_{1\leq j\leq d}\left|\frac{2}{n}\sum_{i=1}^n \xi_i X_{ij}\right| > L\|\boldsymbol{\xi}_n\|_{1/\alpha,1}\sqrt{\frac{\log d}{n}}\right)+\mathfrak{o}(1).\end{aligned}$$

By Lemma 2.4.6, we can apply Theorem 2.2.1 with $\mathcal{F}_1 = \cdots = \mathcal{F}_n \equiv \{\pi_j : \mathbb{R}^d \to \mathbb{R}, j = 1,\ldots,d\}$ where $\pi_j(x) = x_j$ for any $x = (x_l)_{l=1}^d \in \mathbb{R}^d$, and

$$\psi_n(k) \equiv C\left(k^\alpha\left(\log^3 d\cdot\left(M_4\vee\log^2 d\right)\right)^{1/4}+k^{1/2}\sqrt{\log d}\right)$$

for any $1/4 \leq \alpha \leq 1/2$ such that (2.4.4) holds and $\|\boldsymbol{\xi}_n\|_{1/\alpha,1} < \infty$, to conclude that

$$\begin{aligned}\mathbb{E}\max_{1\leq j\leq d}\left|\sum_{i=1}^n \xi_i X_{ij}\right| &\lesssim n^\alpha\left(\log^3 d\cdot\left(M_4\vee\log^2 d\right)\right)^{1/4}\|\boldsymbol{\xi}_n\|_{1/\alpha,1}+n^{1/2}\sqrt{\log d}\|\boldsymbol{\xi}_n\|_{2,1}\\ &\lesssim\left(n^\alpha\left(\log^3 d\cdot\left(M_4\vee\log^2 d\right)\right)^{1/4}+n^{1/2}\sqrt{\log d}\right)\|\boldsymbol{\xi}_n\|_{1/\alpha,1}.\end{aligned}$$

By Markov's inequality, (2.4.6) can be further bounded (up to constants) by

$$\frac{1}{L}\left(\frac{\log d\cdot(M_4\vee\log^2 d)}{n^{2-4\alpha}}\vee 1\right)^{1/4}+\mathfrak{o}(1).$$

The claim of Theorem 2.4.1 therefore follows from the assumption (2.4.4). □

2.5 Proofs for the main results: main steps

In this section, we outline the main steps for the proofs of our main theorems. Proofs for many technical lemmas will be deferred to later sections.

2.5.1 Preliminaries

Let

$$J(\delta, \mathcal{F}, L_2) \equiv \int_0^\delta \sup_Q \sqrt{1 + \log \mathcal{N}(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, \mathrm{d}\varepsilon \tag{2.5.1}$$

denote the *uniform* entropy integral, where the supremum is taken over all discrete probability measures, and

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) \equiv \int_0^\delta \sqrt{1 + \log \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)}\, \mathrm{d}\varepsilon \tag{2.5.2}$$

denote the *bracketing* entropy integral. The following local maximal inequalities for the empirical process play a key role throughout the proof.

Proposition 2.5.1. *Suppose that $\mathcal{F} \subset L_\infty(1)$, and $X_1, \ldots, X_n$'s are i.i.d. random variables with law P. Then with $\mathcal{F}(\delta) \equiv \{f \in \mathcal{F} : Pf^2 < \delta^2\}$,*

1. *If the uniform entropy integral (3.4.2) converges, then*

$$\mathbb{E}\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}(\delta)} \lesssim \sqrt{n} J(\delta, \mathcal{F}, L_2) \left(1 + \frac{J(\delta, \mathcal{F}, L_2)}{\sqrt{n}\delta^2 \|F\|_{P,2}}\right) \|F\|_{P,2}. \tag{2.5.3}$$

2. *If the bracketing entropy integral (2.5.2) converges, then*

$$\mathbb{E}\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}(\delta)} \lesssim \sqrt{n} J_{[]}(\delta, \mathcal{F}, L_2(P)) \left(1 + \frac{J_{[]}(\delta, \mathcal{F}, L_2(P))}{\sqrt{n}\delta^2}\right). \tag{2.5.4}$$

Proof. (2.5.3) follows from [161]; see also Section 3 of [59], or Theorem 3.5.4 of [62]. (4.5.1) follows from Lemma 3.4.2 of [162]. □

We will primarily work with $F \equiv 1$ in the above inequalities. A two-sided estimate for the empirical process will be important for proving lower bounds in Theorems 2.2.6 and 2.3.7. The following definition is from [59], page 1167.

Definition 2.5.2. A function class $\mathcal{F}$ is *α-full* $(0 < \alpha < 2)$ if and only if there exists some constant $K_1, K_2 > 1$ such that both

$$\log \mathcal{N}\big(\varepsilon \|F\|_{L_2(\mathbb{P}_n)}, \mathcal{F}, L_2(\mathbb{P}_n)\big) \le K_1 \varepsilon^{-\alpha}, \qquad a.s.$$

for all $\varepsilon > 0, n \in \mathbb{N}$, and

$$\log \mathcal{N}\big(\sigma \|F\|_{L_2(P)}/K_2, \mathcal{F}, L_2(P)\big) \ge K_2^{-1} \sigma^{-\alpha}$$

hold. Here $\sigma^2 \equiv \sup_{f\in\mathcal{F}} Pf^2$, F denotes the envelope function for $\mathcal{F}$, and $\mathbb{P}_n$ is the empirical measure for i.i.d. samples $X_1, \ldots, X_n$ with law P.

The following lemma, giving a sharp two-sized control for the empirical process under the α-full assumption, is proved in Theorem 3.4 of [59].

Lemma 2.5.3. *Suppose that $\mathcal{F} \subset L_\infty(1)$ is α-full with $\sigma^2 \equiv \sup_{f\in\mathcal{F}} Pf^2$. If $n\sigma^2 \gtrsim_\alpha 1$ and $\sqrt{n}\sigma \left(\frac{\|F\|_{L_2(P)}}{\sigma}\right)^{\alpha/2} \gtrsim_\alpha 1$, then there exists some constant $K > 0$ depending only on α, K_1, K_2 such that*

$$K^{-1}\sqrt{n}\sigma\left(\frac{\|F\|_{L_2(P)}}{\sigma}\right)^{\alpha/2} \le \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} \le K\sqrt{n}\sigma\left(\frac{\|F\|_{L_2(P)}}{\sigma}\right)^{\alpha/2}.$$

Note that the right side of the inequality can also be derived from (2.5.3) (taking supremum over all finitely discrete probability measures only serves to get rid of the random entropy induced by $L_2(\mathbb{P}_n)$ norm therein).

The following lemma guarantees the existence of a particular type of α-full class that serves as the basis of the construction in the proof of Theorems 2.2.6 and 2.3.7. The proof can be found in Section 2.8.

Lemma 2.5.4. *Let $\mathcal{X}, P$ be as in Theorem 2.2.6. Then for each $\alpha > 0$, there exists some function class $\mathcal{F}$ defined on $\mathcal{X}$ which is α-full and contains $\mathcal{G} \equiv \{\mathbf{1}_{[a,b]} : 0 \le a \le b \le 1\}$.*

2.5.2 Proof of Theorem 2.2.1

The key ingredient in the proof of Theorem 2.2.1 is the following, which may be of independent interest.

Proposition 2.5.5. *Suppose Assumption A holds. For any function class $\mathcal{F}$,*

$$\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\|_{\mathcal{F}} \le \mathbb{E}\left[\sum_{k=1}^{n}(|\eta_{(k)}|-|\eta_{(k+1)}|)\mathbb{E}\left\|\sum_{i=1}^{k}\varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right] \tag{2.5.5}$$

where $|\eta_{(1)}| \ge \cdots \ge |\eta_{(n)}| \ge |\eta_{(n+1)}| \equiv 0$ are the reversed order statistics for: (i) (under (A1)) $\{2|\xi_i|\}_{i=1}^n$, (ii) (under (A2)) $\{|\xi_i - \xi_i'|\}_{i=1}^n$ with $\{\xi_i'\}$ being an independent copy of $\{\xi_i\}$.

Proof of Proposition 2.5.5. We drop $\mathcal{F}$ from the notation for supremum norm over $\mathcal{F}$ and write $\|\cdot\|$ for $\|\cdot\|_{\mathcal{F}}$. We first consider the condition (A1). Note that for $(X_1', \ldots, X_n')$ being an independent copy of $(X_1, \ldots, X_n)$, we have

$$\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\| = \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{X}}\left\|\sum_{i=1}^{n}\xi_i\big(f(X_i)-\mathbb{E}_{\boldsymbol{X}'}f(X_i')\big)\right\| \le \mathbb{E}\left\|\sum_{i=1}^{n}\xi_i\big(f(X_i)-f(X_i')\big)\right\|.$$

Here in the first equality we used the centeredness assumption on the function class $\mathcal{F}$ in (A1). Now conditional on $\boldsymbol{\xi}$, for fixed $\varepsilon_1, \ldots, \varepsilon_n$, the map $(X_1, \ldots, X_n, X_1', \ldots, X_n') \mapsto \|\sum_{i=1}^n \xi_i\varepsilon_i\big(f(X_i) - f(X_i')\big)\|$ is a permutation of the original map (without ε_i's). Since $(X_1, \ldots, X_n, X_1', \ldots, X_n')$ is the coordinate projection of a product measure, it follows by taking expectation over $\varepsilon_1, \ldots, \varepsilon_n$ that

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}\left\|\sum_{i=1}^{n}\xi_i\big(f(X_i)-f(X_i')\big)\right\| = \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{X},\boldsymbol{X}'}\left\|\sum_{i=1}^{n}\xi_i\varepsilon_i\big(f(X_i)-f(X_i')\big)\right\|. \tag{2.5.6}$$

This entails that

$$\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\| \le 2\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\varepsilon},\boldsymbol{X}}\left\|\sum_{i=1}^{n}|\xi_i|\mathrm{sgn}(\xi_i)\varepsilon_i f(X_i)\right\| = 2\mathbb{E}\left\|\sum_{i=1}^{n}|\xi_i|\varepsilon_i f(X_i)\right\| \tag{2.5.7}$$

where the equality follows since the random vector $(\mathrm{sgn}(\xi_1)\varepsilon_1, \ldots, \mathrm{sgn}(\xi_n)\varepsilon_n)$ has the same distribution as that of $(\varepsilon_1, \ldots, \varepsilon_n)$ and is independent of $\xi_1, \ldots, \xi_n$. We will simply write $|\xi_i|$ without the absolute value in the sequel for notational convenience. Let π be a permutation over $\{1, \ldots, n\}$ such that $\xi_i = \xi_{(\pi(i))}$. Then the right hand side of (2.5.7) equals

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{n}\xi_{(\pi(i))}\varepsilon_i f(X_i)\right\| &= \mathbb{E}\left\|\sum_{i=1}^{n}\xi_{(i)}\varepsilon_{\pi^{-1}(i)}f(X_{\pi^{-1}(i)})\right\| \quad \text{(by relabelling)} \\ &= \mathbb{E}\left\|\sum_{i=1}^{n}\xi_{(i)}\varepsilon_i f(X_i)\right\| \quad \text{(by invariance of } (P_X \otimes P_\varepsilon)^n\text{)}.\end{aligned} \tag{2.5.8}$$

Now write $\xi_{(i)} = \sum_{k \geq i}(\xi_{(k)} - \xi_{(k+1)})$ where $\xi_{(n+1)} \equiv 0$. The above display can be rewritten as

$$\mathbb{E}\left\| \sum_{i=1}^{n} \sum_{k=i}^{n} (\xi_{(k)} - \xi_{(k+1)}) \varepsilon_i f(X_i) \right\| = \mathbb{E}\left\| \sum_{k=1}^{n} (\xi_{(k)} - \xi_{(k+1)}) \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|. \tag{2.5.9}$$

The claim under (A1) follows by combining (2.5.7)-(2.5.9). For (A2), let ξ_i''s be an independent copy of ξ_i's. Then the analogy of (2.5.7) becomes

$$\begin{aligned} \mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\| &= \mathbb{E}\left\| \sum_{i=1}^{n} (\xi_i - \mathbb{E}\xi_i') f(X_i) \right\| \leq \mathbb{E}\| \sum_{i=1}^{n} (\xi_i - \xi_i') f(X_i) \| \\ &= \mathbb{E}\left\| \sum_{i=1}^{n} \varepsilon_i |\xi_i - \xi_i'| f(X_i) \right\| = \mathbb{E}\left\| \sum_{i=1}^{n} \varepsilon_i |\eta_i| f(X_i) \right\| \end{aligned}$$

where $\eta_i \equiv \xi_i - \xi_i'$. The claim for (A2) follows by repeating the arguments in (2.5.8) and (2.5.9). □

Proof of Theorem 2.2.1. First consider (A1). Using Proposition 2.5.5 we see that,

$$\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} \leq 2\mathbb{E}\left[\sum_{k=1}^{n} (|\xi_{(k)}| - |\xi_{(k+1)}|) \mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_n} \right]. \tag{2.5.10}$$

By the assumption that $\mathcal{F}_k \supset \mathcal{F}_n$ for any $1 \leq k \leq n$,

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_n} \leq \mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_k} \leq \psi_n(k). \tag{2.5.11}$$

Collecting (2.5.10)-(2.5.11), we see that

$$\begin{aligned} \mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} &\leq 2\mathbb{E}\left[\sum_{k=1}^{n} (|\xi_{(k)}| - |\xi_{(k+1)}|) \psi_n(k) \right] = 2\mathbb{E} \sum_{k=1}^{n} \int_{|\xi_{(k+1)}|}^{|\xi_{(k)}|} \psi_n(k) \, \mathrm{d}t \\ &\leq 2\mathbb{E} \int_0^{\infty} \psi_n\left(|\{i : |\xi_i| \geq t\}| \right) \, \mathrm{d}t \leq 2 \int_0^{\infty} \psi_n\left(\sum_{i=1}^{n} \mathbb{P}(|\xi_i| > t) \right) \mathrm{d}t \end{aligned}$$

where the last inequality follows from Fubini's theorem and Jensen's inequality, completing

the proof for the upper bound for (A1). For (A2), mimicking the above proof, we have

$$\begin{aligned}
\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} &\leq \int_0^\infty \psi_n\left(\sum_{i=1}^{n} \mathbb{P}(|\xi_i - \xi_i'| \geq t) \right) \mathrm{d}t \\
&\leq \int_0^\infty \psi_n\left(\sum_{i=1}^{n} \mathbb{P}(|\xi_i| \geq t/2) + \mathbb{P}(|\xi_i'| \geq t/2) \right) \mathrm{d}t \\
&= \int_0^\infty \psi_n\left(2\sum_{i=1}^{n} \mathbb{P}(|\xi_i| \geq t/2) \right) \mathrm{d}t \\
&= 2\int_0^\infty \psi_n\left(2\sum_{i=1}^{n} \mathbb{P}(|\xi_i| > t) \right) \mathrm{d}t.
\end{aligned}$$

The proof of the claim for (A2) is completed by noting that $\psi_n(2x) \leq 2\psi_n(x)$ due to the concavity of ψ_n and $\psi_n(0) = 0$. □

2.5.3 Proof of Theorem 2.2.6

We need the following lemma.

Lemma 2.5.6. *Suppose that $\xi_1, \ldots, \xi_n$ are i.i.d. mean-zero random variables independent of i.i.d. $X_1, \ldots, X_n$. Then*

$$\|\xi_1\|_1 \mathbb{E}\left\| \sum_{i=1}^{n} \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \leq 2\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}}.$$

Proof. The proof follows that of the left hand side inequality in Lemma 2.9.1 of [162], so we omit the details. □

Lemma 2.5.7. *Let $X_1, \ldots, X_n$ be i.i.d. random variables distributed on $[0,1]$ with a probability law P admitting a Lebesgue density bounded away from ∞. Let $\{I_i\}_{i=1}^n$ be a partition of $[0,1]$ such that $I_i \cap I_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^n I_i = [0,1]$, and $L^{-1}n^{-1} \leq |I_i| \leq Ln^{-1}$ for some absolute value $L > 0$. Then there exists some $\tau \equiv \tau_{L,P} \in (0,1)$ such that for n sufficiently large,*

$$\mathbb{P}\left(X_1, \ldots, X_n \text{ lie in at most } \tau n \text{ intervals among } \{I_i\}_{i=1}^n\right) \leq 0.5^{n-1}.$$

The proofs of Lemma 2.5.7 can be found in Section 2.8. Now we are in position to prove Theorem 2.2.6.

Proof of Theorem 2.2.6. The proof will proceed in two steps. The first step aims at establishing a lower bound for the multiplier empirical process on the order of $n^{1/\gamma}$.

Let $\alpha = 2/(\gamma - 1)$, and $\tilde{\mathcal{F}}$ be an α-full class on $\mathcal{X}$ in Lemma 2.5.4. Further let $\delta_k = k^{-1/(2+\alpha)}$ and $\tilde{\mathcal{F}}_k \equiv \tilde{\mathcal{F}}(\delta_k) = \{f \in \tilde{\mathcal{F}} : Pf^2 < \delta_k^2\}$. Then it follows from Lemma 2.5.3 that there exists some constant $K > 0$,

$$K^{-1}k^{\alpha/(2+\alpha)} \leq \mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\tilde{\mathcal{F}}_k} \leq Kk^{\alpha/(2+\alpha)}.$$

Lemma 2.5.6 now guarantees that $\mathbb{E}\|\sum_{i=1}^n \xi_i f(X_i)\|_{\tilde{\mathcal{F}}_n}$ can be bounded from below by a constant multiple of $n^{\alpha/(2+\alpha)} = n^{1/\gamma}$ where the constant depends on $\|\xi_1\|_1$. This completes the first step of the proof.

In the second step, we aim at establishing a lower bound of order $n^{1/p}$. To this end, let $\{I_j\}_{j=1}^n$ be a partition of $\mathcal{X}$ such that $L^{-1}n^{-1} \leq |I_j| \leq Ln^{-1}$. On the other hand, let $f_j \equiv \mathbf{1}_{I_j} \in \tilde{\mathcal{F}}_n$ for $1 \leq j \leq n$ (increase δ_n by constant factors if necessary), and $\mathcal{E}_n$ denote the event that $X_1, \ldots, X_n$ lie in $N \geq \tau n$ sets among $\{I_j\}_{j=1}^n$. Then Lemma 2.5.7 entails that $\mathbb{P}(\mathcal{E}_n) \geq 1 - 0.5^n \geq 1/2$ for n sufficiently large. Furthermore, let $\mathcal{I}_j \equiv \{i : X_i \in I_j\}$ and pick any $X_{\iota(j)} \in I_j$. Note that $\mathcal{I}_j$'s are disjoint, and hence conditionally on $\boldsymbol{X}$ we have

$$\begin{aligned}
\mathbb{E} \max_{1\leq j\leq \tau n} |\xi_j| &\leq \mathbb{E} \max_{1\leq j\leq N} |\xi_{\iota(j)}| \quad \text{(by i.i.d. assumption on } \xi_i\text{'s)} \\
&\leq \mathbb{E} \max_{1\leq j\leq N} \big|\xi_{\iota(j)} + \mathbb{E} \sum_{i\in\mathcal{I}_j\setminus\iota(j)} \xi_i\big| \quad (\mathcal{I}_j\text{'s are disjoint and } \mathbb{E}\xi_i = 0) \\
&\leq \mathbb{E} \max_{1\leq j\leq N} \big|\sum_{i\in\mathcal{I}_j} \xi_i\big| \quad \text{(by Jensen's inequality).}
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\tilde{\mathcal{F}}_n} &\geq \mathbb{E}\big[\max_{1\leq j\leq n} \big| \sum_{i=1}^{n} \xi_i f_j(X_i)\big|\big] \geq \mathbb{E}_{\boldsymbol{X}}\big[\mathbb{E}_{\boldsymbol{\xi}} \max_{1\leq j\leq N} \big|\sum_{i\in\mathcal{I}_j} \xi_i\big|\mathbf{1}_{\mathcal{E}_n}\big] \\
&\geq \mathbb{E}_{\boldsymbol{X}}\big[\mathbb{E}_{\boldsymbol{\xi}} \max_{1\leq j\leq \tau n} |\xi_j|\mathbf{1}_{\mathcal{E}_n}\big] \geq \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}} \max_{1\leq j\leq \tau n} |\xi_j|
\end{aligned}$$

for n sufficiently large. Now the second step follows from the assumption, and hence completing the proof. □

2.5.4 Proof of Theorem 2.3.1

We first prove Proposition 2.3.3.

Proof of Proposition 2.3.3. Let $\mathbb{M}_n f \equiv \frac{2}{n}\sum_{i=1}^n (f-f_0)(X_i)\xi_i - \frac{1}{n}\sum_{i=1}^n (f-f_0)^2(X_i)$, and $Mf \equiv \mathbb{E}\left[\mathbb{M}_n(f)\right] = -P(f-f_0)^2$. Here we used the fact that $\mathbb{E}\xi_i = 0$ and the independence assumption between $\{\xi_i\}$ and $\{X_i\}$. Then it is easy to see that

$$|\mathbb{M}_n f - \mathbb{M}_n f_0 - (Mf - Mf_0)| \le \left|\frac{2}{n}\sum_{i=1}^n (f-f_0)(X_i)\xi_i\right| + |(\mathbb{P}_n - P)(f-f_0)^2|.$$

The first claim (i.e. convergence rate in probability) follows by standard symmetrization and contraction principle for the empirical process indexed by a uniformly bounded function class, followed by an application of Theorem 3.2.5 of [162].

Now assume that $\xi_1, \ldots, \xi_n$ are i.i.d. mean-zero errors with $\|\xi_1\|_p < \infty$ for some $p \ge 2$. Fix $t \ge 1$. For $j \in \mathbb{N}$, let $\mathcal{F}_j \equiv \{f \in \mathcal{F} : 2^{j-1}t\delta_n \le \|f - f_0\|_{L_2(P)} < 2^j t\delta_n\}$. Then by a standard peeling argument, we have

$$\mathbb{P}\left(\|\hat{f}_n - f_0\|_{L_2(P)} \ge t\delta_n\right) \le \sum_{j\ge 1} \mathbb{P}\big(\sup_{f\in\mathcal{F}_j} (\mathbb{M}_n(f) - \mathbb{M}_n(f_0)) \ge 0\big).$$

Each probability term in the above display can be further bounded by

$$\begin{aligned}
&\mathbb{P}\big(\sup_{f\in\mathcal{F}_j} (\mathbb{M}_n(f) - \mathbb{M}_n(f_0) - (Mf - Mf_0)) \ge 2^{2j-2}t^2\delta_n^2\big) \\
&\le \mathbb{P}\bigg(\sup_{f\in\mathcal{F}-f_0:\|f\|_{L_2(P)}\le 2^j t\delta_n} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(X_i)\right| \ge 2^{2j-4}t^2\sqrt{n}\delta_n^2\bigg) \\
&\qquad + \mathbb{P}\bigg(\sup_{f\in\mathcal{F}-f_0:\|f\|_{L_2(P)}\le 2^j t\delta_n} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(f^2(X_i) - Pf^2\right)\right| \ge 2^{2j-3}t^2\sqrt{n}\delta_n^2\bigg).
\end{aligned}$$

By the contraction principle and moment inequality for the empirical process (Lemma 3.4.2),

we have

$$\mathbb{E}\Bigg(\sup_{\substack{f\in\mathcal{F}-f_0:\\ \|f\|_{L_2(P)}\le 2^j t\delta_n}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(X_i) \right|^2 \Bigg) \vee \mathbb{E}\Bigg(\sup_{\substack{f\in\mathcal{F}-f_0:\\ \|f\|_{L_2(P)}\le 2^j t\delta_n}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f^2(X_i) \right|^2 \Bigg)$$
$$\lesssim \left[\phi_n(2^j t\delta_n)\right]^2 + (1\vee\|\xi_1\|_2)^2 2^{2j} t^2 \delta_n^2 + (1\vee\|\xi_1\|_p)^2 n^{-1+2/p}.$$

In the above calculation we used the fact that $\mathbb{E}\max_{1\le i\le n}|\xi_i|^2 \le \|\xi_1\|_p^2 n^{2/p}$ under $\|\xi_1\|_p < \infty$. By Chebyshev's inequality,

$$\mathbb{P}\left(\|\hat{f}_n - f_0\|_{L_2(P)} \ge t\delta_n \right) \le C_\xi \sum_{j\ge 1} \left[\left(\frac{\phi_n(2^j t\delta_n)}{2^{2j} t^2 \sqrt{n}\delta_n^2} \right)^2 \vee \frac{1}{2^{2j} t^2 n \delta_n^2} \vee \frac{1}{2^{4j} t^4 n^{2-2/p} \delta_n^4} \right].$$

Under the assumption that $\delta_n \ge n^{-\frac{1}{2}+\frac{1}{2p}}$, and noting that $\phi_n(2^j t\delta_n) \le 2^j t \phi_n(\delta_n)$ by the assumption that $\delta \mapsto \phi_n(\delta)/\delta$ is non-increasing, the right side of the above display can be further bounded up to a constant by $\sum_{j\ge 1} \left(\frac{\phi_n(\delta_n)}{2^j t\sqrt{n}\delta_n^2} \right)^2 + \frac{1}{t^2} \lesssim \frac{1}{t^2}$ for $t \ge 1$. The expectation bound follows by integrating the tail estimate. □

The following lemma calculates an upper bound for the multiplier empirical process at the target rate in Theorem 2.3.1. The proof can be found in Section 2.8.

Lemma 2.5.8. *Suppose that Assumption A holds with i.i.d. $X_1, \dots, X_n$'s with law P, and $\mathcal{F} \subset L_\infty(1)$ satisfies the entropy condition (F) with $\alpha \in (0,2)$. Further assume for simplicity that ξ_i's have the same marginal distributions with $\|\xi_1\|_{p,1} < \infty$. Then with $\delta_n \equiv n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}$, we have*

$$\mathbb{E} \sup_{Pf^2\le\rho^2\delta_n^2} \left| \sum_{i=1}^n \xi_i f(X_i) \right| \vee \mathbb{E} \sup_{Pf^2\le\rho^2\delta_n^2} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|$$
$$\le \bar{K}_\alpha (\rho^{1-\alpha/2} \vee \rho^{-\alpha}) \begin{cases} n^{\frac{\alpha}{2+\alpha}} \big(1 \vee \|\xi_1\|_{1+2/\alpha,1}\big), & p \ge 1+2/\alpha, \\ n^{\frac{1}{p}} \big(1 \vee \|\xi_1\|_{p,1}\big), & 1 \le p < 1+2/\alpha. \end{cases}$$

Proof of Theorem 2.3.1. The claim follows immediately from Lemma 2.5.8 by noting that the rate δ_n chosen therein corresponds to the condition (2.3.3) in Proposition 2.3.3, along with Proposition 2.5.1 handling (2.3.4). □

2.5.5 Proof of Theorem 2.3.7

We will prove the following slightly more general version of Theorem 2.3.7.

Theorem 2.5.9. *Let $\mathcal{X} = [0,1]$ and P be a probability measure on $\mathcal{X}$ with Lebesgue density bounded away from 0 and ∞. Suppose that $\xi_1, \ldots, \xi_n$ are i.i.d. mean-zero random variables with $\|\xi_1\|_{p,1} < \infty$ for some $p \geq 2$. Then:*

1. *For each $\alpha \in (0,2)$, there exists a function class $\mathcal{F}$ and some $f_0 \in \mathcal{F}$ with $\mathcal{F} - f_0$ satisfying the entropy condition (F), such that for $p \geq 1 + 2/\alpha$, there exists some least squares estimator f_n^* over $\mathcal{F}$ satisfying*

$$\mathbb{E}\|f_n^* - f_0\|_{L_2(P)} \geq \rho \cdot n^{-\frac{1}{2+\alpha}}.$$

   Here $\rho > 0$ is a (small) constant independent of n.

2. *For each $\alpha \in (0,2)$, there exists a function class $\mathcal{F} \equiv \mathcal{F}_n$, some $f_0 \in \mathcal{F}$ with $\mathcal{F} - f_0$ satisfying the entropy condition (F), such that the following holds: suppose $\sqrt{\log n} \leq p \leq (\log n)^{1-\delta}$ for some $\delta \in (0, 1/2)$. Then there exists some law for the error ξ_1 with $\|\xi_1\|_{p,1} \lesssim \log n$, such that for n sufficiently large, there exists some least squares estimator f_n^* over $\mathcal{F}_n$ satisfying*

$$\mathbb{E}\|f_n^* - f_0\|_{L_2(P)} \geq \rho' \cdot n^{-\frac{1}{2}+\frac{1}{2p}} (\log n)^{-2}.$$

   Here $\rho' > 0$ is a (small) constant independent of n.

The strategy

The proof of Theorem 2.5.9 is technically rather involved; here we give a brief outline. There are two main steps:

1. We first show that (see Lemma 2.5.10): the risk of *some* LSE corresponds to the extreme value δ_n^* of the map

$$\delta \mapsto F_n(\delta) \equiv \sup_{f \in \mathcal{F} - f_0 : Pf^2 \leq \delta^2} (\mathbb{P}_n - P)(2\xi f - f^2) - \delta^2 \equiv E_n(\delta) - \delta^2. \tag{2.5.12}$$

This step is similar in spirit to [36, 158] (e.g. Theorem 1.1 of [36]; Lemma 3.1 of [158]); we will deal with the fact that the LSE may not be unique, and the map does not enjoy good geometric properties such as convexity in [36], or uniqueness of the extreme value of the map (2.5.12) as in [158].

2. Step 1 reduces the problem of finding δ_n^* to that of finding $\delta_1 < \delta_2$ such that $E_n(\delta_1) < F_n(\delta_2)$ [This means, $F_n(\delta) < F_n(\delta_2)$ for $\delta \leq \delta_1$, implying that the extreme value $\delta_n^* \geq \delta_1$]. Hence our task will be to find δ_1, δ_2 with matching order such that $E_n(\delta_1)$ is smaller than $F_n(\delta_2)$ up to a constant order under a specific function class. The construction of such an underlying regression function is inspired by the one used in Theorem 2.2.6. The main technical job involves (i) developing a problem-specific approach to derive an upper bound for $E_n(\rho\delta_1)$ for *small* $\rho > 0$ (corresponding to the Poisson (small-sample) domain of the empirical process where general tools fail), (ii) using a Paley-Zygmund moment argument to produce a *sharp* lower bound for $F_n(\delta_2)$ and (iii) handling the delicate fact that the $L_{p,1}$ norm is slightly stronger than the L_p norm.

The reduction scheme

Lemma 2.5.10. *Fix $\varepsilon > 0$. Let $\delta_n^* \equiv \inf\{\delta^* \geq 0 : F_n(\delta^*) \geq \sup_{\delta \in [0,\infty)} F_n(\delta) - \varepsilon\}$. Then there exists a 2ε-approximate LSE f_n^* such that*

$$\|f_n^* - f_0\|_{L_2(P)}^2 \geq (\delta_n^*)^2 - \varepsilon.$$

Proof. Without loss of generality we assume $f_0 = 0$. Let f_n^* be such that $\delta_0^2 \equiv P(f_n^*)^2 \leq (\delta_n^*)^2$ and $E_n(\delta_n^*) \leq (\mathbb{P}_n - P)(2\xi f_n^* - (f_n^*)^2) + \varepsilon$. Note that for any $f \in \mathcal{F}$,

$$\begin{aligned}(\mathbb{P}_n - P)(2\xi f - f^2) - Pf^2 \leq F_n(\|f\|_{L_2(P)}) \leq F_n(\delta_n^*) + \varepsilon = E_n(\delta_n^*) - (\delta_n^*)^2 + \varepsilon \\ \leq (\mathbb{P}_n - P)(2\xi f_n^* - (f_n^*)^2) - P(f_n^*)^2 + 2\varepsilon,\end{aligned}$$

where in the last inequality we used the definition of f_n^* and the fact that $(\delta_n^*)^2 \geq P(f_n^*)^2$. This implies that for any $f \in \mathcal{F}$,

$$\begin{aligned}\|Y - f\|_n^2 &= \mathbb{P}_n\xi^2 - \mathbb{P}_n(2\xi f - f^2) \\ &\geq \mathbb{P}_n\xi^2 - \mathbb{P}_n(2\xi f_n^* - (f_n^*)^2) - 2\varepsilon = \|Y - f_n^*\|_n^2 - 2\varepsilon.\end{aligned}$$

Hence f_n^* is a 2ε-approximate LSE. The claim follows if we can show $\delta_0^2 \geq (\delta_n^*)^2 - \varepsilon$. This is valid: if $(\delta_n^*)^2 > \delta_0^2 + \varepsilon$, then

$$\begin{aligned}F_n(\delta_0) &\geq (\mathbb{P}_n - P)(2\xi f_n^* - (f_n^*)^2) - \delta_0^2 \\ &\geq E_n(\delta_n^*) - \varepsilon - \delta_0^2 = F_n(\delta_n^*) - \varepsilon + ((\delta_n^*)^2 - \delta_0^2) > F_n(\delta_n^*),\end{aligned}$$

a contradiction to the definition of δ_n^* by noting that $\delta_0 \leq \delta_n^*$. □

Since $\varepsilon > 0$ in Lemma 2.5.10 can be arbitrarily small, in the following analysis we will assume without loss of generality that $\varepsilon = 0$.

The following simple observation summarizes the strategy for finding a lower bound on the rate of *some* least squares estimator.

Proposition 2.5.11. *Suppose that $0 < \delta_1 < \delta_2$ are such that $E_n(\delta_1) < F_n(\delta_2)$. Then there exists a LSE f_n^* such that $\|f_n^* - f_0\|_{L_2(P)} \geq \delta_1$.*

Proof. The condition implies that $F_n(\delta) = E_n(\delta) - \delta^2 \leq E_n(\delta) \leq E_n(\delta_1) < F_n(\delta_2)$ for any $\delta \leq \delta_1$ and hence a maximizer δ_n^* of the map $\delta \mapsto F_n(\delta)$ cannot lie in $[0, \delta_1]$, i.e. $\delta_n^* \geq \delta_1$. The claim now follows by Lemma 2.5.10. □

In the next few subsections, we first prove claim (1) of Theorem 2.5.9. The proof of claim (2) follows the similar proof strategy as that of claim (1); the details will be delayed until the last subsection.

Upper bound

The regression function class $\mathcal{F}$ we consider will be the Hölder class constructed in Lemma 2.5.4 with $\mathcal{X} = [0, 1]$, and $f_0 \equiv 0$. We first handle the *upper bound* part of the problem.

Lemma 2.5.8 is awkward in this regard because general tools (Proposition 2.5.1) cannot handle the Poisson (small-sample) domain of the empirical process, and hence the resulting bound is *insensitive* with respect to small $\rho > 0$. The following lemma remedies this for our special function class $\mathcal{F}$.

Lemma 2.5.12. *Suppose that $\mathcal{X} = [0,1]$ and P is a probability measure on $\mathcal{X}$ with Lebesgue density bounded away from 0 and ∞. Suppose further that the ξ_i's are i.i.d. mean-zero and $\|\xi_1\|_{p,1} < \infty$ and $p \geq 1 + 2/\alpha$. Then for any $\rho \in (0,1)$, if*

$$n \geq \min\{n \geq 3 : \rho^2 \geq \log n(n^{-\alpha/(2+\alpha)})\}, \tag{2.5.13}$$

then with $\delta_n \equiv n^{-\frac{1}{2+\alpha}}$ we have,

$$\mathbb{E} \sup_{f \in \mathcal{F}: Pf^2 \leq \rho^2 \delta_n^2} \left| \sum_{i=1}^n \left(2\xi_i f(X_i) - f^2(X_i) + Pf^2\right) \right| \leq \bar{K}_{P,\alpha} \rho^{1-\alpha/2} n^{\frac{\alpha}{2+\alpha}} (1 \vee \|\xi_1\|_{p,1}).$$

Note that in the above lemma we may choose ρ small as long as the sample size condition (2.5.13) is satisfied. The key idea of the proof is to compare $\sup_{f \in \mathcal{F}: Pf^2 < \sigma^2} \mathbb{P}_n f^2$ with σ^2 directly for *(nearly) the whole range of σ^2 including the Poisson (small-sample) domain* by exploiting the geometry of $\mathcal{F}$. Details can be found in Section 2.8.

Lower bound

Next we turn to the *lower* bound part of the problem. We will first consider a lower bound in expectation, then a Paley-Zygmund type argument translates the claim from in expectation to in probability.

Lemma 2.5.13. *Let $\mathcal{X} = [0,1]$, and P be a probability measure on $\mathcal{X}$ with Lebesgue density bounded away from 0 and ∞. Let $\xi_1, \ldots, \xi_n$ be i.i.d. mean-zero random variables such that $\|\xi_1\|_1 > 0$, and $\mathcal{F}$ be the Hölder class constructed in Lemma 2.5.4. Then with $\delta_n \equiv n^{-\frac{1}{2+\alpha}}$, if $\|\xi_1\|_1 > \mathfrak{K}_{\alpha,P}$ and $\vartheta \geq 1$ for some $\mathfrak{K}_{\alpha,P} > 0$ depending only on α, P,*

$$\mathbb{E} \sup_{Pf^2 \leq \vartheta^2 \delta_n^2} \left| \sum_{i=1}^n \left(2\xi_i f(X_i) - f^2(X_i) + Pf^2\right) \right| \geq \underline{K}_{P,\alpha} \|\xi_1\|_1 \vartheta^{1-\alpha/2} \cdot n^{\frac{\alpha}{2+\alpha}}.$$

The proof uses Lemmas 2.5.3 and 2.5.6, and the α-fullness of $\tilde{\mathcal{F}}$; see Section 2.8. The following Paley-Zygmund lower bound is standard.

Lemma 2.5.14 (Paley-Zygmund)**.** *Let Z be any non-negative random variable. Then for any $\varepsilon > 0$, $\mathbb{P}(Z > \varepsilon \mathbb{E}Z) \geq \left(\frac{(1-\varepsilon)\mathbb{E}Z}{(\mathbb{E}Z^q)^{1/q}} \right)^{q'}$, where $q, q' \in (1, \infty)$ are conjugate indices: $1/q + 1/q' = 1$.*

Now we turn the lower bound in expectation in Lemma 2.5.13 to a probability bound by a Paley-Zygmund argument.

Lemma 2.5.15. *Consider the same setup as in Lemma 2.5.13 with $p \geq 2$, and let $Z = \sup_{Pf^2 \leq \vartheta^2 \delta_n^2} |\sum_{i=1}^n 2\xi_i f(X_i) - f^2(X_i) + Pf^2|$. Suppose $p \geq 1 + 2/\alpha$. If $\|\xi_1\|_p < \infty$, $\|\xi_1\|_1 > \mathfrak{K}_\alpha$, $\vartheta \geq 1$ and $1 < q \leq p$. Then*

$$\mathbb{P}\big(Z \geq \frac{1}{2} \underline{K}_{P,\alpha} \|\xi_1\|_1 \vartheta^{1-\alpha/2} \cdot n^{\frac{\alpha}{2+\alpha}}\big) \geq 2^{-q/(q-1)} \bar{L}_{\alpha,\xi,\vartheta,q,P}^{-1/(q-1)} > 0.$$

The constant in the probability estimate is defined below.

Proof. Lemma 2.5.13 entails that $\mathbb{E}Z \geq \underline{K}_{P,\alpha} \|\xi_1\|_1 \vartheta^{1-\alpha/2} \cdot n^{\frac{\alpha}{2+\alpha}}$. By the moment inequality Lemma 3.4.2, if the ξ_i's have finite p-th moments, and $q \leq p$,

$$\begin{aligned}
\frac{\mathbb{E}Z^q}{(\mathbb{E}Z)^q} &\leq C_q \bigg[1 + \frac{(\sqrt{n}(\|\xi_1\|_2 \vee 1)\vartheta\delta_n)^q}{(\mathbb{E}Z)^q} + \frac{1 \vee \mathbb{E}\max_{1 \leq i \leq n} |\xi_i|^q}{(\mathbb{E}Z)^q} \bigg] \\
&\leq C_q \bigg[1 + 2 \left(\underline{K}_{P,\alpha} \|\xi_1\|_1 \wedge 1 \right)^{-q} \\
&\qquad \times \left(\vartheta^{\alpha q/2} (\|\xi_1\|_2 \vee 1)^q n^{-\frac{q\alpha}{2(2+\alpha)}} \vee \vartheta^{q(\alpha/2-1)} \|\xi_1\|_p^q n^{\frac{q}{p} - \frac{q\alpha}{2+\alpha}} \right) \bigg] \\
&\leq C_q \left[1 + 2 \left(\underline{K}_{P,\alpha} \|\xi_1\|_1 \wedge 1 \right)^{-q} \vartheta^{\alpha q/2} (\|\xi_1\|_p \vee 1)^q \right] \equiv \bar{L}_{\alpha,\xi,\vartheta,q,P}.
\end{aligned}$$

In the second inequality we used $\|\max_i |\xi_i|\|_q \leq \|\max_i |\xi_i|\|_p \leq n^{1/p} \|\xi_1\|_p$, and the third inequality follows by noting $\vartheta \geq 1$ and the assumption $p \geq 1 + 2/\alpha$. The proof is complete. □

Putting the pieces together

Proposition 2.5.16. *Suppose* $\|\xi_1\|_{p,1} < \infty$ *for* $p \geq \max\{2, 1+2/\alpha\}$. *If* $\|\xi_1\|_1 > \mathfrak{K}_{\alpha,P}$ *for some (large) constant* $\mathfrak{K}_{\alpha,P} > 0$, *then for* n *sufficiently large, there exist constants* $\rho_{\xi,\alpha,P} < \vartheta_{\xi,\alpha,P}$ *such that on an event with positive probability* $\mathfrak{p}_1 = \mathfrak{p}_1(\alpha, \xi, P) > 0$ *independent of* n,

$$F_n\big(\vartheta_{\xi,\alpha,P} \cdot n^{-\frac{1}{2+\alpha}}\big) > E_n\big(\rho_{\xi,\alpha,P} \cdot n^{-\frac{1}{2+\alpha}}\big).$$

Proof. Lemma 2.5.12 and Markov's inequality entail that for any $\rho > 0$, if $n \geq \min\{n \geq 3 : \rho^2 \geq \log n(n^{-\alpha/(2+\alpha)})\}$, then on an event with probability at least $1 - 1/M$,

$$nE_n(\rho\delta_n) \leq M\bar{C}_{\xi,\alpha,P} \cdot \rho^{1-\alpha/2} n^{\frac{\alpha}{2+\alpha}}. \tag{2.5.14}$$

We will choose ρ, M later on. On the other hand, apply Lemma 2.5.15 with $q = 2$ (since $p \geq 2$) and $\vartheta = (\underline{K}_{P,\alpha}\|\xi_1\|_1/4)^{\frac{2}{2+\alpha}}$ (we may increase $\|\xi_1\|_1$ to ensure $\vartheta \geq 1$ if necessary) we see that on an event with probability at least $2\mathfrak{p}_1 \equiv 2\mathfrak{p}_1(\alpha, \xi, P)$, we have

$$\begin{aligned} nF_n(\vartheta\delta_n) &\geq \frac{1}{2}\underline{K}_{P,\alpha}\|\xi_1\|_1\vartheta^{1-\alpha/2} \cdot n^{\frac{\alpha}{2+\alpha}} - \vartheta^2 n^{\frac{\alpha}{2+\alpha}} \\ &\geq \frac{1}{4}\underline{K}_{P,\alpha}\|\xi_1\|_1\vartheta^{1-\alpha/2} \cdot n^{\frac{\alpha}{2+\alpha}} \equiv \underline{C}_{\xi,\alpha,P} \cdot n^{\frac{\alpha}{2+\alpha}}. \end{aligned} \tag{2.5.15}$$

First we choose $M = 1/\mathfrak{p}_1$ so that with probability at least $\mathfrak{p}_1$, (2.5.14) and (2.5.15) hold simultaneously. Then we choose $\rho = \min\{(\mathfrak{p}_1\underline{C}_{\xi,\alpha,P}/2\bar{C}_{\xi,\alpha,P})^{\frac{2}{2-\alpha}}, \vartheta/2\}$ to conclude $F_n(\vartheta\delta_n) > E_n(\rho\delta_n)$ with probability at least $\mathfrak{p}_1$. □

Now we have completed the program outlined in Proposition 3.4.5.

Proof of Theorem 2.5.9: claim (1). Recall that the regression function class is taken from Lemma 2.5.4 with $f_0 \equiv 0$. Combining the proof outline Proposition 3.4.5, with Proposition 2.5.16, we see that there exists an event with probability at least $\mathfrak{p}_1 = \mathfrak{p}_1(\alpha, \xi, P) > 0$, on which at least one least squares estimator f_n^* over $\mathcal{F}$ satisfies $\|f_n^* - f_0\|_{L_2(P)} \geq \rho \cdot n^{-\frac{1}{2+\alpha}}$, where $\rho > 0$ is a (small) constant independent of n. The claim now follows by bounding the expectation from below on this event. □

Remaining proofs for Theorem 2.3.7

Here we prove the second claim of Theorem 2.5.9. Without loss of generality, we only consider $d = 1$, and the probability measure P is assumed to be uniform for simplicity. To this end, let $\tilde{\mathcal{F}}_n \equiv \{\mathbf{1}_{[a,b]} : a, b \in [0,1] \cap \mathbb{Q}, b - a \geq \delta_n^2\} \cup \{0\}$, and $\tilde{\mathcal{G}}_n \equiv \{g \in C^{1/\alpha}([0,1]) : Pg^2 \geq \delta_n^2\}$, and $f_0 = 0$, where $\delta_n = \rho n^{-1/2+1/2p'}$ with $\frac{1}{p} - \frac{1}{p'} = \varepsilon$ for some numeric constants $\varepsilon, \rho > 0$ to be specified later. Let $\mathcal{F} \equiv \tilde{\mathcal{F}}_n \cup \tilde{\mathcal{G}}_n$.

In the current case $E_n(\delta) = 0$ for $\delta < \delta_n$ and hence our goal is to give a lower bound for $F_n(\delta_n)$. We mimic the proof strategy of the first claim of Theorem 2.5.9 by (i) giving a lower bound for the multiplier empirical process in expectation, and then (ii) using the Paley-Zygmund moment argument to translate the lower bound in probability. The arguments are somewhat delicate due to the fact that the $L_{p,1}$ norm is stronger than the L_p norm. For notational simplicity, let $Z \equiv \sup_{f \in \tilde{\mathcal{F}}_n : Pf^2 \leq \delta_n^2} |\sum_{i=1}^n \xi_i f(X_i)|$, and $\tilde{Z} \equiv \sup_{f \in \tilde{\mathcal{F}}_n : Pf^2 \leq \delta_n^2} |\sum_{i=1}^n 2\xi_i f(X_i) - f^2(X_i) + Pf^2|$.

Lemma 2.5.17. *Suppose that $\xi_1, \ldots, \xi_n$ are i.i.d. symmetric random variables with $\mathbb{P}(|\xi_1| > t) = 1/(1 + |t|^{p'})$. Further suppose that $\rho \leq (64/e^3)^{1/6}$, $\varepsilon^{-1/2} \vee 3 \leq p \leq \log n / \log\log n$ and $n \geq \min\{n \geq 2 : \delta_n = \rho n^{-1/2+1/2p'} \geq \sqrt{\log n / n}\}$. Then there exists some absolute constant $C_1 > 0$, and for any $C_2 > 0$, there exists some constant $C_3 = C_3(C_2) > 0$ such that*

$$\mathbb{P}\big(\tilde{Z} \geq \frac{1}{16} n^{1/p' - 1/(p')^2} - C_3 n^{1/2p'} \sqrt{\log n}\big) \geq C_1 \left((\varepsilon p') \wedge 1\right)^2 n^{-4\varepsilon} - e^{-C_2 \log n}.$$

We need the following before the proof of Lemma 2.5.17.

Lemma 2.5.18. *Suppose $\rho \leq (64/e^3)^{1/6}$ and $\xi_1, \ldots, \xi_n$ are i.i.d. mean-zero random variables. Then for $n \geq 2$, we have $\mathbb{E}Z \geq \frac{1}{2}\mathbb{E}\max_{1 \leq j \leq 4^{-1} n^{1-1/p'}} |\xi_j|$.*

Proof. Let $I_j \equiv [(j-1)\delta_n^2, j\delta_n^2] \subset [0,1]$ for $j = 1, \ldots, N$ where $N = \delta_n^{-2} \leq \rho^{-2} n$. Note that for any $c \in (0,1)$,

$$\begin{aligned}
&\mathbb{P}\left(X_1, \ldots, X_n \text{ lie in at most } cN \text{ intervals among } \{I_j\}_{j=1}^N\right) \\
&= \mathbb{P}\left(\cup_{|\mathcal{I}| = cN} \{X_1, \ldots, X_n \in \cup_{i \in \mathcal{I}} I_i\}\right) \\
&\leq \binom{N}{cN} c^n \leq e^{cN \log(e/c) - n \log(1/c)} \leq e^{\left(c\rho^{-2} \log(e/c) - \log(1/c)\right) n}.
\end{aligned}$$

By choosing $c = \rho^2/4$, the exponent in the above display can be further bounded by $\frac{1}{4}\log(e/c) - \log(1/c) = \frac{1}{4}\log(c^3 e) = \frac{1}{4}\log(e\rho^6/64) \leq -\frac{1}{2}$ where the last inequality follows by the assumption that $\rho \leq (64/e^3)^{1/6}$. Hence we conclude that on an event $\mathcal{E}$ with probability at least $1 - 0.61^n$, the samples $X_1, \ldots, X_n$ must occupy at least $\rho^2 N/4$ many intervals among $\{I_j\}_{j=1}^N$. This implies that

$$\mathbb{E}Z = \mathbb{E} \sup_{f \in \tilde{\mathcal{F}}_n : Pf^2 \leq \delta_n^2} \left| \sum_{i=1}^n \xi_i f(X_i) \right| \geq \mathbb{E} \max_{1 \leq j \leq N} \left| \sum_{i \in I_j} \xi_i \mathbf{1}_{I_j}(X_i) \right| \mathbf{1}_{\mathcal{E}} \geq \frac{1}{2} \mathbb{E} \max_{1 \leq j \leq \rho^2 N/4} |\xi_j|$$

where we used the same arguments as in the proof of Theorem 2.2.6. The claim now follows by noting $\rho^2 N/4 = \rho^2 \delta_n^{-2}/4 = n^{1-1/p'}/4$. □

We also need some auxiliary results.

Lemma 2.5.19. *Suppose that $\xi_1, \ldots, \xi_n$ are i.i.d. symmetric random variables with $p(t) \equiv \mathbb{P}(|\xi_1| > t) = 1/(1 + |t|^p)$. Then for any $1 \leq q < p$, and $n \geq 2$,*

$$\frac{1}{4} n^{1/p} \leq \mathbb{E} \max_{1 \leq i \leq n} |\xi_i| \leq \left(\mathbb{E} \max_{1 \leq i \leq n} |\xi_i|^q \right)^{1/q} \leq \left(\frac{p+q}{p-q} \right)^{1/q} n^{1/p}.$$

We need the following exact characterization concerning the size of maxima of a sequence of independent random variables due to [63], see also Corollary 1.4.2 of [41].

Lemma 2.5.20. *Let $\xi_1, \ldots, \xi_n$ be a sequence of independent non-negative random variables such that $\|\xi_i\|_r < \infty$ for all $1 \leq i \leq n$. For $\lambda > 0$, set $\delta_0(\lambda) \equiv \inf\{t > 0 : \sum_{i=1}^n \mathbb{P}(\xi_i > t) \leq \lambda\}$. Then*

$$\frac{1}{1+\lambda} \sum_{i=1}^n \mathbb{E}\xi_i^r \mathbf{1}_{\xi_i > \delta_0} \leq \mathbb{E} \max_{1 \leq i \leq n} \xi_i^r \leq \frac{1}{1 \wedge \lambda} \sum_{i=1}^n \mathbb{E}\xi_i^r \mathbf{1}_{\xi_i > \delta_0}.$$

Proof of Lemma 2.5.19. For $\lambda \equiv 1$ in Lemma 3.5.3, $\delta_0 = \inf\{t > 0 : np(t) \leq 1\} = (n-1)^{1/p}$.

Lemma 3.5.3 now yields that for $q < p$,

$$\begin{aligned}\mathbb{E}\max_{1\le i\le n}|\xi_i|^q &\le n\mathbb{E}|\xi_1|^q\mathbf{1}_{|\xi_1|>\delta_0}\\ &= n\left[\mathbb{P}(|\xi_1|>\delta_0)\int_0^{\delta_0} qu^{q-1}\,\mathrm{d}u + \int_{\delta_0}^{\infty} qu^{q-1}\mathbb{P}(|\xi_1|>u)\,\mathrm{d}u\right]\\ &\le \frac{n\delta_0^q}{1+\delta_0^p} + qn\int_{\delta_0}^{\infty}\frac{1}{u^{p-q+1}}\,\mathrm{d}u\\ &= (n-1)^{q/p} + \frac{q}{p-q}\frac{n}{n-1}(n-1)^{q/p} \le \frac{p+q}{p-q}n^{q/p}\end{aligned}$$

since $n \ge 2$. For a lower bound for $\mathbb{E}\max_{1\le i\le n}|\xi_i|$, we proceed similarly as above by using $1+u^p \le 2u^p$ on $[\delta_0,\infty)$ for $n \ge 2$:

$$\mathbb{E}\max_{1\le i\le n}|\xi_i| \ge \frac{n}{2}\left[\frac{\delta_0}{1+\delta_0^p} + \int_{\delta_0}^{\infty}\frac{1}{2u^p}\,\mathrm{d}u\right] \ge \frac{(n-1)^{1/p}}{2} \ge \frac{1}{4}n^{1/p}.$$

This completes the proof. □

We also need Talagrand's concentration inequality [145] for the empirical process in the form given by Bousquet [23], recorded as follows.

Lemma 2.5.21. *[Theorem 3.3.9 of [62]] Let $\mathcal{F}$ be a countable class of real-valued measurable functions such that $\sup_{f\in\mathcal{F}}\|f\|_\infty \le b$. Then*

$$\mathbb{P}\big(\sup_{f\in\mathcal{F}}|\mathbb{G}_n f| \ge \mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_n f| + \sqrt{2\bar{\sigma}^2 x} + bx/3\sqrt{n}\big) \le e^{-x},$$

where $\bar{\sigma}^2 \equiv \sigma^2 + 2bn^{-1/2}\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_n f|$ with $\sigma^2 \equiv \sup_{f\in\mathcal{F}}\mathrm{Var}_P f$, and $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$.

In applications, since

$$\begin{aligned}\sqrt{2\bar{\sigma}^2 x} &\le \sqrt{2\sigma^2 x} + \sqrt{4(bx/\sqrt{n})\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_n f|}\\ &\le \sqrt{2\sigma^2 x} + \delta^{-1}(bx/\sqrt{n}) + \delta\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_n f|\end{aligned}$$

by the elementary inequality $2ab \le \delta^{-1}a^2 + \delta b^2$, we have for any $\delta > 0$,

$$\mathbb{P}\big(\sup_{f\in\mathcal{F}}|\mathbb{G}_n f| \ge (1+\delta)\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_n f| + \sqrt{2\sigma^2 x} + (3^{-1}+\delta^{-1})bx/\sqrt{n}\big) \le e^{-x}. \tag{2.5.16}$$

We will mainly use the above form (2.5.16) in the proofs.

Proof of Lemma 2.5.17. The proof is divided into two steps. In the first step, we handle Z, i.e. the multiplier empirical process part. In the second step we handle the residual term, i.e. the purely empirical process part.

(Step 1) We first claim that there exists some absolute constant $C_1 > 0$ such that

$$\mathbb{P}\big(Z \geq \frac{1}{32} n^{1/p'-1/(p')^2}\big) \geq C_1 \left((\varepsilon p') \wedge 1\right)^2 n^{-4\varepsilon}. \tag{2.5.17}$$

Let $\mathcal{G}$ be the class of indicators functions $\mathbf{1}_{[a,b]}$ with $0 \leq a \leq b \leq 1$. Then since $\delta_n \geq \sqrt{\log n/n}$, by local maximal inequalities for the empirical process (Proposition 2.5.1),

$$\mathbb{E} \sup_{g \in \mathcal{G}: Pg^2 \leq \delta_n^2} \left| \sum_{i=1}^n \varepsilon_i g(X_i) \right| \lesssim \sqrt{n}\delta_n \sqrt{\log(1/\delta_n)} \lesssim n^{1/2p'} \sqrt{\log n} \tag{2.5.18}$$

where in the last inequality we used $\rho \lesssim 1$. Applying Theorem 2.2.1 and noting that $p \leq \log n / \log\log n$ implying $n^{1/2p'}\sqrt{\log n} \leq n^{1/2p}\sqrt{\log n} \leq n^{1/p}$, we see that for some absolute constant $C > 0$, $\mathbb{E}Z \leq \mathbb{E}\sup_{g \in \mathcal{G}: Pg^2 \leq \delta_n^2} |\sum_{i=1}^n \xi_i g(X_i)| \leq C n^{1/p} \|\xi_1\|_{p,1}$. By the moment inequality Lemma 3.4.2, we have for any $q \geq 1$,

$$\begin{aligned}
\mathbb{E}Z^q &\leq C_q\big((\mathbb{E}Z)^q + (\sqrt{n}\|\xi_1\|_2\delta_n)^q + \mathbb{E}\max_{1\leq i\leq n}|\xi_i|^q\big) \\
&\leq C_q'\big(n^{q/p}\|\xi_1\|_{p,1}^q + n^{q/2p'}\|\xi_1\|_2^q + \mathbb{E}\max_{1\leq i\leq n}|\xi_i|^q\big) \\
&\leq 2C_q'\big(n^{q/p}(\|\xi_1\|_{p,1} \vee \|\xi_1\|_2)^q + \mathbb{E}\max_{1\leq i\leq n}|\xi_i|^q\big).
\end{aligned}$$

Now using the Paley-Zygmund inequality (Lemma 3.5.2) and Lemma 2.5.18, we see that

$$\begin{aligned}
\mathbb{P}\big(Z > \frac{1}{2}\mathbb{E}Z\big) &\geq 2^{-q'}\left(\frac{\mathbb{E}Z}{(\mathbb{E}Z^q)^{1/q}}\right)^{q'} \\
&\geq C_q''\left(\frac{\mathbb{E}\max_{1\leq j\leq 4^{-1}n^{1-1/p'}}|\xi_j|}{n^{1/p}(\|\xi_1\|_{p,1} \vee \|\xi_1\|_2) + (\mathbb{E}\max_{1\leq i\leq n}|\xi_i|^q)^{1/q}}\right)^{q'}.
\end{aligned}$$

By Lemma 2.5.19, the above display can be further estimated from below by

$$\begin{aligned}
&C_q'''\left(\frac{n^{1/p'-1/(p')^2}}{n^{1/p}(\|\xi_1\|_{p,1} \vee \|\xi_1\|_2) + [(p'+q)/(p'-q)]^{1/q} n^{1/p'}}\right)^{q'} \\
&\geq C'\left(\frac{n^{1/p'-1/(p')^2}}{n^{1/p}(\|\xi_1\|_{p,1} \vee \|\xi_1\|_2) + n^{1/p'}}\right)^2 \quad (\text{choose } q = q' = 2) \\
&\geq C''\left((\varepsilon p') \wedge 1\right)^2 n^{-\frac{2}{p}+\frac{2}{p'}-\frac{2}{(p')^2}} \geq C''\left((\varepsilon p') \wedge 1\right)^2 n^{-4\varepsilon}
\end{aligned}$$

where in the last line we use the following facts: (i) $\frac{1}{p}-\frac{1}{p'}=\varepsilon$, (ii) $p'\geq p\geq \varepsilon^{-1/2}$ and (iii)

$$\begin{aligned}\|\xi_1\|_{p,1}&=\int_0^\infty \mathbb{P}(|\xi_1|>t)^{1/p}\,\mathrm{d}t=\int_0^\infty \frac{1}{(1+t^{p'})^{1/p}}\,\mathrm{d}t\\&\asymp 1+\int_1^\infty \frac{\mathrm{d}t}{t^{p'/p}}=1+\frac{1}{p'/p-1}=1+\frac{1}{\varepsilon p'},\\\|\xi_1\|_2&\leq\|\xi_1\|_{2,1}\asymp 1+\frac{1}{(p'/2)-1}\asymp 1.\end{aligned}$$

The proof of (2.5.17) is complete by noting that

$$\mathbb{E}Z\geq\frac{1}{2}\mathbb{E}\max_{1\leq j\leq 4^{-1}n^{1-1/p'}}|\xi_j|\geq\frac{1}{8}4^{-1/p'}n^{1/p'-1/(p')^2}\geq 16^{-1}n^{1/p'-1/(p')^2}.$$

(Step 2) We next claim that for any $C_2>0$, there exists some absolute constant $C_3>0$ such that

$$\mathbb{P}\Big(\sup_{f\in\tilde{\mathcal{F}}_n:Pf^2\leq\delta_n^2}\big|\sum_{i=1}^n(f^2(X_i)-Pf^2)\big|\leq C_3n^{1/2p'}\sqrt{\log n}\Big)\geq 1-e^{-C_2\log n}. \tag{2.5.19}$$

Note that by contraction principle and (2.5.18),

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}_n:Pf^2\leq\delta_n^2}\left|\sum_{i=1}^n(f^2(X_i)-Pf^2)\right|\lesssim\mathbb{E}\sup_{g\in\mathcal{G}:Pg^2\leq\delta_n^2}\left|\sum_{i=1}^n\varepsilon_ig(X_i)\right|\lesssim n^{1/2p'}\sqrt{\log n}.$$

The claim (2.5.19) now follows from the above display combined with Talagrand's concentration inequality (Lemma 3.4.4) applied with $x=C_2\log n$.

Now the claimed inequality in the lemma follows by considering the event that is the intersection of the events indicated in (2.5.17) and (2.5.19). □

Now we are in a good position to prove the second claim of Theorem 2.5.9.

Proof of Theorem 2.5.9: claim (2). Suppose that $n^{1/(p')^2}\leq 2, p'<\frac{\log n}{2\log(64C_3)+\log\log n}$ and $\rho=1/8$. Then

$$\begin{aligned}&\frac{1}{16}n^{1/p'-1/(p')^2}-C_3n^{1/2p'}\sqrt{\log n}-n\delta_n^2\\&=\big(\frac{1}{16n^{1/(p')^2}}-\rho^2\big)n^{1/p'}-C_3n^{1/2p'}\sqrt{\log n}\geq\frac{1}{64}n^{1/p'}-C_3n^{1/2p'}\sqrt{\log n}>0.\end{aligned}$$

Since $\left\{\tilde{Z} \geq \frac{1}{16}n^{1/p'-1/(p')^2} - C_3 n^{1/2p'}\sqrt{\log n}\right\} \subset \left\{\tilde{Z} - n\delta_n^2 > 0\right\} \subset \{F_n(\delta_n) > E_n(\delta_n/2)\}$, it follows from Lemma 2.5.17 that

$$\mathbb{P}(\mathcal{E}_n) \equiv \mathbb{P}\left(F_n(\delta_n) > E_n(\delta_n/2)\right) \geq C_1\left((\varepsilon p') \wedge 1\right)^2 n^{-4\varepsilon} - e^{-C_2 \log n},$$

provided further $\varepsilon \geq 1/p^2$, $p \geq 3$, $n \geq 2$ and $n^{1/p'} \geq 64 \log n$. Equivalently,

$$\varepsilon \geq 1/p^2, \quad p \geq 3, \quad n \geq 2, \quad \sqrt{\log_2 n} \leq p' \leq \frac{\log n}{\log(64 \log n) + 2\log(64 C_3)}.$$

Furthermore, since $p' = p/(1 - p\varepsilon) \leq 2p$ if $\varepsilon \leq 1/2p$, it suffices to require

$$1/p^2 \leq \varepsilon \leq 1/2p, \quad n \geq n_\delta \vee e^{64C_3^2}, \quad \sqrt{\log_2 n} \leq p \leq (\log n)^{1-\delta}\left[\leq \frac{\log n}{2\log(64 \log n)}\right],$$

where $n_\delta \equiv \min\{n \geq 2 : (\log n)^\delta \geq 2\log(64 \log n)\}$. Hence for n in the indicated range (i.e. sufficiently large depending on δ, C_3), we have

$$\mathbb{P}(\mathcal{E}_n) \geq C_1 (\log n)^{-2} n^{-4\varepsilon} - e^{-C_2 \log n} \geq C (\log n)^{-2} n^{-4\varepsilon},$$

where the first inequality follows from $\varepsilon p' \geq \varepsilon p \geq 1/p \geq 1/\log n$, and the second inequality follows for n sufficiently large by choosing $C_2 = 3$ since $n^{-4\varepsilon} \geq n^{-2}$. By Proposition 3.4.5,

$$\begin{aligned}\mathbb{E}\|f_n^* - f_0\|_{L_2(P)} &\geq \mathbb{E}\|f_n^* - f_0\|_{L_2(P)} \mathbf{1}_{\mathcal{E}_n} \\ &\geq \frac{\delta_n}{2} \cdot C (\log n)^{-2} n^{-4\varepsilon} \geq C' n^{-1/2+1/2p-4.5\varepsilon} (\log n)^{-2}.\end{aligned}$$

For any $1/(1-\delta) < a < 2$ so that $p = (\log n)^{1/a}$, we may choose $\varepsilon = p^{-a}$. The claim then follows by noting $n^{-\varepsilon} = n^{-1/\log n} = e^{-1}$, and $\|\xi_1\|_{p,1} \asymp 1 + (\varepsilon p)^{-1} = 1 + (\log n)^{1-1/a} \lesssim \log n$. □

2.6 Proof of impossibility results

In this section we prove Propositions 2.2.10 and 2.3.12.

Proof of Proposition 2.2.10. Let X_i's be i.i.d. symmetric random variables with the tail probability $\mathbb{P}(|X_1| > x) = x^{-(2+\delta)}$. Let $\xi_i = |X_i|^{2/p}\varepsilon_i$, where ε_i's are Rademacher random

variables independent of all other random variables. Then it is easy to check that (i)-(ii) hold (in particular, $\delta > 0$ guarantees $\|\xi_1\|_{p,1} < \infty$). Now take any $\mathcal{F}$ satisfying the entropy condition (F) with exponent $\alpha = 2/(\gamma - 1) \in (0, 2)$, and let $\mathcal{F}_k \equiv \{f \in \mathcal{F} : Pf^2 < \delta_k^2\} \cup \{\delta_k \mathfrak{e}\}$ where $\delta_k \equiv k^{-1/(2+\alpha)}$ and $\mathfrak{e}(x) = x$. Then by Proposition 2.5.1, we have $\mathbb{E}\|\sum_{i=1}^k \varepsilon_i f(X_i)\|_{\mathcal{F}_k} \lesssim k^{1/\gamma}$. On the other hand, since $p < 4/\delta$, applying Theorem 3.7.2 of [51] we see that for n large enough,

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}_n} &\geq \delta_n \mathbb{E}\left|\sum_{i=1}^n \xi_i \mathfrak{e}(X_i)\right| = \delta_n \mathbb{E}\left|\sum_{i=1}^n \varepsilon_i |X_i|^{1+2/p}\right| \\ &\gtrsim \delta_n n^{\frac{1+2/p}{2+\delta}} \equiv r_n.\end{aligned}$$

The equality in the first line of the above display follows from

$$\xi_i \mathfrak{e}(X_i) = |X_i|^{2/p} \varepsilon_i X_i =_d |X_i|^{2/p} \varepsilon_i \varepsilon_i' |X_i| =_d \varepsilon_i |X_i|^{1+2/p}$$

by symmetry of X_i's (ε_i''s are independent copies of ε_i's).

Now in order that $r_n \gg n^{1/\gamma}$,

$$\begin{aligned}\frac{1+2/p}{2+\delta} - \frac{1}{2+\alpha} > \frac{1}{\gamma} = \frac{\alpha}{2+\alpha} &\Leftrightarrow \frac{1+2/p}{2+\delta} > \frac{1+\alpha}{2+\alpha} \\ &\Leftrightarrow p < \frac{2(1+2/\alpha)}{1+(1+1/\alpha)\delta}.\end{aligned}$$

Hence it suffices to require that $p < 2\gamma/(1+\gamma\delta)$. On the other hand, in order that $r_n \gg n^{1/p}$, it suffices to require that

$$\frac{1+2/p}{2+\delta} - \frac{1}{2+\alpha} > \frac{1}{p} \Leftrightarrow \frac{1}{2+\delta} - \frac{1}{2+\alpha} > \frac{\delta}{2+\delta}\frac{1}{p}.$$

Since $p \geq 2$, we only need to check that

$$\frac{1}{2+\delta} - \frac{1}{2+\alpha} > \frac{\delta/2}{2+\delta},$$

which holds since we choose $\alpha > 4\delta$ (which is equivalent to $\gamma = 1 + 2/\alpha < 1 + 1/(2\delta)$). This completes the proof. □

Proof of Proposition 2.3.12. We only sketch the proof here. Let X_i's, ξ_i's and $\mathfrak{e}$ be defined as in the proof of Proposition 2.2.10, $f_0 = 0$, and $\mathcal{F}$ be any function class defined on $[0,1]$ satisfying the entropy condition (F) with exponent $\alpha \in (0,2)$. Let $\delta_n > 0$ be determined later on, and $\mathcal{F}_n \equiv \{f \in \mathcal{F} : Pf^2 \geq \delta_n^2\} \cup \{\delta_n \mathfrak{e}\} \cup \{0\}$. Then $E_n(\delta_n/2) = 0$ and we only need to show that with positive probability, $F(\delta_n) > 0$. To see this, note that we have shown in Proposition 2.2.10 that $|\sum_{i=1}^n \xi_i \mathfrak{e}(X_i)| \gtrsim n^{\frac{1+2/p}{2+\delta}}$ with positive probability for n large enough. Furthermore, by Talagrand's concentration inequality (cf. Lemma 3.4.4), we have that with overwhelming probability,

$$n^{-1/2} \sup_{f \in \mathcal{F}_n} \left| \sum_{i=1}^n (f^2(X_i) - Pf^2) \right| \lesssim \mathbb{E} \sup_{f \in \mathcal{F}_n} |\mathbb{G}_n f^2| + \delta_n + n^{-1/2} \lesssim_\alpha \delta_n^{1-\alpha/2},$$

since $\mathbb{E} \sup_{f \in \mathcal{F}_n} |\mathbb{G}_n f^2| \lesssim \mathbb{E} \sup_{f \in \mathcal{F}_n} |n^{-1/2} \sum_{i=1}^n \varepsilon_i f(X_i)| \lesssim \delta_n^{1-\alpha/2}$ by using the standard contraction principle and Proposition 2.5.1. Hence if $\delta_n \gtrsim n^{-\frac{1}{2+\alpha}}$, then for n large enough, we have with positive probability,

$$\begin{aligned} nF_n(\delta_n) &\geq \delta_n \left| \sum_{i=1}^n \xi_i \mathfrak{e}(X_i) \right| - \sup_{f \in \mathcal{F}_n} \left| \sum_{i=1}^n (f^2(X_i) - Pf^2) \right| - n\delta_n^2 \\ &\geq C_1 \delta_n n^{\frac{1+2/p}{2+\delta}} - C_2 n^{1/2} \delta_n^{1-\alpha/2} - n\delta_n^2 \\ &\geq C_1 n^\beta (n^{\frac{\alpha}{2+\alpha}} \vee n^{\frac{1}{p}}) - C_2 n^{1/2} \delta_n^{1-\alpha/2} - n\delta_n^2 \end{aligned}$$

for some $\beta \equiv \beta(\delta, \alpha, p) > 0$, where the last inequality follows from the arguments in the proof of Proposition 2.2.10, by assuming that $2 < \gamma < 1 + 1/(2\delta)$ and $2 \leq p < \min\{4/\delta, 2\gamma/(1+\gamma\delta)\}$ where $\gamma = 1 + 2/\alpha$. This condition is equivalent to $4\delta < \alpha < 2$ and $2 \leq p < \min\{4/\delta, (2 + 4/\alpha)/(1 + (1 + 2/\alpha)\delta)\}$. Hence we may choose $\delta_n = C_3 n^{\beta/2}(n^{-\frac{1}{2+\alpha}} \vee n^{-1/2+1/(2p)})$ for some constant $C_3 > 0$ to ensure that the last line of the above display is > 0 for n large enough. □

2.7 Remaining proofs I

2.7.1 Proof of Lemma 2.2.9

Proof of Lemma 2.2.9. Without loss of generality we assume that $a_n \leq 1$ for all $n = 0, 1, \ldots$. For any $\varepsilon \in (0,1)$, since a_n vanishes asymptotically, there exists some N_ε for which $a_n \leq \varepsilon$

as long as $n \geq N_\varepsilon$. Consider

$$\psi_\varepsilon(t) \equiv \begin{cases} \varphi(t), & t \leq N_\varepsilon; \\ (1-\varepsilon)\varphi(N_\varepsilon) + \varepsilon\varphi(t), & t > N_\varepsilon. \end{cases}$$

Then it is easy to verify that ψ_ε is a concave function and majorizes $n \mapsto a_n\varphi(n)$ [since $\psi_\varepsilon(n) \geq \varepsilon\varphi(n) \geq a_n\varphi(n)$ for $n \geq N_\varepsilon$ and $\psi_\varepsilon(n) = \varphi(n) \geq a_n\varphi(n)$ for $n < N_\varepsilon$ by the assumption that $a_n \leq 1$]. Hence by definition of ψ, it follows that $\limsup_{t\to\infty} \psi(t)/\varphi(t) \leq \lim_{t\to\infty} \psi_\varepsilon(t)/\varphi(t) = \varepsilon$. The claim now follows by taking $\varepsilon \to 0$. □

2.7.2 Proof of Lemma 2.4.6

We need some auxiliary lemmas before the proof of Lemma 2.4.6.

Lemma 2.7.1. *Let $F_\beta : \mathbb{R}^d \to \mathbb{R}$ be the soft-max function defined by $F_\beta(x) = \beta^{-1}\log\big(\sum_{i=1}^d \exp(\beta x_i)\big)$. Then $\sup_{x\in\mathbb{R}^d}\sum_{j,k,l,m=1}^d |\partial_{jklm}F_\beta(x)| \leq 25\beta^3$.*

Proof. Let $\pi_j(x) = \partial_j F_\beta(x) = \exp(\beta x_j)/\sum_{i=1}^d \exp(\beta x_i)$ and $\delta_{ij} = \mathbf{1}_{i=j}$. Then it is easy to verify that (see Lemma 4.3 of [40])

$$\begin{aligned} \partial_{jk}F_\beta &= \partial_j(\pi_k) = \beta(\delta_{jk}\pi_j - \pi_j\pi_k), \\ \partial_{jkl}F_\beta &= \beta^2\big[\delta_{jk}\delta_{jl}\pi_l - \delta_{jk}\pi_j\pi_l + 2\pi_j\pi_k\pi_l - (\delta_{jl}+\delta_{kl})\pi_k\pi_l\big]. \end{aligned}$$

Furthermore taking the derivative with respect to x_m, we have

$$\begin{aligned} \partial_{jklm}F_\beta &= \beta^2\big[\delta_{jk}\delta_{jl}\partial_m(\pi_l) - \delta_{jk}\partial_m(\pi_j\pi_l) + 2\partial_m(\pi_j\pi_k\pi_l) - (\delta_{jl}+\delta_{kl})\partial_m(\pi_k\pi_l)\big] \\ &= \beta^2\big[(I) - (II) + 2(III) - (IV)\big], \end{aligned}$$

where $(I) = \beta\delta_{jk}\delta_{jl}\big(\delta_{lm}\pi_l - \pi_l\pi_m\big)$, $(II) = \beta\delta_{jk}\big[(\delta_{jm}+\delta_{lm})\pi_j\pi_l - 2\pi_j\pi_l\pi_m\big]$, $(III) = \beta\big[(\delta_{jm}+\delta_{km}+\delta_{lm})\pi_j\pi_k\pi_l - 3\pi_j\pi_k\pi_l\pi_m\big]$, and $(IV) = \beta(\delta_{jl}+\delta_{kl})\big[(\delta_{jm}+\delta_{km})\pi_j\pi_k - 2\pi_j\pi_k\pi_m\big]$. Now summing over j,k,l,m by noting that $\sum_j \pi_j = 1$ yields the claim. □

Proof of Lemma 2.4.6. In the proof we write $k \equiv n$ to line up with the notation used in [39]. Slightly abusing the notation, we use simply X_i's to denote $\varepsilon_i X_i$'s. Let $Y_1, \ldots, Y_n$ be

centered i.i.d. Gaussian random vectors in $\mathbb{R}^d$ such that $\mathbb{E}Y_1Y_1^\top = \mathbb{E}X_1X_1^\top$. We first claim that it suffices to prove that,

$$|\mathbb{E}F_\beta(X) - F_\beta(Y)| \lesssim \beta^3 n^{-1}\mathbb{E}\big(\max_{1\le j\le d}|X_{1j}|^4 \vee |Y_{1j}|^4\big) \equiv \beta^3 n^{-1}\bar{M}_4, \tag{2.7.1}$$

where $X = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \in \mathbb{R}^d$, $Y = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i \in \mathbb{R}^d$, and $F_\beta : \mathbb{R}^d \to \mathbb{R}$ is defined by $F_\beta(x) = \beta^{-1}\log\big(\sum_{i=1}^d \exp(\beta x_i)\big)$. Once (2.7.1) is proved, we use the inequality $0 \le F_\beta(x) - \max_{1\le j\le d} x_j \le \beta^{-1}\log d$ to obtain that

$$\left|\mathbb{E}\left|\max_{1\le j\le d}\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\right| - \mathbb{E}\left|\max_{1\le j\le d}\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_{ij}\right|\right| \lesssim \beta^3 n^{-1}\bar{M}_4 + \beta^{-1}\log d.$$

The conclusion of Lemma 2.4.6 follows by taking $\beta = \big(n\log d/\bar{M}_4\big)^{1/4}$ and controlling the size of the Gaussian maxima.

The proof of (2.7.1) proceeds in similar lines as in Lemma I.1 of [39] by a fourth moment argument instead of a third moment one used therein. We provide details below. Let $Z(t) = \sqrt{t}X + \sqrt{1-t}Y = \sum_{i=1}^n Z_i(t)$ be the Slepian's interpolation between X and Y, where $Z_i(t) = \frac{1}{\sqrt{n}}(\sqrt{t}X_i + \sqrt{1-t}Y_i)$. Let $Z^{(i)}(t) = Z(t) - Z_i(t)$. Then,

$$\begin{aligned}\mathbb{E}F_\beta(X) - \mathbb{E}F_\beta(Y) &= \mathbb{E}F_\beta(Z(1)) - \mathbb{E}F_\beta(Z(0)) \\ &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}F_\beta(Z(t))\,\mathrm{d}t = \int_0^1 \sum_{i=1}^n\sum_{j=1}^d \mathbb{E}\big[\partial_j F_\beta(Z(t))\dot{Z}_{ij}(t)\big]\,\mathrm{d}t\end{aligned} \tag{2.7.2}$$

where $\dot{Z}_{ij}(t) = \frac{1}{2\sqrt{n}}\big(\frac{1}{\sqrt{t}}X_{ij} - \frac{1}{\sqrt{1-t}}Y_{ij}\big)$. Now using Taylor expansion for $\partial_j F_\beta(\cdot)$ at $Z^{(i)}(t)$, we have

$$\begin{aligned}\partial_j F_\beta(Z(t)) = \partial_j F_\beta(Z^{(i)}(t)) &+ \sum_k \partial_{jk}F_\beta(Z^{(i)}(t))Z_{ik}(t) \\ &+ \sum_{k,l}\partial_{jkl}F_\beta(Z^{(i)}(t))Z_{ik}(t)Z_{il}(t) \\ &+ \sum_{k,l,m}\int_0^1 \partial_{jklm}F_\beta(Z^{(i)}(t)+\tau Z_i(t))Z_{ik}(t)Z_{il}(t)Z_{im}(t)\,\mathrm{d}\tau.\end{aligned} \tag{2.7.3}$$

Hence (2.7.2) can be split into four terms according to (2.7.3). Now the key observation here is that $Z^{(i)}(t)$ is independent of $Z_{i\cdot}, \dot{Z}_{i\cdot}$. Since $\mathbb{E}\dot{Z}_{ij}(t) = 0$, the contribution of the

first order term in (2.7.3) vanishes. Similar observation holds for the second and third order terms. For the second order term, we only need to verify $\mathbb{E}\dot{Z}_{ij}(t)Z_{ik}(t) = 0$; this follows from the construction of Y that matches the second moments of X: $\mathbb{E}\dot{Z}_{ij}(t)Z_{ik}(t) = \frac{1}{2n}\mathbb{E}\left(\frac{1}{\sqrt{t}}X_{ij} - \frac{1}{\sqrt{1-t}}Y_{ij}\right)\left(\sqrt{t}X_{ik} + \sqrt{1-t}Y_{ik}\right) = \frac{1}{2n}\left(\mathbb{E}X_{ij}X_{ik} - \mathbb{E}Y_{ij}Y_{ik}\right) = 0$. For the third order term,

$$
\begin{aligned}
&\mathbb{E}\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t) \\
&= \frac{1}{2n^{3/2}}\mathbb{E}\big(\frac{1}{\sqrt{t}}X_{ij} - \frac{1}{\sqrt{1-t}}Y_{ij}\big)\big(\sqrt{t}X_{ik} + \sqrt{1-t}Y_{ik}\big)\big(\sqrt{t}X_{il} + \sqrt{1-t}Y_{il}\big) \\
&= (2n^{3/2})^{-1}\big(\sqrt{t}\mathbb{E}X_{ij}X_{ik}X_{il} - \sqrt{1-t}\mathbb{E}Y_{ij}Y_{ik}Y_{il}\big).
\end{aligned}
$$

Cross terms in the calculation of the last line in the above display all vanish by the independence and centeredness of X and Y. The first term of the above display is 0 since (recall X_i stands for $\varepsilon_i X_i$ throughout the proof) $\mathbb{E}\varepsilon_i^3 X_{ij}X_{ik}X_{il} = \mathbb{E}\varepsilon_i^3 \cdot \mathbb{E}X_{ij}X_{ik}X_{il} = 0$ by the independence between the Rademacher ε_i and X_i. The second term is also zero by a similar argument: since $Y_i =_d \varepsilon_i Y_i$ for a Rademacher random variable ε_i independent of Y_i, $\mathbb{E}Y_{ij}Y_{ik}Y_{il} = \mathbb{E}\varepsilon_i^3 \cdot \mathbb{E}Y_{ij}Y_{ik}Y_{il} = 0$. Hence the only non-trivial contribution of (2.7.3) in (2.7.2) is the fourth order term:

$$
\begin{aligned}
&|\mathbb{E}F_\beta(X) - F_\beta(Y)| \\
&\le \sum_{i=1}^n \sum_{j,k,l,m=1}^d \int_0^1 \int_0^1 \mathbb{E}|\partial_{jklm}F_\beta(Z^{(i)}(t) + \tau Z_i(t))\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)Z_{im}(t)|\mathrm{d}\tau\ \mathrm{d}t \\
&\le \sum_{i=1}^n \int_0^1 \int_0^1 \mathbb{E}\Big[\sum_{j,k,l,m=1}^d |\partial_{jklm}F_\beta(Z^{(i)}(t) + \tau Z_i(t))| \\
&\qquad\qquad \times \max_{1\le k,l,m\le d} |\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)Z_{im}(t)|\Big]\mathrm{d}\tau\ \mathrm{d}t \\
&\le 25\beta^3 \sum_{i=1}^n \int_0^1 \mathbb{E} \max_{1\le j,k,l,m\le d} |\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)Z_{im}(t)|\ \mathrm{d}t
\end{aligned}
$$

where the last inequality follows from the dimension free property of the third derivatives of

soft max function F_β (Lemma 2.7.1). Now the claim (2.7.1) follows by noting that

$$\begin{aligned}\mathbb{E}\max_{1\le j,k,l,m\le d}|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)Z_{im}(t)| &\le \big(\mathbb{E}\max_{1\le j\le d}|\dot{Z}_{ij}|^4\big)^{1/4}\big(\mathbb{E}\max_{1\le j\le d}|Z_{ij}|^4\big)^{3/4}\\ &\lesssim n^{-2}\bigg(\frac{1}{\sqrt{t}}\vee\frac{1}{\sqrt{1-t}}\bigg)\big(\mathbb{E}\max_{1\le j\le d}|X_{1j}|^4\vee|Y_{1j}|^4\big)\end{aligned}$$

and the fact that the integral $\int_0^1\big(\frac{1}{\sqrt{t}}\vee\frac{1}{\sqrt{1-t}}\big)\,\mathrm{d}t<\infty$ converges. □

2.8 Remaining proofs II

2.8.1 Proof of Lemma 2.5.4

Proof of Lemma 2.5.4. Without loss of generality we assume P is uniform on $\mathcal{X}\equiv[0,1]$. Take $\mathcal{F}=C^{1/\alpha}([0,1])$ to be a $1/\alpha$-Hölder class on $[0,1]$ (see Section 2.7 of [162]). Let $\tilde{\mathcal{F}}\equiv\mathcal{F}\cup\mathcal{G}$. For any discrete probability measure Q on $\mathcal{X}=[0,1]$,

$$\begin{aligned}\mathcal{N}(\varepsilon,\tilde{\mathcal{F}},L_2(Q)) &\le \mathcal{N}(\varepsilon,\mathcal{F},L_2(Q))+\mathcal{N}(\varepsilon,\mathcal{G},L_2(Q))\\ &\le \mathcal{N}(\varepsilon,\mathcal{F},L_\infty([0,1]))+\sup_Q\mathcal{N}(\varepsilon,\mathcal{G},L_2(Q)),\end{aligned}$$

where the last inequality follows from the fact that any ε-cover of $\mathcal{F}$ in L_∞ metric on $[0,1]$ induces an ε-cover on the function class $\mathcal{F}$ under any $L_2(Q)$ on $\mathcal{X}$. Now by Theorem 2.7.1 of [162] and the fact that $\mathcal{G}$ is a bounded VC-subgraph function class (see Section 2.6 of [162]), we have the following entropy estimate:

$$\sup_Q\log\mathcal{N}(\varepsilon,\tilde{\mathcal{F}},L_2(Q))\lesssim\varepsilon^{-\alpha}. \tag{2.8.1}$$

where the supremum is taken over all discrete probability measures supported on $\mathcal{X}$. On the other hand, for some small $c>0$,

$$\mathcal{N}(c\sigma,C^{1/\alpha}([0,1])\cap L_2(\sigma),L_2([0,1]))\gtrsim\exp(c'\sigma^{-\alpha})$$

holds for another constant $c'>0$ for all $\sigma>0$, due to the classical work of [30, 149] in the context of more general Besov spaces. The connection here is $C_1^{1/\alpha}([0,1])=B_{\infty,\infty}^{1/\alpha}(1)$ (in the usual notation for Besov space, see Proposition 4.3.23 of [62]). See also [151], page 103-106

for an explicit construction for a (local) minimax lower bound in L_2 metric for the Hölder class (which is essentially the same problem), where a set of testing functions $\{f_i : i \leq M\}$ is constructed such that $M \geq 2^{m/8}$, $\|f_j - f_k\|_{L_2} \gtrsim m^{-1/\alpha}$ and $\|f_j\|_{L_2} \lesssim m^{-1/\alpha}$. Hence we see that

$$\log \mathcal{N}(c\sigma, \tilde{\mathcal{F}} \cap L_2(\sigma), L_2([0,1])) \gtrsim \sigma^{-\alpha}. \tag{2.8.2}$$

The claim follows by combining (2.8.1) and (2.8.2). □

2.8.2 Proof of Lemma 2.5.7

Proof of Lemma 2.5.7. Note that the event in question equals

$$\cup_{|\mathcal{I}| \leq \tau n} \{X_1, \ldots, X_n \in \cup_{i \in \mathcal{I}} I_i\}.$$

Hence with $K \equiv \sup_{x \in [0,1]} |(\mathrm{d}P/\mathrm{d}\lambda)(x)|$, the probability in question can be bounded by

$$\begin{aligned}
\sum_{k \leq \tau n} \binom{n}{k} \left(k \cdot Ln^{-1} \cdot K\right)^n &\leq \sum_{k \leq \tau n} \exp\left(k \log(en/k)\right) (k/n)^n \cdot (LK)^n \\
&= (LK)^n \sum_{k \leq \tau n} \exp\left(k \log(en/k) - n \log(n/k)\right) \\
&\leq (e^\tau LK)^n \sum_{k \leq \tau n} \exp\left(-(n-k)\log(n/k)\right) \\
&\leq (e^\tau LK)^n \sum_{k \leq \tau n} \exp\left(-(1-\tau)n \log(n/k)\right) \\
&= (e^\tau LK)^n \sum_{k \leq \tau n} \left(\frac{k}{n}\right)^{(1-\tau)n} \\
&\leq (e^\tau LK)^n n \int_0^{\tau + 1/n} x^{(1-\tau)n} \, \mathrm{d}x \\
&= \frac{n}{(1-\tau)n+1} (\tau + 1/n) \cdot \left[e^\tau LK (\tau + 1/n)^{1-\tau}\right]^n \\
&\leq 0.5^{n-1},
\end{aligned}$$

for $\tau < \min\{1/2, 1/8e(LK)^2\}$ and $n \geq \max\{2, 8e(LK)^2\}$. The first line uses the standard inequality $\binom{n}{k} \leq n^k/k! \leq (en/k)^k$, since $e^k = \sum_{i=0}^{\infty} k^i/i! \geq k^k/k!$. The last line follows since

$\frac{n}{(1-\tau)n+1}(\tau+1/n) \le \frac{n}{n/2+1}(1/2+1/2) \le 2$ and $e^\tau LK(\tau+1/n)^{1-\tau} \le \sqrt{e(LK)^2(\tau+1/n)} \le 1/2$ by the conditions on τ and n. □

2.8.3 Proof of Lemma 2.5.8

Proof of Lemma 2.5.8. If $p \ge 1 + 2/\alpha$, then $\delta_n = n^{-\frac{1}{2+\alpha}}$. By local maximal inequalities for empirical processes (see Proposition 2.5.1), we have

$$\begin{aligned} \mathbb{E} \sup_{Pf^2 \le \rho^2 \delta_k^2} \left| \sum_{i=1}^k \varepsilon_i f(X_i) \right| &\le C\sqrt{k}(\rho\delta_k)^{1-\alpha/2} \left(1 \vee \frac{(\rho\delta_k)^{1-\alpha/2}}{\sqrt{k}(\rho\delta_k)^2} \right) \\ &\le Ck^{\frac{\alpha}{2+\alpha}} \rho^{1-\alpha/2} (1 \vee \rho^{-(1+\alpha/2)}) \\ &\le C(\rho^{1-\alpha/2} \vee \rho^{-\alpha}) \cdot k^{\frac{\alpha}{2+\alpha}}. \end{aligned} \tag{2.8.3}$$

Applying Corollary 2.2.2 we see that

$$\mathbb{E} \sup_{Pf^2 \le \rho^2 \delta_n^2} \left| \sum_{i=1}^n \xi_i f(X_i) \right| \le C(\rho^{1-\alpha/2} \vee \rho^{-\alpha}) \cdot n^{\frac{\alpha}{2+\alpha}} \big(1 \vee \|\xi_1\|_{1+2/\alpha,1}\big).$$

If $p < 1 + 2/\alpha$, then $\delta_n = n^{-\frac{1}{2}+\frac{1}{2p}}$. In this case,

$$\begin{aligned} \mathbb{E} \sup_{Pf^2 \le \rho^2 \delta_k^2} \left| \sum_{i=1}^k \varepsilon_i f(X_i) \right| &\le C\sqrt{k}(\rho\delta_k)^{1-\alpha/2} \left(1 \vee \frac{(\rho\delta_k)^{1-\alpha/2}}{\sqrt{k}(\rho\delta_k)^2} \right) \\ &\le C\rho^{1-\alpha/2} \cdot k^{\frac{1}{2}\left(\frac{1}{p}+\frac{\alpha}{2}\cdot\frac{p-1}{p}\right)} \left(1 + \rho^{-(1+\alpha/2)} k^{\frac{1}{2}\left(-\frac{1}{p}+\frac{\alpha}{2}\cdot\frac{p-1}{p}\right)} \right) \\ &\le C(\rho^{1-\alpha/2} \vee \rho^{-\alpha}) k^{\frac{1}{p}} \end{aligned} \tag{2.8.4}$$

where the last inequality follows from $\frac{1}{p} > \frac{\alpha}{2} \cdot \frac{p-1}{p}$ by the assumed relationship between p and α. Now apply Corollary 2.2.2 we have

$$\mathbb{E} \sup_{Pf^2 \le \rho^2 \delta_n^2} \left| \sum_{i=1}^n \xi_i f(X_i) \right| \le C(\rho^{1-\alpha/2} \vee \rho^{-\alpha}) \cdot n^{\frac{1}{p}} (\|\xi_1\|_{p,1} \vee 1)$$

as desired. □

2.8.4 Proof of Lemma 2.5.12

We first need the following.

Lemma 2.8.1. *Let $X_1, \ldots, X_n$ be i.i.d. P on $[0,1]$ where P has Lebesgue density bounded away from 0 and ∞. Set $\gamma_n = \kappa_P \log n/n$ where $\kappa_P \geq 1$ is a constant depending only on P. Let $I_j \equiv [(j-1)\gamma_n, j\gamma_n)$ for $j = 1, \ldots, n/(\kappa_P \log n) \equiv N$. Then for some $c_P > 0, n_P$ sufficiently large depending on P, if $n \geq n_P$, with probability at least $1 - 2n^{-2}$, all intervals $\{I_j\}$ contain at least one and at most $c_P \log n$ samples.*

Proof. Without loss of generality we assume that P is uniform on $[0,1]$. The general case where P has Lebesgue density bounded away from 0 and ∞ follows from minor modification. Let $\mathcal{E}_1(\mathcal{E}_2)$ be the event that all intervals $\{I_j\}$ contain at least one sample(at most $c \log n$ samples). Then for $\kappa_P = 6$,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}\left(\cup_{1\leq j\leq N}\{I_j \text{ contains no samples}\}\right)\\ &\leq N \cdot \left(1 - \frac{\kappa_P \log n}{n}\right)^n \leq N e^{-\kappa_P \log n} \leq n^{-5}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_2^c) &= \mathbb{P}\big(\max_{1\leq j\leq N} \big|\sum_{i=1}^n \mathbf{1}_{I_j}(X_i)\big| > c\log n\big)\\ &\leq \sum_{j=1}^N \mathbb{P}\big(\big|\sum_{i=1}^n \big(\mathbf{1}_{I_j}(X_i) - \gamma_n\big)\big| > (c-6)\log n\big).\end{aligned}$$

Now we use Bernstein inequality in the following form (cf. (2.10) of [22]): for $S = \sum_{i=1}^n (Z_i - \mathbb{E}Z_i)$, $v = \sum_{i=1}^n \mathbb{E}Z_i^2$ where $|Z_i| \leq b$ for all $1 \leq i \leq n$, we have $\mathbb{P}(S > t) \leq \exp\left(-\frac{t^2}{2(v+bt/3)}\right)$. We apply this with $Z_i \equiv \mathbf{1}_{I_j}(X_i)$ and hence $\gamma_n = \mathbb{E}Z_i$ and $v = \sum_{i=1}^n \gamma_n = 6\log n$, $b = 1$, to see that right side of the above display can be further bounded by

$$\sum_{j=1}^N \exp\left(-\frac{(c-6)^2 \log^2 n}{2(6\log n + (c-6)\log n/3)}\right) \leq N e^{-3\log n} \leq n^{-2}$$

by choosing $c = 14$. Combining the two cases completes the proof. □

We also need Dudley's entropy integral bound for sub-Gaussian processes, recorded below for the convenience of the reader.

Lemma 2.8.2 (Theorem 2.3.7 of [62]). *Let (T, d) be a pseudo metric space, and $(X_t)_{t\in T}$ be a sub-Gaussian process such that $X_{t_0} = 0$ for some $t_0 \in T$. Then*

$$\mathbb{E}\sup_{t\in T}|X_t| \le C\int_0^{\mathrm{diam}(T)} \sqrt{\log \mathcal{N}(\varepsilon, T, d)}\ \mathrm{d}\varepsilon.$$

Here C is a universal constant.

Proof of Lemma 2.5.12. By the contraction principle, we only need to handle

$$\mathbb{E}\sup_{Pf^2\le\rho^2\delta_n^2}\left|\sum_{i=1}^n \xi_i f(X_i)\right|, \quad \mathbb{E}\sup_{Pf^2\le\rho^2\delta_n^2}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|.$$

Let $\mathcal{F}$ be the Hölder class constructed in Lemma 2.5.4. We first claim that on an event $\mathcal{E}_n$ with probability at least $1 - 2n^{-2}$, for any $f \in \mathcal{F}$,

$$\mathbb{P}_n f^2 \le \mathfrak{c}_P\left(Pf^2 + \frac{\log n}{n}\right). \tag{2.8.5}$$

By Lemma 2.8.1, we see that on an event $\mathcal{E}_n$ with probability at least $1 - 2n^{-2}$,

$$\frac{1}{n}\sum_{i=1}^n f^2(X_i) = \frac{1}{n}\sum_{j=1}^N \sum_{X_i\in I_j} f^2(X_i) \le \frac{1}{n}\sum_{j=1}^N c_P \log n \cdot \max_{X_i\in I_j} f^2(X_i).$$

Here $N = n/(\kappa_P \log n)$ is the number of intervals $\{I_j\}$. The trick now is to observe that since f is at least 1/2-Hölder, we have $\max_{X_i\in I_j} f(X_i) \le \min_{x\in I_j} f(x) + \sqrt{\gamma_n}$, where $\gamma_n = \kappa_P \log n/n$ is the length for each interval I_j. Hence on the same event as above, the right side of the above display can be further bounded by

$$\begin{aligned}\frac{2c_P \log n}{n}\sum_{j=1}^N \big(\min_{x\in I_j} f^2(x) + \gamma_n\big) &= \frac{2c_P}{\kappa_P}\sum_{j=1}^N \gamma_n \min_{x\in I_j} f^2(x) + \frac{2c_P \log n}{n} \\ &\le \frac{2c_P}{\kappa_P}\int_0^1 f^2(x)\ \mathrm{d}x + \frac{2c_P \log n}{n},\end{aligned}$$

where the inequality follows from the definition of Riemann integral. The claim (2.8.5) is thus proven by noting that the intergal in the above display is equivalent to Pf^2 up to a

constant depending on P only. Now using Dudley's entropy integral (see Lemma 4.5.11) and (2.8.5), we have for the choice $\sigma_n = \rho(n^{-\frac{1}{2+\alpha}}) \geq \sqrt{\log n/n}$ [the inequality holds when $n \geq \min\{n \geq 3 : \rho^2 \geq \log n(n^{-\alpha/(2+\alpha)})\}$],

$$\begin{aligned}
&\mathbb{E} \sup_{f\in\mathcal{F}: Pf^2 \leq \sigma_n^2} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \\
&\leq C\mathbb{E} \int_0^{2\sqrt{\sup_{f\in\mathcal{F}} \mathbb{P}_n f^2}} \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{F}, L_2(\mathbb{P}_n))}\, \mathrm{d}\varepsilon \\
&\lesssim \int_0^{2\sqrt{\mathfrak{c}_P(\sigma_n^2 + \log n/n)}} \varepsilon^{-\alpha/2}\, \mathrm{d}\varepsilon + J(\infty, \mathcal{F}, L_2)\mathbb{P}(\mathcal{E}_n^c) \lesssim_{P,\alpha} \left(\sigma_n^{1-\alpha/2} + n^{-2}\right).
\end{aligned}$$

Since $\sqrt{n}\sigma_n^{1-\alpha/2} = \rho^{1-\alpha/2} n^{\frac{1}{2} - \frac{1}{2}\frac{2-\alpha}{2+\alpha}} = \rho^{1-\alpha/2} n^{\frac{\alpha}{2+\alpha}}$ and $\sqrt{n} \cdot n^{-2} \leq n^{-1} \leq \frac{\rho^2}{n^{\frac{2}{2+\alpha}} (\log n)^2} \leq \rho^2 \leq \rho^{1-\alpha/2} n^{\frac{\alpha}{2+\alpha}}$, in this case Corollary 2.2.2 along with the assumption $p \geq 1 + 2/\alpha$ yields that

$$\mathbb{E} \sup_{f\in\mathcal{F}: Pf^2 \leq \sigma_n^2} \left| \sum_{i=1}^n \xi_i f(X_i) \right| \lesssim_{P,\alpha} \rho^{1-\alpha/2} n^{\frac{\alpha}{2+\alpha}} \|\xi_1\|_{1+2/\alpha,1} \leq \rho^{1-\alpha/2} n^{\frac{\alpha}{2+\alpha}} \|\xi_1\|_{p,1}.$$

The proof is complete. □

2.8.5 Proof of Lemma 2.5.13

Proof of Lemma 2.5.13. By Lemmas 2.5.3 and 2.5.6, and the α-fullness of $\tilde{\mathcal{F}}$, we have

$$\mathbb{E} \sup_{Pf^2 \leq \vartheta^2 \delta_n^2} \left| \sum_{i=1}^n \xi_i f(X_i) \right| \geq \frac{1}{2} \|\xi_1\|_1 \mathbb{E} \sup_{Pf^2 \leq \vartheta^2 \delta_n^2} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \geq C_1 \|\xi_1\|_1 \vartheta^{1-\alpha/2} n^{\frac{\alpha}{2+\alpha}}.$$

On the other hand, during the proof of Lemma 2.5.8 (see (2.8.3)) we see that $\mathbb{E} \sup_{Pf^2 \leq \vartheta^2 \delta_n^2} |\sum_{i=1}^n \varepsilon_i f^2(X_i)| \leq C_2(\vartheta^{1-\alpha/2} \vee 1) \cdot n^{\frac{\alpha}{2+\alpha}}$. By de-symmetrization,

$$\mathbb{E} \sup_{Pf^2 \leq \vartheta^2 \delta_n^2} \left| \sum_{i=1}^n (f^2(X_i) - Pf^2) \right| \leq 2C_2(\vartheta^{1-\alpha/2} \vee 1) \cdot n^{\frac{\alpha}{2+\alpha}}.$$

Here C_1, C_2 are constants depending on α, P only. Now for $\|\xi_1\|_1 \geq 2C_2/C_1$, since $\vartheta \geq 1$, by the triangle inequality we see that

$$\mathbb{E} \sup_{Pf^2 \leq \vartheta^2 \delta_n^2} \left| \sum_{i=1}^n \left(2\xi_i f(X_i) - f^2(X_i) + Pf^2\right) \right| \geq C_1 \|\xi_1\|_1 \vartheta^{1-\alpha/2} n^{\frac{\alpha}{2+\alpha}},$$

as desired. □

Chapter 3

LEAST SQUARES ESTIMATION WITH HEAVY-TAILED ERRORS II: AN ENVELOPE PERSPECTIVE

3.1 Introduction

3.1.1 Overview

Suppose we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the regression model

$$Y_i = f_0(X_i) + \xi_i, \quad 1 \leq i \leq n. \tag{3.1.1}$$

where the X_i's are independent and identically distributed $\mathcal{X}$-valued covariates with law P, and the ξ_i's are mean-zero errors independent of X_i's. The goal is to recover the true signal f_0 based on the observed data $\{(X_i, Y_i)\}_{i=1}^n$.

In the canonical setting where the errors ξ_i's are Gaussian, perhaps the simplest estimation procedure for the regression model (3.1.1) is the *least squares estimator* (LSE) $\hat{f}_n$ defined by

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2, \tag{3.1.2}$$

where $\mathcal{F}$ is a model chosen by the user. The use of the LSE in the Gaussian regression model has been theoretically justified in the 1990s and the early 2000s, cf. [16, 17, 21, 84, 85, 108, 153, 160, 162]:

Theorem 3.1.1. *Suppose that:*

(E) the errors $\{\xi_i\}$ are sub-Gaussian (or at least sub-exponential);

(F) the model $\mathcal{F}$ satisfies an entropy condition with exponent $\alpha \in (0,2)$[1].

Then

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2+\alpha}}\big). \tag{3.1.3}$$

Furthermore, the rate (3.1.3) is unimprovable under the entropy conditions (F) in a minimax sense, see e.g. [169].

Although the condition (F) is widely accepted in the literature as a complexity measurement of the model $\mathcal{F}$, it is far from clear if the light-tailed condition on the errors (E) is necessary for the theory. Recently, we showed [76] that the condition (E) is actually more than a mere technicality:

Theorem 3.1.2. *Suppose that condition (E) in Theorem 3.1.1 is replaced by*

(E') the errors $\{\xi_i\}$ have a finite $L_{p,1}$ moment ($p \geq 1$)

and (F) holds. Then

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big). \tag{3.1.4}$$

We also showed [76] that the rate (3.1.4) cannot be improved under (F) alone. Comparing with (3.1.3), the rate in (3.1.4) clearly indicates that if the model $\mathcal{F}$ only satisfies (F), the best possible moment condition on the errors to guarantee the same rate of convergence of the LSE as in the case of Gaussian errors is $p \geq 1 + 2/\alpha$.

The starting point for this paper originates from a remarkable result due to Cun-Hui Zhang [172] in the context of isotonic regression. Zhang [172] showed that the L_2 loss of the isotonic LSE achieves the usual worst-case (minimax) $\mathcal{O}_{\mathbf{P}}(n^{-1/3})$ rate, and the adaptive rate $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$ if the true signal is, say, f_0 equals a constant, under only a second moment assumption on the errors.

---

[1] $\mathcal{F}$ satisfies an entropy condition with exponent $\alpha \in (0,2)$ if either (i) $\sup_Q \log \mathcal{N}(\varepsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)) \lesssim \varepsilon^{-\alpha}$, where the supremum is over all finitely discrete measures Q on $(\mathcal{X}, \mathcal{A})$; or (ii) $\log \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_2(P)) \lesssim \varepsilon^{-\alpha}$.

We view the first of these two properties established by Zhang as a "robustness property" of the LSE with respect to the distribution of the errors $\{\xi_i\}$. We formalize this with the following definition:

Definition 3.1.3. We will say that the estimator sequence $\{\hat{f}_n\}$ is *L_2-robust with respect to the errors $\{\xi_i\}$ in the model $\mathcal{F}$* (or just L_2-robust), if $\hat{f}_n$ converges to f_0 in $L_2(P)$ at the same rate for zero mean 0 errors with $\|\xi_i\|_2 < \infty$ as for errors $\{\xi_i\}$ that are Gaussian or sub-Gaussian. Similarly, if the same rate holds for zero mean errors with $\|\xi_i\|_{2,1} < \infty$, we say that $\{\hat{f}_n\}$ is *$L_{2,1}$-robust with respect to the errors $\{\xi_i\}$ in the model $\mathcal{F}$.*

Similarly, we view the second of the two properties established by Zhang as an "adaptation property" of the LSE with respect to the model $\mathcal{F}$:

Definition 3.1.4. We will say that the estimator sequence $\{\hat{f}_n\}$ is *adaptive to a subset $\mathcal{G}_m$ of the model $\mathcal{F}$* if it achieves a nearly (up to factors of $\log n$) parametric rate of convergence at all points $f \in \mathcal{G}_m$.

For the shape-constrained models we consider here the subsets $\mathcal{G}_m$ of $\mathcal{F}$ are natural subclasses of extreme points of the class $\mathcal{F}$: in the isotonic model $\mathcal{F}$ the collections $\mathcal{G}_m$ consisting of m constant non-decreasing pieces, and in the convex regression model $\mathcal{G}_m$ can be taken to be the piecewise linear (convex) functions with at most m linear pieces.

Zhang's work [172] has generated intensive research interest in further understanding the adaptation properties of the isotonic and other shape-restricted LSEs in recent years, cf. [18, 33, 34, 35, 72, 73, 75]. These papers share a common theme: the shape-restricted LSEs are adaptive to certain subsets $\{\mathcal{G}_m\}$ of the model $\mathcal{F}$ under a sub-gaussian assumption on the distribution of the errors in the regression model.

Despite substantial progress in the adaptation properties of various shape-restricted LSEs, there remains little progress in further understanding their L_2-robustness properties beyond the isotonic model studied by Zhang [172]. Indeed, the challenges involved here were noted in Guntuboyina and Sen [73] (page 30) as follows: "*.....However the existing proof techniques*

for these risk bounds strongly rely on the assumption of sub-Gaussianity. It will be very interesting to prove risk bounds in these problems without Gaussianity. We believe that new techniques will need to be developed for this". One of the goals of this paper is to provide new approaches and insights concerning the L_2(or $L_{2,1}$)-robustness of various shape-restricted LSEs.

Initially we had hoped to study this problem by appealing to the general Theorem 3.1.2. However, the theory in Theorem 3.1.2 requires at least a third moment (note that here $\alpha = 1$ for the isotonic model). This implies that the isotonic shape constraint must contain more information than that provided by the entropic structure alone, so that Theorem 3.1.2 fails to fully capture the L_2-robustness of the isotonic LSE.

One particular useful feature of the isotonic model is an explicit min-max formula for the isotonic LSE in terms of partial sum processes; see e.g. [129]. Zhang's techniques [172] make full use of the min-max representation, and are therefore substantially of an analytic flavor. Similar techniques have also been used in [33, 56], but have apparently not yet successful in dealing with any other shape constrained models. The rigidity in this analytic approach naturally motivates the search for other 'softer' properties of the isotonic shape constrained model that explain the robustness of the LSE. These considerations lead to the following question.

Question 3.1.5. *What geometric aspects of the isotonic shape constrained model give rise to the L_2(or $L_{2,1}$)-robustness property of the LSE?*

To put this question into a more general setting, note that Theorem 3.1.2 implies that the LSE can converge as slowly as $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ for certain hard models when the errors only have a second moment, while in the aforementioned isotonic regression case, it is possible that the LSE converges at a nearly parametric rate $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$ for certain special isotonic functions. Therefore it seems more promising to search for a characterization of the convergence rate of the L_2 loss of the LSE in terms of some geometric feature of the model $\mathcal{F}$, when the errors have only an L_2(or $L_{2,1}$) moment.

The first main contribution of this paper is to shed light on Question 3.1.5 from an 'envelope' perspective at this general level. Roughly speaking, the size of the 'localized envelopes' of the model $\mathcal{F}$ determines the convergence rate of the L_2 loss of the LSE when the errors only have an $L_{2,1}$ moment. More specifically, let $F_0(\delta)$ be the envelope for $\mathcal{F}_0(\delta) \equiv \{f \in \mathcal{F}_0 : Pf^2 \leq \delta^2\}$ where $\mathcal{F}_0 \equiv \mathcal{F} - f_0$. We show that (cf. Theorem 3.2.1), under a certain uniform entropy condition on the function class, if for some $0 \leq \gamma \leq 1$, the localized envelopes have the growth rate

$$\|F_0(\delta)\|_{L_2(P)} \sim \delta^\gamma : \tag{3.1.5}$$

then the convergence rate of the LSE in the L_2 loss is no worse than

$$\mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2(2-\gamma)}}\big). \tag{3.1.6}$$

Furthermore, the rate (3.1.6) cannot be improved under the condition (3.1.5), cf. Theorem 3.2.5. It is easily seen from (3.1.6) that, as the size of the localized envelopes increases, the rate of the L_2 loss of the LSE deteriorates from the parametric rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ to the worst-case rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ as suggested by Theorem 3.1.2. For isotonic regression, we will see that the localized envelopes of the model are small in the sense that $\gamma \approx 1$ (up to logarithmic factors) when $f_0 = 0$, and hence the LSE converges at a nearly parametric rate under an $L_{2,1}$ moment assumption on the errors. For the hard models identified in [76] (cf. Example 3.2.9 below), the localized envelopes are big in the sense that $\gamma = 0$ so the LSE can only converge at the worst-case rate.

Addressing Question 3.1.5 from a geometric point of view is not only of interest in its own right, but also serves as an important step in better understanding the robustness properties of other shape constrained models. This is the context of the second main contribution of this paper: we aim at improving our understanding of the $L_{2,1}$-robustness property of shape-restricted LSEs, by providing a systematic approach to proving oracle inequalities in the random design regression setting for these LSEs under an $L_{2,1}$ moment condition on the errors. This goal is achieved by exploiting the idea of small envelopes from the solution to

Question 3.1.5. The formulation of the oracle inequality follows its fixed-design counterparts that highlight the automatic rate-adaptive behavior of the LSE, cf. [18, 33]. More specifically, we first prove the following oracle inequality that holds for the canonical isotonic and convex LSEs in the simple regression models (cf. Theorem 3.3.1): Suppose that $\|f_0\|_\infty < \infty$ and the errors $\{\xi_i\}$ are i.i.d. mean-zero with $\|\xi_1\|_{2,1} < \infty$. Then for any $\delta \in (0,1)$, there exists some constant $c > 0$ such that with probability $1 - \delta$,

$$\|\hat{f}_n - f_0^*\|_{L_2(P)}^2 \leq c \inf_{m \in \mathbb{N}} \left(\inf_{f_m \in \mathcal{G}_m} \|f_m - f_0^*\|_{L_2(P)}^2 + \frac{m}{n} \cdot \log^2 n \right), \tag{3.1.7}$$

where f_0^* is the $L_2(P)$-projection of f_0 onto the space of square integrable monotonic non-decreasing (resp. convex) functions, and $\mathcal{G}_m$ is the class of piecewise constant non-decreasing (resp. linear convex) functions on $[0,1]$ with at most m pieces in the isotonic (resp. convex) model. The oracle inequality (3.1.7) is further verified for the shape-restricted LSEs in the additive model (cf. Theorem 3.3.5), where now f_0 is the marginal L_2 projection of the true regression function. One striking message of the oracle inequality for the shape-restricted LSEs in the additive model is the following: both the adaptation and $L_{2,1}$-robustness properties of the LSE can be preserved, for estimating the shape-constrained proxy of the marginal L_2 projection of the true regression function, *essentially regardless of whether or not the additive structure is correctly specified*.

The proofs in this paper rely heavily on the new empirical process tools and proof techniques developed in [76]. Although we will list relevant results, readers are referred to [76] for more discussion of the new tools. Along the way we also resolve the stochastic boundedness issue of convexity shape-restricted LSEs at the boundary, which may be of independent interest (this problem is in fact an open problem in the field, cf. [73]).

3.2 Convergence rate of the LSE: the envelope characterization

3.2.1 Upper and lower bounds

Our first main result is the following.

Theorem 3.2.1. *Suppose that $\xi_1, \dots, \xi_n$ are i.i.d. mean-zero errors independent of i.i.d. covariates $X_1, \dots, X_n$ with law P such that $\|\xi_1\|_{2,1} < \infty$. Further suppose that $\mathcal{F}_0 \equiv \mathcal{F} - f_0$ is a VC-subgraph class, and the envelopes $F_0(\delta)$ of $\mathcal{F}_0(\delta) \equiv \{f \in \mathcal{F}_0 : Pf^2 \leq \delta^2\}$ satisfy the growth condition*

$$\|F_0(\delta)\|_{L_2(P)} \leq c \cdot \delta^\gamma, \quad \text{for all } \delta > 0 \tag{3.2.1}$$

for some constants $0 \leq \gamma \leq 1$ and $c > 0$. If $\|\hat{f}_n - f_0\|_\infty = \mathcal{O}_\mathbf{P}(1)$, then

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_\mathbf{P}\big(n^{-\frac{1}{2(2-\gamma)}}\big).$$

Remark 3.2.2. Some technical remarks are in order.

1. Proving stochastic boundedness $\|\hat{f}_n - f_0\|_\infty = \mathcal{O}_\mathbf{P}(1)$ is often the first step in deriving convergence rate of least squares type estimators, cf. [160]; see also [120] (e.g. Theorem 6.1) for related techniques in the context of model selection. If instead of $\|\hat{f}_n - f_0\|_\infty = \mathcal{O}_\mathbf{P}(1)$ it is assumed that $\mathcal{F}_0 \subset L_\infty(1)$, then the conclusion of Theorem 3.2.1 can be strengthened to an expectation: $\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}\big(n^{-\frac{1}{2(2-\gamma)}}\big)$.

2. Condition (3.2.1) on the size of the localized envelopes can be modified to incorporate logarithmic factors. In particular, if

   $$\|F_0(\delta)\|_{L_2(P)} \leq c \cdot \delta^\gamma \log^\tau(1/\delta),$$

   then we may slightly modify the proof of Theorem 3.2.1 to see that the convergence rate of the L_2 loss of the LSE is given by

   $$\mathcal{O}_\mathbf{P}\big(n^{-\frac{1}{2(2-\gamma)}} \log^{\frac{\tau}{2-\gamma}} n\big).$$

3. We assume that the errors are identically distributed for simplicity: the case of mean-zero, independent but not necessarily identically distributed errors follows from a minor modification of the proof.

Remark 3.2.3. Theorem 3.2.1 is actually proved for $\mathcal{F}_0$ under a more general *uniform VC-type* condition: $\mathcal{F}_0$ is said to be of uniform VC-type if there exists some $\alpha \in [0, 2)$ and $\beta \in [0, \infty)$ [2] such that for any probability measure Q, and any $\varepsilon \in (0, 1), \delta > 0$,

$$\log \mathcal{N}\left(\varepsilon \|F_0(\delta)\|_{L_2(Q)}, \mathcal{F}_0(\delta), L_2(Q)\right) \lesssim \varepsilon^{-\alpha} \log^{\beta}(1/\varepsilon). \tag{3.2.2}$$

The most significant examples for uniform VC-type classes are the VC-subgraph classes ($\alpha = 0, \beta = 1$). Other important examples include the VC-major classes, which satisfy (3.2.2) up to a logarithmic factor (cf. Lemma 3.4.14). As we will see in Section 3.3, the canonical examples of VC-major classes that satisfy (3.2.2) considered in this paper are the classes of bounded monotonic non-decreasing and convex functions on $[0, 1]$.

Remark 3.2.4. From a purely probabilistic point of view, the condition (3.2.1) is related to Alexander's capacity function [3, 4, 5] defined for VC class of sets that gives relatively sharp asymptotic local moduli of weighted empirical processes indexed by such classes. Results in a similar vein can be found in [59] who generalized this notion to bounded VC-subgraph function classes.

So far we have derived an upper bound for the convergence rate of the L_2 loss of the LSE under the condition (3.2.1). It is natural to wonder if such an upper bound is sharp in an appropriate sense.

Theorem 3.2.5. *Let P be the uniform distribution on $[0, 1]$. For any $\gamma \in (0, 1]$, there exists some uniformly bounded VC-subgraph class $\tilde{\mathcal{F}}$ on $[0, 1]$ and some $f_0 \in \tilde{\mathcal{F}}$ such that $\tilde{\mathcal{F}}_0 \equiv \tilde{\mathcal{F}} - f_0$ satisfies (3.2.1), and the following property holds: for each $\varepsilon \in (0, 1/2)$, there exist some constants $c_{\varepsilon,\gamma} > 0$, $\mathfrak{p} > 0$ and some law for ξ_1 with $\|\xi_1\|_{2(1-\varepsilon)} < \infty$ such that, for n large enough depending on ε, γ, there exists a LSE f_n^* whose L_2 loss satisfies*

$$\|f_n^* - f_0\|_{L_2(P)} \geq c_{\varepsilon,\gamma} \cdot n^{-\frac{1}{2(2-\gamma)} - c'_\gamma \varepsilon}$$

with probability at least $\mathfrak{p} > 0$. The constant c'_γ can be taken to be $2/\gamma$.

---

[2] We can also allow $\alpha = 2, \beta < -2$ but we are not aware of any such examples.

Theorem 3.2.5 shows that our upper bound Theorem 3.2.1 cannot be improved substantially under (3.2.1): the size of the localized envelopes drives the convergence rate of the L_2 loss of the LSE over VC-subgraph models (or more generally, models of uniform VC-type) in the heavy-tailed regression setting where the errors only admit (roughly) a second moment. Since the median regression estimator over VC-subgraph models achieves a nearly parametric rate $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$ at least when the errors are symmetric and admit smooth densities; cf. Section 3.4.4 of [162], Theorem 3.2.5 illustrates a genuine deficiency of the LSE in VC-subgraph models when the envelopes of the model are not small. We remark that the case $\gamma = 0$ is excluded mainly for simplicity of presentation; similar conclusions hold under a slightly weaker formulation, cf. Theorem 5 of [76].

The proofs of Theorems 3.2.1 and 3.2.5 are based on recent developments on the *equivalence* between the convergence rate of the L_2 loss of the LSE and the size of the multiplier empirical process, cf. [36, 76, 158]. For the upper bound, our proofs rely heavily on a new multiplier inequality developed in [76]. The lower bound, on the other hand, is based on an explicit construction of $\tilde{\mathcal{F}}$ that witnesses the desired rate within uniformly bounded VC-subgraph classes satisfying (3.2.1).

3.2.2 Examples

In this section, we use Theorem 3.2.1 to examine the convergence rate of the L_2 loss of the LSE in several important examples.

Example 3.2.6 (Linear model)**.** Let $\mathcal{F} \equiv \{f_\beta(x) \equiv \beta^\top x : \beta \in \mathbb{R}^d\}$ and let P be the uniform distribution on $[0,1]^d$. This is the simplest linear regression model. A second moment assumption on the errors ξ_i's yields a closed-form LSE with a parametric convergence rate: $\|\hat{f}_n - f_0\|_{L_2(P)} \asymp \|\hat{\beta}_n - \beta_0\|_2 = \mathcal{O}_{\mathbf{P}}(n^{-1/2})$. This rate is obviously much faster than the worst-case rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ as suggested by Theorem 3.1.2. Thus the LSE sequence $\{\hat{f}_n\}$ is L_2-robust for the model $\mathcal{F}$ by a direct argument while our Theorem 1 very nearly recovers this: it shows that $\{\hat{f}_n\}$ is $L_{2,1}$-robust for the model $\mathcal{F}$.

For simplicity of discussion, we assume $d = 1$ in the sequel. We may also restrict the model to be $\{f_\beta : \beta \in [-1, 1]\}$; this is viable since the LSE *localizes* in the sense that $\|\hat{f}_n\|_\infty = |\hat{\beta}_n| = \mathcal{O}_{\mathbf{P}}(1)$. Moreover, it is clear that the model is a VC-subgraph class. For any $\delta > 0$, $\|f_\beta\|_{L_2(P)} \leq \delta$ implies that $|\beta| \leq \sqrt{3}\delta$, and thus

$$F(\delta)(x) = \sup_{\beta \in [-\sqrt{3}\delta, \sqrt{3}\delta]} |\beta x| = \sqrt{3}\delta|x|,$$

which in turn yields $\|F(\delta)\|_{L_2(P)} = \delta$. Hence Theorem 3.2.1 applies with $\gamma = 1$ to recover the usual parametric rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ for the L_2 loss of the LSE.

Our approach here should be compared with the common practice of using local entropy to recovery the exact parametric rate for parametric models—but the latter does not extend directly to the heavy-tailed regression setting, cf. pages 152-153 of [160].

Example 3.2.7 (Isotonic model)**.** Let $\mathcal{F}$ be the class of monotonic non-decreasing functions on $[0, 1]$ and let P be the uniform distribution on $[0, 1]$. It is shown in a related fixed design setting (cf. [33, 56, 172]) that a second moment condition on the errors ξ_i is sufficient for the isotonic LSE to achieve the nearly parametric adaptive rate $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$ in the discrete ℓ_2 loss, when the true signal is $f_0 = 0$. This naturally suggests a similar rate for the L_2 loss of the isotonic LSE in the random design setting. Apparently, this (suggested) nearly parametric rate is far from the worst-case rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$.

In this model, since the univariate isotonic LSE localizes in L_∞ norm (cf. Lemma 3.4.10), we may assume without loss of generality that $\mathcal{F} \equiv \{f : \text{non-decreasing}, \|f\|_\infty \leq 1\}$. The entropy condition (3.2.2) can be verified using the VC-major property of $\mathcal{F}$ up to a logarithmic factor (cf. Lemma 3.4.14). On the other hand, for any $\delta > 0$, by monotonicity and the L_2 constraint, we can take

$$F(\delta)(x) \equiv \delta \cdot \left(x^{-1/2} \vee (1 - x)^{-1/2}\right) \wedge 1.$$

Evaluating the integral we see that $\|F(\delta)\|_{L_2(P)} \lesssim \delta\sqrt{\log(1/\delta)}$. Then an application of Theorem 3.2.1 along with Remarks 3.2.2 (2) and 3.2.3, we see that the L_2 loss of the LSE $\hat{f}_n$ converges at a parametric rate up to logarithmic factors when the truth f_0 is a constant

function and the errors are $L_{2,1}$. The observation concerning the role of the localized envelopes in the isotonic model here is the starting point for a systematic development of oracle inequalities for shape-restricted LSEs in Section 3.3.

Example 3.2.8 (Single change-point model)**.** Let $\mathcal{F} \equiv \{\mathbf{1}_{[a,1]} : a \in [0,1]\}$ be the model containing signals on $[0,1]$ with a single change point. Let P be the uniform distribution on $[0,1]$.

This model is contained in the isotonic model—from here we already know by Example 3.2.7 that the localized envelopes of $\mathcal{F}$ are small, and hence the LSE converges at a rate no worse than a nearly parametric rate under an $L_{2,1}$ moment assumption on the errors. We can do better: since the localized envelopes are exactly given by $F(\delta) = \mathbf{1}_{[1-\delta^2,1]}$, it follows that $\|F(\delta)\|_{L_2(P)} = \delta$, and hence by Theorem 3.2.1 with $\gamma = 1$ we see that the LSE converges exactly at the parametric rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ even if the errors only admit an $L_{2,1}$ moment. This is in stark contrast with the *multiple change-points model* detailed below.

Example 3.2.9 (Multiple change-points model)**.** Consider the following multiple change-points model:

$$\mathcal{F}_k \equiv \Bigg\{ \sum_{i=1}^{k} c_i \mathbf{1}_{[x_{i-1},x_i]} : |c_i| \leq 1, \\ 0 \leq x_0 < x_1 < \ldots < x_{k-1} < x_k \leq 1 \Bigg\}, k \geq 1.$$

It is shown in [76] that the L_2 loss of the LSE over (a subset of) $\mathcal{F}_k$ cannot converge at a rate faster than $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ for some errors ξ_i with only (roughly) a second moment. The LSE fails to be rate-optimal in this model: if the errors are Gaussian (or even bounded), the convergence rate of the L_2 loss of the LSE (over VC-subgraph classes) is no worse than $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$.

Note that in this model, the localized envelopes are given by $F(\delta) \equiv 1$ for any $\delta > 0$ and hence $\|F(\delta)\|_{L_2(P)} = 1$. Applying Theorem 3.2.1 with $\gamma = 0$ recovers the correct rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ for the L_2 loss of the LSE in this model.

Example 3.2.10 (Unimodal model)**.** Let $\mathcal{F}$ contain all (bounded) unimodal functions on $[0,1]$, i.e. all $f:[0,1]\to\mathbb{R}$ such that there exists some $x^*\in[0,1]$ with $f|_{[0,x^*]}$ non-decreasing and $f|_{[x^*,1]}$ non-increasing. [35] and [18] considered the performance of the LSE in a fixed-design unimodal Gaussian regression setting, where similar adaptive behavior as in the isotonic case (cf. [172]) is derived. Since the class of (bounded) unimodal functions on $[0,1]$ contains the class of multiple change-points model $\mathcal{F}_1$ as studied in Example 3.2.9, our results here imply that *the unimodal shape constraint does not inherit* the L_2 (or $L_{2,1}$)-robustness property as in the isotonic shape constraint in Example 3.2.7: the worst-case $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ is attained by the LSE in the unimodal regression model for some errors ξ_i's with (roughly) a second moment.

3.3 Shape-restricted regression problems

As briefly mentioned in the Introduction, it is well-known that in the fixed design regression setting, the isotonic least squares estimator (LSE) only requires a second moment condition on the errors to enjoy an oracle inequality, cf. [33, 56, 172]. The proof techniques used therein rely crucially on (i) some form of representation of the isotonic LSE in terms of partial sum processes, and (ii) martingale inequalities. Unfortunately, such an explicit representation does not exist beyond the isotonic LSE, and hence these techniques do not readily extend to other problems.

Our goal here is to give a systematic treatment of the robustness properties of shape-restricted LSEs in a random design setting, up to error distributions with an $L_{2,1}$ moment. The examples we examine are (i) the canonical isotonic and convex regression models, and (ii) additive regression models with monotonicity and convexity shape constraints. As we will see, the 'smallness' of the localized envelopes, along with their special geometric properties, play a central role in our approach.

Henceforth, the isotonic (resp. convex) model refers to the regression model based on the class of monotonic non-decreasing (resp. convex) functions on $[0,1]$.

3.3.1 Prologue: the canonical problems

We start by considering the 'canonical' problems in the area of shape-restricted regression: the isotonic and convex regression problems. Note that a generic LSE $\hat{f}_n$ in (3.1.2) is only well-defined on the design points $X_1, \ldots, X_n$. Our results below hold for the *canonical LSEs*: for the isotonic (respectively convex) model, $\hat{f}_n$ is defined to be the unique left-continuous piecewise constant (resp. linear) function on $[0,1]$ with jumps (respectively kinks) at (potentially a subset of) $\{\hat{f}_n(X_i)\}_{i=1}^n$.

Some further notation: let $\mathcal{M}_m \equiv \mathcal{M}_m([0,1])$ (respectively $\mathcal{C}_m \equiv \mathcal{C}_m([0,1])$) be the class of all non-decreasing piecewise constant functions (respectively convex piecewise linear functions) on $[0,1]$ with at most m pieces. Let P denote the uniform distribution on $[0,1]$ for simplicity of exposition.

Theorem 3.3.1. *Consider the regression model (3.1.1). Let $\mathcal{F}$ be either the isotonic or convex model. Suppose that $\|f_0\|_\infty < \infty$, and the errors are i.i.d. mean-zero with $\|\xi_1\|_{2,1} < \infty$. Then for any $\delta \in (0,1)$, there exists $c \equiv c(\delta, \|\xi\|_{2,1}, \|f_0\|_\infty, \mathcal{F}) > 0$ such that with probability $1-\delta$, the canonical LSE $\hat{f}_n$ defined above satisfies*

$$\|\hat{f}_n - f_0^*\|_{L_2(P)}^2 \leq c \inf_{m \in \mathbb{N}} \left(\inf_{f_m \in \mathcal{G}_m} \|f_m - f_0^*\|_{L_2(P)}^2 + \frac{m}{n} \cdot \log^2 n \right),$$

where $f_0^* = \mathrm{argmin}_{g \in \mathcal{F} \cap L_2(P)} \|f_0 - g\|_{L_2(P)}$, and $\mathcal{G}_m = \mathcal{M}_m$ for the isotonic model and $\mathcal{G}_m = \mathcal{C}_m$ for the convex model.

The isotonic regression problem, included here mainly for sake of later development in the additive model, is a benchmark example in the family of shape-restricted regression problems. Even in this simplest case, the above oracle inequality in $L_2(P)$ loss seems new[3].

For the more interesting convex regression problem, our oracle inequality here confirms for the first time both the adaptation and robustness properties of the convex LSE up to

[3]An oracle inequality in $L_2(\mathbb{P}_n)$ loss follows immediately from [33] (with a second moment assumption on the errors) since the monotone cone does *not* change with the design points. See [75] for different techniques in the multivariate isotonic regression problem when the errors are Gaussian.

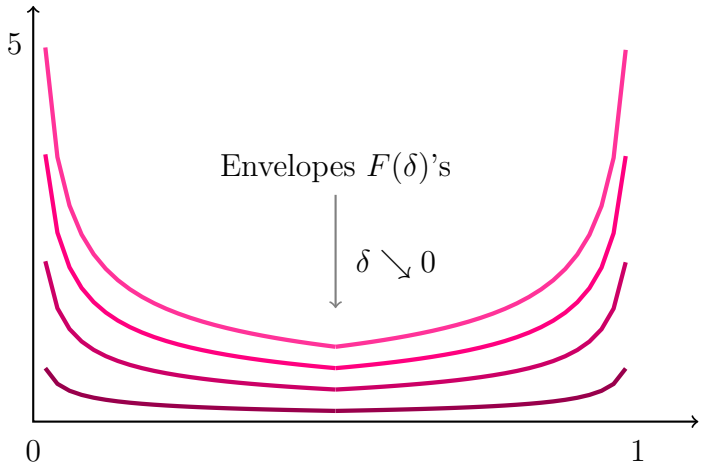

Figure 3.1: Envelopes for isotonic model with $c = 1$ in (3.3.2). From top to bottom: $\delta = 0.7, 0.5, 0.3, 0.1$.

error distributions with an $L_{2,1}$ moment. Previous oracle inequalities for the convex LSE exclusively focused on the fixed-design setting under a (sub-)Gaussian assumption on the errors [18, 33]; see also Section 3 of [73] for a review.

Remark 3.3.2. Two technical comments on the formulation of the oracle inequality in Theorem 3.3.1:

1. The oracle inequality holds for the projection f_0^* of f_0 to $\mathcal{F} \cap L_2(P)$ and hence allows for model mis-specification: the only assumption on f_0 is boundedness: $\|f_0\|_\infty < \infty$. The same comment also applies to the oracle inequality in the additive model below.

2. The oracle inequality cannot be strengthened to an expectation, in view of a counterexample discovered in [13] in the convex model: the convex LSE $\hat{f}_n$ has infinite L_2 risk in estimating $f_0 = 0$ even if the errors are bounded: $\mathbb{E}\|\hat{f}_n - 0\|^2_{L_2(P)} = \infty$.

The proof of Theorem 3.3.1 contains two major steps.

(Step 1) We first localize the shape-restricted LSEs in L_∞ norm. This step requires some understanding of the boundary behavior of the shape-restricted LSEs under a second moment

assumption on the errors. The case for isotonic regression is relatively straightforward, while the case for convex regression is much more difficult. Here we resolve this issue in Lemma 3.4.10.

(Step 2) After the localization in Step 1, the problem essentially reduces to controlling a multiplier empirical process of the form

$$\mathbb{E} \sup_{f \in \mathcal{F}: f - f_0^* \in L_2(\delta_n) \cap L_\infty(B)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i (f - f_0^*)(X_i) \right|. \tag{3.3.1}$$

A sharp bound for (3.3.1) is inspired by the observation in Example 3.2.7, where the (untruncated) localized envelopes of the isotonic model take the form

$$F(\delta)(x) \equiv c\delta \cdot \left(x^{-1/2} \vee (1-x)^{-1/2} \right) \tag{3.3.2}$$

for some absolute constant $c > 0$. The envelopes for the convex model also take the same form (3.3.2), cf. Lemma 3.4.15. On the other hand, the localized envelopes (3.3.2) are centered at 0, while the multiplier empirical process (3.3.1) in question is centered at f_0^*. By exploiting the exact form of (3.3.2), we perform a 'change-of-center argument' on (3.3.1) by shifting f_0^* to an arbitrary piecewise simple signal $f_m \in \mathcal{G}_m \cap L_\infty(\|f_0^*\|_\infty)$, cf. Lemma 3.4.12, thereby reducing the control of (3.3.1) to control of several multiplier empirical processes centered at 0. The effect of the heavy-tailed ξ_i's is then accounted for, via the multiplier inequality developed in [76], by a uniform estimate for the corresponding empirical processes in terms of the L_2 size of the localized envelopes (3.3.2).

The envelope technique used in Step 2 is related to a recent work [75] in the context of multivariate isotonic regression with Gaussian errors. There the envelope of the class of multivariate block-decreasing functions is useful in obtaining a sharp estimate for the (symmetrized) empirical process indexed by such a class localized at 0.

Remark 3.3.3. Currently our oracle inequality comes with a $\log^2 n$ term. It is known in (i) the fixed design isotonic model with a second moment assumption on the errors, and (ii)

the fixed design convex model with sub-Gaussian errors, that the power of the logarithmic factor can be reduced to 1. The additional logarithmic factor in Theorem 3.3.1 occurs due to the use of VC-major property for the isotonic and convex models in the random design setting: the entropy estimate of bounded VC-major classes comes with logarithmic factors that involve the L_2 size of the envelopes (cf. Lemma 3.4.14).

3.3.2 Additive regression model with shape constraints

Consider fitting $(x, z) \mapsto \phi_0(x, z)$, the conditional mean of the regression model

$$Y_i = \phi_0(X_i, Z_i) + \xi_i, \quad 1 \le i \le n, \tag{3.3.3}$$

by additive models of the form $\{(x, z) \mapsto f(x) + h(z)\}_{f\in\mathcal{F},h\in\mathcal{H}}$, where $\mathcal{F}, \mathcal{H}$ are two function classes on $[0, 1]$. To capture the mathematical essence of the problem, we assume that the covariates $\{(X_i, Z_i)\}_{i=1}^n$ are i.i.d. from the uniform law P on $[0, 1]^2$ and are independent of the errors $\{\xi_i\}$. We use P_X, P_Z to denote the marginal distributions of P. For identifiability we assume that $\mathcal{H}$ is centered.

Additive models of the type have a long history; see e.g. [78, 139]. When the additive model is well specified (i.e. $\phi_0(x, z) = f_0(x) + h_0(z)$ with $f_0 \in \mathcal{F}, h_0 \in \mathcal{H}$), and the non-parametric components enjoy smoothness assumptions, standard methods such as iterative backfitting, e.g. [98] and penalized LSE (smooth spline), e.g. [164], can be used to estimate f_0 and h_0.

Instead of computational issues, we will be interested here in certain structural aspects of the additive LSE $\hat{f}_n$ defined via:

$$(\hat{f}_n, \hat{h}_n) \in \operatorname*{argmin}_{(f,h)\in\mathcal{F}\times\mathcal{H}} \sum_{i=1}^{n} \big(Y_i - f(X_i) - h(Z_i)\big)^2. \tag{3.3.4}$$

Since the true regression function ϕ_0 need not have an additive structure, one may naturally expect that $\hat{f}_n$ and $\hat{h}_n$ estimate the marginal L_2 projections $x \mapsto f_0(x) \equiv P_Z\phi_0(x, Z)$ and $z \mapsto h_0(z) \equiv P_X\phi_0(X, z) - P\phi_0$ (cf. Appendix 4, page 439 of [19]). Our primary *structural*

question on the behavior of the additive LSE $\hat{f}_n$ concerns the situation in which the model $\mathcal{F}$ involves shape constraints:

Question 3.3.4. *Does the additive LSE $\hat{f}_n$ over the shape constrained model $\mathcal{F}$ enjoy similar robustness and adaptation properties as in the univariate case (treated in Theorem 3.3.1)?*

The next theorem gives an affirmative answer to Question 3.3.4.

Theorem 3.3.5. *Suppose that (X_i, Z_i, Y_i), $i = 1, \ldots, n$, are i.i.d. with values in $[0,1] \times [0,1] \times \mathbb{R}$ and satisfy (3.3.3) where $\|\phi_0\|_\infty < \infty$, and the errors $\{\xi_i\}$ are i.i.d. mean zero with $\|\xi_1\|_{2,1} < \infty$. Let $\mathcal{F}$ be either the isotonic or convex model. Further suppose that $\mathcal{H} \subset L_\infty(2\|\phi_0\|_\infty)$ satisfies the following L_∞ covering bound: for some $\gamma \in (0,2)$*

$$\log \mathcal{N}(\varepsilon, \mathcal{H}, L_\infty) \lesssim \varepsilon^{-\gamma}, \text{ for all } \varepsilon \in (0,1). \tag{3.3.5}$$

Then for any $\delta \in (0,1)$, there exists $c \equiv c(\delta, \|\xi\|_{2,1}, \|\phi_0\|_\infty, \mathcal{F}, \mathcal{H}) > 0$ such that with probability $1-\delta$, the canonical LSE $\hat{f}_n$ in (3.3.4) satisfies

$$\|\hat{f}_n - f_0^*\|_{L_2(P)}^2 \le c \inf_{m \in \mathbb{N}} \left(\inf_{f_m \in \mathcal{G}_m} \|f_m - f_0^*\|_{L_2(P)}^2 + \frac{m}{n} \cdot \log^2 n \right),$$

where $f_0^* = \operatorname{argmin}_{g \in \mathcal{F} \cap L_2(P)} \|f_0 - g\|_{L_2(P)}$ with $f_0 = P_Z \phi_0(\cdot, Z)$, and $\mathcal{G}_m = \mathcal{M}_m$ for the isotonic model and $\mathcal{G}_m = \mathcal{C}_m$ for the convex model.

There is very limited theoretical understanding of the properties of shape-restricted estimators when additive models are used. [116] investigated identifiability issue for the additive LSE in the fixed design setting. [101] considered *pointwise* performance of the LSE where both $\mathcal{F}$ and $\mathcal{H}$ are monotonic with errors admitting exponential moments. [38] gives an extension to a semiparametric setting assuming the same moment condition on the errors, still considering pointwise performance of the LSEs for the isotonic components. [37] proved consistency of the MLEs for a generalized class of additive and index models with shape constraints, without rate considerations. A common feature of all these works is that the model is required to be well-specified.

To the best knowledge of the authors, Theorem 3.3.5 is the first oracle inequality for shape-restricted LSEs in regression using an additive model, and moreover, allowing for model mis-specification: not only the regression function class $\mathcal{F}$ can be mis-specified, but the additive model itself may also be mis-specified. Our result here therefore gives a strong positive answer to Question 3.3.4: both the adaptation and robustness properties of additive shape-restricted LSEs can be preserved in estimating the shape constrained proxy of the marginal L_2 projection of the true regression function, up to error distributions with an $L_{2,1}$ moment, *essentially regardless of whether or not the additive structure is correctly specified*.

Examples under correct specification of the additive structure

Now we consider the important situation when ϕ_0 has an additive structure:

$$\phi_0(x, z) \equiv f_0(x) + h_0(z).$$

In such a scenario, our result here is related to the recent work [157], who asserted that the rate optimality nature of the (penalized) LSE over $\mathcal{F}$ in the Gaussian regression setting can be preserved regardless of the smoothness level of $\mathcal{H}$. Our Theorem 3.3.5 reveals a further structural property of the LSEs: the robustness and adaptation merits due to shape constraints can also be preserved, regardless of the choice of $\mathcal{H}$ under the entropy condition (3.3.5).

To further illustrate this point, we consider some examples.

- (*Parametric model*) $\mathcal{H} \equiv \{f_\beta(z) \equiv \beta(z - 1/2) : \beta \in [-1, 1]\}$. In this case (3.3.3) becomes the semiparametric partially linear model.

- (*Smooth model*) $\mathcal{H}$ is the class of centered uniformly bounded α-Hölder ($\alpha > 1/2$) continuous functions on $[0, 1]$ with uniformly bounded derivatives (cf. Theorem 2.7.1 of [162]).

- (*Shape constrained model*) $\mathcal{H}$ is the class of centered uniformly Lipschitz convex functions on $[0, 1]$ (cf. Corollary 2.7.10 of [162]).

Proof strategy of Theorem 3.3.5

The basic strategy in our proof of Theorem 3.3.5 is similar to that of Theorem 3.3.1. First, we need to localize the LSEs in L_∞ norm under a second moment assumption on the errors and $P_Z H^2 < \infty$, cf. Lemma 3.4.20. Next, in addition to the multiplier empirical process (3.3.1), the major additional empirical process we need to control is

$$\mathbb{E} \sup_{\substack{f\in\mathcal{F}: f-f_0^* \in L_2(\delta_n)\cap L_\infty(B) \\ h\in\mathcal{H}}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f-f_0^*)(X_i)(h-(\phi_0-f_0))(X_i,Z_i) \right|. \tag{3.3.6}$$

where the ε_i's are i.i.d. Rademacher random variables. One notable feature in (3.3.6) is that the supremum over $\mathcal{H}$ need *not* be localized when the interest is in the behavior of $\hat{f}_n$, cf. Proposition 3.4.16. In other words, no apriori information on the behavior of $\hat{h}_n$ (other than the assumption (3.3.5)) is needed in order to understand the behavior of $\hat{f}_n$.

The entropy condition (3.3.5) serves as a sufficient condition for a sharp estimate for (3.3.6) (and thereby for the oracle inequality in Theorem 3.3.5), but is apparently not necessary; we make such a choice here to cover the above common examples. A case-by-case study is possible as long as (3.3.6) can be well-controlled. For instance, it is not hard to verify a similar bound for (3.3.6) as in Lemma 3.4.17 (and hence the oracle inequality for shape-restricted LSEs $\hat{f}_n$) when the additive structure is correctly specified, and $\mathcal{H}$ is the class of centered indicator functions over closed intervals on $[0,1]$ and $h_0 = 0$ (note that this class fails to satisfy (3.3.5) since $\mathcal{H}$ is not totally bounded in L_∞). This is a difficult case: although the L_2 loss of the LSE $\hat{h}_n$ is known to converge at a worst-case rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ (cf. Example 3.2.9), Theorem 3.3.5 tells us that the bad behavior of $\hat{h}_n$ has no effect on the good (robust and adaptive) performance of $\hat{f}_n$, at least under reasonable assumption on the distribution of the covariates (X, Z).

3.4 Proofs of the main results

In this section we outline the main steps in proving the main results of the paper, namely:

1. Theorems 3.2.1 and 3.2.5 characterizing the geometric feature of the model that deter-

mines the actual convergence rate of the L_2 loss of the least squares estimator, and

2. Theorems 3.3.1 and 3.3.5 highlighting oracle inequalities in shape-restricted regression models with a $L_{2,1}$ moment assumption on the errors.

Proofs of many technical intermediate results will be deferred to Section 3.5.

3.4.1 Preliminaries

In this subsection we collect the empirical process tools that will be needed in the proofs to follow. Our first ingredient is a sharp multiplier inequality proved in [76].

Lemma 3.4.1 (Theorem 1 in [76]). *Suppose that $\xi_1, \dots, \xi_n$ are i.i.d. mean-zero random variables independent of i.i.d. $X_1, \dots, X_n$. Let $\mathcal{F}_1 \supset \cdots \supset \mathcal{F}_n$ be a non-increasing sequence of function classes. Assume further that there exist non-decreasing concave functions $\{\psi_n\} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\psi_n(0) = 0$ such that*

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_k} \leq \psi_n(k) \tag{3.4.1}$$

holds for all $1 \leq k \leq n$. Then

$$\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} \leq 4 \int_0^\infty \psi_n\big(n \cdot \mathbb{P}(|\xi_1| > t)\big)\, \mathrm{d}t.$$

Lemma 4.5.6 controls the first moment of the multiplier empirical process. For higher moments, the following moment inequality is useful.

Lemma 3.4.2 (Proposition 3.1 of [61]). *Suppose $X_1, \dots, X_n$ are i.i.d. with law P and $\xi_1, \dots, \xi_n$ are i.i.d. mean-zero random variables with $\|\xi_1\|_2 < \infty$. Let $\mathcal{F}$ be a class of measurable functions such that $\sup_{f \in \mathcal{F}} Pf^2 \leq \sigma^2$. Then for any $q \geq 1$,*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \xi_i f(X_i) \right|^q \leq K^q \Bigg[\left(\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \xi_i f(X_i) \right| \right)^q + q^{q/2} (\sqrt{n} \|\xi_1\|_2 \sigma)^q + q^q \mathbb{E} \max_{1 \leq i \leq n} |\xi_i|^q \sup_{f \in \mathcal{F}} |f(X_i)|^q \Bigg].$$

Here $K > 0$ is a universal constant.

To use Lemma 4.5.6, we need to control the size of the empirical process. Let

$$J(\delta, \mathcal{F}, L_2) \equiv \int_0^\delta \sup_Q \sqrt{1 + \log \mathcal{N}\big(\varepsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)\big)}\, \mathrm{d}\varepsilon \tag{3.4.2}$$

denote the *uniform entropy integral*, where the supremum is taken over all discrete probability measures.

We will frequently use the following Koltchinskii-Pollard maximal inequality.

Lemma 3.4.3 (Theorem 2.14.1 of [162])**.** *Let $\mathcal{F}$ be a class of measurable functions with measurable envelope F, and $X_1, \ldots, X_n$ are i.i.d. random variables with law P. Then*

$$\mathbb{E}\left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \lesssim \sqrt{n} J(1, \mathcal{F}, L_2) \|F\|_{L_2(P)}.$$

Our last technical ingredient is Talagrand's concentration inequality [145] for the empirical process in the form given by [106]:

Lemma 3.4.4. *Let $\mathcal{F}$ be a class of measurable functions such that $\sup_{f\in\mathcal{F}} \|f\|_\infty \le b$. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\mathbb{G}_n f| \ge 2\mathbb{E} \sup_{f\in\mathcal{F}} |\mathbb{G}_n f| + \sqrt{8\sigma^2 x} + 34.5 b \frac{x}{\sqrt{n}} \right) \le e^{-x},$$

where $\sigma^2 \equiv \sup_{f\in\mathcal{F}} \mathrm{Var}_P f$, and $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$.

3.4.2 Proof of Theorem 3.2.1

Proof of Theorem 3.2.1. We only prove the case $\mathcal{F}_0 \subset L_\infty(1)$ as in Remark 3.2.2 (1). The proof for the case $\|\hat{f}_n - f_0\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$ follows with only minor modifications. We also work with the more general uniform VC-type condition as in Remark 3.2.3. Let $\delta_n \equiv n^{-\frac{1}{2(2-\gamma)}}$. By the proof of Proposition 2 of [76], we only need to estimate for each $t \ge 1$, with $\mathcal{F}_0(r) = \{f \in \mathcal{F} - f_0 : \|f\|_{L_2(P)} \le r\}$,

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}_0(2^j t\delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(X_i) \right| \right)^2, \quad \mathbb{E}\left(\sup_{f\in\mathcal{F}_0(2^j t\delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f^2(X_i) \right| \right)^2.$$

By the contraction principle for Rademacher processes and the moment inequality Lemma 3.4.2, we only need to estimate the sum of

$$(I) \equiv \left(\mathbb{E} \sup_{f \in \mathcal{F}_0(2^j t \delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(X_i) \right| \right)^2 + \left(\mathbb{E} \sup_{f \in \mathcal{F}_0(2^j t \delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right)^2 \tag{3.4.3}$$

and

$$(II) \equiv \left(2^j t \delta_n (\|\xi_1\|_2 \vee 1) \right)^2 + n^{-1} \cdot \mathbb{E} \max_{1 \le i \le n} (|\xi_i| \vee 1)^2 \cdot \|F_0(2^j t \delta_n)\|_{L_2(P)}^2. \tag{3.4.4}$$

For the first summand (3.4.3), by the Koltchinskii-Pollard maximal inequality for empirical processes (cf. Lemma 3.4.3), since $\mathcal{F}$ is of uniform VC-type, it follows that

$$\max_{1 \le k \le n} \mathbb{E} \sup_{f \in \mathcal{F}_0(2^j t \delta_n)} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i f(X_i) \right| \le C_{\mathcal{F}} \|F_0(2^j t \delta_n)\|_{L_2(P)} \le C'_{\mathcal{F}} (2^j t)^\gamma \delta_n^\gamma.$$

We may apply the multiplier inequality Lemma 4.5.6 with $\psi_n(k) \equiv \sqrt{k} C'_{\mathcal{F}} (2^j t)^\gamma \delta_n^\gamma$ to see that

$$\mathbb{E} \sup_{f \in \mathcal{F}_0(2^j t \delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(X_i) \right| \le 4 C'_{\mathcal{F}} (2^j t)^\gamma \|\xi_1\|_{2,1} \delta_n^\gamma.$$

Hence,

$$(3.4.3) \le C_{\mathcal{F},\xi} (2^j t \delta_n)^{2\gamma}. \tag{3.4.5}$$

(3.4.4) is easy to handle by noting that $\mathbb{E} \max_{1 \le i \le n} (|\xi_i| \vee 1)^2 \lesssim n$ under the assumption that $\|\xi_1\|_2 < \infty$, which entails that

$$(3.4.4) \le C_\xi \left((2^j t \delta_n)^2 + (2^j t \delta_n)^{2\gamma} \right). \tag{3.4.6}$$

Combining (3.4.5) and (3.4.6) and the arguments in the proof of Proposition 2 of [76], we have

$$\begin{aligned} \mathbb{P}\big(\|\hat{f}_n - f_0\|_{L_2(P)} \ge t \delta_n \big) &\le C_{\mathcal{F},\xi} \sum_{j \ge 0 : 2^j t \delta_n \le 2} \frac{(2^j t \delta_n)^2 + (2^j t \delta_n)^{2\gamma}.}{\left(2^{2j} t^2 \sqrt{n} \delta_n^2\right)^2} \\ &\le C'_{\mathcal{F},\xi} \big(n \delta_n^{2(2-\gamma)} \big)^{-1} \sum_{j \ge 0} \frac{1}{(2^j t)^{4-2\gamma}} \le C''_{\mathcal{F},\xi} t^{-2}, \end{aligned}$$

where the last inequality follows from the choice of δ_n. Now the claim of the theorem (in the form of Remark 3.2.2 (1) and under the more general condition as in Remark 3.2.3) follows by integrating the above tail estimate. □

3.4.3 Proof of Theorem 3.2.5

The basic device we will use to derive a lower bound for the risk of the least squares estimator is the following.

Proposition 3.4.5 (Proposition 6 of [76])**.** *Let*

$$F_n(\delta) \equiv \sup_{f \in \mathcal{F}-f_0 : Pf^2 \le \delta^2} (\mathbb{P}_n - P)(2\xi f - f^2) - \delta^2 \equiv E_n(\delta) - \delta^2.$$

Suppose that $0 < \delta_1 < \delta_2$ *are such that* $E_n(\delta_1) < F_n(\delta_2)$*. Then there exists a LSE* f_n^* *such that* $\|f_n^* - f_0\|_{L_2(P)} \ge \delta_1$*.*

The key ingredient in applying the above device is the following.

Proposition 3.4.6. *For any* $\gamma \in (0, 1]$*, there exists some VC-subgraph class* $\tilde{\mathcal{F}}$ *satisfying (3.2.1) with the following property: for each* $\varepsilon \in (0, 1/2)$*, there exists some law for* ξ_1 *with* $\|\xi_1\|_{2(1-\varepsilon)} < \infty$ *such that*

1. *for any* $\vartheta \ge 4$*, there exists some* $\mathfrak{p} > 0$*, with* $\delta_2 \equiv \vartheta n^{-\frac{1}{2(2-\gamma)}}$*,*

$$\mathbb{P}\left(F_n(\delta_2) \ge \frac{1}{2} c_1 \vartheta^\gamma n^{-\frac{1}{2-\gamma}} \tau_n(\varepsilon, \gamma) \right) \ge 2\mathfrak{p},$$

   holds for n *large enough depending on* $\varepsilon, \vartheta, \gamma$*. Here* c_1 *depends on* ε, γ*, and* $\tau_n(\varepsilon, \gamma) \equiv n^{\frac{1-\gamma}{2-\gamma} \cdot \frac{\varepsilon}{2-\varepsilon}}$*.*

2. *for any* $\rho > 0$*, with* $\delta_1 \equiv \rho n^{-\frac{1}{2(2-\gamma)} - \beta_\varepsilon}$*,*

$$\mathbb{P}\left(E_n(\delta_1) \le \mathfrak{p}^{-1} C_{\varepsilon,\xi} \rho^\gamma n^{-\frac{1}{2-\gamma}} \omega_n(\varepsilon, \gamma) \right) \ge 1 - \mathfrak{p}.$$

   Here $\omega_n(\varepsilon, \gamma) = n^{-\gamma\beta_\varepsilon + \frac{\varepsilon}{2(1-\varepsilon)}}$*.*

In (1)-(2) above, $F_n(\delta) \equiv \sup_{f \in \tilde{\mathcal{F}} : Pf^2 \le \delta^2} (\mathbb{P}_n - P)(2\xi f - f^2) - \delta^2 \equiv E_n(\delta) - \delta^2$*.*

The proof of Proposition 3.4.6 relies on a delicate construction of a tree-structured $\tilde{\mathcal{F}}$, and a sequence of technical arguments including concentration of empirical processes, the

Paley-Zygmund moment argument, and an exact characterization of the size of the maxima of summations. To ease reading, a formal proof of Proposition 3.4.6 will be given in Section 3.5.

Proof of Theorem 3.2.5. Let $f_0 = 0$. In order to apply Proposition 3.4.5, we only need to require an order in the exponent of $\tau_n(\cdot, \cdot)$ and $\omega_n(\cdot, \cdot)$ in Proposition 3.4.6, by making a good choice of β_ε. To this end, it suffices to require

$$-\gamma\beta_\varepsilon + \frac{\varepsilon}{2(1-\varepsilon)} < \frac{1-\gamma}{2-\gamma}\frac{\varepsilon}{2-\varepsilon} \Leftrightarrow \beta_\varepsilon > \frac{\varepsilon}{\gamma}\left[\frac{2-\varepsilon\gamma}{(2-\varepsilon)(2-\gamma)(2-2\varepsilon)}\right].$$

Since $\varepsilon \in (0, 1/2)$ and $\gamma \in (0, 1]$, we may choose $\beta_\varepsilon = (2/\gamma) \cdot \varepsilon$, along with any $\vartheta \geq 4$ and $\rho > 0$ small enough to conclude. □

3.4.4 Proof of Theorem 3.3.1

The proof of Theorem 3.3.1 follows from a more principled oracle inequality presented below—it captures the essential geometric property in the model that accounts for both the adaptation and robustness property of the shape-restricted LSE up to error distributions with an $L_{2,1}$ moment.

3.4.5 The general oracle inequality

First some definitions.

Definition 3.4.7. $\mathcal{F}$ is said to satisfy a *convexity-based shape constraint* (under P) if $\mathcal{F}$ is convex, and $\mathcal{F}(\delta) = \{f \in \mathcal{F} : Pf^2 \leq \delta^2\}$ admits a convex envelope $F(\delta)$.

Definition 3.4.8. $\mathcal{G} \subset \mathcal{F}$ is said to be a *basic adaptive subset* of $\mathcal{F}$ if $\mathcal{F} - \mathcal{G} \subset \mathcal{F}$. $\mathcal{G}_m$ is said to be an *m-th order adaptive subset* of $\mathcal{F}$ if for any $g_m \in \mathcal{G}_m$, there is an interval partition $\{I_j\}_{j=1}^m$ of $\mathcal{X} = [0, 1]$ and elements $\tilde{g}_j \in \mathcal{G}$ such that $g_m = \sum_{i=1}^m \mathbf{1}_{I_j}\tilde{g}_j \in \mathcal{F}$.

Before stating the general oracle inequality, recall that a function class $\mathcal{F}$ defined on $\mathcal{X} = [0, 1]$ is called VC-major if the sets $\{x \in \mathcal{X} : f(x) \geq t\}$ with f ranging over $\mathcal{F}$ and t over $\mathbb{R}$ form a VC-class of sets.

Theorem 3.4.9. *Consider the regression model (3.1.1) and the LSE $\hat{f}_n$ in (3.1.2). Suppose that $\|f_0\|_\infty \vee \|f_0^*\|_\infty < \infty$, and that $\xi_1, \ldots, \xi_n$ are mean zero errors independent of i.i.d. covariates X_i's with $\|\xi_1\|_{2,1} < \infty$. Further assume that: (i) $\mathcal{F}$ satisfies a convexity-based shape constraint, and $\mathcal{F} \cap L_\infty(B)$ is a VC-major class for any $B > 0$, and (ii) $\|\hat{f}_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$. Then for any $\delta \in (0,1)$, there exists $c \equiv c(\delta, \|\xi\|_{2,1}, \mathcal{F}, \|f_0\|_\infty, \|f_0^*\|_\infty) > 0$ such that with probability $1-\delta$,*

$$\|\hat{f}_n - f_0^*\|_{L_2(P)}^2 \leq c \inf_{m \in \mathbb{N}} \left(\inf_{f_m \in \mathcal{G}_m \cap L_\infty(\|f_0^*\|_\infty)} \|f_m - f_0^*\|_{L_2(P)}^2 + \frac{m}{n} \cdot \log^2 n \right),$$

where $f_0^* = \operatorname{argmin}_{g \in \mathcal{F} \cap L_2(P)} \|f_0 - g\|_{L_2(P)}$, and $\mathcal{G}_m$ is an m-th order adaptive subset of $\mathcal{F}$.

The proof of Theorem 3.4.9 will be deferred to the next subsection. We first use it to prove Theorem 3.3.1. To this end, we only need to check: (i) the convexity-based shape constraint and VC-major condition of the isotonic and convex models; and (ii) the stochastic boundedness condition for the corresponding LSEs $\hat{f}_n$.

Proof of Theorem 3.3.1. For the isotonic model $\mathcal{F}$, $\mathcal{F}$ is clearly convex, and (3.3.2) is an envelope for $\mathcal{F}(\delta)$ by the L_2 constraint and monotonicity of the function class. Furthermore, it is clear by definition that $\mathcal{F} \cap L_\infty(B)$ is VC-major. Similarly we can verify that the convex model satisfies both the convexity-based shape constraint with the envelope (3.3.2) (cf. Lemma 3.4.15) and the VC-major condition.

The stochastic boundedness of the isotonic and convex LSEs is established in the following lemma:

Lemma 3.4.10. *If $\|f_0\|_\infty < \infty$ and $\|\xi_1\|_2 < \infty$, then both the canonical isotonic and convex LSEs are stochastically bounded: $\|\hat{f}_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$.*

For the isotonic LSE, we use an explicit min-max representation (cf. [129]) to prove this lemma, while for the convex LSE, the explicit characterization of the convex LSE derived in [70] plays a crucial role. The details of the proof of this lemma can be found in Section 3.5. Now the claim of Theorem 3.3.1 follows from Theorem 3.4.9, by noting that $\|f_0^*\|_\infty < \infty$ under $\|f_0\|_\infty < \infty$, and that $\inf_{f_m \in \mathcal{G}_m \cap L_\infty(\|f_0^*\|_\infty)} \|f_m - f_0^*\|_{L_2(P)}^2 = \inf_{f_m \in \mathcal{G}_m} \|f_m - f_0^*\|_{L_2(P)}^2$ for

isotonic model, and the same holds for the convex model when $L_\infty(\|f_0^*\|_\infty)$ is replaced by $L_\infty(C\|f_0^*\|_\infty)$ for some large enough $C > 0$. □

3.4.6 Proof of Theorem 3.4.9

The first ingredient of the proof is the following proposition relating the convergence rate of $\hat{f}_n$ to the size of localized empirical processes.

Proposition 3.4.11. *Consider the regression model (3.1.1) and the least squares estimator $\hat{f}_n$ in (3.1.2). Suppose that $\xi_1, \dots, \xi_n$ are mean-zero random variables independent of $X_1, \dots, X_n$, and $\mathcal{F}$ is convex with $\mathcal{F} - f_0^* \subset L_\infty(1)$. Further assume that*

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}: \|f - f_0^*\|_{L_2(P)} \le \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i (f - f_0^*)(X_i) \right| &\lesssim \phi_n(\delta), \\
\mathbb{E} \sup_{f \in \mathcal{F}: \|f - f_0^*\|_{L_2(P)} \le \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f - f_0^*)(X_i) \right| &\lesssim \phi_n(\delta), \\
\mathbb{E} \sup_{f \in \mathcal{F}: \|f - f_0^*\|_{L_2(P)} \le \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f - f_0^*)(X_i)(f_0 - f_0^*)(X_i) \right| &\lesssim \phi_n(\delta),
\end{aligned} \tag{3.4.7}$$

hold for some ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta$ is non-increasing. Then $\|\hat{f}_n - f_0^*\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}(\delta_n)$ holds for any δ_n such that $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$.

Proof. This is a special case of Proposition 3.4.16, the proof of which will be given therein. □

By Proposition 3.4.11, we only need to control the size of the empirical processes (3.4.7) centered at f_0^*. The following lemma will be useful in this regard by approximating f_0^* via arbitrary $f_m \in \mathcal{G}_m$.

Lemma 3.4.12. *Suppose that the hypotheses of Theorem 3.4.9 hold. Let $\{\delta_n\}_{n \in \mathbb{N}}$ be a sequence of positive real numbers such that $\delta_n \ge 1/n$. Then for any $f_m \in \mathcal{G}_m \cap L_\infty(\|f_0^*\|_\infty)$*

and $B > 0$,

$$\max\Bigg\{\mathbb{E}\sup_{f\in\mathcal{F}: f-f_0^*\in L_2(\delta_n)\cap L_\infty(B)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i(f-f_0^*)(X_i)\right|,$$
$$\mathbb{E}\sup_{f\in\mathcal{F}: f-f_0^*\in L_2(\delta_n)\cap L_\infty(B)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i(f-f_0^*)(X_i)\right|,$$
$$\mathbb{E}\sup_{f\in\mathcal{F}: f-f_0^*\in L_2(\delta_n)\cap L_\infty(B)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i(f-f_0^*)(X_i)(f_0-f_0^*)(X_i)\right|\Bigg\}$$
$$\le C_{\mathcal{F},\|f_0\|_\infty,\|f_0^*\|_\infty,B}\cdot\|\xi_1\|_{2,1}\sqrt{\log(1/\delta_n)}\bar{L}_n\cdot(\delta_n\vee\|f_m-f_0^*\|_{L_2(P)})\sqrt{m},$$

where $\bar{L}_n \equiv \sqrt{\log n}$.

To prove Lemma 3.4.12, we need the following form of a multiplier inequality proved in Proposition 1 of [76].

Lemma 3.4.13. *Suppose that* $\xi_1,\dots,\xi_n$ *are i.i.d. mean-zero random variables independent of i.i.d.* $X_1,\dots,X_n$. *Then for any function class* $\mathcal{F}$,

$$\mathbb{E}\left\|\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}} \le \mathbb{E}\left[\sum_{k=1}^n(|\eta_{(k)}|-|\eta_{(k+1)}|)\mathbb{E}\left\|\sum_{i=1}^k \varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right] \tag{3.4.8}$$

where $|\eta_{(1)}|\ge\cdots\ge|\eta_{(n)}|\ge|\eta_{(n+1)}|\equiv 0$ *are the reversed order statistics for* $\{|\xi_i-\xi_i'|\}_{i=1}^n$ *with* $\{\xi_i'\}$ *being an independent copy of* $\{\xi_i\}$.

The following entropy estimate for bounded VC-major classes will be useful.

Lemma 3.4.14. *Let* $\mathcal{F}_0\subset L_\infty(1)$ *be a VC-major class defined on* $\mathcal{X}$. *Then there exists some constant* $C\equiv C_{\mathcal{F}_0}>0$ *such that for any* $\mathcal{F}\subset\mathcal{F}_0$, *and any probability measure* Q, *the entropy estimate*

$$\log\mathcal{N}\big(\varepsilon\|F\|_{L_2(Q)},\mathcal{F},L_2(Q)\big)\le\frac{C}{\varepsilon}\log\left(\frac{C}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon\|F\|_{L_2(Q)}}\right),\ \textit{for all}\ \varepsilon\in(0,1)$$

holds for any envelope F *of* $\mathcal{F}$.

The proof of this lemma essentially follows from page 1171-1172 of [59] with a minor modification. We include some details in Section 3.5 for the convenience of the reader.

We also need the following lemma concerning the envelope of a convex function given constraints on its L_2 size. The proof can be found in Lemma 7.3 of [72].

Lemma 3.4.15. *If f is a convex function on $[0,1]$ with $\int_0^1 |f(x)|^2 \, \mathrm{d}x \le 1$, then $|f(x)| \le 2\sqrt{3}\big(x^{-1/2} \vee (1-x)^{-1/2}\big)$ for all $x \in (0,1)$.*

Proof of Lemma 3.4.12. In the proof we omit the dependence on $L_\infty(B)$ if there is no confusion. All three empirical processes can be handled in essentially the same way so we focus on the most difficult first one (with ξ_i's only admitting a $L_{2,1}$ moment). We will apply Lemma 3.4.13 in the following form:

$$\begin{aligned} &\mathbb{E} \sup_{f\in\mathcal{F}: f-f_0^* \in L_2(\delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i (f-f_0^*)(X_i) \right| \\ &\qquad \le 3\|\xi_1\|_{2,1} \max_{1\le k\le n} \mathbb{E} \sup_{f\in\mathcal{F}: f-f_0^* \in L_2(\delta_n)} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i (f-f_0^*)(X_i) \right|. \end{aligned} \tag{3.4.9}$$

To see this, note that the right hand side of (4.5.8) can be bounded by

$$\mathbb{E}\Big[\sum_{k=1}^n \sqrt{k}(|\eta_{(k)}| - |\eta_{(k+1)}|) \Big] \cdot \max_{1\le k\le n} \mathbb{E} \Big\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i f(X_i) \Big\|_{\mathcal{F}}$$

where $\mathbb{E}\big[\sum_{k=1}^n \sqrt{k}(|\eta_{(k)}| - |\eta_{(k+1)}|)\big] \le \sqrt{n}\|\eta_1\|_{2,1} \le 3\sqrt{n}\|\xi_1\|_{2,1}$. The first inequality follows from similar lines as in the proof of Theorem 1 of [76] and the second inequality uses Problem 2 on page 186 of [162]. This proves (3.4.9). Note that any $f_m \in \mathcal{G}_m$ has a representation $f_m = \sum_{j=1}^m g_j \mathbf{1}_{I_j}$, where $\{I_j = [x_j, x_{j+1}]\}_{j=1}^m$ is a partition of $\mathcal{X} = [0,1]$ with $x_1 = 0, x_{m+1} = 1$ and $g_j \in \mathcal{G}$. Then for any $f_m \in \mathcal{G}_m$, the empirical process localized at f_0^* can be controlled via

$$\begin{aligned} &\mathbb{E} \sup_{f\in\mathcal{F}: f-f_0^* \in L_2(\delta_n)\cap L_\infty(B)} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i (f-f_0^*)(X_i) \right| \\ &\le \mathbb{E} \sup_{\substack{f\in\mathcal{F}: \|f-f_m\|_{L_2(P)} \le \delta_n + \|f_m - f_0^*\|_{L_2(P)}, \\ \|f\|_\infty \le B + \|f_0^*\|_\infty}} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i (f-f_m)(X_i) \right| + \|f_0^* - f_m\|_{L_2(P)}, \end{aligned} \tag{3.4.10}$$

where the second term holds because the collection $\{f_0^* - f_m\}$ consists of just one element. The first term in the above term can be further bounded by

$$\sum_{j=1}^{m} \mathbb{E}\left[\frac{\sqrt{k_j}}{\sqrt{k}}\mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}:\|f-f_m\|_{L_2(P)}\le\delta_n+\|f_m-f_0^*\|_{L_2(P)},\\ \|f\|_\infty\le B+\|f_0^*\|_\infty}}\left|\frac{1}{\sqrt{k_j}}\sum_{X_i\in I_j}\varepsilon_i(f-g_j)(X_i)\right| \right.\right. \tag{3.4.11}$$

$$\left.\left. \Big| k_j(\boldsymbol{X}) = k_j\right]\right]$$

$$\le \sum_{j=1}^{m} \mathbb{E}\left[\frac{\sqrt{k_j}}{\sqrt{k}}\mathbb{E}\left[\sup_{\substack{f|_{I_j}\in\mathcal{F}|_{I_j}:\\ \|f\|_\infty\le B+2\|f_0^*\|_\infty,\\ Pf^2\le(\delta_n+\|f_m-f_0^*\|_{L_2(P)})^2}}\left|\frac{1}{\sqrt{k_j}}\sum_{X_i\in I_j}\varepsilon_i f|_{I_j}(X_i)\right| \Big| k_j(\boldsymbol{X}) = k_j\right]\right]$$

where $k_j(\boldsymbol{X}) = \sum_{i=1}^{k} \mathbf{1}_{I_j}(X_i)$, and in the second line we used the definition of a basic adaptive subset (cf. Definition 3.4.8). From now on we write $\tilde{\delta}_n \equiv \delta_n + \|f_m - f_0^*\|_{L_2(P)}$ and $B_0 \equiv B + 2\|f_0^*\|_\infty$ for notational convenience. Since $\big(\mathcal{F}\cap L_\infty(B_0)\big)|_{I_j}$ is VC-major, so is its subset $\mathcal{F}_{I_j}(\tilde{\delta}_n) \equiv \{f|_{I_j} \in \big(\mathcal{F}\cap L_\infty(B_0)\big)|_{I_j} : Pf^2 \le \tilde{\delta}_n^2\}$. It follows by Lemma 3.4.14 that there exists some $C \equiv C_{\mathcal{F},B_0} > 0$ such that for any probability measure Q on I_j, and any $\varepsilon \in (0,1)$,

$$\log \mathcal{N}\left(\varepsilon\|F_{I_j}(\tilde{\delta}_n)\|_{L_2(Q)}, \mathcal{F}_{I_j}(\tilde{\delta}_n), L_2(Q)\right) \le \frac{C}{\varepsilon}\log\left(\frac{C}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon\|F_{I_j}(\tilde{\delta}_n)\|_{L_2(Q)}}\right),$$

where $F_{I_j}(\delta)$ is any envelope for $\mathcal{F}_{I_j}(\delta)$. This enables us to apply the Koltchinskii-Pollard maximal inequality to see that the summand (=conditional expectation) in the second line of (3.4.11) can be bounded by (further conditioning on which X_i's lie in the interval I_j, each case corresponds to i.i.d. uniforms on I_j)

$$\int_0^1 \sqrt{\frac{C}{\varepsilon}\log\left(\frac{C}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon \inf_Q\|F_{I_j}(\tilde{\delta}_n)\|_{L_2(Q)}}\right)}\,\mathrm{d}\varepsilon \cdot \sqrt{P_{I_j}F_{I_j}^2(\tilde{\delta}_n)}, \tag{3.4.12}$$

where P_{I_j} is the uniform distribution on I_j.

In order to evaluate (3.4.12), note that by the definition of convexity-based shape constraint and Lemma 3.4.15, the envelopes $F_{I_j}(\delta)$'s can be taken as the restrictions of the global envelope

$$F(\delta)(x) \equiv \left(\frac{\delta}{\sqrt{x}} \vee \frac{\delta}{\sqrt{1-x}}\right) \wedge B_0$$

to the I_j's. Without loss of generality we assume: (i) $B_0 = 1$, (ii) $\tilde{\delta}_n^2 < 1/2$ and (iii) $\tilde{\delta}_n^2$ and $1 - \tilde{\delta}_n^2$ are one of the endpoints of some intervals in $\{I_j\}$ (otherwise, we may take an alternative representation of $f_m \in \mathcal{G}_{m+2}$ by adding these two points).

Note that $\inf_Q \|F_{I_j}(\tilde{\delta}_n)\|_{L_2(Q)} \geq \sqrt{2}\tilde{\delta}_n > 1/n$ by the assumption $\delta_n \geq 1/n$, and hence the integral term in (3.4.12) can be bounded by

$$\int_0^1 \sqrt{\frac{C}{\varepsilon} \log\left(\frac{C}{\varepsilon}\right) \log\left(\frac{n}{\varepsilon}\right)} \, \mathrm{d}\varepsilon \lesssim \sqrt{\log n} \equiv \bar{L}_n.$$

To handle the $\sqrt{P_{I_j} F_{I_j}^2(\tilde{\delta}_n)}$ term in (3.4.12), define the index sets $\mathcal{J}_1 \equiv \{1 \leq j \leq m : I_j \subset [0, \tilde{\delta}_n^2]\}, \mathcal{J}_2 \equiv \{1 \leq j \leq m : I_j \subset [\tilde{\delta}_n^2, 1 - \tilde{\delta}_n^2]\}$ and $\mathcal{J}_3 \equiv \{1 \leq j \leq m : I_j \subset [1 - \tilde{\delta}_n^2, 1]\}$. It is easy to see that $\mathcal{J}_1 \cup \mathcal{J}_2 \cup \mathcal{J}_3 = \{1, \ldots, m\}$. Clearly for $j \in \mathcal{J}_1 \cup \mathcal{J}_3$,

$$P_{I_j} F_{I_j}^2(\tilde{\delta}_n) = |I_j|^{-1} \int_{I_j} F_{I_j}^2(\tilde{\delta}_n)(x) \, \mathrm{d}x \leq 1,$$

and for $j \in \mathcal{J}_2$,

$$\begin{aligned} P_{I_j} F_{I_j}^2(\tilde{\delta}_n) &\leq |I_j|^{-1} \tilde{\delta}_n^2 \int_{x_j}^{x_{j+1}} \left(\frac{1}{x} \vee \frac{1}{1-x}\right) \, \mathrm{d}x \\ &\leq |I_j|^{-1} \tilde{\delta}_n^2 \left[\log\left(\frac{x_{j+1}}{x_j}\right) \vee \log\left(\frac{1-x_j}{1-x_{j+1}}\right)\right]. \end{aligned}$$

Summarizing the above discussion shows that we can further bound (3.4.11) by a $\mathcal{O}(\bar{L}_n)$ multiple of

$$\begin{aligned} &\sum_{j \in \mathcal{J}_1 \cup \mathcal{J}_3} \mathbb{E}\left[\sqrt{\frac{k_j}{k}} \cdot 1\right] + \sum_{j \in \mathcal{J}_2} \tilde{\delta}_n \cdot \mathbb{E}\left[\sqrt{\frac{k_j}{k}} \cdot \sqrt{\frac{\log(x_{j+1}) - \log(x_j)}{x_{j+1} - x_j}}\right] \\ &\qquad + \sum_{j \in \mathcal{J}_2} \tilde{\delta}_n \cdot \mathbb{E}\left[\sqrt{\frac{k_j}{k}} \cdot \sqrt{\frac{\log(1-x_j) - \log(1-x_{j+1})}{(1-x_j) - (1-x_{j+1})}}\right] \\ &\equiv (I) + (II) + (III). \end{aligned} \tag{3.4.13}$$

The first term of (3.4.13) is easy to handle: by the Cauchy-Schwarz inequality,

$$(I) \leq \sqrt{k^{-1}\left(\mathbb{E} \sum_{j \in \mathcal{J}_1 \cup \mathcal{J}_3} k_j(\boldsymbol{X})\right) \cdot |\mathcal{J}_1 \cup \mathcal{J}_3|} \leq \sqrt{\sum_{j \in \mathcal{J}_1 \cup \mathcal{J}_3} |I_j|} \cdot \sqrt{m} \lesssim \tilde{\delta}_n \sqrt{m}.$$

The second and third terms of (3.4.13) can be handled in a similar fashion; we only consider the second term of (3.4.13). Again by the Cauchy-Schwarz inequality,

$$\begin{aligned}(II) &\le \tilde{\delta}_n\sqrt{m}\cdot\sqrt{\mathbb{E}\Big[\sum_{j\in\mathcal{J}_2}\frac{k_j(\boldsymbol{X})}{k}\cdot\frac{\log(x_{j+1})-\log(x_j)}{x_{j+1}-x_j}\Big]}\\ &= \tilde{\delta}_n\sqrt{m}\sqrt{\sum_{j\in\mathcal{J}_2}\big(\log(x_{j+1})-\log(x_j)\big)} \lesssim \sqrt{m}\cdot\tilde{\delta}_n\sqrt{\log(1/\tilde{\delta}_n)}.\end{aligned}$$

Collecting the above estimates, we see that (3.4.11) can be bounded by a constant multiple of $\sqrt{m}\cdot\tilde{\delta}_n\sqrt{\log(1/\tilde{\delta}_n)}\bar{L}_n$. Thus, (3.4.10) yields that

$$\max_{1\le k\le n}\mathbb{E}\sup_{f\in\mathcal{F}:f-f_0^*\in L_2(\delta_n)}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\varepsilon_i(f-f_0^*)(X_i)\right| \le C'\sqrt{m}\cdot\tilde{\delta}_n\sqrt{\log(1/\tilde{\delta}_n)}\bar{L}_n.$$

Combined with (3.4.9), the claim of the lemma follows. □

Proof of Theorem 3.4.9. The proof follows easily from the reduction scheme Proposition 3.4.16 and Lemma 3.4.12 by solving a quadratic inequality. We provide some details below. Abusing notation, we let $f_m \in \operatorname{argmin}_{g_m\in\mathcal{G}_m}\|g_m - f_0^*\|_{L_2(P)}$ and m be the index attaining the infimum of the oracle inequality in the statement of the theorem. We only need to choose δ_n such that

$$\sqrt{m}(\delta_n + \|f_m - f_0^*\|_{L_2(P)})\sqrt{\log(1/\delta_n)}\bar{L}_n \le c_{\delta,\mathcal{F},\|f_0^*\|_\infty,\|\xi\|_{2,1}}\sqrt{n}\delta_n^2.$$

Suppose $\log(1/\delta_n)\lesssim\log n$. Then we can easily solve for the zeros for quadratic forms to see that the inequality in the last display holds if

$$\delta_n^2 \gtrsim \frac{m\bar{L}_n^2\log n}{n} + \sqrt{\frac{m\bar{L}_n^2\log n}{n}}\|f_m - f_0^*\|_{L_2(P)}.$$

The assumption $\log(1/\delta_n)\lesssim\log n$ apparently holds. The right hand side of the above display can be further bounded up to a constant by $\frac{m\bar{L}_n^2\log n}{n} + \|f_m - f_0^*\|_{L_2(P)}^2$ by the basic inequality $ab \le (a^2+b^2)/2$, thereby completing the proof of Theorem 3.4.9. □

3.4.7 Proof of Theorem 3.3.5

The proof of Theorem 3.3.5 follows a similar strategy as that of Theorem 3.4.9. First we need the following reduction scheme.

Proposition 3.4.16. *Consider the additive model (3.3.3) and the least squares estimator $\hat{f}_n$ in (3.3.4). Suppose that $\xi_1, \dots, \xi_n$ are mean-zero random variables independent of $(X_1, Z_1), \dots, (X_n, Z_n)$, and $\mathcal{F}$ is convex with $\mathcal{F} - f_0^* \subset L_\infty(1)$. Further assume that all three parts of (3.4.7) and*

$$\mathbb{E} \sup_{\substack{f \in \mathcal{F}: \|f - f_0^*\|_{L_2(P)} \le \delta \\ h \in \mathcal{H}}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f - f_0^*)(X_i)(h - (\phi_0 - f_0))(X_i, Z_i) \right| \lesssim \phi_n(\delta), \tag{3.4.14}$$

hold for some ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta$ is non-increasing. Then $\|\hat{f}_n - f_0^*\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}(\delta_n)$ holds for any δ_n such that $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$.

Proof. Recall that $f_0 = P_Z \phi_0(\cdot, Z)$. By the definition of the LSE,

$$\begin{aligned}
&\mathbb{P}_n(\phi_0 + \xi - \hat{f}_n - \hat{h}_n)^2 \le \mathbb{P}_n(\phi_0 + \xi - f_0^* - \hat{h}_n)^2 \\
\Leftrightarrow \quad & \mathbb{P}_n(f_0^* - \hat{f}_n)\big(2\phi_0 + 2\xi - \hat{f}_n - f_0^* - 2\hat{h}_n\big) \le 0 \\
\Leftrightarrow \quad & \mathbb{P}_n(f_0^* - \hat{f}_n)^2 + 2\mathbb{P}_n(f_0^* - \hat{f}_n)\big(\phi_0 + \xi - f_0^* - \hat{h}_n\big) \le 0 \\
\Leftrightarrow \quad & -\mathbb{P}_n(f_0^* - \hat{f}_n)^2 - 2\mathbb{P}_n(f_0^* - \hat{f}_n)\xi - 2\mathbb{P}_n(f_0^* - \hat{f}_n)(f_0 - f_0^*) \\
& \qquad - 2\mathbb{P}_n(f_0^* - \hat{f}_n)(\phi_0 - f_0 - \hat{h}_n) \ge 0 \\
\Leftrightarrow \quad & -(\mathbb{P}_n - P)\Big[(f_0^* - \hat{f}_n)^2 - 2\xi(f_0^* - \hat{f}_n)\Big] - P(f_0^* - \hat{f}_n)^2 \\
& \qquad - 2(\mathbb{P}_n - P)(f_0^* - \hat{f}_n)(f_0 - f_0^*) - 2P(f_0^* - \hat{f}_n)(f_0 - f_0^*) \\
& \qquad\qquad - 2(\mathbb{P}_n - P)(f_0^* - \hat{f}_n)(\phi_0 - f_0 - \hat{h}_n) \ge 0.
\end{aligned}$$

The last equivalence holds since

$$\begin{aligned}
& P(f_0^* - \hat{f}_n)(X)(\phi_0 - f_0 - \hat{h}_n)(X, Z) \\
&= P\Big[(f_0^* - \hat{f}_n)(X) P\big[(\phi_0 - f_0 - \hat{h}_n)(X, Z)\big|X\big]\Big] \\
&= P\Big[(f_0^* - \hat{f}_n)(X)\big(P\big[\phi_0(X, Z)\big|X\big] - f_0(X) - P\hat{h}_n(Z)\big)\Big] = 0,
\end{aligned}$$

where we used (i) $P[\phi_0(X,Z)|X] = f_0(X)$, and (ii) $Ph = 0$ for all $h \in \mathcal{H}$. Now since $f_0^* \in \text{argmin}_{g \in \mathcal{F} \cap L_2(P)} \|f_0 - g\|_{L_2(P)}$, it follows from the convexity of $\mathcal{F}$ that $P(f_0^* - \hat{f}_n)(f_0 - f_0^*) \geq 0$ [more specifically, for each $\varepsilon > 0$, since $(1-\varepsilon)f_0^* + \varepsilon \hat{f}_n^* \in \mathcal{F} \cap L_2(P)$ by convexity of $\mathcal{F}$, the definition of f_0^* yields that $P(f_0 - f_0^*)^2 \leq P(f_0 - (1-\varepsilon)f_0^* - \varepsilon \hat{f}_n)^2 = P(f_0 - f_0^* + \varepsilon(f_0^* - \hat{f}_n))^2$. The claim follows by expanding the square and taking $\varepsilon \to 0$]. This implies that, with $S_j(\delta_n) \equiv \{f \in \mathcal{F} : 2^{j-1}\delta_n < \|f - f_0^*\|_{L_2(P)} \leq 2^j \delta_n\}$, on the event $\{2^{j-1}\delta_n < \|\hat{f}_n - f_0^*\|_{L_2(P)} \leq 2^j \delta_n\}$, it holds that

$$
\begin{aligned}
&\sup_{f \in S_j(\delta_n)} |(\mathbb{P}_n - P)(f - f_0^*)^2| + 2 \sup_{f \in S_j(\delta_n)} |(\mathbb{P}_n - P)\xi(f - f_0^*)| \\
&\quad + 2 \sup_{f \in S_j(\delta_n)} |(\mathbb{P}_n - P)(f - f_0^*)(f_0 - f_0^*)| \\
&\quad\quad + 2 \sup_{f \in S_j(\delta_n), h \in \mathcal{H}} |(\mathbb{P}_n - P)(f - f_0^*)(h - (\phi_0 - f_0))| \\
&\geq -(\mathbb{P}_n - P)\left[(f_0^* - \hat{f}_n)^2 - 2\xi(f_0^* - \hat{f}_n)\right] \\
&\quad - 2(\mathbb{P}_n - P)(f_0^* - \hat{f}_n)(f_0 - f_0^*) - 2(\mathbb{P}_n - P)(f_0^* - \hat{f}_n)(\phi_0 - f_0 - \hat{h}_n) \\
&\geq 2^{2j-2}\delta_n^2.
\end{aligned}
$$

Hence by symmetrization, the contraction principle for Rademacher processes and the as-

sumptions we see that

$$
\begin{aligned}
&\mathbb{P}\big(\|\hat{f}_n - f_0^*\|_{L_2(P)} > 2^{M-1}\delta_n\big) \\
&\leq \sum_{j\geq M} \mathbb{P}\bigg(\sup_{f\in S_j(\delta_n)} |(\mathbb{P}_n - P)(f-f_0^*)^2| + 2 \sup_{f\in S_j(\delta_n)} |(\mathbb{P}_n - P)\xi(f-f_0^*)| \\
&\qquad + 2 \sup_{f\in S_j(\delta_n)} |(\mathbb{P}_n - P)(f-f_0^*)(f_0-f_0^*)| \\
&\qquad + 2 \sup_{f\in S_j(\delta_n), h\in\mathcal{H}} |(\mathbb{P}_n - P)(f-f_0^*)(h-(\phi_0-f_0))| \geq 2^{2j-2}\delta_n^2 \bigg) \\
&\lesssim \sum_{j\geq M} \big(2^{2j}\sqrt{n}\delta_n^2\big)^{-1} \bigg(\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_0(2^j\delta_n)} \vee \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_0(2^j\delta_n)\otimes\xi} \\
&\qquad \vee \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_0(2^j\delta_n)\otimes(f_0-f_0^*)} \vee \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_0(2^j\delta_n)\otimes(\mathcal{H}-(\phi_0-f_0))} \bigg) \\
&\leq C \sum_{j\geq M} \frac{\phi_n(2^j\delta_n)}{2^{2j}\sqrt{n}\delta_n^2} \leq C \sum_{j\geq M} \frac{\phi_n(\delta_n)}{2^j\sqrt{n}\delta_n^2} \lesssim \sum_{j\geq M} 2^{-j} \to 0
\end{aligned}
$$

as $M \to \infty$. Here we denote $\mathcal{F}_0 \equiv \mathcal{F} - f_0^*$, and in the last sequence of inequalities we used the assumption that $\delta \mapsto \phi_n(\delta)/\delta$ is non-decreasing and the definition of δ_n. This completes the proof. □

By Proposition 3.4.16, apart from the empirical processes in Lemma 3.4.12, we also need to control the empirical process (3.4.14) indexed by a suitably localized subset of $\mathcal{F}\otimes(\mathcal{H}-(\phi_0-f_0)) \equiv \{f(x)\big(h(z)-\phi_0(x,z)-f_0(x)\big) : f\in\mathcal{F}, h\in\mathcal{H}\}$. In a related work, [155] derived bounds for similar empirical processes under L_∞-type entropy conditions for both $\mathcal{F}$ and $\mathcal{H}$ (cf. Theorem 3.1 of [155]), which apparently fail for shape constrained classes.

Lemma 3.4.17. *Suppose that the hypotheses of Theorem 3.3.5 hold. Let $\{\delta_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers such that $\delta_n \geq 1/n$. Then for any $f_m \in \mathcal{G}_m \cap L_\infty(\|f_0^*\|_\infty)$, and $B > 0$,*

$$
\begin{aligned}
&\mathbb{E} \sup_{\substack{f\in\mathcal{F}: f-f_0^*\in L_2(\delta_n)\cap L_\infty(B) \\ h\in\mathcal{H}}} \bigg| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f-f_0^*)(X_i)(h-h_0)(X_i,Z_i) \bigg| \\
&\qquad \leq C_{\mathcal{H},\mathcal{F},\|\phi_0\|_\infty,\|f_0^*\|_\infty,B} \cdot \sqrt{\log(1/\delta_n)}\bar{L}_n \cdot (\delta_n \vee \|f_m - f_0^*\|_{L_2(P)})\sqrt{m}.
\end{aligned}
$$

Here $\bar{L}_n \equiv \sqrt{\log n}$.

We need some technical lemmas. Recall P_X, P_Z are the marginal probability distributions of (X, Z), i.e. uniform distribution on $[0, 1]$.

Lemma 3.4.18. *Let* $\mathcal{H}$ *be a class of measurable functions defined on* $[0, 1]$*, and let* $f \in L_2(P_X), g \in L_2(P)$*. Then for any probability measure* Q *on* $[0, 1]^2$*,*

$$\mathcal{N}\big(\varepsilon\|f \otimes 1\|_{L_2(Q)}, f \otimes \big(\mathcal{H} - g\big), L_2(Q)\big) \le \mathcal{N}\big(\varepsilon, \mathcal{H}, L_\infty\big).$$

Lemma 3.4.19. *Suppose the conditions on* $\mathcal{H}$ *in Theorem 3.3.5 hold and* $\mathcal{F}$ *is the class of monotonic non-decreasing or convex functions on* $[0, 1]$*. Then for any* $\mathcal{F}' \subset \mathcal{F} \cap L_\infty(1)$ *and any probability measure* Q *on* $[0, 1]^2$*, the entropy estimate*

$$\begin{aligned}&\log \mathcal{N}\big(\varepsilon\|F' \otimes 1\|_{L_2(Q)}, \mathcal{F}' \otimes (\mathcal{H} - (\phi_0 - f_0)), L_2(Q)\big)\\ &\quad\lesssim \frac{1}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon\|F' \otimes 1\|_{L_2(Q)}}\right) \vee \varepsilon^{-\gamma},\ \textit{for all } \varepsilon \in (0, 1)\end{aligned}$$

holds for any envelope F' *of* $\mathcal{F}'$*. The constant in the above estimate does not depend on the choice of* $\mathcal{F}'$ *or* Q*.*

The proofs of Lemmas 3.4.18 and 3.4.19 are standard. We include the details in Section 3.5 for completeness.

Proof of Lemma 3.4.17. The proof follows the same strategy as that of Lemma 3.4.12. We only prove the isotonic case $\mathcal{G}_m = \mathcal{M}_m$; the convex case follows by similar arguments. As in the proof of Lemma 3.4.12, we will omit the explicit dependence on $L_\infty(B)$ if no confusion arises. Note that

$$\begin{aligned}&\mathbb{E}\sup_{f\in\mathcal{F}: f - f_0^* \in L_2(\delta_n), h\in\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i (f - f_0^*)(X_i)(h - (\phi_0 - f_0))(X_i, Z_i)\right| \qquad (3.4.15)\\ &\le \mathbb{E}\sup_{\substack{f\in\mathcal{F}: \|f - f_m\|_{L_2(P)} \le \delta_n + \|f_m - f_0^*\|_{L_2(P)},\\ h\in\mathcal{H}}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i (f - f_m)(X_i)(h - (\phi_0 - f_0))(X_i, Z_i)\right|\\ &\qquad + \mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i (f_m - f_0^*)(X_i)(h - (\phi_0 - f_0))(X_i, Z_i)\right| \equiv (I) + (II).\end{aligned}$$

We first handle (II) in (3.4.15). The entropy assumption on $\mathcal{H}$ coupled with Lemma 3.4.18 entails that the uniform entropy integral for the class $(f_m - f_0^*) \otimes \big(\mathcal{H} - (\phi_0 - f_0)\big)$ converges. By Theorem 2.14.1 of [162], we have the following estimate

$$(II) \leq C_{\mathcal{H}} \|f_m - f_0^*\|_{L_2(P)}.$$

For the first term (I) in (3.4.15), we mimic the proof strategy in Lemma 3.4.12: any piecewise constant $f_m \in \mathcal{M}_m$ has a representation $f_m = \sum_{j=1}^m g_j \mathbf{1}_{I_j}$, where $\{I_j = [x_j, x_{j+1}]\}_{j=1}^m$ is a partition of $[0,1]$ with $x_1 = 0, x_{m+1} = 1$ and g_j takes constant values on the intervals I_j. Then for such $f_m \in \mathcal{M}_m$, write $\tilde{I}_j = I_j \times [0,1]$, we have

$$\begin{aligned}
&\sum_{j=1}^m \mathbb{E}\bigg[\frac{\sqrt{n_j}}{\sqrt{n}} \mathbb{E}\bigg[\sup_{\substack{f \in \mathcal{F}: f - f_m \in L_2(\tilde{\delta}_n) \\ h \in \mathcal{H}}} \bigg| \frac{1}{\sqrt{n_j}} \sum_{(X_i, Z_i) \in \tilde{I}_j} \varepsilon_i (f - g_j)(X_i)(h - (\phi_0 - f_0))(X_i, Z_i) \bigg| \\
&\qquad\qquad \bigg| n_j(\boldsymbol{X}, \boldsymbol{Z}) = n_j \bigg] \bigg] \\
&\leq \sum_{j=1}^m \mathbb{E}\bigg[\frac{\sqrt{n_j}}{\sqrt{n}} \mathbb{E}\bigg[\sup_{\substack{f|_{I_j} \in \mathcal{F}|_{I_j}: \\ \|f\|_\infty \leq B + 2\|f_0^*\|_\infty, \\ P_X f^2 \leq \tilde{\delta}_n^2, h \in \mathcal{H}}} \bigg| \frac{1}{\sqrt{n_j}} \sum_{(X_i, Z_I) \in \tilde{I}_j} \varepsilon_i f \otimes (h - (\phi_0 - f_0))|_{\tilde{I}_j}(X_i, Z_i) \bigg| \\
&\qquad\qquad \bigg| n_j(\boldsymbol{X}, \boldsymbol{Z}) = n_j \bigg] \bigg]
\end{aligned}$$

where $\tilde{\delta}_n \equiv \delta_n + \|f_m - f_0^*\|_{L_2(P)}$. Here $n_j(\boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^n \mathbf{1}_{\tilde{I}_j}(X_i, Z_i)$ and in the second line we used the fact that $(f - f_m)|_{I_j} \in \mathcal{F}|_{I_j}$. By Lemma 3.4.19 and the Koltchinskii-Pollard maximal inequality, each summand of the above display can be bounded up to a constant (depending on $\mathcal{F}, \mathcal{H}, \|\phi_0\|_\infty$) by

$$\begin{aligned}
&\int_0^1 \sqrt{\frac{1}{\varepsilon} \log\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon \|F_{I_j}(\tilde{\delta}_n) \otimes 1\|_{L_2(Q)}}\right) \vee \varepsilon^{-\gamma}}\, \mathrm{d}\varepsilon \\
&\qquad \times \Big(P_{\tilde{I}_j} F_{I_j}^2(\tilde{\delta}_n)\Big)^{1/2} \lesssim \bar{L}_n \cdot \sqrt{P_{I_j} F_{I_j}^2(\tilde{\delta}_n)},
\end{aligned}$$

where $P_{\tilde{I}_j}$ is the uniform distribution on $\tilde{I}_j$ and $F_{I_j}(\delta)$ is the envelope for $\big(\mathcal{F} \cap L_\infty(B + 2\|f_0^*\|_\infty) \cap L_2(\delta)\big)|_{I_j}$, and the inequality in the above display follows from similar arguments as in the proof of Lemma 3.4.12. From here the proof proceeds along the same lines as that of the proof for Lemma 3.4.12. □

Proof of Theorem 3.3.5. The proof of Theorem 3.3.5 follows the arguments of the proof of Theorem 3.4.9 by using Proposition 3.4.16 along with Lemmas 3.4.12 and 3.4.17, combined with the stochastic boundedness of the LSE:

Lemma 3.4.20. *Suppose that the hypotheses of Theorem 3.3.5 hold (except that $\mathcal{H}$ is only required to have a continuously square integrable envelope $P_Z H^2 < \infty$). Then both the canonical isotonic and convex LSEs in the additive regression model (3.3.4) are stochastically bounded: $\|\hat{f}_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$.*

The proof of this lemma will be detailed in Section 3.5, and hence completes the proof of Theorem 3.3.5. □

3.5 Proofs of technical results

In this section, we collect the proofs for technical results in three groups:

1. the key Proposition 3.4.6 used in the proof of Theorem 3.2.5;

2. entropy results in Lemmas 3.4.14, 3.4.18 and 3.4.19;

3. stochastic boundedness for shape-restricted LSEs in Lemmas 3.4.10 and 3.4.20.

3.5.1 Proof of Proposition 3.4.6

In the next few subsections, we will prove Proposition 3.4.6 step by step.

Construction of $\tilde{\mathcal{F}}$

First consider the case $\gamma \in (0,1)$. We will do the construction iteratively. For $l = 1$, since $[0,1]$ contains $\left\lfloor 2^{\frac{1}{1-\gamma}} \right\rfloor$ many equal-length intervals (with length $\left(2^{\frac{1}{1-\gamma}}\right)^{-1}$), we can pick 2 intervals among them; this is denoted $\tilde{\mathscr{C}}_1$. For $l = 2$, each interval in $\tilde{\mathscr{C}}_1$ contains $\left\lfloor 2^{\frac{1}{1-\gamma}} \right\rfloor$ many equal-length subintervals with length $\left(2^{\frac{1}{1-\gamma}}\right)^{-2}$, we can pick 2 subintervals among each of the interval; this is denoted $\tilde{\mathscr{C}}_2$. In this way we can define iteratively $\tilde{\mathscr{C}}_l$ for any $l \in \mathbb{N}$. Let $\tilde{\mathcal{F}}_l \equiv \{\mathbf{1}_I : I \in \tilde{\mathscr{C}}_l\}$. Clearly $|\tilde{\mathcal{F}}_l| = 2^l$ and contains indicators over intervals in $[0,1]$ with

length $\left(2^{\frac{1}{1-\gamma}}\right)^{-l}$. Now let $\tilde{\mathcal{F}} \equiv \cup_{l\in\mathbb{N}}\tilde{\mathcal{F}}_l \cup \{\mathbf{0}\}$ where $\mathbf{0}$ denotes a mapping taking identical value 0. Next, for $\gamma = 1$, let $\tilde{\mathcal{F}} \equiv \{\mathbf{1}_{[0,\delta]} : 0 \le \delta \le 1\}$.

We show that the constructed $\tilde{\mathcal{F}}$ satisfies the desired growth condition (3.2.1). Recall P is the uniform distribution on $[0, 1]$.

Lemma 3.5.1. *It holds that*

$$\|\tilde{F}(\delta)\|_{L_2(P)} \le \sqrt{2}\delta^{\gamma},$$

where $\tilde{F}(\delta)$ denotes the envelope for $\tilde{\mathcal{F}}(\delta)$.

Proof. The claim is trivial for $\gamma = 1$. For $\gamma \in (0, 1)$, since each element in $\tilde{\mathscr{C}}_{l+1}$ is contained in some element in $\tilde{\mathscr{C}}_l$, we only need to count the number of intervals for the smallest level $l(\delta)$ such that the length of intervals in $\tilde{\mathcal{F}}_{l(\delta)}$ is no more than δ^2. In other words, $l(\delta)$ is the integer for which

$$\left(2^{\frac{1}{1-\gamma}}\right)^{-l(\delta)} \le \delta^2, \quad \left(2^{\frac{1}{1-\gamma}}\right)^{-l(\delta)+1} > \delta^2.$$

Hence the number of intervals in $\tilde{\mathcal{F}}_{l(\delta)}$ is $N(\delta) = 2^{l(\delta)} \in [\delta^{-(2-2\gamma)}, 2\delta^{-(2-2\gamma)}]$, from which the claim of the lemma holds. □

Proof of claim (1) of Proposition 3.4.6

The following standard Paley-Zygmund lower bound will be used.

Lemma 3.5.2 (Paley-Zygmund)**.** *Let Z be any non-negative random variable. Then for any $\varepsilon > 0$, $\mathbb{P}(Z > \varepsilon\mathbb{E}Z) \ge \left(\frac{(1-\varepsilon)\mathbb{E}Z}{(\mathbb{E}Z^q)^{1/q}}\right)^{q'}$, where $q, q' \in (1, \infty)$ are conjugate indices: $1/q + 1/q' = 1$.*

We need the following exact characterization concerning the size of maxima of a sequence of independent random variables due to [63], see also Corollary 1.4.2 of [41].

Lemma 3.5.3. *Let $\xi_1, \ldots, \xi_n$ be a sequence of independent non-negative random variables such that $\|\xi_i\|_r < \infty$ for all $1 \le i \le n$. For $\lambda > 0$, set $\delta_0(\lambda) \equiv \inf\{t > 0 : \sum_{i=1}^n \mathbb{P}(\xi_i > t) \le \lambda\}$. Then*

$$\frac{1}{1+\lambda}\sum_{i=1}^n \mathbb{E}\xi_i^r \mathbf{1}_{\xi_i > \delta_0} \le \mathbb{E}\max_{1\le i\le n}\xi_i^r \le \frac{1}{1\wedge\lambda}\sum_{i=1}^n \mathbb{E}\xi_i^r \mathbf{1}_{\xi_i > \delta_0}.$$

Proof of Proposition 3.4.6, claim (1). **(Case 1: $0 < \gamma < 1$).** Recall $\delta_2 \equiv \vartheta n^{-\frac{1}{2(2-\gamma)}}$. Then by the proof of Lemma 3.5.1, we see that there exists some level $l(\delta_2) \in \mathbb{N}$ such that the $N(\delta_2)$ many intervals $\{I_l\}_{l=1}^{N(\delta_2)}$ in $\tilde{\mathcal{F}}_{l(\delta_2)}$ have length at most δ_2^2 and at least $2^{-1/(1-\gamma)}\delta_2^2$, while the number of intervals satisfies $\vartheta^{-(2-2\gamma)} n^{\frac{1-\gamma}{2-\gamma}} \le N(\delta_2) \le 2\vartheta^{-(2-2\gamma)} n^{\frac{1-\gamma}{2-\gamma}}$. Let $\mathcal{E}_n$ be the event that all intervals $\{I_l\}_{l=1}^{N(\delta_2)}$ contain at least $2^{-\frac{2-\gamma}{1-\gamma}}\vartheta^2 n^{\frac{1-\gamma}{2-\gamma}}$ of the X_i's and at most $\frac{5}{4}\vartheta^2 n^{\frac{1-\gamma}{2-\gamma}}$ of the X_i's. Then by a union bound and Bernstein's inequality (cf. (2.10) of [22]),

$$\begin{aligned} \mathbb{P}(\mathcal{E}_n^c) &\le \mathbb{P}\bigg(\max_{1\le l\le N(\delta_2)} \bigg| \sum_{i=1}^n \mathbf{1}_{I_l}(X_i) - n|I_l| \bigg| > 2^{-\frac{2-\gamma}{1-\gamma}}\vartheta^2 n^{\frac{1-\gamma}{2-\gamma}} \bigg) \\ &\le 2\vartheta^{-(2-2\gamma)} n^{\frac{1-\gamma}{2-\gamma}} \exp\big(- c_\gamma \vartheta^2 n^{\frac{1-\gamma}{2-\gamma}} \big). \end{aligned} \tag{3.5.1}$$

Let $\mathcal{I}_l \equiv \{X_i \in I_l\}$ for $1 \le l \le N(\delta_2)$ and $\{\xi_i^{(l)}\}_{i,l\ge 1}$ be i.i.d. random variables with the same law as ξ_1. Then for some $t_n > 0$ to be determined later,

$$\begin{aligned} \mathbb{P}\bigg(\sup_{f\in\tilde{\mathcal{F}}: Pf^2 \le \delta_2^2} \bigg| \sum_{i=1}^n \xi_i f(X_i) \bigg| \ge t_n \bigg) &\ge \mathbb{E}_{\boldsymbol{X}} \bigg[\mathbb{P}_{\boldsymbol{\xi}} \bigg(\max_{1\le l\le N(\delta_2)} \bigg| \sum_{i=1}^n \xi_i \mathbf{1}_{I_l}(X_i) \bigg| \ge t_n \bigg) \mathbf{1}_{\mathcal{E}_n} \bigg] \\ &= \mathbb{E}_{\boldsymbol{X}} \bigg[\mathbb{P}_{\boldsymbol{\xi}} \bigg(\max_{1\le l\le N(\delta_2)} \bigg| \sum_{i=1}^{|\mathcal{I}_l|} \xi_i^{(l)} \bigg| \ge t_n \bigg) \mathbf{1}_{\mathcal{E}_n} \bigg]. \end{aligned} \tag{3.5.2}$$

Our goal now is to make a good choice of the law for $\xi^{(\cdot)}$'s so that we may obtain a good estimate for t_n and thereby using the Paley-Zygmund argument. Let ξ_1 be distributed according to the symmetric $\alpha_\varepsilon \equiv 2 - \varepsilon$ stable law, i.e. the characteristic function of ξ_1 is $\varphi_{\xi_1}(t) = \exp(-|t|^{\alpha_\varepsilon})$. Apparently, $k^{-1/\alpha_\varepsilon}\sum_{i=1}^k \xi_i^{(l)}$ has the same law as that of ξ_1, and hence we can take

$$t_n = \frac{1}{2}\bigg(2^{-\frac{2-\gamma}{1-\gamma}}\vartheta^2 n^{\frac{1-\gamma}{2-\gamma}} \bigg)^{1/\alpha_\varepsilon} \mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|. \tag{3.5.3}$$

Then the conditional probability in the last line of (3.5.2) can be bounded from below by

$$\mathbb{P}_{\boldsymbol{\xi}} \bigg(\max_{1\le l\le N(\delta_2)} |\xi_l| \ge \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l| \bigg) \ge \bigg(\frac{\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|}{2\left(\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|^r \right)^{1/r}} \bigg)^{r'} \tag{3.5.4}$$

for some conjugate indices $(r, r') \in (1,\infty)^2$. (3.5.2) and (3.5.4) suggest that we need to derive a lower bound for $\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|$ and an upper bound for $\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|^r$. This can

be done via the help of Lemma 3.5.3: since $\mathbb{P}(|\xi_1| > t) \asymp \frac{C_\varepsilon}{1+t^{\alpha_\varepsilon}}$ (cf. Property 1.2.15, page 16 of [132]), we can choose $\lambda \equiv 1$ and $\delta_0 \asymp_\varepsilon N(\delta_2)^{1/\alpha_\varepsilon}$ to see that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|^r &\asymp \sum_{l=1}^{N(\delta_2)} \mathbb{E}|\xi_l|^r \mathbf{1}_{\xi_l>\delta_0} \\ &= N(\delta_2)\bigg(\mathbb{P}\left(|\xi_1|>\delta_0\right)\int_0^{\delta_0} ru^{r-1}\,\mathrm{d}u \\ &\qquad + \int_{\delta_0}^{\infty} ru^{r-1}\mathbb{P}\left(|\xi_1|>u\right)\,\mathrm{d}u\bigg) \\ &\asymp_{\varepsilon,r} N(\delta_2)^{r/\alpha_\varepsilon}.\end{aligned}$$

Now as long as $\varepsilon < 1/2$, we may choose $r > 1$ close enough to 1, e.g. $r = 1.1$, to conclude that there exists $\mathfrak{p}_1 \in (0, 1/8)$ that only depends on ε such that

$$\text{Left hand side of (3.5.4)} \ge 8\mathfrak{p}_1. \tag{3.5.5}$$

Combining (3.5.1), (3.5.2) and (3.5.5), and the fact that $t_n = c_1\big(\vartheta^\gamma n^{\frac{1-\gamma}{2-\gamma}}\big)^{2/\alpha_\varepsilon}$ for some constant c_1 depending on ε, γ only, we have that for n large enough depending on ϑ, γ,

$$\mathbb{P}\bigg(\sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_2^2}\bigg|\sum_{i=1}^n \xi_i f(X_i)\bigg| \ge c_1\big(\vartheta^\gamma n^{\frac{1-\gamma}{2-\gamma}}\big)^{2/\alpha_\varepsilon}\bigg) \ge 4\mathfrak{p}_1. \tag{3.5.6}$$

On the other hand, by Talagrand's concentration inequality (cf. Lemma 3.4.4) and the contraction principle for Rademacher processes, we have with probability at least $1 - 2\mathfrak{p}_1$,

$$\begin{aligned}\sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_2^2} |\mathbb{G}_n(f^2)| &\le C\big(\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_2^2} |\mathbb{G}_n f| + \delta_2\sqrt{\log(1/2\mathfrak{p}_1)} + \log(1/2\mathfrak{p}_1)/\sqrt{n}\big) \\ &\le C_\varepsilon\cdot\delta_2\sqrt{\log(1/\delta_2)} \le C_{\varepsilon,\gamma}\vartheta n^{-\frac{1}{2(2-\gamma)}}\sqrt{\log n}.\end{aligned} \tag{3.5.7}$$

Combining (3.5.6)-(3.5.7), we see that with probability at least $2\mathfrak{p}_1$,

$$\begin{aligned}&\sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_2^2} (\mathbb{P}_n - P)(2\xi f - f^2) \\ &\ge 2c_1\big(\vartheta^\gamma n^{\frac{1-\gamma}{2-\gamma}}\big)^{2/\alpha_\varepsilon}\cdot n^{-1} - C_{\varepsilon,\gamma}\vartheta n^{-\frac{1}{2(2-\gamma)}-\frac{1}{2}}\sqrt{\log n} \\ &\ge 2c_1\vartheta^\gamma n^{-\frac{1}{2-\gamma}}\cdot\tau_n(\varepsilon,\gamma) - C_{\varepsilon,\gamma}\vartheta n^{-\frac{(3-\gamma)/2}{(2-\gamma)}}\sqrt{\log n} \ge c_1\vartheta^\gamma n^{-\frac{1}{2-\gamma}}\cdot\tau_n(\varepsilon,\gamma)\end{aligned}$$

for n large enough depending on $\varepsilon, \vartheta, \gamma$, where $\tau_n(\varepsilon,\gamma) \equiv n^{\frac{1-\gamma}{2-\gamma}\cdot\frac{\varepsilon}{2-\varepsilon}}$. Hence with the same probability estimate,

$$\begin{aligned} F_n(\delta_2) &= \sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_2^2} (\mathbb{P}_n - P)(2\xi f - f^2) - \delta_2^2 \\ &\ge c_1\vartheta^\gamma n^{-\frac{1}{2-\gamma}}\tau_n(\varepsilon,\gamma) - \vartheta^2 n^{-\frac{1}{2-\gamma}} \ge \frac{1}{2}c_1\vartheta^\gamma n^{-\frac{1}{2-\gamma}}\tau_n(\varepsilon,\gamma) \end{aligned}$$

holds for n large enough depending on $\varepsilon, \vartheta, \gamma$, completing the proof for the claim for $0 < \gamma < 1$.

(Case 2: $\gamma = 1$). Recall $\delta_2 = \vartheta n^{-1/2}$, and there exists one interval I with length δ_2^2. It is easy to see that $\mathbb{P}\big(|\sum_{i=1}^n \mathbf{1}_I(X_i) - \vartheta^2| > \vartheta^2/2\big) \le 2\exp(-\vartheta^2/10)$. For $\vartheta \ge 4$, we see that with probability at least 0.5, there are $\mathcal{O}(1)$ points $X_i \in I$. Denote this event $\mathcal{E}_1$. Let

$$Z_n \equiv \sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_2^2} \left(2\sum_{i=1}^n \xi_i f(X_i) - n\left(\mathbb{P}_n - P\right)(f^2)\right).$$

Note we can use the absolute value in the suprema in the above display. Since $\mathbb{E}|(\mathbb{P}_n - P)(\mathbf{1}_I)|^2 \le \vartheta^2 n^{-2}$, we see that on an event with probability at least 0.96, $|n(\mathbb{P}_n - P)(\mathbf{1}_I)| \le 25\vartheta$. Denote this event by $\mathcal{E}_2$. Then for any ξ such that $\mathbb{E}|\xi| \ge 25\vartheta$, let $t = \mathbb{E}|\xi| - 25\vartheta$, and $N_I \equiv \sum_{i=1}^n \mathbf{1}_I(X_i)$,

$$\begin{aligned} \mathbb{P}\left(Z_n \ge t\right) &\ge \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{P}_{\boldsymbol{\xi}}\left(\left|2\sum_{i=1}^n \xi_i \mathbf{1}_I(X_i) - n\left(\mathbb{P}_n - P\right)(\mathbf{1}_I)\right| \ge t\right)\mathbf{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] \\ &\ge \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{P}_{\boldsymbol{\xi}}\left(\left|\sum_{i=1}^{N_I}\xi_i\right| > (t+25\vartheta)/2\right)\mathbf{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] \\ &\ge \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{P}_{\boldsymbol{\xi}}\left(\left|\sum_{i=1}^{N_I}\xi_i\right| > \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left|\sum_{i=1}^{N_I}\xi_i\right|\right)\mathbf{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] \end{aligned}$$

where in the last inequality we used Jensen's inequality. Let η be a symmetric random variable given by $\mathbb{P}(|\eta| > t) = 1/(1+t^2)$, then it is easy to calculate that $\mathbb{E}|\eta| = \pi/2$, and $\mathbb{E}|\eta|^r \equiv c_r < \infty$ for $r < 2$. Let $\xi \equiv 50\vartheta^2 \cdot \eta$. Then $\mathbb{E}|\xi| = 25\pi\vartheta^2 > 25\vartheta$, and hence choosing $r > 1$ close enough to 1 in the Paley-Zygmund Lemma 3.5.2 yields that

$$\mathbb{P}\left(Z_n \ge 25\pi\vartheta^2 - 25\vartheta\right) \ge 2\mathfrak{p}_2$$

for some constant $\mathfrak{p}_2 > 0$ depending only on ϑ (through the estimate on N_I on the event $\mathcal{E}_1$). Hence with probability at least $2\mathfrak{p}_2$,

$$F_n(\delta_2) \geq (25\pi\vartheta^2 - 25\vartheta)n^{-1} - \vartheta^2 n^{-1} \geq 285\vartheta n^{-1}.$$

This completes the proof. □

Proof of claim (2) of Proposition 3.4.6

Proof of Proposition 3.4.6, claim (2). Recall $\delta_1 \equiv \rho n^{-\frac{1}{2(2-\gamma)}-\beta_\varepsilon}$. Note that by Koltchinskii-Pollard maximal inequality for empirical processes (cf. Theorem 2.14.1 of [162]), we have

$$\max_{1\leq k\leq n} \mathbb{E} \sup_{f\in\tilde{\mathcal{F}}: Pf^2\leq\delta_1^2} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \varepsilon_i f(X_i) \right| \lesssim \|\tilde{F}(\delta_1)\|_{L_2(P)} \leq C_1\delta_1^\gamma.$$

Hence we may take $\psi_n(k) \equiv C_1 k^{1/(2-2\varepsilon)}\delta_1^\gamma$ in the multiplier inequality Lemma 4.5.6 to see that

$$\begin{aligned} \mathbb{E} \sup_{f\in\tilde{\mathcal{F}}: Pf^2\leq\delta_1^2} \left| \sum_{i=1}^{n} \xi_i f(X_i) \right| &\leq 4\int_0^\infty \psi_n\big(n\mathbb{P}(|\xi_1| > t)\big)\, \mathrm{d}t \\ &\leq 4C_1\delta_1^\gamma n^{1/2(1-\varepsilon)} \|\xi_1\|_{2(1-\varepsilon),1}. \end{aligned}$$

On the other hand, again by the Koltchinskii-Pollard maximal inequality and the contraction principle for Rademacher processes,

$$\mathbb{E} \sup_{f\in\tilde{\mathcal{F}}: Pf^2\leq\delta_1^2} |\mathbb{G}_n(f^2)| \lesssim \mathbb{E} \sup_{f\in\tilde{\mathcal{F}}: Pf^2\leq\delta_1^2} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \lesssim \delta_1^\gamma.$$

Combining the above estimates, we arrive at

$$\begin{aligned} \mathbb{E}E_n(\delta_1) &\leq 8C_1\delta_1^\gamma n^{-1} \cdot n^{1/2(1-\varepsilon)} \|\xi_1\|_{2(1-\varepsilon),1} + C_2 n^{-1/2}\delta_1^\gamma \\ &\leq C_{\varepsilon,\xi}\rho^\gamma n^{-\frac{1}{2-\gamma}}\omega_n(\varepsilon,\gamma). \end{aligned}$$

The claim (2) of Proposition 3.4.6 now follows from Markov's inequality and hence the proof of Theorem 3.2.5 is complete. □

3.5.2 Proof of entropy results

Proof of Lemma 3.4.14. Let $t_j \equiv (1+\varepsilon)^{-j}$ and $m(\varepsilon)$ be the smallest integer j such that $t_j \leq \varepsilon \|F\|_{L_2(Q)}$. Now for any $f \in \mathcal{F}$, define

$$f_\varepsilon \equiv \sum_{j=1}^{m(\varepsilon)} \big(t_j \mathbf{1}_{t_j < f \leq t_{j-1}} + (-t_{j-1}) \mathbf{1}_{-t_{j-1} < f \leq -t_j} \big).$$

Then if $x \in \mathcal{X}$ is such that

1. $t_j < f(x) \leq t_{j-1}$ for some $j \leq m(\varepsilon)$,

$$0 \leq f(x) - f_\varepsilon(x) \leq t_{j-1} - t_j \leq \varepsilon t_j \leq \varepsilon f(x) \leq \varepsilon F(x).$$

2. $-t_{j-1} < f(x) \leq -t_j$ for some $j \leq m(\varepsilon)$,

$$0 \leq f(x) - f_\varepsilon(x) \leq -t_j - (-t_{j-1}) \leq \varepsilon t_j \leq \varepsilon \big(-f(x) \big) \leq \varepsilon F(x).$$

3. $-t_{m(\varepsilon)} < f(x) \leq t_{m(\varepsilon)}$,

$$|f(x) - f_\varepsilon(x)| \leq t_{m(\varepsilon)} \leq \varepsilon \|F\|_{L_2(Q)}.$$

Combining the above discussion we arrive at $\|f - f_\varepsilon\|^2_{L_2(Q)} \leq 3\varepsilon^2 \|F\|^2_{L_2(Q)}$. Let $\mathcal{F}_\varepsilon \equiv \{f_\varepsilon : f \in \mathcal{F}\}$. Then since the sets

$$\begin{aligned} \big\{(x,t) : f_\varepsilon(x) \geq t\big\} = \cup_{j=1}^{m(\varepsilon)} \big\{x : f(x) \geq t_j\big\} \times (t_j, t_{j-1}] \\ \bigcup \cup_{j=1}^{m(\varepsilon)} \big\{x : f(x) \geq -t_{j-1}\big\} \times (-t_{j-1}, -t_j] \\ \bigcup \big\{x : f(x) \geq -t_{m(\varepsilon)}\big\} \times (-t_{m(\varepsilon)}, t_{m(\varepsilon)}] \end{aligned}$$

as f_ε ranges over $\mathcal{F}_\varepsilon$ is the union of at most $2m(\varepsilon)+1$ VC-classes with disjoint supports, and hence the VC-dimension of $\mathcal{F}_\varepsilon$ is no larger than $Vm(\varepsilon)$, where $V \in (0, \infty)$ only depends on $\mathcal{F}_0$. The rest of the proof proceeds along the same lines as in page 1172 of [59]. □

Proof of Lemma 3.4.18. Let $\{h_i\}_{i=1}^N$ be a minimal ε-covering set of $\mathcal{H}$ under L_∞. For any probability measure Q on $[0,1]^2$, and any $f \otimes (h-g) \in f \otimes \big(\mathcal{H} - g\big)$, take h_i such that $\|h - h_i\|_\infty \le \varepsilon$. Then

$$\|f \otimes (h_i - g) - f \otimes (h-g)\|_{L_2(Q)}^2 \le \|f\|_{L_2(Q)}^2 \varepsilon^2 = \|f \otimes 1\|_{L_2(Q)}^2 \varepsilon^2.$$

completing the proof. □

Proof of Lemma 3.4.19. Since $\mathcal{F}' \subset \mathcal{F} \cap L_\infty(1)$ is VC-major, Lemma 3.4.14 yields that for any probability measure Q_x on $[0,1]$ and any $\varepsilon > 0$,

$$\log \mathcal{N}\big(\varepsilon \|F'\|_{L_2(Q_x)}, \mathcal{F}', L_2(Q_x)\big) \le \frac{C}{\varepsilon} \log\left(\frac{C}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon \|F'\|_{L_2(Q_x)}}\right).$$

Now for any discrete probability measure $Q = n^{-1}\sum_{i=1}^n \delta_{(x_i,z_i)}$ on $[0,1]^2$, let $Q_x \equiv n^{-1}\sum_{i=1}^n \delta_{x_i}$ be the (marginal) probability measure on $[0,1]$. Take a minimal $\varepsilon\|F'\|_{L_2(Q_x)}$-cover of $\mathcal{F}'$ under $L_2(Q_x)$, namely $\{f_k\}$, the log-cardinality of which is no more than

$$\frac{C}{\varepsilon} \log\left(\frac{C}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon \|F'\|_{L_2(Q_x)}}\right).$$

Further take a minimal ε-cover of $\mathcal{H}$ under L_∞, namely $\{h_l\}$, the log-cardinality of which is at most a constant multiple of $\varepsilon^{-\gamma}$. Consider the set $\{f_k \otimes h_l\}$, the log-cardinality of which is at most a constant multiple of

$$\frac{1}{\varepsilon} \log\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon \|F' \otimes 1\|_{L_2(Q)}}\right) \vee \varepsilon^{-\gamma}.$$

For every $f \otimes (h - (\phi_0 - f_0)) \in \mathcal{F}' \otimes (\mathcal{H} - (\phi_0 - f_0))$, let $\tilde{f}_k, \tilde{h}_l$ be such that $\|f - \tilde{f}_k\|_{L_2(Q_x)} \le \varepsilon \|F'\|_{L_2(Q_x)}$ and $\|h - \tilde{h}_l\|_\infty \le \varepsilon$. Then

$$\begin{aligned}
&\|f \otimes (h - (\phi_0 - f_0)) - \tilde{f}_k \otimes (\tilde{h}_l - (\phi_0 - f_0))\|_{L_2(Q)}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \Big(f(x_i)\big(h(z_i) - (\phi_0(x_i,z_i) - f_0(x_i))\big) \\
&\qquad\qquad - \tilde{f}_k(x_i)\big(\tilde{h}_l(z_i) - (\phi_0(x_i,z_i) - f_0(x_i))\big)\Big)^2 \\
&\lesssim \|h - \tilde{h}_l\|_\infty^2 \|F'\|_{L_2(Q_x)}^2 + \|\phi_0\|_\infty^2 \|f - \tilde{f}_k\|_{L_2(Q_x)}^2 \\
&\lesssim (1 \vee \|\phi_0\|_\infty^2)\varepsilon^2 \|F'\|_{L_2(Q_x)}^2 = (1 \vee \|\phi_0\|_\infty^2)\varepsilon^2 \|F' \otimes 1\|_{L_2(Q)}^2,
\end{aligned}$$

as desired. □

3.5.3 Proof of stochastic boundedness of shape-restricted LSEs

Proof of Lemma 3.4.10, isotonic case. The isotonic least squares estimator $\hat{f}_n$ has a well-known min-max representation [129]:

$$\hat{f}_n(X_j) = \min_{v \geq j} \max_{u \leq j} \frac{1}{v-u+1} \sum_{i=u}^{v} Y_i \tag{3.5.8}$$

where we slightly abuse the notation X_i's so that $X_1 \leq \cdots \leq X_n$ denote the ordered covariates and Y_i denotes the corresponding observed response at X_i. Since $\hat{f}_n$ is non-decreasing, we only need to consider

$$\alpha_n \equiv \hat{f}_n(X_1) = \min_{v \geq 1} \frac{1}{v} \sum_{i=1}^{v} Y_i, \quad \beta_n \equiv \hat{f}_n(X_n) = \max_{u \leq n} \frac{1}{n-u+1} \sum_{i=u}^{n} Y_i.$$

Note that

$$\mathbb{E}|\alpha_n| \vee \mathbb{E}|\beta_n| \leq \mathbb{E} \max_{k \leq n} \left| \frac{1}{k} \sum_{i=1}^{k} \xi_i \right| + \|f_0\|_\infty.$$

The first term is $\mathcal{O}(1)$ by a simple blocking argument and a Lévy-type maximal inequality due to Montgomery-Smith [118] (see also Theorem 1.1.5 of [41]); we include some details for the convenience of the reader: suppose without loss of generality that $\log_2 n$ is an integer, then for any $t \geq 1$,

$$\begin{aligned}
\mathbb{P}\left(\max_{1 \leq k \leq n} \left| \frac{1}{k} \sum_{i=1}^{k} \xi_i \right| > t \right) &\leq \sum_{j=1}^{\log_2 n} \mathbb{P}\left(\max_{2^{j-1} \leq k < 2^j} \left| \frac{1}{k} \sum_{i=1}^{k} \xi_i \right| > t \right) + \mathbb{P}\left(\left| \sum_{i=1}^{n} \xi_i \right| > nt \right) \\
&\leq \sum_{j=1}^{\log_2 n} \mathbb{P}\left(\max_{2^{j-1} \leq k < 2^j} \left| \sum_{i=1}^{k} \xi_i \right| > 2^{j-1} t \right) + \frac{\|\xi_1\|_2^2}{nt^2} \\
&\leq 9 \sum_{j=1}^{\log_2 n} \mathbb{P}\left(\left| \sum_{i=1}^{2^j} \xi_i \right| > 2^{j-1} t/30 \right) + \frac{\|\xi_1\|_2^2}{nt^2} \\
&\leq C \|\xi_1\|_2^2 \left(\sum_{j=1}^{\log_2 n} \frac{1}{2^j t^2} + \frac{1}{nt^2} \right) \leq C' \|\xi_1\|_2^2 t^{-2},
\end{aligned}$$

completing the proof. □

The proof of stochastic boundedness of the convex least squares estimator crucially uses the characterization developed in Lemma 2.6 of [70]. Note that the characterization is purely deterministic.

Lemma 3.5.4. *$\hat{f}_n$ is a convex least squares estimator if and only if for all $j = 2, \ldots, n$,*

$$\sum_{k=1}^{j-1} R_k(X_{k+1} - X_k) \geq \sum_{k=1}^{j-1} S_k(X_{k+1} - X_k),$$

with inequality holds if and only if $\hat{f}_n$ has a kink at X_j. Here $R_k = \sum_{i=1}^{k} \hat{f}_n(X_i)$ and $S_k = \sum_{i=1}^{k} Y_i$, where we abuse the notation X_i's for the ordered covariates such that $X_1 \leq \ldots \leq X_n$, and Y_i's are the corresponding observed responses at X_i.

Proof of Lemma 3.4.10, convex case. By symmetry we only consider the behavior of $\hat{f}_n(0)$. Let τ_n denote the first kink of $\hat{f}_n$ away from 0. Then it follows from the characterization Lemma 3.5.4 that

$$\sum_{k=1}^{\tau_n - 2} R_k(X_{k+1} - X_k) \geq \sum_{k=1}^{\tau_n - 2} S_k(X_{k+1} - X_k),$$
$$\sum_{k=1}^{\tau_n - 1} R_k(X_{k+1} - X_k) = \sum_{k=1}^{\tau_n - 1} S_k(X_{k+1} - X_k).$$

The above two (in)equalities necessarily entail that

$$R_{\tau_n - 1}(X_{\tau_n} - X_{\tau_n - 1}) \leq S_{\tau_n - 1}(X_{\tau_n} - X_{\tau_n - 1}).$$

Hence with probability 1 we have $R_{\tau_n - 1} \leq S_{\tau_n - 1}$, i.e.

$$\sum_{i=1}^{\tau_n - 1} \hat{f}_n(X_i) \leq \sum_{i=1}^{\tau_n - 1} Y_i. \tag{3.5.9}$$

Since $\hat{f}_n$ is linear on $[0, X_{\tau_n}]$, we can write

$$\hat{f}_n(x) = \left(1 - \frac{x}{X_{\tau_n}}\right) \hat{f}_n(0) + \frac{x}{X_{\tau_n}} \hat{f}_n(X_{\tau_n}). \tag{3.5.10}$$

Combining (3.5.9) and (3.5.10) we see that

$$\Big[\sum_{i=1}^{\tau_n-1}\Big(1-\frac{X_i}{X_{\tau_n}}\Big)\Big]\hat{f}_n(0)+\Big[\sum_{i=1}^{\tau_n-1}\frac{X_i}{X_{\tau_n}}\Big]\hat{f}_n(X_{\tau_n})\le\sum_{i=1}^{\tau_n-1}Y_i,$$

and hence

$$\hat{f}_n(0)\le\Big(\frac{1}{1-\beta_{\tau_n}}\Big)\cdot\frac{\sum_{i=1}^{\tau_n-1}Y_i}{\tau_n-1}+\frac{\beta_{\tau_n}}{1-\beta_{\tau_n}}\big|\inf_{x\in[0,1]}\hat{f}_n(x)\big|, \tag{3.5.11}$$

where

$$\beta_k=\Big(\frac{1}{k-1}\sum_{i=1}^{k-1}X_i\Big)\cdot\frac{1}{X_k}.$$

By (3.5.11), we need to handle three terms:

(i) $(1-\beta_{\tau_n})^{-1}$,

(ii) $\frac{\sum_{i=1}^{\tau_n-1}Y_i}{\tau_n-1}$, and

(iii) $|\inf_{x\in[0,1]}\hat{f}_n(x)|$.

We first handle term (i). We claim that for some universal constant $C>0$, it holds that

$$\mathbb{P}\big(\max_{2\le k\le n}(1-\beta_k)^{-1}\ge t\big)\le Ct^{-1}. \tag{3.5.12}$$

To see this, note that for each $k\le n$, conditional on X_k, $X_1/X_k,\ldots,X_{k-1}/X_k$ are distributed as the order statistics for $k-1$ uniform random variables on $[0,1]$. Let $U_1,\ldots,U_n$ be an i.i.d. sequence of uniformly distributed random variables on $[0,1]$, and $0\le U_{(1)}^n\le\ldots\le U_{(n)}^n\le 1$ be their associated order statistics. Then by using a union bound, the probability in (3.5.12) is bounded by

$$\sum_{k=2}^{n}\mathbb{P}\Big(\frac{1}{k-1}\sum_{i=1}^{k-1}\frac{X_i}{X_k}\ge 1-t^{-1}\Big)\le\sum_{k=1}^{n-1}\mathbb{E}\Big[\mathbb{P}\Big(\frac{1}{k}\sum_{j=1}^{k}U_{(j)}^k\ge 1-t^{-1}\Big)\Big|X_{k+1}\Big].$$

For $t \geq 3$, the probability in the bracket equals $\mathbb{P}\big(\sum_{j=1}^{k} U_j \leq kt^{-1}\big) = \frac{(kt^{-1})^k}{k!}$ by volume computation: $|\{\sum_{j=1}^{k} x_j \leq a\}| = a^k/k!$. Now combining the probability estimates we arrive at

$$\mathbb{P}\big(\max_{2\leq k\leq n}(1-\beta_k)^{-1} \geq t\big) \leq \sum_{k\geq 1} \frac{(kt^{-1})^k}{k!} \leq \sum_{k\geq 1} \frac{(kt^{-1})^k}{(k/e)^k} \leq Ct^{-1},$$

proving the claim (3.5.12) for $t \geq 3$. For $t < 3$, it suffices to increase C.

The second term (ii) can be handled along the same lines as in the proof for the isotonic model, assuming $\|f_0\|_\infty < \infty$ and $\|\xi_1\|_2 < \infty$.

Finally we consider the third term (iii) $|\inf_{x\in[0,1]} \hat{f}_n(x)|$. We claim that with probability 1,

$$\limsup_{n\to\infty} \sup_{x\in[1/4,3/4]} |\hat{f}_n(x)| \leq C_{\xi,f_0}. \tag{3.5.13}$$

The claim will be verified in the proof of Lemma 3.4.20 below in a more general setting. In particular, (3.5.13) implies that $\sup_{x\in[1/4,3/4]}|\hat{f}_n(x)| = \mathcal{O}_{\mathbf{P}}(1)$. Hence for any $\varepsilon > 0$, there exists a constant $K_\varepsilon > 0$ such that for all n large enough, with probability at least $1-\varepsilon$, $\sup_{x\in[1/4,3/4]}|\hat{f}_n(x)| \leq K_\varepsilon$. This event is denoted $\mathcal{E}_\varepsilon$. Now by convexity of $\hat{f}_n$, it follows that $|\inf_{x\in[0,1]} \hat{f}_n(x)| \leq 2K_\varepsilon$ on $\mathcal{E}_\varepsilon$. To see this, we only need to consider the case where the minimum of $\hat{f}_n$ is attained in, say, $[0, 1/4]$: then the line connecting $(1/4, \hat{f}_n(1/4))$ and $(3/4, \hat{f}_n(3/4))$ minorizes $\hat{f}_n$ on $[0, 1/4]$, which is bounded from below by $-2K_\varepsilon$ and hence the same lower bound holds for $\inf_{x\in[0,1]} \hat{f}_n(x)$ on the event $\mathcal{E}_\varepsilon$. An upper bound for $\inf_{x\in[0,1]} \hat{f}_n(x)$ is trivial: $\inf_{x\in[0,1]} \hat{f}_n(x) \leq \sup_{x\in[1/4,3/4]} \hat{f}_n(x) \leq K_\varepsilon$ on $\mathcal{E}_\varepsilon$. These arguments complete the proof for $|\inf_{x\in[0,1]} \hat{f}_n(x)| = \mathcal{O}_{\mathbf{P}}(1)$.

The claim that $\|\hat{f}_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$ follows by combining the discussion of the three terms above and (3.5.11) which proved $|\hat{f}_n(0)| \vee |\hat{f}_n(1)| = \mathcal{O}_{\mathbf{P}}(1)$ and $|\inf_{x\in[0,1]} \hat{f}_n(x)| = \mathcal{O}_{\mathbf{P}}(1)$. □

Proof of Lemma 3.4.20, isotonic case. The proof essentially follows the isotonic case of Lemma 3.4.10 by noting that the least squares estimator $\hat{f}_n$ for $\mathcal{F}$ in the additive model has the fol-

lowing representation:

$$\hat{f}_n(X_j) = \min_{v \geq j} \max_{u \leq j} \frac{1}{v-u+1} \sum_{i=u}^{v} \big(Y_i - \hat{h}_n(Z_i)\big)$$

where $X_1 \leq \cdots \leq X_n$ denote the ordered X_i's, Y_i's are the observed responses at the corresponding X_i's, and Z_i's are the corresponding Z_i's following the ordering of the X_i's. The rest of the proof proceeds along the same lines as in the isotonic case of Lemma 3.4.10 by noting that

$$\begin{aligned} &\max_{1 \leq k \leq n} \sup_{h \in \mathcal{H}} \left| \frac{1}{k} \sum_{i=1}^{k} \big(\phi_0(X_i, Z_i) - h(Z_i)\right| \\ &\leq \|\phi_0\|_\infty + \max_{1 \leq k \leq n} \left(\frac{1}{k} \sum_{i=1}^{k} H(Z_i) \right) = \mathcal{O}_{\mathbf{P}}(1), \end{aligned} \tag{3.5.14}$$

where the stochastic boundedness follows from the same arguments using Lévy-type maximal inequality as in the isotonic case of Lemma 3.4.10, since we have assumed $P_Z H^2 < \infty$. □

Proof of Lemma 3.4.20, convex case. We use the same strategy as the convex case of Lemma 3.4.10 by replacing Y_i with $Y_i - \hat{h}_n(Z_i)$, and handling terms (i), (ii) and (iii) as in the proof of the convex case of Lemma 3.4.10 . Term (i) can be handled using the same arguments as in the proof of the convex case of Lemma 3.4.10 ; term (ii) can be handled similar to (3.5.14). Hence it remains to handle (iii). Let $\hat{\phi}_n(x,z) \equiv \hat{f}_n(x) + \hat{h}_n(z)$. We claim that there exists some $M > 0$ such that

$$\mathbb{P}\left(\inf_{(x,z) \in [1/4,3/4]^2} |\hat{\phi}_n(x,z) - \phi_0(x,z)| > M \text{ i.o.} \right) = 0. \tag{3.5.15}$$

Once (3.5.15) is proved, the event $\mathcal{E} \equiv \cup_{m \geq 1} \cap_{n \geq m} \{\inf_{x \in [1/4,3/4]} |\hat{f}_n(x)| \leq \bar{M}\}$ happens with probability 1, where $\bar{M} \equiv M + \sup_{(x,z) \in [1/4,3/4]^2} H(z) + \|\phi_0\|_\infty < \infty$. Let $x_n \in \operatorname{argmin}_{x \in [1/4,3/4]} \hat{f}_n(x)$ and $M_n \equiv |\hat{f}_n(x_n)|$. On the event $\mathcal{E}$, for all n large enough, there exists $x_n^* \in [1/4, 3/4]$ such that $|\hat{f}_n(x_n^*)| \leq 2\bar{M}$. The key observation is the following: if $M_n > 10\bar{M}$, then

$$\inf_{x \in [1/16,1/8]} \hat{f}_n(x) \vee \inf_{x \in [7/8,15/16]} \hat{f}_n(x) \geq \frac{1}{4}\big(M_n - 10\bar{M}\big). \tag{3.5.16}$$

To see this, we only consider the case $1/4 \le x_n < x_n^* \le 3/4$, and derive a lower bound for $\inf_{x\in[7/8,15/16]} \hat{f}_n(x)$; the other case follows from similar arguments. Note that the line L connecting $(x_n, \hat{f}_n(x_n))$ and $(x_n^*, \hat{f}_n(x_n^*))$ minorizes $\hat{f}_n$ on $[7/8, 15/16]$. Since $M_n > 10\bar{M} > 2\bar{M}$, $\hat{f}_n(x_n) < 0$ and hence the line L has a positive slope s_L bounded below by $(M_n - 2\bar{M})/(3/4 - 1/4) = 2(M_n - 2\bar{M})$. This implies that for any $x \in [7/8, 15/16]$,

$$\begin{aligned}\hat{f}_n(x) \ge \hat{f}_n(7/8) \ge L(7/8) &= L(x_n^*) + s_L(7/8 - x_n^*) \\ &\ge \hat{f}_n(x_n^*) + 2(M_n - 2\bar{M}) \cdot (7/8 - 3/4) \\ &\ge (-2\bar{M}) + \frac{1}{4}(M_n - 2\bar{M}) = \frac{1}{4}\big(M_n - 10\bar{M}\big),\end{aligned}$$

proving (3.5.16). Now we assume without loss of generality that $\inf_{x\in[1/16,1/8]} \hat{f}_n(x) \ge (M_n - 10\bar{M})/4$. Let $I \equiv [1/16, 1/8] \times [0, 1]$. Since

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \hat{\phi}_n(X_i, Z_i)\big)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\big(\xi_i + \phi_0(X_i, Z_i) - \hat{h}_n(Z_i) - \hat{f}_n(X_i)\big)^2 \\ &\ge \frac{1}{n}\sum_{(X_i,Z_i)\in I}\big(\hat{f}_n(X_i) - \big(H(Z_i) + \|\phi_0\|_\infty + |\xi_i|\big)\big)_+^2 \\ &\ge \frac{1}{2n}\sum_{(X_i,Z_i)\in I}\hat{f}_n^2(X_i) - \frac{1}{n}\sum_{(X_i,Z_i)\in I}\big(3H^2(Z_i) + 3\|\phi_0\|_\infty^2 + 3\xi_i^2\big) \\ &\ge \left(\frac{(M_n - 10\bar{M})^2}{32} - 3\|\phi_0\|_\infty^2\right)\frac{|\{i \in [1:n] : (X_i, Z_i) \in I\}|}{n} \\ &\qquad - \frac{3}{n}\sum_{(X_i,Z_i)\in I} H^2(Z_i) - \frac{3}{n}\sum_{(X_i,Z_i)\in I}\xi_i^2.\end{aligned}$$

Hence by the law of large numbers, on an event with probability 1, if $M_n > 10\bar{M}$,

$$\begin{aligned}&\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \hat{\phi}_n(X_i, Z_i)\big)^2 \\ &\qquad \ge \frac{(\limsup_{n\to\infty} M_n - 10\bar{M})^2}{16 \cdot 32} - \frac{3}{16}\big(\|\phi_0\|_\infty^2 + P_Z H^2 + \mathbb{E}\xi_1^2\big).\end{aligned} \tag{3.5.17}$$

On the other hand, since $\hat{\phi}_n$ is the least squares estimator, for any $h' \in \mathcal{H}$,

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \hat{\phi}_n(X_i, Z_i)\big)^2 \tag{3.5.18}$$
$$\leq \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - h'(Z_i)\big)^2 \leq 3\mathbb{E}\xi_1^2 + 3\|\phi_0\|_\infty^2 + 3P_Z H^2.$$

Combining (3.5.17) and (3.5.18), it follows that on an event with probability 1,

$$\limsup_{n\to\infty} M_n \leq C\big(\|\xi_1\|_2 + \|H\|_{L_2(P_Z)} + \|\phi_0\|_\infty + \bar{M}\big),$$

holds for some absolute constant $C > 0$, thus proving that with probability 1,

$$\limsup_{n\to\infty} \big| \inf_{x\in[1/4,3/4]} \hat{f}_n(x)\big| \leq C_{\xi,H,\phi_0,M}.$$

That

$$\limsup_{n\to\infty} \big| \sup_{x\in[1/4,3/4]} \hat{f}_n(x)\big| \leq C'_{\xi,H,\phi_0,M}$$

with probability 1 can be proved in a completely similar manner by noting that the supremum of $\hat{f}_n$ over $[1/4, 3/4]$ is taken either at $1/4$ or $3/4$. These claims show that with probability 1,

$$\limsup_{n\to\infty} \sup_{x\in[1/4,3/4]} |\hat{f}_n(x)| \leq C''_{\xi,H,\phi_0,M}.$$

Note that we have also verified the announced claim (3.5.13) in the convex case of Lemma 3.4.10 by taking $\phi_0(x, z) \equiv f_0(x)$ and $\mathcal{H} \equiv \{0\}$. The rest of proof for handling term (iii) proceeds along the same lines as in the proof of the convex case of Lemma 3.4.10, modulo the unproved claim (3.5.15). Below we prove that (3.5.15) holds for $M > \sqrt{32\big(\|\xi_1\|_2^2 + \|\phi_0\|_\infty^2 + P_Z H^2\big)}$. To this end, first we prove

$$\mathbb{P}\bigg(\mathcal{E}_1 \equiv \big\{ \inf_{(x,z)\in[1/4,3/4]^2} \big(\hat{\phi}_n(x, z) - \phi_0(x, z)\big) > M \text{ i.o.}\big\}\bigg) = 0. \tag{3.5.19}$$

On the event $\mathcal{E}_1$ intersecting a probability-one event, there exists a subsequence $\{n_k\}_{k\geq 1}$ such that

$$\begin{aligned}
&\liminf_{k\to\infty} \frac{1}{n_k} \sum_{i=1}^{n_k} \big(Y_i - \hat{\phi}_{n_k}(X_i, Z_i)\big)^2 \qquad (3.5.20)\\
&\geq \liminf_{k\to\infty} \frac{1}{2n_k} \sum_{(X_i,Z_i)\in[1/4,3/4]^2} \big(\phi_0 - \hat{\phi}_{n_k}\big)^2(X_i, Z_i) - \lim_{k\to\infty} \frac{1}{n_k} \sum_{i=1}^{n_k} \xi_i^2\\
&\geq M^2/8 - \mathbb{E}\xi_1^2,
\end{aligned}$$

and thus by (3.5.18), $M^2 \leq 32\big(\|\xi_1\|_2^2 + \|\phi_0\|_\infty^2 + P_Z H^2\big)$. Hence $\mathcal{E}_1$ must be a probability-zero event, which proves (3.5.19). Using the same arguments we can prove

$$\mathbb{P}\bigg(\sup_{(x,z)\in[1/4,3/4]^2} \big(\hat{\phi}_n(x,z) - \phi_0(x,z)\big) < -M \text{ i.o.} \bigg) = 0. \qquad (3.5.21)$$

The claim (3.5.15) now follows from (3.5.19) and (3.5.21). This completes the proof. □

Chapter 4

LEAST SQUARES ESTIMATION IN NON-DONSKER MODELS I: ESTIMATION INVOLVING SETS

4.1 Introduction

4.1.1 Overview

In this paper we are interested in various behavior related to the empirical process indexed by a class of measurable sets. Such an object has been a classical topic in the history of empirical process theory, when the indexing set is a *Donsker* class, cf. [3, 4, 5]. A particular fruitful complexity measurement of the indexing set is based on the VC-dimension, a notion that can be easily generalized to the setting where the index set is a function class [59].

Somewhat surprisingly, such a simple object has not received much attention when the indexing set is *non-Donsker*. The only works in this regard we are aware of are [47, 1], who obtained global upper bounds in probability for the empirical process.

Out interest in the behavior of the empirical process indexed by non-Donsker class of sets, beyond the obvious theoretical desire to better understand the behavior of a natural object in a complicated setting, lies in the deep connection related to statistical applications discovered by Birgé and Massart, who showed in their seminal work [21] that the convergence rate of the empirical risk minimization procedures(='minimum contrast estimators' coined in [21]) over a non-Donsker class can be strictly *sub-optimal* (i.e. off the optimal rate by a multiplicative polynomial factor of n), while the rate in the Donsker regime is typical *optimal*.

Such a strange phenomenon has a far-reaching influence on the development of 'efficient' estimation procedures in statistical theory. One generic way of getting around the issue is to design 'sieved' estimators [166, 160] that mimics a maximal packing set of the parameter

space which achieves the minimax rate of convergence [169]. The sieving idea [68] has been used frequently in statistical theory to get optimal rates for non-Donsker classes, cf. [99, 71, 26], to name a few.

At the empirical process level, the main technical reason for this to happen is the discrepancy between Sudakov-minorization-type lower bound and the Dudley-entropy-integral-type upper bound. The former, intrinsically connected with a minimax approach to the estimation problem, usually leads to optimal rates in classical statistical experiments(e.g. regression, density estimation, etc.). The latter, on the other hand, only matches with the former when the entropy integral converges, i.e. the parameter space cannot be non-Donsker.

It is now well understood that the convergence rate of (a large class of) ERM procedures can be completely characterized by the behavior of the underlying empirical process [36, 158]. The curious phenomenon observed in [21] naturally suggests that general tools for controlling the empirical process indexed by non-Donsker classes, are unfortunately *generically* sharp from a worst-case perspective. On the other hand, the lack of matching lower bounds for the empirical process indexed by non-Donsker classes precludes determination if the worst case analysis presented in [21] is an exotic phenomenon, or a genuinely unfortunate feature of ERM procedures for 'most' massive parameter spaces. This point of view naturally raises the following question:

Question 4.1.1. *Can we give sharp upper and lower bounds for the empirical process indexed by some natural non-Donsker classes?*

A closely related question that asks for a 'best-case' analysis rather than the worst-case presented in [21] is the following:

Question 4.1.2. *Can we find a natural subset of non-Donsker classes over which global ERM procedures are rate-optimal?*

Our main goal of this paper is to give a solution to Questions 4.1.1 and 4.1.2 in the context of the class of indicators over measurable sets. Although our main interests lie in

non-Donsker classes of sets, we adopt a unified perspective: we show that, if the L_2-size of the indexing set $\mathscr{C}$ is not too small ($\sigma^2 \gtrsim n^{-1/(\alpha+1)}$),

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} |\mathbb{G}_n(C)| \asymp \max\{\sigma^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\}, \tag{4.1.1}$$

where $\mathscr{C}(\sigma) \equiv \{C \in \mathscr{C} : P(C) \leq \sigma^2\}$, and α is an entropy measurement of the complexity of $\mathscr{C}$ in the sense that an (appropriate) entropy of $\mathscr{C}$ scales as $\mathcal{O}(\varepsilon^{-\alpha})$. Here $\mathbb{G}_n(C) \equiv \sqrt{n}(\mathbb{P}_n - P)(C)$ is the empirical process. The unified perspective (4.1.1) helps us to identify an important phase transition phenomenon: the size of the empirical process indexed by Donsker sets is determined by its L_2-size and along with its entropy complexity, while for non-Donsker sets only the complexity of class matters.

It should also be mentioned that both the proofs for the upper and lower bounds in (4.1.1) require novel techniques. The upper bound (for the non-Donsker regime) utilizes a special chaining technique tailored to the geometry of indicators that prevents the chaining steps from being terminated too early, as in the case for a general non-Donsker class of functions. The proof of the lower bound makes crucial use of a sharp multiplier inequality proved in the authors' previous work [76] that allows Gaussianization of the empirical process without inducing additional costs. The lower bound technique extends to empirical processes indexed by general classes of bounded functions without difficulty, which may be of independent interest.

Using the sharp bounds (4.1.1), we further investigate the behavior of various ratio-type empirical processes, complementing the results of [5, 60, 59] to the class of sets satisfying bracketing entropy conditions instead of the VC-type(uniform) entropy conditions, both in the Donsker and the non-Donsker regimes. In particular, we obtain the exact order of the normalizing factor for these ratio-type empirical processes, instead of only an upper bound as in [5, 60, 59]. The local asymptotic moduli (originated in [5]) for the empirical process indexed by class of sets satisfying bracketing entropy conditions is also examined.

As mentioned earlier, the sharp bounds (4.1.1) translate to a rate of convergence for ERM procedures that involve indicators over sets. In particular, we consider global ERM

procedures for the edge estimation problem in both additive and multiplicative regression models (cf. [87, 86]), and the binary classification problem in the learning theory (cf. [42, 100, 150, 108, 84, 89]), where all the estimators are shown to achieve the optimal rate of convergence when the indexing sets are non-Donsker. We thereby answer Question 4.1.2 by identifying a number of natural statistical problems for which the global ERM procedures are indeed rate-optimal for non-Donsker classes, while general empirical process tools fail to obtain sharp rates.

We end the introduction with a quotation from Van de Geer's book [160] page 121-122:

> *... Because there do exist other estimators with a better rate of convergence, the conclusion is that one should not use the maximum likelihood estimator when the entropy integral diverges.*

Our results here show that, the problem of determining whether a global ERM procedure in a non-Donsker problem should be used, is a more subtle issue: the specific property of the indexing function class, beyond its complexity measurement in terms of the metric(bracketing) entropy numbers, must be taken into account.

Although the focus of the current paper is on the classical case for classes of measurable sets, we hope our approach here, along with a case study for the class of multivariate block-increasing functions conducted in [75], can serve as useful starting points for future contributions on the size of the empirical processes indexed by general classes of measurable functions and related ERM procedures in the non-Donsker regime.

4.2 Empirical processes indexed by sets

Let $X_1, \ldots, X_n$ be i.i.d. with distribution P on a sample space $(\mathcal{X}, \mathcal{A})$, and $\mathscr{C}$ a collection of measureable sets contained in $\mathcal{X}$. For any $\sigma > 0$, let $\mathscr{C}(\sigma) \equiv \{C \in \mathscr{C} : P(C) \leq \sigma^2\}$. Following the notation in [48] (page 270, (7.4)), let $\mathcal{N}_I(\varepsilon, \mathscr{C}, P)$ be the ε-bracketing number for $\mathscr{C}$ under P, i.e. the smallest integer m such that there exist $\{C_i \subset D_i\}_{i=1}^m \subset \mathcal{A}$ with the following property: for any $C \in \mathscr{C}$, there exists some $i \in \{1, \ldots, m\}$ such that $C_i \subset C \subset D_i$,

and $P(D_i \setminus C_i) \leq \varepsilon$. $\mathcal{N}(\varepsilon, \mathscr{C}, P)$ will be used for the standard ε-covering number for $\mathscr{C}$ under P.

Assumption B. Fix some $\alpha > 0$ and consider the following entropy conditions:

(E1) $\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \leq L\varepsilon^{-\alpha}$.

(E2) $\log \mathcal{N}(\varepsilon/2, \mathscr{C}(\sqrt{\varepsilon}), P) \geq L^{-1}\varepsilon^{-\alpha}$.

For examples satisfying the above entropy conditions, see [48]. L will be a large enough absolute constant throughout the article, the dependence of which will not be explicitly stated in the theorems.

For $0 < \alpha < 1$, the bracketing condition in (E1) can also be replaced by a uniform entropy condition:

$$\sup_Q \log \mathcal{N}(\varepsilon, \mathscr{C}, Q) \leq L\varepsilon^{-\alpha},$$

where the supremum is taken over all finitely discrete probability measures Q. This case is essentially covered in [59]. Our proof techniques also apply to this case; see remarks after the proofs of Theorems 4.2.1 and 4.2.3.

4.2.1 Upper and lower bounds

We first state the upper bound.

Theorem 4.2.1. *Suppose (E1) holds and* $\sigma^2 \gtrsim n^{-1/(\alpha+1)}$.

1. *Let* $0 < \alpha < 1$. *Then we have*

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} |\mathbb{G}_n(C)| \lesssim_\alpha \sigma^{1-\alpha}.$$

2. *Let* $\alpha > 1$. *Then we have*

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} |\mathbb{G}_n(C)| \leq \mathbb{E} \sup_{C \in \mathscr{C}} |\mathbb{G}_n(C)| \lesssim_\alpha n^{(\alpha-1)/2(\alpha+1)}.$$

Global upper bounds in probability are obtained in [47, 1]. Here we strengthen their results to an expectation bound in the sense that, from a modern perspective the first moment of a bounded empirical process contains *all* probabilistic information in view of the celebrated Talagrand's concentration inequality [145].

The upper bounds derived in [59] take into account the size of the envelope function of the localized function class. Such a consideration is especially suited to the study of VC-type classes, where the entropy bound naturally incorporates the information of the size of the envelope function. We refer the reader to the discussion in [59] page 1171 for more details. Since we are not aware of any specific example of class of sets that has a *bracketing entropy estimate* taking into account the size of the envelope function, we will not address this point hereafter.

We now comment on the proof technique of Theorem 4.2.1, which is based on a specialized chaining technique. The classical chaining argument for Donsker indexing function classes iterates the chaining step infinitely many times(=construct a chain that approximates any given element in the function class with *arbitrary* precision) and then a union bound is used to control uniformly the errors occured in each iteration. This technique, originated in Dudley's entropy integral [46], further developed by [122], leads to an undesirable, yet generically sharp, entropy-integral-type upper bound for non-Donsker classes since the chaining *has to be terminated* before reaching the Poisson domain of the empirical process, which optimally matches with Sudakov-type lower bounds. The special feature of the class of indicator functions allows us to run the chaining argument *exactly* until the empirical process is localized near its Poisson domain, thereby maintaining sharp estimates. The success of this approach lies in the crucial geometric property of the indicator functions: a neighborhood in L_1 metric is the same as the one in *squared* L_2 metric.

Remark 4.2.2. To compare our Theorem 4.2.1 with general empirical process results, let us appeal to the usual local maximal inequality for the empirical process under bracketing

condition: by Lemma 2.14.3 of [162], we have

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)|\lesssim\inf_{\gamma>0}\left\{\sqrt{n}\gamma+\int_\gamma^\sigma\sqrt{\log\mathcal{N}_I(\varepsilon^2,\mathscr{C},P)}\,\mathrm{d}\varepsilon\right\}.$$

For $\sigma^2=c^2n^{-1/(\alpha+1)}$ and the entropy condition of Theorem 4.2.1 with $\alpha>1$, the above bound reduces to

$$\mathbb{E}\sup_{C\in\mathscr{C}(cn^{-1/2(\alpha+1)})}|\mathbb{G}_n(C)|\lesssim\inf_{\gamma>0}\left\{\sqrt{n}\gamma+\gamma^{1-\alpha}\right\}-\sigma^{1-\alpha}\asymp n^{(\alpha-1)/2\alpha}.$$

Similarly we can compute

$$\mathbb{E}\sup_{C\in\mathscr{C}}|\mathbb{G}_n(C)|\lesssim\inf_{\gamma>0}\left\{\sqrt{n}\gamma+\gamma^{1-\alpha}\right\}\asymp n^{(\alpha-1)/2\alpha}.$$

Compared with the bounds obtained in Theorem 4.2.1, we see that current general empirical process tools lead to strictly sub-optimal bounds in the non-Donsker regime.

Our next result asserts matching lower bounds under additional lower bounds on the metric entropy.

Theorem 4.2.3. *Suppose both (E1)-(E2) hold, and* $\sigma^2\gtrsim n^{-1/(\alpha+1)}$.

1. *Let* $0<\alpha<1$. *Then we have*

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)|\gtrsim_\alpha\sigma^{1-\alpha}. \tag{4.2.1}$$

2. *Let* $\alpha>1$. *Then we have*

$$\mathbb{E}\sup_{C\in\mathscr{C}}|\mathbb{G}_n(C)|\geq\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)|\gtrsim_\alpha n^{(\alpha-1)/2(\alpha+1)}. \tag{4.2.2}$$

Remark 4.2.4. Our proof shows that in the Donsker regime (i.e. $0<\alpha<1$), thw lower bound (4.2.1) holds under a weaker entropy condition: we only need to require that

$$\log\mathcal{N}(\varepsilon/4,\mathscr{C}(\sqrt{\varepsilon}),P)\leq L\varepsilon^{-\alpha},\quad\log\mathcal{N}(\varepsilon/2,\mathscr{C}(\sqrt{\varepsilon}),P)\geq L^{-1}\varepsilon^{-\alpha}$$

hold for some large $L>0$.

Compared with upper bounds for empirical processes, much less is understood for lower bounds. [59] provided a lower bound under: (i) an upper uniform entropy condition; (ii) a lower metric entropy condition in the *Donsker* regime. Here our entropy conditions are based on the bracketing entropy that applies to classical set examples (cf. [48]), and the lower bound also applies to the non-Donsker regime. Another set of case-by-case lower bounds are provided in Section 11.1 of [48], where the lower bounds therein holds for all $\mathbb{P}_n$. The strong statement helps to identify that the studied cases therein are non-Donsker at the borderline. This certainly comes at a price: no generality is provided, we must do a case-by-case analysis.

We briefly mention our proof technique for the lower bounds, which is based on Gaussianization of the empirical process, followed by Sudakov minorization, coupled with a sharp multiplier inequality proved in [76] to remove the effect of Gaussian multipliers without incurring the worst-case cost (which is a multiplicative factor of $\mathcal{O}(\sqrt{\log n})$). The Gaussianization trick is classically adopted to prove asymptotic (equicontinuity) statements, cf. [64, 65, 66], while previous lower bound techniques (cf. [59]) are based on the corresponding lower bounds for Rademacher processes (cf. [93]), which inevitably involves uniform entropy conditions.

Our lower bound technique holds more generally for empirical processes indexed by classes of bounded measurable functions. For sake of completeness, we state this result without a formal proof.

Theorem 4.2.5. *Suppose that $\mathcal{F} \subset L_\infty(1)$ is a class of measurable functions, and that the following entropy estimate holds for some $\alpha > 0$:*

$$\log \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_2(P)) \leq L\varepsilon^{-2\alpha}, \quad \log \mathcal{N}(\varepsilon/2, \mathcal{F}(\varepsilon), L_2(P)) \geq L^{-1}\varepsilon^{-2\alpha}. \tag{4.2.3}$$

Here $\mathcal{F}(\varepsilon) \equiv \{f \in \mathcal{F} : Pf^2 \leq \varepsilon^2\}$. Then for $\sigma^2 \gtrsim n^{-1/(\alpha+1)}$, we have:

1. *Let $0 < \alpha < 1$. Then,*

$$\mathbb{E} \sup_{f \in \mathcal{F}(\sigma)} |\mathbb{G}_n(f)| \gtrsim_\alpha \sigma^{1-\alpha}. \tag{4.2.4}$$

2. Let $\alpha > 1$. Then,

$$\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)| \geq \mathbb{E}\sup_{f\in\mathcal{F}(\sigma)}|\mathbb{G}_n(f)| \gtrsim_\alpha n^{(\alpha-1)/2(\alpha+1)}. \tag{4.2.5}$$

The analogy of Remark 4.2.4 applies here as well. It is easy to see from the local maximal inequality for the empirical process (cf. Lemma 4.5.2) that (4.2.4) is sharp. On the other hand, as the counterexample suggested in [21], the lower bound (4.2.5) is *not* sharp for all $\mathcal{F}$.

Combining Theorems 4.2.1 and 4.2.3, we see one important message in these bounds:

1. In the Donsker regime ($0 < \alpha < 1$), the local size of the empirical process

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)| \asymp \sigma^{1-\alpha}$$

   is determined by the L_2-radius σ and the complexity of the class of sets $\mathscr{C}$.

2. In the non-Donsker regime ($\alpha > 1$), the size of the empirical process

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)| \asymp \mathbb{E}\sup_{C\in\mathscr{C}}|\mathbb{G}_n(C)| \asymp n^{(\alpha-1)/2(\alpha+1)}$$

   is determined by the complexity of the set $\mathscr{C}$, *regardless* of the local L_2-radius σ.

To write Theorems 4.2.1 and 4.2.3 in a more compact form, we have the following.

Corollary 4.2.6. *Suppose both (E1)-(E2) hold for some $\alpha \neq 1$. Then for $\sigma^2 \gtrsim n^{-1/(\alpha+1)}$, it holds that*

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)| \asymp \max\{\sigma^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\}.$$

4.2.2 Equivalence of sizes of the empirical process $\mathbb{G}_n$ and the limiting P-bridge process G_P

Let G_P be the standard P-Brownian bridge process. By Donsker theorem, we know that $\mathbb{G}_n \to_d G_P$ in $\ell_\infty(\mathscr{C})$ for a Donsker class of set $\mathscr{C}$, and hence

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)| \to \mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|G_P(C)|$$

as $n \to \infty$ for a *fixed* L_2-radius $\sigma > 0$. Here we are interested in the following question: *Can we assert the 'closeness' between the size of the empirical process* $\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)|$ *and its limiting counterpart* $\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|G_P(C)|$ *in a non-asymptotic setting (in particular,* $\sigma = \sigma_n$ *may change with* n*)?*

The following theorem is a rather immediate consequence of Theorems 4.2.1 and 4.2.3.

Theorem 4.2.7. *Suppose both (E1)-(E2) hold for some* $0 < \alpha < 1$. *Then for* $\sigma^2 \geq cn^{-1/(\alpha+1)}$, *there exists some* $K \equiv K(c,\alpha) > 0$ *such that*

$$K^{-1}\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|G_P(C)| \leq \mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)| \leq K\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|G_P(C)|.$$

To the best knowledge of the authors, Theorem 4.2.7 is the first result in the literature asserting the equivalence (up to multiplicative constants) of the sizes of the empirical process and its limiting Brownian bridge process in a non-asymptotic setting. Although far from being satisfactory, such a result can also be viewed as a first step towards a 'Gaussian characterization' for the size of empirical processes in analogy to Talagrand's Majorizing Measure Theorem [146] for the size of Gaussian processes.

Remark 4.2.8. Some technical remarks in order:

1. Using same proof techniques, Theorem 4.2.7 also holds for empirical processes indexed by a general class of functions: if both entropy estimates in (4.2.3) hold for a uniformly bounded class $\mathcal{F}$, then

   $$\mathbb{E}\sup_{f\in\mathcal{F}(\sigma)}|\mathbb{G}_n(f)| \asymp_{c,\alpha} \mathbb{E}\sup_{f\in\mathcal{F}(\sigma)}|G_P(f)|$$

   holds for $\sigma^2 \geq cn^{-1/(\alpha+1)}$. We remark again that the bracketing entropy condition can be replaced by a uniform entropy condition.

2. The condition $\sigma^2 \gtrsim n^{-1/(\alpha+1)}$ requires the empirical process is in its Gaussian domain where the L_2-radius of $\mathscr{C}$ dominates the size of the empirical process. Such a condition cannot be relaxed in general: if the empirical process is in its Poisson domain, the L_∞

size of $\mathscr{C}$ can be dominating. As an illustrating example, we may take $\mathscr{C} \equiv \{[a,b] : 0 \le a \le b \le 1\}$ and $n = 1$. Then $\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_1(C)| \asymp 1$ for all (small) $\sigma > 0$, while $\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|G_P(C)| \lesssim \sigma\sqrt{\log(1/\sigma)}$.

3. By Talagrand's concentration inequality for empirical processes and Borell-Sudakov-Tsirelson concentration inequality for Gaussian processes, it can be easily deduced that the conclusion of Theorem 4.2.7 also holds in its $L_p (1 \le p < \infty)$ version:

$$\mathbb{E}\left(\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)|\right)^p \asymp_{p,\alpha} \mathbb{E}\left(\sup_{C\in\mathscr{C}(\sigma)}|G_P(C)|\right)^p.$$

4.3 Ratio-type empirical processes indexed by sets

4.3.1 Ratio-type empirical processes

In this subsection, we will be interested in the following prototypical question in the theory of ratio-type empirical process:

Question 4.3.1. *For what sequences of $r_n \to 0$ and $\gamma_n \to \infty$, is the sequence*

$$\gamma_n \sup_{C\in\mathscr{C}: r_n^2 \le P(C) \le 1} \frac{|\mathbb{P}_n(C) - P(C)|}{\sqrt{P(C)}}, \quad n \in \mathbb{N} \tag{4.3.1}$$

bounded away from 0 and ∞, and does the sequence

$$\sup_{C\in\mathscr{C}: r_n^2 \le P(C) \le 1} \left|\frac{\mathbb{P}_n(C)}{P(C)} - 1\right|, \quad n \in \mathbb{N} \tag{4.3.2}$$

converge to 0 almost surely?

Such a question, initiated addressed in [165, 136, 142, 104, 143] for uniform empirical processes on (subsets of) $\mathbb{R}$ (or $\mathbb{R}^d$), is further investigated in [5] for VC classes of sets, and extended by [60, 59] who studied more general VC-subgraph classes. These authors work with classes satisfying uniform entropy conditions, and the class of sets (or functions) need be Donsker apriori. The lack of corresponding results for non-Donsker class of sets are mainly due to the lack of sharp upper and lower bounds for the behavior of empirical process. Here we fill in this gap by using our Theorems 4.2.1 and 4.2.3.

Theorem 4.3.2. *Consider the sequence (4.3.1). Let* $r_n^2 \gtrsim n^{-1/(\alpha+1)}$ *and*

$$\gamma_n = \begin{cases} n^{1/2} r_n^{\alpha}, & \alpha \in (0,1); \\ n^{\frac{1}{2(\alpha+1)}}, & \alpha > 1. \end{cases}$$

Then we have the following:

1. *If (E1) holds,*

$$\limsup_{n\to\infty} \gamma_n \sup_{C\in\mathscr{C}: r_n^2 \le P(C) \le 1} \frac{|\mathbb{P}_n(C) - P(C)|}{\sqrt{P(C)}} < \infty \qquad a.s. \tag{4.3.3}$$

2. *If (E1)-(E2) hold,*

$$\liminf_{n\to\infty} \gamma_n \sup_{C\in\mathscr{C}: r_n^2 \le P(C) \le 1} \frac{|\mathbb{P}_n(C) - P(C)|}{\sqrt{P(C)}} > 0 \qquad a.s. \tag{4.3.4}$$

Theorem 4.3.3. *Consider the sequence (4.3.2). Then there exists some large constant* $K_\alpha > 0$ *such that:*

1. *If (E1) holds and*

$$\liminf_{n\to\infty} r_n^2 \cdot n^{1/(\alpha+1)} \ge \underline{\rho}$$

for some $\underline{\rho} \in (K_\alpha, \infty]$*, then*

$$\limsup_{n\to\infty} \sup_{C\in\mathscr{C}: r_n^2 \le P(C) \le 1} \left|\frac{\mathbb{P}_n(C)}{P(C)} - 1\right| \le \mathcal{O}\left(\underline{\rho}^{-\left(1\wedge\frac{1+\alpha}{2}\right)}\right), \qquad a.s.$$

2. *If (E1)-(E2) hold and*

$$\limsup_{n\to\infty} r_n^2 \cdot n^{1/(\alpha+1)} \le \bar{\rho}$$

for some $\bar{\rho} \in (K_\alpha, \infty]$*, then*

$$\liminf_{n\to\infty} \sup_{C\in\mathscr{C}: r_n^2 \le P(C) \le 1} \left|\frac{\mathbb{P}_n(C)}{P(C)} - 1\right| \ge \mathcal{O}\left(\bar{\rho}^{-\left(1\wedge\frac{1+\alpha}{2}\right)}\right), \qquad a.s.$$

Remark 4.3.4. We make some technical remarks in order.

1. The analogy of Remark 4.2.4 holds for Theorems 4.3.2 and 4.3.3.

2. Theorem 4.3.2 reveals a similar phenomenon as have seen in Theorems 4.2.1 and 4.2.3: the normalizing factor γ_n for the standardized empirical process is determined by the smallest L_2-radius of the element when the indexing class is Donsker, while such a radius plays no role in γ_n for a non-Donsker class of sets.

3. An interesting corollary of Theorem 4.3.3 is that under entropy conditions (E1)-(E2), the sequence in (4.3.2) converges to 0 as $n \to \infty$ almost surely if and only if $r_n^2 \cdot n^{1/(\alpha+1)} \to \infty$.

4. Theorems 4.3.2 and 4.3.3 are also valid in their $L_p (1 \leq p < \infty)$ versions (which can be seen by integrating the tail estimates in the proofs). For instance, if (E1) holds, then

$$\limsup_{n\to\infty} \| \gamma_n \sup_{C\in\mathscr{C}: r_n^2 \leq P(C) \leq 1} \frac{|\mathbb{P}_n(C) - P(C)|}{\sqrt{P(C)}} \|_{L_p(P^{\otimes n})} < \infty,$$

and

$$\limsup_{n\to\infty} \| \sup_{C\in\mathscr{C}: r_n^2 \leq P(C) \leq 1} |\frac{\mathbb{P}_n(C)}{P(C)} - 1| \|_{L_p(P^{\otimes n})} \leq \mathcal{O}\left(\underline{\rho}^{-\left(1\wedge\frac{1+\alpha}{2}\right)}\right).$$

Remark 4.3.5. We may consider more general weighting functions of form $\phi(\sqrt{P(C)})$ as in [59] rather than the special cases $\phi_1(t) = t$ in Theorem 4.3.2 and $\phi_2(t) = t^2$ in Theorem 4.3.3. Here we make these choices mainly due to the fact that ϕ_1, ϕ_2 are of special interest in the history of empirical process theory [165, 136, 142, 104, 143, 5], and the corresponding results for more general cases follow from minor modifications of the proofs.

[59] derived similar theorems (cf. Theorem 4.1 and Theorem 4.6 therein) as our Theorems 4.3.2 and 4.3.3 in terms of the size of a normalized version of the localized empirical processes. In concrete applications, the results of [59] are especially suited for VC classes of sets, with a special focus on recovering the classical results for uniform empirical processes on (subsets of) $\mathbb{R}$ and $\mathbb{R}^d$. Our results here complement the results in [59] by allowing non VC-type classes of sets, both Donsker and non-Donsker.

4.3.2 Local asymptotic moduli

In this subsection we study *local asymptotic moduli* of the empirical process, which has been considered historically for VC-type classes of sets and function classes in [5, 60, 59].

Definition 4.3.6. A *local asymptotic moduli* of the empirical process indexed by a class of sets $\mathscr{C}$ is an increasing function $\omega(\cdot)$ for which there exist some $r_n \ll \delta_n \leq 1/2$ both non-increasing with $\sqrt{n}\delta_n$ non-decreasing, such that

$$\limsup_{n\to\infty} \sup_{r_n^2 < P(C) \leq \delta_n^2} \frac{|\mathbb{G}_n(C)|}{\omega(\sqrt{P(C)})} < \infty \quad \text{a.s.} \tag{4.3.5}$$

Such a local asymptotic moduli is called *sharp* if furthermore

$$\liminf_{n\to\infty} \sup_{r_n^2 < P(C) \leq \delta_n^2} \frac{|\mathbb{G}_n(C)|}{\omega(\sqrt{P(C)})} > 0 \quad \text{a.s.} \tag{4.3.6}$$

An important message of Theorems 4.2.1 and 4.2.3 is that *a local asymptotic moduli of the empirical process does not exist for a non-Donsker indexing class of sets*, and hence we will only consider Donsker classes of sets in the following theorem.

Theorem 4.3.7. *Let* $0 < \alpha < 1$. *If (E1) holds, then* $\omega(t) = t^{1-\alpha}$ *is a local asymptotic moduli, i.e. (4.3.5) holds. If furthermore (E2) holds, then such a moduli is sharp, i.e. (4.3.6) holds.*

We remark that our Theorem 4.3.7 extends directly to uniformly bounded function class with entropy conditions (E1)-(E2) replaced by (4.2.3). The local asymptotic moduli for the empirical process indexed by a class of measurable functions can be defined similarly as Definition 4.3.6, where $\sqrt{P(C)}$ therein is replaced by $\sigma_P(f) \equiv \sqrt{\mathrm{Var}_P(f)} = \sqrt{P(f - Pf)^2}$.

The classical results concerning the local asymptotic moduli focus on VC class of sets (cf. [5]), or more generally, VC-type class of functions (cf. [60, 59]). This roughly corresponds to the case $\alpha \approx 0$ in (E1) with the bracketing entropy condition replaced by a uniform entropy condition. Indeed, it is known that for a VC-subgraph class of functions $\mathcal{F}$, $\omega(t) = t\sqrt{\log\log(1/t) + \log(\|F_t\|_{L_2(P)}/t)}$ is a local asymptotic moduli for $\mathcal{F}$, where F_t is a measurable envelope for $\{f \in \mathcal{F} : Pf^2 \leq t^2\}$. As such, the local asymptotic moduli varies

from $\sim t\sqrt{\log\log(1/t)}$ to $\sim t\sqrt{\log(1/t)}$ for VC-subgraph classes, depending on the size of the localized envelope. As mentioned before, since we are not aware of any example for which the size of the envelope is explicitly adjusted in the bracketing entropy estimates, we do not illustrate such improvements here (in fact, we do not know any such example beyond VC-type classes). We also note that none of [5, 60, 59] examined the sharpness of the local asymptotic moduli in the sense of (4.3.6) in our Definition 4.3.6.

Remark 4.3.8. Similar to (3) of Remark 4.3.4, under respective conditions in Theorem 4.3.7, (4.3.5) and (4.3.6) also hold in their L_p versions.

4.4 Statistical applications

In this section, we apply the sharp bounds in Theorem 4.2.1 and 4.2.3 to several non-Donsker statistical problems including edge estimation problem in additive and multiplicative regression model, and the binary classification problem. The common theme is that global empirical risk minimization procedures in these non-Donsker problems converge at an *optimal* rate rather than a strictly sub-optimal rate obtained in [21].

4.4.1 Least squares estimation: additive errors

Let $X_1, \ldots, X_n$ be i.i.d. samples from law P on a sample space $(\mathcal{X}, \mathcal{A})$. In this subsection we consider the regression model with additive errors:

$$Y_i = \mathbf{1}_{C_0}(X_i) + \xi_i, \quad i = 1, \ldots, n. \tag{4.4.1}$$

This model has been historically considered by [86, 87] and more recently by [26]. We assume for simplicity that the ξ_i's are Gaussian errors with variance 1, and are independent of X_i's. Let $\mathscr{C}$ be a collection of measurable sets in $\mathcal{X}$, and we will fit the regression model by $\{\mathbf{1}_C : C \in \mathscr{C}\}$. Our interest will be the behavior of the least squares estimator $\hat{C}_n$ defined by

$$\hat{C}_n \in \arg\min_{C \in \mathscr{C}} \sum_{i=1}^{n} \left(Y_i - \mathbf{1}_C(X_i)\right)^2. \tag{4.4.2}$$

We assume that $\hat{C}_n$ is well-defined without loss of generality.

Theorem 4.4.1. *Suppose that for some $\alpha \neq 1$, we have the entropy estimate*

$$\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \leq L\varepsilon^{-\alpha}.$$

Then

$$\sup_{C_0 \in \mathscr{C}} \mathbb{E}_{C_0} |\hat{C}_n \Delta C_0| \lesssim n^{-1/(\alpha+1)}.$$

By the seminal work [169], the rate $n^{-1/(\alpha+1)}$ cannot be improved in a minimax sense if furthermore a lower bound on the metric entropy on the same order as that of the upper bound is available.

As a straightforward corollary of the above Theorem 4.4.1, let $\mathscr{C}_d$ be the collection of all convex bodies contained in the unit ball in $\mathbb{R}^d$ and P the uniform distribution on the unit ball.

Corollary 4.4.2. *Fix $d \geq 4$. Then*

$$\sup_{C_0 \in \mathscr{C}_d} \mathbb{E}_{C_0} |\hat{C}_n \Delta C_0| \lesssim n^{-2/(d+1)}.$$

Proof. The claim essentially follows from Theorem 8.25 and Corollary 8.26 of [48], asserting that we can take $\alpha = (d-1)/2$ in Theorem 4.4.1. □

The corollary shows that we can use a global least squares estimator rather than a sieved least squares estimator (cf. [26]) to achieve the optimal rate of convergence.

4.4.2 Least squares estimation: multiplicative errors

In this subsection we consider the regression model with multiplicative errors as in [87, 99]:

$$Y_i = f_{C_0}(X_i)\eta_i \tag{4.4.3}$$

where $f_{C_0}(x) = 2\mathbf{1}_{C_0}(x) - 1$ and η_i's are i.i.d. random variables such that $\mathbb{P}(\eta_i = 1) = 1/2 + a$ and $\mathbb{P}(\eta_i = -1) = 1/2 - a$ for some $a \in (0, 1/2)$. Such a model is motivated by estimation of

sets in multi-dimensional 'black and white' pictures, where $Y_i = 1$ is interpreted as observing black, and $Y_i = -1$ is white. We refer the reader to [99] for more motivation for this model. Apparently, the model (4.4.3) can be rewritten as

$$Y_i = 2af_{C_0}(X_i) + \xi_i \tag{4.4.4}$$

where $\xi_i = f_{C_0}(X_i)(\eta_i - 2a)$'s are bounded errors. An important property for these errors is that $\mathbb{E}[\xi_i|X_i] = 0$ for all $i = 1, \ldots, n$. Note here ξ_i is *not* independent of X_i and hence a different analysis is needed. Now consider the least squares estimator

$$\hat{C}_n \equiv \arg\min_{C\in\mathscr{C}} \sum_{i=1}^{n} (Y_i - 2af_C(X_i))^2. \tag{4.4.5}$$

Theorem 4.4.3. *Suppose that for some $\alpha \neq 1$, we have the following entropy estimate:*

$$\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \leq L\varepsilon^{-\alpha}.$$

Then

$$\sup_{C_0\in\mathscr{C}} \mathbb{E}_{C_0}|\hat{C}_n \Delta C_0| \lesssim n^{-1/(\alpha+1)}.$$

Compared with Theorem 4.1 in [99], we use an unsieved least squares estimator to achieve the optimal rate, rather than their theoretical 'sieved' estimator. This provides another example for which the simple least squares estimator can be rate-optimal for non-Donsker function classes in a natural setting.

4.4.3 Binary classification: excess risk bounds

In this subsection we consider the binary classification problem in learning theory, cf. [150, 108]. Suppose one observes i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$ with law P, where X takes values in $\mathcal{X}$, and the responses $Y_i \in \{0, 1\}$. A classifier $g : \mathcal{X} \to \{0, 1\}$ over a class $\mathcal{G}$ has a generalized error $P(Y \neq g(X))$. The excess risk for a classifier g under a law P is given by

$$\mathcal{E}_P(g) \equiv P(Y \neq g(X)) - \inf_{g'\in\mathcal{G}} P(Y \neq g'(X)).$$

It is known that for a given law P on (X, Y), the minimal generalized error is attained by a Bayes classifier $g_0(x) \equiv \mathbf{1}_{\eta(x)\geq 1/2}$ where $\eta(x) \equiv \mathbb{E}[Y|X = x]$, cf. [42]. It is then natural to consider an estimator of g_0 by minimizing the empirical training error:

$$\hat{g}_n \equiv \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_i\neq g(X_i)}.$$

The quality of the estimator $\hat{g}_n$ is measured by the excess risk:

$$\mathcal{E}_P(\hat{g}_n) \equiv P(Y \neq \hat{g}_n(X)) - P(Y \neq g_0(X)).$$

Let Π be the marginal distribution of X under P. We assume the following 'Tsybakov's margin(low noise) condition' (cf. [100, 150]): there exists some $c > 0$ such that for all $g \in \mathcal{G}$,

$$\mathcal{E}_P(g) \geq c\big(\Pi(g(X) \neq g_0(X))\big) = c\|g - g_0\|^2_{L_2(P)}. \tag{4.4.6}$$

Here we have assumed that the margin condition holds with $\kappa = 1$. For more general margin condition $\kappa \geq 1$, faster rates are possible, cf. [100, 150]. We will not go into this direction to avoid distraction from our main points.

Below is the main result in this subsection, the formulation of which follows that of [84, 59].

Theorem 4.4.4. *Suppose $\mathcal{G} \equiv \{\mathbf{1}_C : C \in \mathscr{C}\}$ satisfies the following entropy condition: there exists some $\alpha \neq 1$ such that for all $\varepsilon > 0$,*

$$\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \leq L\varepsilon^{-\alpha}.$$

If $r_n^2 \gtrsim n^{-1/(\alpha+1)}$, then

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}:\mathcal{E}_P(g)\geq r_n^2} \left|\frac{\mathcal{E}_{\mathbb{P}_n}(g)}{\mathcal{E}_P(g)} - 1\right| \geq \left(\frac{1}{2} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right)\right) \leq K' \exp(-s/K'),$$

holds for some constants $K, K' > 0$. In particular,

$$\mathbb{P}\left(\mathcal{E}_P(\hat{g}_n) \geq r_n^2\right) \leq K' \exp(-nr_n^2/K'')$$

holds for some constant $K'' > 0$.

[150] considered the same problem under the working assumption $\alpha \in (0,1)$ (cf. Assumption A2, page 140). [108] used ratio-type empirical process techniques to give a more unified treatment of deriving risk bounds for this problem, when the class of classifiers satisfies a Donsker bracketing entropy condition (i.e. $0 < \alpha < 1$), or a Donsker uniform entropy condition. [59] further improved the result of [108] in the Donsker regime under a uniform entropy condition, by taking into account the size of the localized envelopes. See also [84] page 2618, [89] page 1706 for similar Donsker conditions. To the best knowledge of the authors, our Theorem 4.4.4 gives a first result for the global ERM to be rate-optimal in a non-Donsker setting in the classification problem.

4.5 Proofs of the main results

In this section we prove Theorems 4.2.1-4.3.7. Theorem 4.2.5 follows line by line from the proof of Theorem 4.2.3 so we omit the details.

4.5.1 Proof of Theorem 4.2.1

Proposition 4.5.1. *Suppose there exists some $\alpha > 1$ such that for $\varepsilon > 0$, we have the entropy estimate*

$$\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \leq L\varepsilon^{-\alpha}.$$

Then

$$\mathbb{E} \sup_{C \in \mathscr{C}} |\mathbb{G}_n(C)| \leq C_{L,\alpha} n^{(\alpha-1)/2(\alpha+1)}.$$

Proof. Let k_n be such that $2^{k_n} = n^{1/(\alpha+1)}$ (by slightly ignoring the rounding issue for notational convenience). For any $1 \leq k \leq k_n$, let $N_k \equiv \mathcal{N}_I(2^{-k}, \mathscr{C}, P) \leq \exp(L2^{k\alpha})$, and let $\{A_{k,i} \subset B_{k,i}\}_{i=1}^{N_k}$ denote a minimal bracketing set such that $P(B_{k,i} \setminus A_{k,i}) \leq 2^{-k}$. Then for any $A \in \mathscr{C}$, there exists some $i \equiv i(k, A) \in [1 : N_k]$ such that $A_{k,i} \subset A \subset B_{k,i}$. Note that for any $k \geq 0$,

$$\begin{aligned} P\left(A_{k,i(k,A)} \Delta A_{k-1,i(k-1,A)}\right) &\leq P\left(A_{k,i(k,A)} \Delta A\right) + P\left(A_{k-1,i(k-1,A)} \Delta A\right) \\ &\leq 2^{-k} + 2^{-k+1} \leq 2^{-k+2}. \end{aligned}$$

Consider the set

$$\mathcal{Q}_k \equiv \{C \in \mathcal{A} : C \in \{A_{k,i} \setminus A_{k-1,j}, A_{k-1,j} \setminus A_{k,i}, B_{k,i} \setminus A_{k,i}\}_{1\le i\le N_k, 1\le j\le N_{k-1}}, \\ P(C) \le 2^{-k+2}\}.$$

Then $|\mathcal{Q}_k| \le 2N_{k-1}N_k + N_k \le 3\exp(2L\cdot 2^{k\alpha})$. By Bernstein's inequality (cf. page 36 of [22]), for any $1 \le k \le k_n$,

$$\mathbb{E}\max_{C\in\mathcal{Q}_k}|\mathbb{G}_n(C)| \lesssim \frac{\log|\mathcal{Q}_k|}{\sqrt{n}} + 2^{-k/2}\sqrt{\log|\mathcal{Q}_k|} \\ \lesssim_L \frac{2^{k\alpha}}{\sqrt{n}} + 2^{k(\alpha-1)/2}.$$

For any $A \in \mathscr{C}$, consider the following chaining:

$$\mathbf{1}_{A_{k_n,i(k_n,A)}} = \sum_{r=1}^{k_n}\left(\mathbf{1}_{A_{r,i(r,A)}} - \mathbf{1}_{A_{r-1,i(r-1,A)}}\right)$$

where we define $A_{0,\cdot} = \emptyset$. Then,

$$\begin{aligned}
&\mathbb{E}\sup_{A\in\mathscr{C}}|\mathbb{G}_n(A_{k_n,i(k_n,A)})| \\
&\le \sum_{r=1}^{k_n}\mathbb{E}\sup_{A\in\mathscr{C}}|\mathbb{G}_n(A_{r,i(r,A)} \setminus A_{r-1,i(r-1,A)})| + \sum_{r=1}^{k_n}\mathbb{E}\sup_{A\in\mathscr{C}}|\mathbb{G}_n(A_{r-1,i(r-1,A)} \setminus A_{r,i(r,A)})| \\
&\le 2\sum_{r=1}^{k_n}\mathbb{E}\max_{C\in\mathcal{Q}_r}|\mathbb{G}_n(C)| \lesssim_L \sum_{r=1}^{k_n}\left(\frac{2^{r\alpha}}{\sqrt{n}} + 2^{r(\alpha-1)/2}\right) \\
&\lesssim_{L,\alpha} \frac{2^{k_n\alpha}}{\sqrt{n}} + 2^{k_n(\alpha-1)/2} \le C_{L,\alpha}n^{(\alpha-1)/2(\alpha+1)}
\end{aligned}$$

by our choice of k_n such that $2^{k_n} = n^{1/(\alpha+1)}$. On the other hand,

$$\begin{aligned}
&\mathbb{E}\sup_{A\in\mathscr{C}}|\mathbb{G}_n(A \setminus A_{k_n,i(k_n,A)})| \\
&\le \mathbb{E}\sup_{A\in\mathscr{C}}\sqrt{n}\mathbb{P}_n(B_{k_n,i(k_n,A)} \setminus A_{k_n,i(k_n,A)}) + \sup_{A\in\mathscr{C}}\sqrt{n}P(B_{k_n,i(k_n,A)} \setminus A_{k_n,i(k_n,A)}) \\
&\le \mathbb{E}\sup_{C\in\mathcal{Q}_{k_n}}|\mathbb{G}_n(C)| + 2\sup_{A\in\mathscr{C}}\sqrt{n}P(B_{k_n,i(k_n,A)} \setminus A_{k_n,i(k_n,A)}) \\
&\lesssim \frac{2^{k_n\alpha}}{\sqrt{n}} + 2^{k_n(\alpha-1)/2} + \sqrt{n}2^{-k_n} \lesssim_{L,\alpha} n^{(\alpha-1)/2(\alpha+1)}.
\end{aligned}$$

This entails that

$$\mathbb{E}\sup_{A\in\mathscr{C}}|\mathbb{G}_n(A)| \le \mathbb{E}\sup_{A\in\mathscr{C}}|\mathbb{G}_n(A\setminus A_{k_n,i(k_n,A)})| + \mathbb{E}\sup_{A\in\mathscr{C}}|\mathbb{G}_n(A_{k_n,i(k_n,A)})|$$
$$\lesssim_{L,\alpha} n^{(\alpha-1)/2(\alpha+1)},$$

as desired. □

We shall note that the above chaining argument used $L_1 \asymp L_2^2$ in the end. This property explains why such a chaining method is particularly well suited for the class of indicator functions and does not extend to a general function class.

Lemma 4.5.2 (Lemma 3.4.2 of [162]). *Suppose that $\mathcal{F} \subset L_\infty(1)$, and $X_1, \ldots, X_n$'s are i.i.d. random variables with law P. Then with $\mathcal{F}(\delta) \equiv \{f \in \mathcal{F} : Pf^2 < \delta^2\}$. Then*

$$\mathbb{E}\|\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}(\delta)} \lesssim \sqrt{n} J_{[]}(\delta, \mathcal{F}, L_2(P))\left(1 + \frac{J_{[]}(\delta, \mathcal{F}, L_2(P))}{\sqrt{n}\delta^2}\right). \tag{4.5.1}$$

Here

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) \equiv \int_0^\delta \sqrt{1 + \log \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)}\, \mathrm{d}\varepsilon.$$

Proof of Theorem 4.2.1. First consider $0 < \alpha < 1$. Upper bounds in this regime follow directly from the local maximal inequality as in Lemma 4.5.2 by noting that $\mathcal{N}_{[]}(\varepsilon, \mathscr{C}, L_2(P)) = \mathcal{N}_I(\varepsilon^2, \mathscr{C}, P)$, and hence

$$J_{[]}(\sigma, \mathscr{C}, L_2(P)) \lesssim \int_0^\sigma \varepsilon^{-\alpha}\, \mathrm{d}\varepsilon \lesssim_\alpha \sigma^{1-\alpha}.$$

For $\sigma^2 \gtrsim n^{-1/(\alpha+1)}$, we then have

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)}|\mathbb{G}_n(C)| \lesssim_\alpha \sigma^{1-\alpha}.$$

Next consider $\alpha > 1$. A global upper bound follows from Proposition 4.5.1:

$$\mathbb{E}\sup_{C\in\mathscr{C}}|\mathbb{G}_n(C)| \lesssim_\alpha n^{(\alpha-1)/2(\alpha+1)},$$

completing the proof. □

Remark 4.5.3. From the proof of Theorem 4.2.1, we see that for $0 < \alpha < 1$, the bracketing entropy condition can be replaced by the uniform entropy condition–then we only need to use the corresponding local maximal inequality for empirical processes instead of Lemma 4.5.2 used here, cf. [161] or Section 3 of [59].

4.5.2 Proof of Theorem 4.2.3

We first prove the lower bound for Gaussianized empirical process.

Proposition 4.5.4. *For any $\sigma \geq 15n^{-1/2}$ such that*

$$\log \mathcal{N}(\sigma/4, \mathscr{C}(\sigma), L_2(P)) \leq n\sigma^2/400,$$

we have

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \mathbf{1}_C(X_i)| \gtrsim \sigma \sqrt{\log \mathcal{N}(\sigma/2, \mathscr{C}(\sigma), L_2(P))}.$$

Lemma 4.5.5 (Sudakov minorization [144])**.** *Let $(X_t)_{t \in T}$ be a centered separable Gaussian process, and $\|t - s\|^2 := \mathbb{E}(X_t - X_s)^2$. Then*

$$\mathbb{E} \sup_{t \in T} X_t \geq C^{-1} \sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathcal{N}(\varepsilon, T, \|\cdot\|)}.$$

Here C is a universal constant.

Proof of Proposition 4.5.4. By Sudakov minorization (cf. Lemma 4.5.5), for any $\sigma > 0$,

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \mathbf{1}_C(X_i)| \gtrsim \mathbb{E}\sigma \sqrt{\log \mathcal{N}(\sigma/10, \mathscr{C}(\sigma), L_2(\mathbb{P}_n))}. \tag{4.5.2}$$

We claim that for any $\sigma > 0$ such that $\log \mathcal{N}(\sigma/4, \mathscr{C}(\sigma), L_2(P)) \leq n\sigma^2/400$,

$$\mathbb{P}\left(\mathcal{N}(\sigma/10, \mathscr{C}(\sigma), L_2(\mathbb{P}_n)) \geq \mathcal{N}(\sigma/2, \mathscr{C}(\sigma), L_2(P))\right) \geq 1 - e^{-n\sigma^2/200}. \tag{4.5.3}$$

To see this, let $C_1, \ldots, C_N$ be a maximal $\sigma/2$-packing set in the $L_2(P)$ metric, i.e. for $i \neq j$, $P(\mathbf{1}_{C_i} - \mathbf{1}_{C_j})^2 \geq \sigma^2/4$. Since $P(\mathbf{1}_{C_i} - \mathbf{1}_{C_j})^4 = P(\mathbf{1}_{C_i} - \mathbf{1}_{C_j})^2 \leq 2\sigma^2$, we apply Bernstein's inequality followed by a union bound to see that with probability at least $1 - N^2 \exp(-t)$,

$$\max_{1 \leq i \neq j \leq N} \left(nP(\mathbf{1}_{C_i} - \mathbf{1}_{C_j})^2 - \sum_{k=1}^{n} (\mathbf{1}_{C_i} - \mathbf{1}_{C_j})^2(X_k) \right) \leq \frac{2t}{3} + \sqrt{4tn\sigma^2}.$$

With $t = cn\sigma^2$ with a constant $c > 0$ to be specified below, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\min_{1\le i\ne j\le N}\frac{1}{n}\sum_{k=1}^{n}(\mathbf{1}_{C_i}-\mathbf{1}_{C_j})^2(X_k) \ge \sigma^2\left(\frac{1}{4}-\frac{2c}{3}-\sqrt{4c}\right)\right)\\
&\ge 1-\exp\left(2\log\mathcal{D}(\sigma/2,\mathscr{C}(\sigma),L_2(P))-cn\sigma^2\right)\\
&\ge 1-\exp\left(2\log\mathcal{N}(\sigma/4,\mathscr{C}(\sigma),L_2(P))-cn\sigma^2\right).
\end{aligned}$$

By choosing $c = 0.01$ and $\log\mathcal{N}(\sigma/4,\mathscr{C}(\sigma),L_2(P)) \le n\sigma^2/400$, we have

$$\mathbb{P}\left(\min_{1\le i\ne j\le N}\frac{1}{n}\sum_{k=1}^{n}(\mathbf{1}_{C_i}-\mathbf{1}_{C_j})^2(X_k) \ge 0.04\sigma^2\right) \ge 1-\exp(-n\sigma^2/200).$$

The event in the probability on the left side of the above display is

$$\mathcal{D}(\sigma/5,\mathscr{C}(\sigma),L_2(\mathbb{P}_n)) \ge N \equiv \mathcal{D}(\sigma/2,\mathscr{C}(\sigma),L_2(P)).$$

Hence for any $\sigma > 0$ such that $\log\mathcal{N}(\sigma/4,\mathscr{C}(\sigma),L_2(P)) \le n\sigma^2/400$, with probability at least $1-e^{-n\sigma^2/200}$,

$$\begin{aligned}
\mathcal{N}(\sigma/10,\mathscr{C}(\sigma),L_2(\mathbb{P}_n)) &\ge \mathcal{D}(\sigma/5,\mathscr{C}(\sigma),L_2(\mathbb{P}_n))\\
&\ge \mathcal{D}(\sigma/2,\mathscr{C}(\sigma),L_2(P)) \ge \mathcal{N}(\sigma/2,\mathscr{C}(\sigma),L_2(P)),
\end{aligned}$$

completing the proof of (4.5.3). Hence for any $\sigma \ge 15n^{-1/2}$ such that the entropy $\log\mathcal{N}(\sigma/4,\mathscr{C}(\sigma),L_2(P)) \le n\sigma^2/400$, the claim of the proposition follows from (4.5.2) and (4.5.3). □

Next we eliminate the effect of the Gaussian multiplier. We need the following form of a multiplier inequality to handle lower bounds in the Donsker regime.

Lemma 4.5.6 (Theorem 1 in [76]). *Suppose that $\xi_1,\dots,\xi_n$ are i.i.d. mean-zero random variables independent of i.i.d. $X_1,\dots,X_n$. Let $\mathcal{F}_1 \supset \cdots \supset \mathcal{F}_n$ be a non-increasing sequence of function classes. Assume further that there exist non-decreasing concave functions $\{\psi_n\}$: $\mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ with $\psi_n(0) = 0$ such that*

$$\mathbb{E}\|\sum_{i=1}^{k}\varepsilon_i f(X_i)\|_{\mathcal{F}_k} \le \psi_n(k) \tag{4.5.4}$$

holds for all $1 \le k \le n$. *Then*

$$\mathbb{E}\|\sum_{i=1}^{n} \xi_i f(X_i)\|_{\mathcal{F}_n} \le 4 \int_0^\infty \psi_n\big(n \cdot \mathbb{P}(|\xi_1| > t)\big)\, \mathrm{d}t. \tag{4.5.5}$$

Proposition 4.5.7. *Suppose that the following entropy estimates*

$$\log \mathcal{N}(\varepsilon/4, \mathscr{C}(\sqrt{\varepsilon}), P) \le L\varepsilon^{-\alpha}, \quad \log \mathcal{N}(\varepsilon/2, \mathscr{C}(\sqrt{\varepsilon}), P) \ge L^{-1}\varepsilon^{-\alpha}$$

hold for some $\alpha \in (0,1)$ *and* $L > 0$ *large enough. Then for* $\sigma_n^2 \ge cn^{-1/(\alpha+1)}$ *with some constant* $c > 0$,

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma_n)} |\mathbb{G}_n(C)| \gtrsim_\alpha \sigma_n^{1-\alpha}.$$

Proof. By Proposition 4.5.4, the Gaussianized empirical process satisfies

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma_n)} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i \mathbf{1}_C(X_i)| \gtrsim \sigma_n \sqrt{\log \mathcal{N}(\sigma_n/2, \mathscr{C}(\sigma_n), L_2(P))} \ge C_1^{-1} \sigma_n^{1-\alpha}.$$

Suppose that $\sigma_n^2 \le c$. Without loss of generality, we assume that $\sigma_n^2 \equiv \sigma_n(\gamma)^2 = cn^{-\gamma}$ for some $0 < \gamma \le 1/(\alpha+1)$. We first prove the following claim: there exists some $c_1 > 0$ such that for any $0 \le \gamma \le 1/(\alpha+1)$,

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma_n)} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \mathbf{1}_C(X_i)| \ge c_1 \sigma_n^{1-\alpha}. \tag{4.5.6}$$

Suppose the contrary, then for any $c_2 > 0$, we can find some $\gamma' \in [0, 1/(\alpha+1)]$ such that

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma_n)} |\sum_{i=1}^{n} \varepsilon_i \mathbf{1}_C(X_i)| < c_2 \sqrt{n} \sigma_n(\gamma')^{1-\alpha} = c_2 c^{(1-\alpha)/2} n^\beta,$$

where $\beta \equiv \beta(\alpha, \gamma') \equiv \frac{1}{2}\big(1-(1-\alpha)\gamma'\big) \in [\alpha/(1+\alpha), 1/2]$. Now apply the multiplier inequality Lemma 4.5.6, we see that

$$\begin{aligned} c^{(1-\alpha)/2} n^\beta = \sqrt{n}\sigma_n^{1-\alpha} &\le C_1 \cdot \mathbb{E} \sup_{C \in \mathscr{C}(\sigma_n)} |\sum_{i=1}^{n} g_i \mathbf{1}_C(X_i)| \\ &\le 4c_2 c^{(1-\alpha)/2} n^\beta \int_0^\infty \big(\mathbb{P}(|g_1| > t)\big)^\beta \, \mathrm{d}t \\ &\le 4c_2 G_\alpha \cdot c^{(1-\alpha)/2} n^\beta, \end{aligned}$$

where $G_\alpha = \int_0^\infty \big(\mathbb{P}(|g_1| > t)\big)^{\alpha/(1+\alpha)}\, \mathrm{d}t < \infty$ since Gaussian random variables have finite moments of any order. The above display will be invalid as soon as $c_2 < 1/4G_\alpha$. This proves our claim (4.5.6). Now by de-symmetrization inequality (cf. Lemma 2.3.6 of [162]), we have that

$$\mathbb{E} \sup_{C\in\mathscr{C}(\sigma_n)} |\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \mathbf{1}_C(X_i)| \le 2\mathbb{E} \sup_{C\in\mathscr{C}(\sigma_n)} |\mathbb{G}_n(C)| + 2\sigma_n. \tag{4.5.7}$$

For $\sigma_n \le (1/4c_1)^{1/\alpha} \wedge c^{1/2}$, the claim of the proposition follows from (4.5.6) and (4.5.7). On the other hand, the claim is trivial for $\sigma_n > (1/4c_1)^{1/\alpha} \wedge c^{1/2}$. □

The following alternative formulation of the multiplier inequality, proved in Proposition 1 of [76], will be useful in handling the lower bounds in the non-Donsker regime.

Lemma 4.5.8. *Let $\xi_1, \ldots, \xi_n$ be i.i.d. symmetric mean-zero multipliers independent of i.i.d. samples $X_1, \ldots, X_n$. For any function class $\mathcal{F}$,*

$$\mathbb{E}\|\sum_{i=1}^n \xi_i f(X_i)\|_{\mathcal{F}} \le \mathbb{E}\left[\sum_{k=1}^n (|\xi_{(k)}| - |\xi_{(k+1)}|)\mathbb{E}\|\sum_{i=1}^k \varepsilon_i f(X_i)\|_{\mathcal{F}}\right] \tag{4.5.8}$$

where $|\xi_{(1)}| \ge \cdots \ge |\xi_{(n)}| \ge |\xi_{(n+1)}| \equiv 0$ are the reversed order statistics for $\{\xi_i\}_{i=1}^n$.

Proposition 4.5.9. *Suppose that the following entropy estimates*

$$\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \le L\varepsilon^{-\alpha}, \quad \log \mathcal{N}(\varepsilon/2, \mathscr{C}(\sqrt{\varepsilon}), P) \ge L^{-1}\varepsilon^{-\alpha}$$

hold for some $\alpha > 1$. Then for $\sigma_n^2 \equiv cn^{-1/(\alpha+1)}$ with some constant $c > 0$,

$$\mathbb{E} \sup_{C\in\mathscr{C}(\sigma_n)} |\mathbb{G}_n(C)| \gtrsim n^{(\alpha-1)/2(\alpha+1)}.$$

Proof. Proposition 4.5.4 shows that

$$\mathbb{E} \sup_{C\in\mathscr{C}(\sigma_n)} |\frac{1}{\sqrt{n}} \sum_{i=1}^n g_i \mathbf{1}_C(X_i)| \gtrsim \sigma_n \sqrt{\log \mathcal{N}(\sigma_n/2, \mathscr{C}(\sigma_n), L_2(P))} \gtrsim n^{(\alpha-1)/2(\alpha+1)}.$$

Now applying Lemma 4.5.8 in the following form,

$$\mathbb{E} \sup_{C\in\mathscr{C}(\sigma_n)} |\frac{1}{\sqrt{n}} \sum_{i=1}^n g_i \mathbf{1}_C(X_i)| \lesssim \max_{1\le k\le n} \mathbb{E} \sup_{C\in\mathscr{C}(\sigma_n)} |\frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i \mathbf{1}_C(X_i)|,$$

we see that

$$\max_{1\le k\le n}\mathbb{E}\sup_{C\in\mathscr{C}(\sigma_n)}|\mathbb{G}_k(C)|\gtrsim n^{(\alpha-1)/2(\alpha+1)}.$$

On the other hand, Proposition 4.5.1 entails that

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma_k)}|\mathbb{G}_k(C)|\lesssim_\alpha k^{(\alpha-1)/2(\alpha+1)},$$

and hence

$$\max_{1\le k\le n}\mathbb{E}\sup_{C\in\mathscr{C}(\sigma_k)}|\mathbb{G}_k(C)|\lesssim_\alpha n^{(\alpha-1)/2(\alpha+1)}$$

by the assumption $\alpha>1$. Combining the upper and lower estimates we see that there exists some $K\equiv K_\alpha>1$ such that

$$\begin{aligned}K^{-1}n^{(\alpha-1)/2(\alpha+1)}&\le\max_{1\le k\le n}\mathbb{E}\sup_{C\in\mathscr{C}(\sigma_n)}|\mathbb{G}_k(C)|\\&\le\max_{1\le k\le n}\mathbb{E}\sup_{C\in\mathscr{C}(\sigma_k)}|\mathbb{G}_k(C)|\le Kn^{(\alpha-1)/2(\alpha+1)}.\end{aligned}$$

Now we will argue that the max operator can be 'eliminated'. To this end, let $a_k\equiv\mathbb{E}\sup_{C\in\mathscr{C}(\sigma_n)}|\mathbb{G}_k(C)|$ and $\beta\equiv(\alpha-1)/2(\alpha+1)$ for notational convenience. Let $k_n\equiv\arg\max_{1\le k\le n}a_k$. We claim that $k_n\in[cn,n]$ where $c=K^{-2/\beta}\in(0,1)$. To see this, we only need to note

$$K^{-1}n^\beta\le\max_{1\le k\le n}a_k=a_{k_n}\le Kk_n^\beta$$

which entails $k_n^\beta\ge K^{-2}n^\beta$. Hence

$$K^{-1}n^{(\alpha-1)/2(\alpha+1)}\le\mathbb{E}\sup_{C\in\mathscr{C}(\sigma_n)}|\mathbb{G}_{k_n}(C)|\le\frac{1}{\sqrt{c}}\mathbb{E}\sup_{C\in\mathscr{C}(\sigma_n)}|\mathbb{G}_n(C)|$$

where the last inequality follows from Jensen's inequality, proving the claim. □

Proof of Theorem 4.2.3. The claims follow by combining Propositions 4.5.7 and 4.5.9. □

Remark 4.5.10. For $0<\alpha<1$, Proposition 4.5.7 only requires a metric entropy condition, which is weaker than a uniform entropy condition (cf. Problem 2.5.1 of [162]). Hence our proof technique applies to the uniform entropy assumption case as studied in [59].

4.5.3 Proof of Theorem 4.2.7

We need Dudley's entropy integral bound, recorded below for the convenience of the reader.

Lemma 4.5.11 ([46]). *Let $(X_t)_{t\in T}$ be a centered separable Gaussian process, and $\|t-s\|^2 := \mathbb{E}(X_t - X_s)^2$. Then*

$$\mathbb{E}\sup_{t\in T} X_t \le C\int_0^{\mathrm{diam}(T)} \sqrt{\log \mathcal{N}(\varepsilon, T, \|\cdot\|)}\ \mathrm{d}\varepsilon.$$

Here C is a universal constant.

Proof of Theorem 4.2.7. Let g be a standard normal random variable independent of G_P, and let $Z_P(C) \equiv G_P(C) + g\cdot P(C)$. The covariance structure of Z_P is then given by $\mathbb{E}Z_P(C)Z_P(C') = P\mathbf{1}_C\mathbf{1}_{C'} = P(C\cap C')$, and hence the canonical distance associated with Z_P is $d_Z(C,C') \equiv \sqrt{\mathbb{E}(Z_P(C)-Z_P(C'))^2} = \sqrt{P(\mathbf{1}_C-\mathbf{1}_{C'})^2} = \sqrt{P(C\Delta C')}$. By Sudakov minorization (cf. Lemma 4.5.5), we have

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |Z_P(C)| \gtrsim \sigma\sqrt{\log \mathcal{N}(\sigma^2/2, \mathscr{C}(\sigma), P)} \gtrsim \sigma^{1-\alpha}.$$

Dudley's entropy integral (cf. Lemma 4.5.11) entails that

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |Z_P(C)| \lesssim \int_0^{\sigma} \sqrt{\log \mathcal{N}(\varepsilon^2, \mathscr{C}, P)}\ \mathrm{d}\varepsilon \lesssim \sigma^{1-\alpha}.$$

Now Theorems 4.2.1 and 4.2.3 entail that

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |\mathbb{G}_n(C)| \asymp_\alpha \mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |Z_P(C)|.$$

On the other hand, by definition of Z_P, we have

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |Z_P(C)| \le \mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |G_P(C)| + \sup_{C\in\mathscr{C}} P(C) \lesssim \sigma^{1-\alpha} + \sigma \lesssim \sigma^{1-\alpha}$$

since $\sigma \le 1$, and

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |Z_P(C)| \ge \mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |G_P(C)| - \sup_{C\in\mathscr{C}} P(C) \ge c_1\sigma^{1-\alpha} - \sigma \ge (c_1/2)\sigma^{1-\alpha}$$

if $\sigma \le (c_1/2)^{1/\alpha}$. Thus we have shown that

$$\mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |\mathbb{G}_n(C)| \asymp_\alpha \mathbb{E}\sup_{C\in\mathscr{C}(\sigma)} |G_P(C)|.$$

holds for $\sigma \le (c_1/2)^{1/\alpha}$. The case for $\sigma > (c_1/2)^{1/\alpha}$ holds trivially from the ordinary Donsker theorem. □

4.5.4 Proofs of Theorems 4.3.2-4.3.7

Before proving Theorems 4.3.2-4.3.7, we need some preparation. We will investigate the behavior of ratio-type empirical processes in a more general setting as in [59]. Let ϕ be a continuous and strictly increasing function with $\phi(0) = 0$. Let $\mathscr{C}(r) \equiv \{C \in \mathscr{C} : P(C) \leq r^2\}$ and $\mathscr{C}(r, s] \equiv \mathscr{C}(s) \setminus \mathscr{C}(r)$. Fix $0 < r < \delta \leq 1$. For a real number $1 < q \leq 2$, let $l \equiv l_{r,\delta,q}$ be the smallest integer no smaller than $\log_q(\delta/r)$. Put for any $\boldsymbol{s} \equiv (s_1, \ldots, s_l) \in \mathbb{R}^l_{\geq 0}$,

$$\begin{aligned} \beta_{n,q}(r, \delta) &\equiv \max_{1 \leq j \leq l} \frac{\mathbb{E}\|\mathbb{G}_n\|_{\mathscr{C}(rq^{j-1}, rq^j)}}{\phi(rq^j)}, \\ \tau_{n,q}(r, \delta, \boldsymbol{s}) &\equiv \max_{1 \leq j \leq l} \frac{rq^j\sqrt{s_j} + s_j/\sqrt{n}}{\phi(rq^j)}. \end{aligned} \tag{4.5.9}$$

The following result is essentially due to [59]. We state a somewhat simplified and easier-to-use version.

Proposition 4.5.12. *Assume that ϕ is continuous, strictly increasing and satisfies $\sup_{r \leq x \leq 1} \phi(qx)/\phi(x) = \kappa_{r,q} < \infty$ for some $1 < q \leq 2$. Then for any $\boldsymbol{s} \equiv (s_1, \ldots, s_l) \in \mathbb{R}^l_{\geq 0}$, both the probabilities*

$$\mathbb{P}\left[\sup_{C \in \mathscr{C}: r^2 < P(C) \leq \delta^2} \frac{|\mathbb{G}_n(C)|}{\phi\left(\sqrt{P(C)}\right)} \geq K\kappa_{r,q}\left(\beta_{n,q}(r, \delta) + \tau_{n,q}(r, \delta, \boldsymbol{s})\right)\right]$$

and

$$\mathbb{P}\left[\sup_{C \in \mathscr{C}: r^2 < P(C) \leq \delta^2} \frac{|\mathbb{G}_n(C)|}{\phi\left(\sqrt{P(C)}\right)} \leq K\left(\beta_{n,q}(r, \delta) - \tau_{n,q}(r, \delta, \boldsymbol{s})\right)\right]$$

can be bounded by

$$K\sum_{j=1}^{l} \exp\big(-s_j/K\big).$$

Here $K > 0$ is a universal constant.

Proof of Proposition 4.5.12. We only prove the first claim; the second follows from similar arguments. The proof is a simple application of Talagrand's concentration inequality combined with a peeling device. Write $\mathscr{C}_j \equiv \mathscr{C}(rq^{j-1}, rq^j]$ and $\phi_q(u) \equiv \phi(rq^j)$ if $u \in (rq^{j-1}, rq^j]$ for notational convenience. By Talagrand's concentration inequality,

$$\mathbb{P}\left[\|\mathbb{G}_n\|_{\mathscr{C}_j} \geq K\left(\mathbb{E}\|\mathbb{G}_n\|_{\mathscr{C}_j} + \sqrt{\sigma_j^2 s_j} + \frac{s_j}{\sqrt{n}}\right)\right] \leq K\exp\big(-s_j/K\big)$$

where $\sigma_j^2 = \sup_{f\in\mathscr{C}_j} P(C) = r^2q^{2j}$. Hence by a union bound we see that with probability at least $1 - \sum_{j=1}^{l} K\exp(-s_j/K)$, it holds that

$$\left(\sup_{C\in\mathscr{C}:r^2<P(C)\le\delta^2} \frac{|\mathbb{G}_n(C)|}{\phi_q\left(\sqrt{P(C)}\right)} - K\beta_{n,q}(r,\delta)\right)_+$$
$$\le \max_{1\le j\le l}\left(\frac{\|\mathbb{G}_n\|_{\mathscr{C}_j}}{\phi(rq^j)} - \frac{K\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}(rq^{j-1},rq^j)}}{\phi(rq^j)}\right)_+ \le K\max_{1\le j\le l}\frac{rq^j\sqrt{s_j}+s_j/\sqrt{n}}{\phi(rq^j)}.$$

Now the conclusion follows from $\sup_{r\le x\le 1}\phi(qx)/\phi(x) < \infty$. □

The next lemma, due to Lemma 7.2 of [5], provides a convenient device to derive almost sure results for ratio-type empirical processes.

Lemma 4.5.13. *Let c_n, u_n be such that $c_n/n \searrow$ and $u_n \searrow$, and assume that $r_n \searrow$ and $\sqrt{n}\delta_n \nearrow$. For a centered function class $\mathcal{F} \subset L_2(P)$, put the events*

$$A_n \equiv \{|\mathbb{G}_n f| \ge c_n\phi(\sigma_P f) + u_n \text{ for some } f\in\mathcal{F}, r_n \le \sigma_P f \le \delta_n\},$$

and

$$A_n^\varepsilon \equiv \Big\{|\mathbb{G}_n f| \ge (1-\varepsilon)\big(c_n\phi(\sigma_P f) + u_n\big) \text{ for some } f\in\mathcal{F},$$
$$r_n \le \sigma_P f \le \sqrt{1+\varepsilon}\cdot\delta_n\Big\}.$$

Assume that $\inf_{n\ge1,t\in[r_n,\delta_n]} c_n\frac{\phi(t)}{t} > 0$. *Then if*

$$\mathbb{P}(A_n^\varepsilon) = \mathcal{O}(1/(\log n)^{1+\theta})$$

holds for some $\varepsilon,\theta > 0$, we have

$$\mathbb{P}(A_n \text{ } i.o.) = 0.$$

Proof of Theorem 4.3.2 . Consider the first claim. Note that for $0 < \alpha < 1$,

$$\beta_{n,q} \lesssim \max_{1\le j\le l}\frac{(r_nq^j)^{1-\alpha}}{r_nq^j} \asymp r_n^{-\alpha},$$

while for $\alpha > 1$,

$$\beta_{n,q} \lesssim \max_{1\le j\le l}\frac{n^{(\alpha-1)/2(\alpha+1)}}{r_nq^j} \le n^{\frac{\alpha}{2(\alpha+1)}}.$$

For $s_j \equiv s + 2K \log j$, we have

$$\begin{aligned}\tau_{n,q} &\lesssim \max_{1\le j\le l}\left(\sqrt{s+2K\log j} + \frac{s+2K\log j}{\sqrt{n} r_n q^j}\right)\\ &\lesssim \sqrt{s \vee \log\log(1/r_n)} + (s\vee 1) n^{-\frac{\alpha}{2(\alpha+1)}},\end{aligned}$$

and the probability estimate

$$K\sum_{j=1}^{l} \exp(-s_j/K) = Ke^{-s}\sum_{j=1}^{l} j^{-2} \le K' e^{-s}.$$

This proves that

$$\begin{aligned}&\mathbb{P}\Bigg(\sup_{C\in\mathscr{C}: r_n^2 \le P(C)\le 1} \frac{|\mathbb{G}_n(C)|}{\sqrt{P(C)}} \\ &\qquad \ge K\left(\beta_{n,q} + \sqrt{s\vee\log\log(1/r_n)} + (s\vee 1) n^{-\frac{\alpha}{2(\alpha+1)}}\right)\Bigg) \qquad (4.5.10)\\ &\le K' e^{-s}.\end{aligned}$$

The first claim of (1) follows from Lemma 4.5.13 by setting $s \asymp \log\log n$. The second claim follows from similar lines by observing that under (E2), Theorem 4.2.3 yields that for $0 < \alpha < 1$,

$$\beta_{n,q} \gtrsim \max_{1\le j\le l} \frac{(r_n q^j)^{1-\alpha}}{r_n q^j} \asymp r_n^{-\alpha},$$

while for $\alpha > 1$,

$$\beta_{n,q} \gtrsim \max_{1\le j\le l} \frac{n^{(\alpha-1)/2(\alpha+1)}}{r_n q^j} \asymp n^{\frac{\alpha}{2(\alpha+1)}},$$

and $\tau_{n,q}$ can be estimated from above using the same arguments. □

Proof of Theorem 4.3.3. The proof of Theorem 4.3.3 uses a similar strategy as that of Theorem 4.3.2. For convenience of the reader we provide some details. Consider the first claim. Note that for $0 < \alpha < 1$,

$$\beta_{n,q} \lesssim \max_{1\le j\le l} \frac{(r_n q^j)^{1-\alpha}}{r_n^2 q^{2j}} \asymp r_n^{-(1+\alpha)}$$

while for $\alpha > 1$,

$$\beta_{n,q} \lesssim \max_{1\le j\le l} \frac{n^{(\alpha-1)/2(\alpha+1)}}{r_n^2 q^{2j}} \le r_n^{-2} n^{\frac{\alpha-1}{2(\alpha+1)}}.$$

For $s_j \equiv s + 2K \log j$, we have

$$\begin{aligned}\tau_{n,q} &\lesssim \max_{1\le j\le l}\left(r_n^{-1}\sqrt{s+2K\log j} + \frac{s+2K\log j}{\sqrt{n}r_n^2 q^{2j}}\right)\\ &\lesssim \sqrt{r_n^{-2}\big(s \vee \log\log(1/r_n)\big)} + (s\vee 1)(\sqrt{n}r_n^2)^{-1}.\end{aligned}$$

This shows that, for

$$\bar\gamma_n \equiv \left(r_n^{-2}n^{-\frac{1}{\alpha+1}}\right)^{1\wedge\frac{1+\alpha}{2}} = \begin{cases} n^{-1/2}r_n^{-(1+\alpha)}, & \alpha\in(0,1);\\ r_n^{-2}n^{-\frac{1}{\alpha+1}}, & \alpha>1.\end{cases}$$

we have

$$\begin{aligned}&\mathbb{P}\Bigg(\sup_{C\in\mathscr{C}: r_n^2\le P(C)\le 1}\frac{|\mathbb{P}_n(C)-P(C)|}{P(C)}\\ &\qquad\ge K\left(\bar\gamma_n + \sqrt{(nr_n^2)^{-1}\big(s\vee\log\log(1/r_n)\big)} + (s\vee 1)(nr_n^2)^{-1}\right)\Bigg) \qquad (4.5.11)\\ &\le K'e^{-s}.\end{aligned}$$

The first claim of the theorem follows by taking $s \asymp \log\log n$, applying Lemma 4.5.13 and noting that $\limsup_n \bar\gamma_n \le \underline{\rho}^{-1}$ by the assumption. The second claim follows similarly by estimating $\beta_{n,q}$ from below, up to a multiplicative constant, by $\bar\gamma_n$ and then repeat the arguments as above. □

Proof of Theorem 4.3.7. Let $r_n^2 \asymp n^{-1/(\alpha+1)}$ and $\delta_n^2 = \mathfrak{o}\big((\log\log n)^{-1/\alpha}\big)$. We will apply Proposition 4.5.12 with $q=2$. Note that $\sup_{r_n\le x\le 1}\omega(2x)/\omega(x) = 2^{1-\alpha} < \infty$, satisfying the assumption of Proposition 4.5.12. If (E1) holds, we see by Theorem 4.2.1 that $\beta_{n,q}(r_n,\delta_n) \le C_1$ for some $C_1 > 0$. Choose $s_j \equiv 3K\log\log n$, we have that

$$\begin{aligned}\tau_{n,q}(r_n,\delta_n,\boldsymbol{s}) &\asymp \max_{1\le j\le l}\frac{r_n 2^j\sqrt{\log\log n} + \log\log n/\sqrt{n}}{r_n^{1-\alpha}2^{j(1-\alpha)}}\\ &\lesssim \delta_n^\alpha\sqrt{\log\log n} + \frac{\log\log n}{\sqrt{n}r_n^{1-\alpha}} = \mathfrak{o}(1)\end{aligned}$$

for our chosen r_n and δ_n. Now the claim follows from the first inequality of Proposition 4.5.12 and Lemma 4.5.13. The second claim follows by noting that $\beta_{n,q}(r_n,\delta_n) \ge C_2$ for some $C_2 > 0$ in view of Theorem 4.2.3, and using the same estimate for $\tau_{n,q}(r_n,\delta_n,\boldsymbol{s})$ as above. □

4.6 Proofs of the applications

In this section we prove Theorems 4.4.1-4.4.4.

4.6.1 Proof of Theorem 4.4.1

Lemma 4.6.1 (Proposition 2 of [76])**.** *Consider the regression model (4.4.1) and the least squares estimator $\hat{C}_n$ in (4.4.2). Suppose that $\xi_1, \dots, \xi_n$ are mean-zero random variables independent of $X_1, \dots, X_n$. Further assume that*

$$\mathbb{E} \sup_{C \in \mathscr{C}: |C \Delta C_0| \le \delta^2} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i (\mathbf{1}_C - \mathbf{1}_{C_0})(X_i)| \lesssim \phi_n(\delta),$$

$$\mathbb{E} \sup_{C \in \mathscr{C}: |C \Delta C_0| \le \delta^2} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (\mathbf{1}_C - \mathbf{1}_{C_0})(X_i)| \lesssim \phi_n(\delta),$$

hold for some ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta$ is non-increasing. Then $\mathbb{E}|\hat{C}_n \Delta C_0| = \mathcal{O}(\delta_n^2)$ holds for any δ_n such that $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$.

Proof of Theorem 4.4.1. By Lemma 4.6.1 the risk of the least squares estimator

$$\delta_n^2 \equiv \sup_{C_0 \in \mathscr{C}} \mathbb{E}_{C_0} |\hat{C}_n \Delta C_0| = \sup_{C_0 \in \mathscr{C}} \mathbb{E}_{C_0} (\mathbf{1}_{\hat{C}_n} - \mathbf{1}_{C_0})^2$$

can be solved by

$$\sup_{C_0 \in \mathscr{C}} \mathbb{E} \sup_{C \in \mathscr{C}: |C \Delta C_0| \le \delta_n^2} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta_i (\mathbf{1}_C - \mathbf{1}_{C_0})(X_i)| \lesssim \sqrt{n}\delta_n^2,$$

$$\sup_{C_0 \in \mathscr{C}} \mathbb{E} \sup_{C \in \mathscr{C}: |C \Delta C_0| \le \delta_n^2} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (\mathbf{1}_C - \mathbf{1}_{C_0})(X_i)| \lesssim \sqrt{n}\delta_n^2.$$

Since the global entropy estimate is invariant over shift, by Theorem 4.2.3, we obtain

$$\sup_{C_0 \in \mathscr{C}} \mathbb{E} \sup_{C \in \mathscr{C}: |C \Delta C_0| \le \delta_n^2} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta_i (\mathbf{1}_C - \mathbf{1}_{C_0})(X_i)| \lesssim \max\{\delta_n^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\}.$$

It is now easy to see that the choice $\delta_n^2 \asymp n^{-1/(\alpha+1)}$ leads to an upper bound of the the above display on the desired order $\sqrt{n}\delta_n^2$. The Gaussian randomized empirical process can be handled via the multiplier inequality Lemma 4.5.6 by letting $\psi_n(t) \equiv \psi(t) \equiv t^{\alpha/(\alpha+1)}$ and then use the fact that Gaussian random variables have infinitely many moments. We omit the details. □

4.6.2 Proof of Theorem 4.4.3

We need the following analogy of Lemma 4.6.1 before proving Theorem 4.4.3.

Proposition 4.6.2. *Consider the regression model (4.4.4) and the least squares estimator $\hat{C}_n$ in (4.4.5). Further assume that*

$$\mathbb{E} \sup_{C\in\mathscr{C}:|C\Delta C_0|\le\delta^2} |\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\mathbf{1}_C-\mathbf{1}_{C_0})(X_i)| \lesssim \phi_n(\delta),$$

holds for some ϕ_n such that $\delta\mapsto\phi_n(\delta)/\delta$ is non-increasing. Then $\mathbb{E}|\hat{C}_n\Delta C_0| = \mathcal{O}(\delta_n^2)$ holds for any δ_n such that $\phi_n(\delta_n)\le\sqrt{n}\delta_n^2$.

Proof. Let

$$\mathbb{M}_n f = \frac{2}{n}\sum_{i=1}^{n}(f-f_0)(X_i)\xi_i - \frac{1}{n}\sum_{i=1}^{n}(f-f_0)^2(X_i), Mf = -P(f-f_0)^2.$$

Here we used the fact that $\mathbb{E}[\xi_i|X_i]=0$. Then it is easy to see that

$$|\mathbb{M}_n f-\mathbb{M}_n f_0-(Mf-Mf_0)| \le |\frac{2}{n}\sum_{i=1}^{n}(f-f_0)(X_i)\xi_i| + |(\mathbb{P}_n-P)(f-f_0)^2|.$$

The first term can be handled by a conditioning argument with contraction principle for the empirical process so that ξ_i's are eliminated (i.e. the dependence structure of ξ_i's does not matter), while the second term is standard, and then we use the proof strategy of Proposition 2 in [76]. □

Proof of Theorem 4.4.3. The proof follows by Proposition 4.6.2 and similar arguments as the proof of Theorem 4.4.1. □

4.6.3 Proof of Theorem 4.4.4

We need some further notations. For any $g\in\mathcal{G}$, write $f_g(x,y)\equiv\mathbf{1}_{y\neq g(x)}$. Let $\mathcal{G}(\delta)\equiv\{g\in\mathcal{G}:\mathcal{E}_P(g)\le\delta\}$. Let l be the smallest integer such that $r_n^2 2^l\ge 1$, and for any $1\le j\le l$, let $\mathcal{F}_j\equiv\{f_{g_1}-f_{g_2}: g_1,g_2\in\mathcal{G}(r_n^2 2^j)\}$.

Lemma 4.6.3. *Suppose $\mathcal{G} \equiv \{\mathbf{1}_C : C \in \mathscr{C}\}$ satisfies the following entropy condition: there exists some $\alpha \neq 1$ such that for all $\varepsilon > 0$,*

$$\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \leq L\varepsilon^{-\alpha}.$$

Then

$$\mathbb{P}\left(\max_{1\leq j\leq l} \frac{\sup_{f\in\mathcal{F}_j}|\mathbb{P}_n(f) - P(f)|}{r_n^2 2^j} \geq c\left(\frac{1}{4} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right)\right) \leq K' \exp(-s/K')$$

holds for some constants $K, K' > 0$ provided $r_n^2 \cdot n^{1/(\alpha+1)} \gg 1$.

Proof. The proof is similar to that of Theorem 4.5.12. By Talagrand's concentration inequality, we see that

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}_j}|\mathbb{G}_n(f)| \geq K\left(\mathbb{E}\sup_{f\in\mathcal{F}_j}|\mathbb{G}_n(f)| + \sqrt{\sigma_j^2 s_j} + \frac{s_j}{\sqrt{n}}\right)\right] \leq K\exp(-s_j/K).$$

Let $\mathscr{S} \equiv \{S : f_g = \mathbf{1}_S, g \in \mathcal{G}\}$. Note that for $g_1 = \mathbf{1}_{C_1}, g_2 = \mathbf{1}_{C_2} \in \mathcal{G}$, where $C_1, C_2 \in \mathscr{C}$, we have $f_{g_1} = \mathbf{1}_{S_1}, f_{g_2} = \mathbf{1}_{S_2}$, and hence

$$P(S_1 \Delta S_2) = P(f_{g_1} - f_{g_2})^2 \leq P(g_1 - g_2)^2 = P(C_1 \Delta C_2).$$

This shows that $\mathcal{N}_I(\varepsilon, \mathscr{S}, P) \leq \mathcal{N}_I(\varepsilon, \mathscr{C}, P)$. Furthermore, for any $g \in \mathcal{G}(r_n^2 2^j)$, let $S \in \mathscr{S}$ be such that $f_g = \mathbf{1}_S$. Then similar to the above display, we have

$$P(S\Delta S_0) \leq \|g - g_0\|_{L_2(P)}^2 \leq c^{-1} r_n^2 2^j,$$

where the last inequality follows from the margin condition. Now by Theorem 4.2.1, we obtain

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}_j}|\mathbb{G}_n(f)| &\lesssim \mathbb{E}\sup_{g\in\mathcal{G}(r_n^2 2^j)}|\mathbb{G}_n(f_g)| \\
&\leq \mathbb{E}\sup_{S\in\mathscr{S}: P(S\Delta S_0)\leq c^{-1}r_n^2 2^j}|\mathbb{G}_n(S)| \\
&\lesssim \max\{(r_n^2 2^j)^{(1-\alpha)/2}, n^{(\alpha-1)/2(\alpha+1)}\}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}\sigma_j^2 &\equiv \sup_{f\in\mathcal{F}_j}\|f\|_{L_2(P)}^2 = \sup_{g_1,g_2\in\mathcal{G}(r_n^2 2^j)}\|f_{g_1}-f_{g_2}\|_{L_2(P)}^2 \\ &\le 4\sup_{g\in\mathcal{G}(r_n^2 2^j)}\|g-g_0\|_{L_2(P)}^2 \le 4c^{-1}\sup_{g\in\mathcal{G}(r_n^2 2^j)}\mathcal{E}_P(g) \le 4c^{-1}r_n^2 2^j.\end{aligned}$$

This implies that with $s_j = s2^j$,

$$\begin{aligned}&\mathbb{P}\Bigg[\frac{\sup_{f\in\mathcal{F}_j}|\mathbb{P}_n(f)-P(f)|}{r_n^2 2^j} \\ &\qquad \ge K_c r_n^{-2}2^{-j}\Bigg(\max\left\{n^{-1/2}(r_n^2 2^j)^{(1-\alpha)/2}, n^{-1/(\alpha+1)}\right\} \\ &\qquad\qquad + n^{-1/2}(r_n^2 2^j)^{1/2}\sqrt{s}2^{j/2} + n^{-1}s2^j\Bigg)\Bigg] \\ &\le K\exp(-s2^j/K).\end{aligned}$$

Note that

$$\begin{aligned}&r_n^{-2}2^{-j}\left(\max\left\{n^{-1/2}(r_n^2 2^j)^{(1-\alpha)/2}, n^{-1/(\alpha+1)}\right\} + n^{-1/2}(r_n^2 2^j)^{1/2}\sqrt{s}2^{j/2} + n^{-1}s2^j\right) \\ &\le \max\left\{\frac{1}{\sqrt{n}r_n^{\alpha+1}}, \frac{1}{r_n^2 n^{1/(\alpha+1)}}\right\} + \sqrt{\frac{s}{nr_n^2}} + \frac{s}{nr_n^2} \\ &\le \frac{c}{4K_c} + \sqrt{\frac{s}{nr_n^2}} + \frac{s}{nr_n^2}\end{aligned}$$

under the assumption. Now a union bound leads to the desired claim. □

Proof of Theorem 4.4.4. Given the estimate in Lemma 4.6.3, the proof of the theorem closely follows that of Theorem 7.1 of [59]. We provide some details for the convenience of the reader. Then on the event

$$E \equiv \left\{\max_{1\le j\le l}\frac{\sup_{f\in\mathcal{F}_j}|\mathbb{P}_n(f)-P(f)|}{r_n^2 2^j} \le c\left(\frac{1}{4} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right)\right\},$$

we have for any $g \in \mathcal{G}(r_n^2 2^j) \setminus \mathcal{G}(r_n^2 2^{j-1})$ and $g' \in \mathcal{G}(\sigma)$ for some $0 < \sigma < r_n^2 2^j$,

$$\begin{aligned}
\mathcal{E}_P(g) &= P(f_g - f_{g'}) + \left[P(f_{g'}) - Pf_{g_0}\right] \\
&\leq P(f_g - f_{g'}) + \sigma \\
&\leq \mathbb{P}_n(f_g - f_{g'}) + \sigma + \|\mathbb{P}_n - P\|_{\mathcal{F}_j} \\
&\leq \mathcal{E}_{\mathbb{P}_n}(g) + \sigma + c\left(\frac{1}{4} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right) r_n^2 2^j \\
&\leq \mathcal{E}_{\mathbb{P}_n}(g) + \sigma + \left(\frac{1}{4} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right) 2\mathcal{E}_P(g).
\end{aligned}$$

Since $\sigma > 0$ is taken arbitrarily, we see that on the event E, it holds that

$$\frac{\mathcal{E}_{\mathbb{P}_n}(g)}{\mathcal{E}_P(g)} \geq 1 - \left(\frac{1}{2} + 2K\sqrt{\frac{s}{nr_n^2}} + 2K\frac{s}{nr_n^2}\right) \tag{4.6.1}$$

for all $g \in \mathcal{G}$ such that $\mathcal{E}_P(g) \geq r_n^2$. Furthermore, the above display entails that on the event E, we necessarily have $\mathcal{E}_P(\hat{g}_n) < r_n^2$ for n large enough. Hence for any $g \in \mathcal{G}(r_n^2 2^j) \setminus \mathcal{G}(r_n^2 2^{j-1})$, we have

$$\begin{aligned}
\mathcal{E}_{\mathbb{P}_n}(g) &= \mathbb{P}_n(f_g) - \mathbb{P}_n(f_{\hat{g}_n}) \\
&\leq Pf_g - Pf_{\hat{g}_n} + \|\mathbb{P}_n - P\|_{\mathcal{F}_j} \\
&\leq \mathcal{E}_P(g) + \left(\frac{1}{4} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right) 2\mathcal{E}_P(g).
\end{aligned}$$

This entails that

$$\frac{\mathcal{E}_{\mathbb{P}_n}(g)}{\mathcal{E}_P(g)} \leq 1 + \left(\frac{1}{2} + 2K\sqrt{\frac{s}{nr_n^2}} + 2K\frac{s}{nr_n^2}\right). \tag{4.6.2}$$

The proof of the claim is complete by combining (4.6.1)-(4.6.2) along with Lemma 4.6.3. □

4.7 Talagrand's concentration inequality

We frequently use Talagrand's concentration inequality [145] for the empirical process in the form given by [106] in the proofs. For sake of completeness, we record it as follows.

(Talagrand's concentration inequality) Let $\mathcal{F}$ be a countable class of real-valued measurable functions such that $\sup_{f\in\mathcal{F}}\|f\|_\infty \leq b$. Then

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}|\mathbb{G}_n f| \geq 2\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_n f| + \sqrt{8\sigma^2 x} + 34.5b\frac{x}{\sqrt{n}}\right) \leq e^{-x},$$

where $\sigma^2 \equiv \sup_{f \in \mathcal{F}} \mathrm{Var}_P f$, and $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$.

Chapter 5

LEAST SQUARES ESTIMATION IN NON-DONSKER MODELS II: MULTIVARIATE ISOTONIC MODEL

5.1 Introduction

Isotonic regression is perhaps the simplest form of shape-constrained estimation problem, and has wide applications in a number of fields. For instance, in medicine, the expression of a leukaemia antigen has been modelled as a monotone function of white blood cell count and DNA index [134], while in education, isotonic regression has been used to investigate the dependence of college grade point average on high school ranking and standardised test results [53]. A further application area for isotonic regression approaches has recently emerged in genetic heritability studies, where it is often generally accepted that phenotypes such as height, fitness or disease depend in a monotone way on genetic factors [102, 131, 97]. In these latter contexts, as an initial simplifying structure, it is natural to ignore potential genetic interactions and consider additive isotonic regression models; however, these have been found to be inadequate in several instances [135, 67, 54]. Alternative simplifying interaction structures have also been explored, including those based on products [55], logarithms [133] and minima [148], but the form of genetic interaction between factors is not always clear and may vary between phenotypes [102, 97].

Motivated by these considerations, we note that a general class of isotonic functions, which includes all of the above structures as special cases, is the class of block increasing functions

$$\mathcal{F}_d := \big\{ f : [0,1]^d \to \mathbb{R},\ f(x_1, \ldots, x_d) \leq f(x'_1, \ldots, x'_d) \\ \text{when } x_j \leq x'_j \text{ for } j = 1, \ldots, d \big\}.$$

In this paper, we suppose that we observe data $(X_1, Y_1), \ldots, (X_n, Y_n)$, with $n \geq 2$, satisfying

$$Y_i = f_0(X_i) + \xi_i, \quad i = 1, \ldots, n, \tag{5.1.1}$$

where $f_0 : [0,1]^d \to \mathbb{R}$ is Borel measurable, $\xi_1, \ldots, \xi_n$ are independent $N(0,1)$ noise, and the covariates $X_1, \ldots, X_n$, which take values in the set $[0,1]^d$, can either be fixed or random (independent of $\xi_1, \ldots, \xi_n$). Our goal is to study the performance of the least squares isotonic regression estimator $\hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}_d} \sum_{i=1}^n \{Y_i - f(X_i)\}^2$ in terms of its empirical risk

$$R_n(\hat{f}_n, f_0) := \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n \{\hat{f}_n(X_i) - f_0(X_i)\}^2\bigg]. \tag{5.1.2}$$

Note that this loss function only considers the errors made at the design points $X_1, \ldots, X_n$, and these points naturally induce a directed acyclic graph $G_X = (V(G_X), E(G_X))$ with $V(G_X) = \{1, \ldots, n\}$ and $E(G_X) = \{(i, i') : (X_i)_j \leq (X_{i'})_j \ \forall\, j = 1, \ldots, d\}$. It is therefore natural to restate the problem in terms of isotonic vector estimation on directed acyclic graphs. Recall that given a directed acyclic graph $G = (V(G), E(G))$, we may define a partially ordered set $(V(G), \leq)$, where $u \leq v$ if and only if there exists a directed path from u to v. We define the class of isotonic vectors on G by

$$\mathcal{M}(G) := \{\theta \in \mathbb{R}^{V(G)} : \theta_u \leq \theta_v \text{ for all } u \leq v\}.$$

Hence, for a signal vector $\theta_0 = ((\theta_0)_i)_{i=1}^n := (f_0(X_i))_{i=1}^n \in \mathcal{M}(G_X)$, the least squares estimator $\hat{\theta}_n = ((\hat{\theta}_n)_i)_{i=1}^n := (\hat{f}_n(X_i))_{i=1}^n$ can be seen as the projection of $(Y_i)_{i=1}^n$ onto the polyhedral convex cone $\mathcal{M}(G_X)$. Such a geometric interpretation means that least squares estimators for isotonic regression, in general dimensions or on generic directed acyclic graphs, can be efficiently computed using convex optimisation algorithms (see, e.g., [52, 88, 140]).

In the special case where $d = 1$, model (5.1.1) reduces to the univariate isotonic regression problem that has a long history [e.g. [27, 163, 15, 153, 154, 45, 21, 115, 50, 49, 167]]. See [69] for a general introduction. Since the risk only depends on the ordering of the design points in the univariate case, fixed and random designs are equivalent for $d = 1$ under the empirical risk function (5.1.2). It is customary to write $R_n(\hat{\theta}_n, \theta_0)$ in place of $R_n(\hat{f}_n, f_0)$ for

model (5.1.1) with fixed design points. When $(\theta_0)_1 \leq \cdots \leq (\theta_0)_n$ (i.e. $X_1 \leq \cdots \leq X_n$), [172] proved that for $d = 1$ there exists a universal constant $C > 0$ such that

$$R_n(\hat{\theta}_n, \theta_0) \leq C\bigg\{\bigg(\frac{(\theta_0)_n - (\theta_0)_1}{n}\bigg)^{2/3} + \frac{\log n}{n}\bigg\},$$

which shows in particular that the risk of the least squares estimator is no worse than $O(n^{-2/3})$ for signals θ_0 of bounded uniform norm. In recent years, there has been considerable interest and progress in studying the automatic rate-adaptation phenomenon of shape-constrained estimators. This line of study was pioneered by [172] in the context of univariate isotonic regression, followed by [33] and most recently [18], who proved that

$$R_n(\hat{\theta}_n, \theta_0) \leq \inf_{\theta \in \mathcal{M}(G_X)} \bigg\{\frac{\|\theta - \theta_0\|_2^2}{n} + \frac{k(\theta)}{n} \log\bigg(\frac{en}{k(\theta)}\bigg)\bigg\}, \tag{5.1.3}$$

where $k(\theta)$ is the number of constant pieces in the isotonic vector θ. The inequality (5.1.3) is often called a *sharp oracle inequality*, with the sharpness referring to the fact that the approximation error term $n^{-1}\|\theta - \theta_0\|_2^2$ has leading constant 1. The bound (5.1.3) shows nearly parametric adaptation of the least squares estimator in univariate isotonic regression when the underlying signal has a bounded number of constant pieces. Other examples of adaptation in univariate shape-constrained problems include the maximum likelihood estimator of a log-concave density [83], and the least squares estimator in unimodal regression [35].

Much less is known about the rate of convergence of the least squares estimator in the model (5.1.1), or indeed the adaptation phenomenon in shape-restricted problems more generally, in multivariate settings. The only work of which we are aware in the isotonic regression case is [34], which deals with the fixed, lattice design case when $d = 2$. For a general dimension d, and for $n_1, \ldots, n_d \in \mathbb{N}$, we define this lattice by $\mathbb{L}_{d,n_1,\ldots,n_d} := \prod_{j=1}^d \{1/n_j, 2/n_j, \ldots, 1\}$; when $n_1 = \ldots = n_d = n^{1/d}$ for some $n \in \mathbb{N}$, we also write $\mathbb{L}_{d,n} := \mathbb{L}_{d,n_1,\ldots,n_d}$ as shorthand. When $\{X_1, \ldots, X_n\} = \mathbb{L}_{2,n_1,n_2}$, [34] showed that there exists a universal constant $C > 0$ such that

$$R_n(\hat{\theta}_n, \theta_0) \leq C\bigg\{\frac{((\theta_0)_{n_1,n_2} - (\theta_0)_{1,1}) \log^4 n}{n^{1/2}} + \frac{\log^8 n}{n}\bigg\},$$

with a corresponding minimax lower bound of order $n^{-1/2}$ over classes of uniformly bounded signals. They also provided a sharp oracle inequality of the form

$$R_n(\hat{\theta}_n, \theta_0) \leq \inf_{\theta \in \mathcal{M}(\mathbb{L}_{2,n_1,n_2})} \left(\frac{\|\theta - \theta_0\|_2^2}{n} + \frac{Ck(\theta) \log^8 n}{n} \right), \tag{5.1.4}$$

where $k(\theta)$ is the minimal number of rectangular blocks into which $\mathbb{L}_{2,n_1,n_2}$ may be partitioned such that θ is constant on each rectangular block.

A separate line of work has generalised the univariate isotonic regression problem to multivariate settings by assuming an additive structure (see e.g. [10, 119, 101, 37]). In the simplest setting, these works investigate the regression problem (5.1.1), where the signal f_0 belongs to

$$\mathcal{F}_d^{\mathrm{add}} := \left\{ f \in \mathcal{F}_d : f(x_1, \ldots, x_d) = \sum_{j=1}^{d} f_j(x_j), f_j \in \mathcal{F}_1, \|f_j\|_\infty \leq 1 \right\}.$$

The additive structure greatly reduces the complexity of the class; indeed, it can be shown that the least squares estimator over $\mathcal{F}_d^{\mathrm{add}}$ attains the univariate risk $n^{-2/3}$, up to multiplicative constants depending on d [e.g. Theorem 9.1 [160]].

The main contribution of this paper is to provide risk bounds for the isotonic least squares estimator when $d \geq 3$, both from a worst-case perspective and an adaptation point of view. Specifically, we show that in the fixed lattice design case, the least squares estimator satisfies

$$\sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}), \|\theta_0\|_\infty \leq 1} R_n(\hat{\theta}_n, \theta_0) \leq C n^{-1/d} \log^4 n, \tag{5.1.5}$$

for some universal constant $C > 0$. This rate turns out to be the minimax risk up to poly-logarithmic factors in this problem. Furthermore, we establish a sharp oracle inequality: there exists a universal constant $C > 0$ such that for every $\theta_0 \in \mathbb{R}^{\mathbb{L}_{d,n}}$,

$$R_n(\hat{\theta}_n, \theta_0) \leq \inf_{\theta \in \mathcal{M}(\mathbb{L}_{d,n})} \left\{ \frac{\|\theta - \theta_0\|_2^2}{n} + C \left(\frac{k(\theta)}{n} \right)^{2/d} \log^8 \left(\frac{en}{k(\theta)} \right) \right\}, \tag{5.1.6}$$

where $k(\theta)$ is the number of constant hyperrectangular pieces in θ. This reveals an adaptation rate of nearly $(k/n)^{2/d}$ for signals that are close to an element of $\mathcal{M}(\mathbb{L}_{d,n})$ that has at most

k hyperrectangular blocks. A corresponding lower bound is also provided, showing that the least squares estimator cannot adapt faster than the $n^{-2/d}$ rate implied by (5.1.6) even for constant signal vectors. Some intuition for this rate is provided by the notion of *statistical dimension*, which can be thought of as a measure of complexity of the underlying parameter space; see (5.2.2) below for a formal definition. A key step in the proof of (5.1.6) is to observe that for $d \geq 2$, the statistical dimension of $\mathcal{M}(\mathbb{L}_{d,n})$ is of order $n^{1-2/d}$ up to poly-logarithmic factors; see Table 5.1. The adaptation rate in (5.1.6), at least in the constant signal case, can therefore be understood as the ratio of the statistical dimension to the sample size. This reasoning is developed and discussed in greater detail at the end of Section 5.2.

We further demonstrate that analogues of the worst-case bounds and oracle inequalities (5.1.5) and (5.1.6), with slightly different poly-logarithmic exponents, remain valid for random design points $X_1, \ldots, X_n$ sampled independently from a distribution P on $[0,1]^d$ with a Lebesgue density bounded away from 0 and ∞. Such random design settings arguably occur more frequently in practice (cf. the examples given at the beginning of this introduction) and are particularly natural in high dimensions, where sampling design points on a fixed lattice is rarely feasible or even desirable. Nevertheless, we are not aware of any previous works on isotonic regression with random design even for $d = 2$; this is undoubtedly due to the increased technical challenges (described in detail after the statement of Theorem 5.3.2 in Section 5.3) that arise in handling the relevant empirical processes.

In addition to the risk $R_n(\hat{f}_n, f_0)$ in (5.1.2), for random designs we also study the natural population squared risk

$$R(\hat{f}_n, f_0) := \mathbb{E}\|\hat{f}_n - f_0\|^2_{L_2(P)} = \mathbb{E}\big[\big\{\hat{f}_n(X) - f_0(X)\big\}^2\big],$$

where $(X, Y) \overset{d}{=} (X_1, Y_1)$ and is independent of $(X_1, Y_1), \ldots, (X_n, Y_n)$. We note that the quantity $\mathbb{E}\big[\big\{Y - \hat{f}_n(X)\big\}^2\big]$, often referred to as the generalisation error for squared error loss in the machine learning literature, is simply equal to $1 + R(\hat{f}_n, f_0)$ in our context. Both our upper and lower bounds for the $R(\hat{f}_n, f_0)$ are broadly similar to the $R_n(\hat{f}_n, f_0)$ setting, though the proofs are very different (and quite intricate), and we incur an additional multiplicative

factor of order $\log n$ for the approximation error term in the oracle inequality.

Our results in both the fixed and random design settings are surprising in particular with regard to the following two aspects:

1. The negative results of [21] have spawned a heuristic belief that one should not use global empirical risk minimisation procedures[1] when the entropy integral for the corresponding function class diverges (e.g. [160] page 121-122, [127]). It is therefore of particular interest to see that in our isotonic regression function setting, the global least squares estimator is still rate optimal (up to poly-logarithmic factors). See also the discussion after Corollary 5.2.3.

2. Sharp adaptive behaviour for shape-constrained estimators has previously only been shown when the adaptive rate is nearly parametric (see, e.g., [72, 33, 18, 83]). On the other hand, our results here show that the least squares estimator in the d-dimensional isotonic regression problem necessarily adapts at a strictly nonparametric rate. Clearly, the minimax optimal rate for constant functions is parametric. Hence, the least squares estimator in this problem adapts at a strictly suboptimal rate while at the same time being nearly rate optimal from a worst-case perspective.

In both the fixed lattice design and the more challenging random design cases, our analyses are based on a novel combination of techniques from empirical process theory, convex geometry and combinatorics. We hope these methods can serve as a useful starting point towards understanding the behaviour of estimators in other multivariate shape-restricted models.

The rest of the paper is organised as follows. In Section 5.2, we state the main results for the fixed lattice design model. Section 5.3 describes corresponding results in the random design case. Proofs of all main theoretical results are contained in Sections 5.4-5.7.

---

[1]The term 'global' refers here to procedures that involve minimisation over the entire function class, as opposed to only over a sieve; cf. [160].

5.1.1 Some further notation

For any Borel measurable $\mathcal{R} \subseteq \mathcal{X}$, we write $\|f\|_{L_p(P;\mathcal{R})} := \left(\int_{\mathcal{R}} |f|^p \, \mathrm{d}P\right)^{1/p}$. For $r \geq 0$, we write $B_p(r, P) := \{f : \mathcal{X} \to \mathbb{R}, \|f\|_{L_p(P)} \leq r\}$ and $B_\infty(r) := \{f : \mathcal{X} \to \mathbb{R}, \|f\|_\infty \leq r\}$. We will abuse notation slightly and also write $B_p(r) := \{v \in \mathbb{R}^n : \|v\|_p \leq r\}$ for $p \in [1, \infty]$. The Euclidean inner product on $\mathbb{R}^d$ is denoted by $\langle \cdot, \cdot \rangle$. For $x, y \in \mathbb{R}^d$, we write $x \preceq y$ if $x_j \leq y_j$ for all $j = 1, \ldots, d$.

Throughout the article $\xi_1, \ldots, \xi_n$ and $\{\xi_w : w \in \mathbb{L}_{d,n_1,\ldots,n_d}\}$ denote independent standard normal random variables and $\varepsilon_1, \ldots, \varepsilon_n$ denote independent Rademacher random variables, both independent of all other random variables. For two probability measures P and Q defined on the same measurable space $(\mathcal{X}, \mathcal{A})$, we write $d_{\mathrm{TV}}(P, Q) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$ for their total variation distance, and $d^2_{\mathrm{KL}}(P, Q) := \int_{\mathcal{X}} \log \frac{\mathrm{d}P}{\mathrm{d}Q} \, \mathrm{d}P$ for their Kullback–Leibler divergence.

5.2 Fixed lattice design

In this section, we focus on the model (5.1.1) in the case where the set of design points forms a finite cubic lattice $\mathbb{L}_{d,n}$, defined in the introduction. In particular, we will assume in this section that $n = n_1^d$ for some $n_1 \in \mathbb{N}$. We use the same notation $\mathbb{L}_{d,n}$ both for the set of points and the directed acyclic graph on these points with edge structure arising from the natural partial ordering induced by $\preceq$. Thus, in the case $d = 1$, the graph $\mathbb{L}_{1,n}$ is simply a directed path, and this is the classical univariate isotonic regression setting. The case $d = 2$ is studied in detail in [34]. Our main interest lies in the cases $d \geq 3$.

5.2.1 Worst-case rate of the least squares estimator

Our first result provides an upper bound on the risk of the least squares estimator $\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$ of $\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n})$.

Theorem 5.2.1. *Let $d \geq 2$. There exists a universal constant $C > 0$ such that*

$$\sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} R_n(\hat{\theta}_n, \theta_0) \leq C n^{-1/d} \log^4 n.$$

Theorem 5.2.1 reveals that, up to a poly-logarithmic factor, the empirical risk of the least squares estimator converges to zero at rate $n^{-1/d}$. The upper bound in Theorem 5.2.1 is matched, up to poly-logarithmic factors, by the following minimax lower bound.

Proposition 5.2.2. *There exists a constant $c_d > 0$, depending only on d, such that for $d \geq 2$,*

$$\inf_{\tilde{\theta}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} R_n(\tilde{\theta}_n, \theta_0) \geq c_d n^{-1/d},$$

where the infimum is taken over all estimators $\tilde{\theta}_n = \tilde{\theta}_n(Y_1, \ldots, Y_n)$ of θ_0.

Recall that, given a directed acyclic graph $G = (V, E)$, a *chain* in G of cardinality L is a directed path of the form $(i_1, \ldots, i_L)$, where $(i_j, i_{j+1}) \in E$ for each $j = 1, \ldots, L-1$; an *antichain* in G of cardinality L is a subset $\{i_1, \ldots, i_L\}$ of V such that for each distinct $j, j' \in \{1, \ldots, L\}$ there is no chain containing both i_j and $i_{j'}$. A key observation in the proof of Proposition 5.2.2 is that $\mathbb{L}_{d,n}$ contains a large antichain of size $L \gtrsim_d n^{1-1/d}$. As design points in the antichain are mutually incomparable, an intuitive explanation for the lower bound in Proposition 5.2.2 comes from the fact that we have L unconstrained parameters in $[-1, 1]$ to estimate from n observations, which translates to a rate at least of order L/n. From Theorem 5.2.1 and Proposition 5.2.2, together with existing results mentioned in the introduction for the case $d = 1$, we see that the worst-case risk $n^{-\min\{2/(d+2), 1/d\}}$ (up to poly-logarithmic factors) of the least squares estimator exhibits different rates of convergence in dimension $d = 1$ and dimensions $d \geq 3$, with $d = 2$ being a transitional case. From the proof of Proposition 5.2.2, we see that it is the competition between the cardinality of the maximum chain in G_X and the cardinality of the maximum antichain in G_X that explains the different rates. Similar transitional behaviour was recently observed by [82] in the context of log-concave density estimation, though there it is the tension between estimating the density in the interior of its support and estimating the support itself that drives the transition.

The two results above can readily be translated into bounds for the rate of convergence for estimation of a block monotonic function with a fixed lattice design. Recall that $\mathcal{F}_d$ is the class of block increasing functions. Suppose that for some $f_0 \in \mathcal{F}_d$, and at each $x \in \mathbb{L}_{d,n}$, we observe $Y(x) \sim N(f_0(x), 1)$ independently. Define $P_n := n^{-1} \sum_{x \in \mathbb{L}_{d,n}} \delta_x$ and let $\mathcal{A}$ denote the set of hypercubes of the form $A = \prod_{j=1}^d A_j$, where either $A_j = [0, \frac{1}{n_1}]$ or $A_j = (\frac{i_j - 1}{n_1}, \frac{i_j}{n_1}]$ for some $i_j \in \{2, \ldots, n_1\}$. Now let $\mathcal{H}$ denote the set of functions $f \in \mathcal{F}_d$ that are piecewise constant on each $A \in \mathcal{A}$, and set $\hat{f}_n := \operatorname{argmin}_{f \in \mathcal{H}} \sum_{x \in \mathbb{L}_{d,n}} \{Y(x) - f(x)\}^2$. The following is a fairly straightforward corollary of Theorem 5.2.1 and Proposition 5.2.2.

Corollary 5.2.3. *There exist $c_d, C_d > 0$, depending only on d, such that for $Q = P_n$ or Lebesgue measure on $[0,1]^d$, we have*

$$\begin{aligned} c_d n^{-1/d} &\leq \inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E} \|\tilde{f}_n - f_0\|_{L_2(Q)}^2 \\ &\leq \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E} \|\hat{f}_n - f_0\|_{L_2(Q)}^2 \leq C_d n^{-1/d} \log^4 n, \end{aligned}$$

where the infimum is over all measurable functions of $\{Y(x) : x \in \mathbb{L}_{d,n}\}$.

This corollary is surprising for the following reason. [57] Theorem 1.1 proved that when $d \geq 3$ and Q denotes Lebesgue measure on $[0,1]^d$,

$$\log N\big(\varepsilon, \mathcal{F}_d \cap B_\infty(1), \|\cdot\|_{L_2(Q)}\big) \asymp_d \varepsilon^{-2(d-1)}. \tag{5.2.1}$$

In particular, for $d \geq 3$, the classes $\mathcal{F}_d \cap B_\infty(1)$ are massive in the sense that the entropy integral $\int_\delta^1 \log^{1/2} N(\varepsilon, \mathcal{F}_d \cap B_\infty(1), \|\cdot\|_{L_2(Q)}) \, \mathrm{d}\varepsilon$ diverges at a polynomial rate in δ^{-1} as $\delta \searrow 0$. To the best of our knowledge, this is the first example of a setting where a global empirical risk minimisation procedure has been proved to attain (nearly) the minimax rate of convergence over such massive parameter spaces.

5.2.2 Sharp oracle inequality

In this subsection, we consider the adaptation behaviour of the least squares estimator in dimensions $d \geq 2$ (again, the $d = 2$ case is covered in [34]). Our main result is the sharp

oracle inequality in Theorem 5.2.4 below. We call a set in $\mathbb{R}^d$ a hyperrectangle if it is of the form $\prod_{j=1}^d I_j$ where $I_j \subseteq \mathbb{R}$ is an interval for each $j = 1, \ldots, d$. If $A = \prod_{j=1}^d [a_j, b_j]$ where $|\{j : b_j = a_j\}| \geq d - 2$, then we say A is a *two-dimensional sheet*. A two-dimensional sheet is therefore a special type of hyperrectangle whose intrinsic dimension is at most two. For $\theta \in \mathcal{M}(\mathbb{L}_{d,n})$, let $K(\theta)$ denote the smallest K such that $\mathbb{L}_{d,n} \subseteq \sqcup_{\ell=1}^K A_\ell$, where $A_1, \ldots, A_K$ are disjoint two-dimensional sheets and the restricted vector $\theta_{A_\ell \cap \mathbb{L}_{d,n}}$ is constant for each $\ell = 1, \ldots, K$.

Theorem 5.2.4. *Let $d \geq 2$. There exists a universal constant $C > 0$ such that for every $\theta_0 \in \mathbb{R}^{\mathbb{L}_{d,n}}$,*

$$R_n(\hat{\theta}_n, \theta_0) \leq \inf_{\theta \in \mathcal{M}(\mathbb{L}_{d,n})} \left\{ \frac{\|\theta - \theta_0\|_2^2}{n} + \frac{CK(\theta)}{n} \log_+^8 \left(\frac{n}{K(\theta)} \right) \right\}.$$

We remark that Theorem 5.2.4 does not imply (nearly) parametric adaptation when $d \geq 3$. This is because even when θ_0 is constant on $\mathbb{L}_{d,n}$ for every n, we have $K(\theta_0) = n^{(d-2)/d} \to \infty$ as $n \to \infty$. The following corollary of Theorem 5.2.4 gives an alternative (weaker) form of oracle inequality that offers easier comparison to lower dimensional results given in (5.1.3) and (5.1.4). Let $\mathcal{M}^{(k)}(\mathbb{L}_{d,n})$ be the collection of all $\theta \in \mathcal{M}(\mathbb{L}_{d,n})$ such that there exist disjoint hyperrectangles $\mathcal{R}_1, \ldots, \mathcal{R}_k$ with the properties that $\mathbb{L}_{d,n} \subseteq \sqcup_{\ell=1}^k \mathcal{R}_\ell$ and that for each ℓ, the restricted vector $\theta_{\mathcal{R}_\ell \cap \mathbb{L}_{d,n}}$ is constant.

Theorem 5.2.5. *Let $d \geq 2$. There exists a universal constant $C > 0$ such that for every $\theta_0 \in \mathbb{R}^{\mathbb{L}_{d,n}}$,*

$$R_n(\hat{\theta}_n, \theta_0) \leq \inf_{k \in \mathbb{N}} \left\{ \inf_{\theta \in \mathcal{M}^{(k)}(\mathbb{L}_{d,n})} \frac{\|\theta - \theta_0\|_2^2}{n} + C \left(\frac{k}{n} \right)^{2/d} \log_+^8 \left(\frac{n}{k} \right) \right\}.$$

It is important to note that both Theorems 5.2.4 and 5.2.5 allow for model misspecification, as it is not assumed that $\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n})$. For signal vectors θ_0 that are piecewise constant on k hyperrectangles, Theorem 5.2.5 provides an upper bound of the risk of order $(k/n)^{2/d}$ up to poly-logarithmic factors. The following proposition shows that even for a constant signal vector, the adaptation rate of $n^{-2/d}$ given in Theorem 5.2.5 cannot be improved.

Proposition 5.2.6. *Let $d \geq 2$. There exists a constant $c_d > 0$, depending only on d, such that for any $\theta_0 \in \mathcal{M}^{(1)}(\mathbb{L}_{d,n})$,*

$$R_n(\hat{\theta}_n, \theta_0) \geq c_d \begin{cases} n^{-1} \log^2 n & \text{if } d = 2 \\ n^{-2/d} & \text{if } d \geq 3. \end{cases}$$

The case $d = 2$ of this result is new, and reveals both a difference with the univariate situation, where the adaptation rate is of order $n^{-1} \log n$ [18], and that a poly-logarithmic penalty relative to the parametric rate is unavoidable for the least squares estimator. Moreover, we see from Proposition 5.2.6 that for $d \geq 3$, although the least squares estimator achieves a faster rate of convergence than the worst-case bound in Theorem 5.2.1 on constant signal vectors, the rate is not parametric, as would have been the case for a minimax optimal estimator over the set of constant vectors. This is in stark contrast to the nearly parametric adaptation results established in (5.1.3) and (5.1.4) for dimensions $d \leq 2$.

Another interesting aspect of these results relates to the notion of *statistical dimension*, defined for an arbitrary cone C in $\mathbb{R}^n$ by[2]

$$\delta(C) := \int_{\mathbb{R}^n} \|\Pi_C(x)\|_2^2 (2\pi)^{-n/2} e^{-\|x\|_2^2/2} \, \mathrm{d}x, \tag{5.2.2}$$

where Π_C is the projection onto the set C [6]. The proofs of Theorem 5.2.5 and Proposition 5.2.6 reveal a type of phase transition phenomenon for the statistical dimension $\delta(\mathcal{M}(\mathbb{L}_{d,n})) = R_n(\hat{\theta}_n, 0)$ of the monotone cone (cf. Table 5.1).

The following corollary of Theorem 5.2.4 gives another example where different adaptation behaviour is observed in dimensions $d \geq 3$, in the sense that the $n^{-2/d} \log^8 n$ adaptive rate achieved for constant signal vectors is actually available for a much wider class of isotonic signals that depend only on $d - 2$ of all d coordinates of $\mathbb{L}_{d,n}$. For $r = 0, 1, \ldots, d$, we say a vector $\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n})$ is a *function of r variables*, written $\theta_0 \in \mathcal{M}_r(\mathbb{L}_{d,n})$, if there exists $\mathcal{J} \subseteq \{1, \ldots, d\}$, of cardinality r, such that $(\theta_0)_{(x_1,\ldots,x_d)} = (\theta_0)_{(x'_1,\ldots,x'_d)}$ whenever $x_j = x'_j$ for all $j \in \mathcal{J}$.

[2]Our reason for defining the statistical dimension via an integral rather than as $\mathbb{E}\|\Pi_C(\xi)\|_2^2$ is because, in the random design setting, the cone C is itself random, and in that case $\delta(C)$ is a random quantity.

Table 5.1: Bounds* for $\delta\big(\mathcal{M}(\mathbb{L}_{d,n})\big)$.

d	upper bound	lower bound
1	$\sum_{i=1}^{n} i^{-1}$ †	$\sum_{i=1}^{n} i^{-1}$ †
2	$\lesssim \log^8 n$ ‡	$\gtrsim \log^2 n$
≥ 3	$\lesssim n^{1-2/d} \log^8 n$	$\gtrsim_d n^{1-2/d}$

* Entries without a reference are proved in this paper.

† [6]

‡ [34]

Corollary 5.2.7. *For $d \geq 2$, there exists constant $C_d > 0$, depending only on d, such that*

$$\sup_{\theta_0 \in \mathcal{M}_r(\mathbb{L}_{d,n}) \cap B_\infty(1)} R_n(\hat{\theta}_n, \theta_0) \leq C_d \begin{cases} n^{-2/d} \log^8 n & \text{if } r \leq d-2 \\ n^{-4/(3d)} \log^{16/3} n & \text{if } r = d-1 \\ n^{-1/d} \log^4 n & \text{if } r = d. \end{cases}$$

If the signal vector θ_0 belongs to $\mathcal{M}_r(\mathbb{L}_{d,n})$, then it is intrinsically an r-dimensional isotonic signal. Corollary 5.2.7 demonstrates that the least squares estimator exhibits three different levels of adaptation when the signal is a function of $d, d-1, d-2$ variables respectively. However, viewed together with Proposition 5.2.6, Corollary 5.2.7 shows that no further adaptation for the least squares estimator is available when the intrinsic dimension of the signal vector decreases further. Moreover, if we let $\tilde{n} = n^{2/d}$ denote the maximum cardinality of the intersection of $\mathbb{L}_{d,n}$ with a two-dimensional sheet, then the three levels of adaptive rates in Corollary 5.2.7 are $\tilde{n}^{-1}$, $\tilde{n}^{-2/3}$ and $\tilde{n}^{-1/2}$ respectively, up to poly-logarithmic factors, matching the two-dimensional 'automatic variable adaptation' result described in [34]

Theorem 2.4. In this sense, the adaptation of the isotonic least squares estimator in general dimensions is essentially a two-dimensional phenomenon.

5.3 Random design

In this section, we consider the setting where the design points $X_1, \ldots, X_n$ are independent and identically distributed from some distribution P supported on the unit cube $[0,1]^d$. We will assume throughout that P has Lebesgue density p_0 such that $0 < m_0 \leq \inf_{x \in [0,1]^d} p_0(x) \leq \sup_{x \in [0,1]^d} p_0(x) \leq M_0 < \infty$. Since the least squares estimator $\hat{f}_n$ is only well-defined on $X_1, \ldots, X_n$, for definiteness, we extend $\hat{f}_n$ to $[0,1]^d$ by defining $\hat{f}_n(x) := \min\big(\{\hat{f}_n(X_i) : 1 \leq i \leq n, X_i \succeq x\} \cup \{\max_i \hat{f}_n(X_i)\}\big)$. If we let $\mathbb{P}_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$, then we can consider the empirical and population risks $R_n(\hat{f}_n, f_0) = \mathbb{E}\|\hat{f}_n - f_0\|^2_{L_2(\mathbb{P}_n)}$ and $R(\hat{f}_n, f_0) = \mathbb{E}\|\hat{f}_n - f_0\|^2_{L_2(P)}$.

The main results of this section are the following two theorems, establishing respectively the worst-case performance and the adaptation behaviour for the least squares estimator in the random design setting. We write $\mathcal{F}_d^{(k)}$ for the class of functions in $\mathcal{F}_d$ that are piecewise constant on k hyperrectangular pieces. In other words, if $f \in \mathcal{F}_d^{(k)}$, then there exists a partition $[0,1]^d = \sqcup_{\ell=1}^k \mathcal{R}_\ell$, such that each $\mathcal{R}_\ell$ is a hyperrectangle and f is a constant function when restricted to each $\mathcal{R}_\ell$. Let $\gamma_2 := 9/2$ and $\gamma_d := (d^2+d+1)/2$ for $d \geq 3$.

Theorem 5.3.1. *Let $d \geq 2$. There exists $C_{d,m_0,M_0} > 0$, depending only on d, m_0 and M_0, such that*

$$\sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \max\big\{R(\hat{f}_n, f_0)\,,\, R_n(\hat{f}_n, f_0)\big\} \leq C_{d,m_0,M_0} n^{-1/d} \log^{\gamma_d} n.$$

Theorem 5.3.2. *Fix $d \geq 2$, and a Borel measurable function $f_0 : [0,1]^d \to \mathbb{R}$. There exists $C_{d,m_0,M_0} > 0$, depending only on d, m_0 and M_0, such that*

$$R_n(\hat{f}_n, f_0) \leq \inf_{k \in \mathbb{N}} \bigg\{ \inf_{f \in \mathcal{F}_d^{(k)}} \|f - f_0\|^2_{L_2(P)} + C_{d,m_0,M_0} \bigg(\frac{k}{n}\bigg)^{2/d} \log_+^{2\gamma_d}\bigg(\frac{n}{k}\bigg) \bigg\}.$$

On the other hand, if we also have $\|f_0\|_\infty \leq 1$, then there exists a universal constant $C > 0$ such that

$$R(\hat{f}_n, f_0) \leq \inf_{k \in \mathbb{N}} \bigg\{ C \log n \inf_{f \in \mathcal{F}_d^{(k)}} \|f - f_0\|^2_{L_2(P)} + C_{d,m_0,M_0} \bigg(\frac{k}{n}\bigg)^{2/d} \log^{2\gamma_d} n \bigg\}.$$

To the best of our knowledge, the bound in $L_2(\mathbb{P}_n)$ risk in Theorem 5.3.2 is the first sharp oracle inequality in the shape-constrained regression literature with random design. The different norms on the left- and right-hand sides for the $R_n(\hat{f}_n, f_0)$ bound arise from the observation that $\mathbb{E}\|f - f_0\|^2_{L_2(\mathbb{P}_n)} = \|f - f_0\|^2_{L_2(P)}$ for $f \in \mathcal{F}_d^{(k)}$. For the $R(\hat{f}_n, f_0)$ bound, the norms on both sides are the same, but we pay a price of a multiplicative factor of order $\log n$ for the approximation error.

The proofs of Theorems 5.3.1 and 5.3.2 are considerably more involved than those of the corresponding Theorems 5.2.1 and 5.2.4 in Section 5.2. We briefly mention two major technical difficulties:

1. The size of $\mathcal{F}_d$, as measured by its entropy, is large when $d \geq 3$, even after L_∞ truncation (cf. (5.2.1)). As rates obtained from the entropy integral [e.g. [160] Theorem 9.1] do not match those from Sudakov lower bounds for such classes, standard entropy methods result in a non-trivial gap between the minimax rates of convergence, which typically match the Sudakov lower bounds [e.g. [169] Proposition 1], and provable risk upper bounds for least squares estimators when $d \geq 3$.

2. In the fixed lattice design case, our analysis circumvents the difficulties of standard entropy methods by using the fact that a d-dimensional cubic lattice can be decomposed into a union of lower-dimensional pieces. This crucial property is no longer valid when the design is random.

We do not claim any optimality of the power in the poly-logarithmic factor in Theorems 5.3.1 and 5.3.2. On the other hand, similar to the fixed, lattice design case, the worst-case rate of order $n^{-1/d}$ up to poly-logarithmic factors cannot be improved, as can be seen from the proposition below.

Proposition 5.3.3. *Let $d \geq 2$. There exists a constant $c_{d,m_0,M_0} > 0$, depending only on d, m_0 and M_0, such that,*

$$\inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \min\big\{R(\tilde{f}_n, f_0)\,,\, R_n(\tilde{f}_n, f_0)\big\} \geq c_{d,m_0,M_0} n^{-1/d},$$

where the infimum is taken over all measurable functions $\tilde{f}_n$ of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

We can also provide lower bounds on the adaptation rate risks for the least squares estimator when f_0 is constant.

Proposition 5.3.4. *Let $d \geq 2$. There exists a constant $c_{d,M_0} > 0$, depending only on d and M_0, such that for any $f_0 \in \mathcal{F}_d^{(1)}$,*

$$R_n(\hat{f}_n, f_0) \geq c_{d,M_0} n^{-2/d}.$$

On the other hand, when $d \geq 2$, there exist a universal constant $c_2 > 0$ and $c_{d,m_0,M_0} > 0$ for $d \geq 3$, depending only on d, m_0 and M_0, such that for any $f_0 \in \mathcal{F}_d^{(1)}$,

$$R(\hat{f}_n, f_0) \geq \begin{cases} c_2 n^{-1} & \text{for } d = 2 \\ c_{d,m_0,M_0} n^{-2/d} \log^{-2\gamma_d} n & \text{for } d \geq 3. \end{cases}$$

A key step in proving the first part of Proposition 5.3.4 is to establish that with high probability, the cardinality of the maximum antichain in G_X is at least of order $n^{1-1/d}$. When $d = 2$, the distribution of this maximum cardinality is the same as the distribution of the length of the longest decreasing subsequence of a uniform permutation of $\{1, \ldots, n\}$, a famous object of study in probability and combinatorics. See [130] and references therein.

5.4 Proofs of results in Section 5.2

Throughout this section, $\xi = (\xi_w)_{w \in \mathbb{L}_{d,n_1,\ldots,n_d}}$ denotes a vector of independent standard normal random variables. It is now well understood that the risk of the least squares estimator in the Gaussian sequence model is completely characterised by the size of a localised Gaussian process; cf. [36]. The additional cone property of $\mathcal{M}(\mathbb{L}_{d,n})$ makes the reduction even simpler: we only need to evaluate the Gaussian complexity of $\mathcal{M}(\mathbb{L}_{d,n}) \cap B_2(1)$, where the *Gaussian complexity* of $T \subseteq \mathbb{R}^{\mathbb{L}_{d,n_1,\ldots,n_d}}$ is defined as $w_T := \mathbb{E} \sup_{\theta \in T} \langle \xi, \theta \rangle$. Thus the result in the following proposition constitutes a key ingredient in analysing the risk of the least squares estimator.

Proposition 5.4.1. *There exists a universal constant $C > 0$ such that for $d \geq 2$ and every $1 \leq n_1 \leq \cdots \leq n_d$ with $\prod_{j=1}^d n_j = n$, we have*

$$\frac{\sqrt{2/\pi}}{(d-1)^{d-1}} n_1^{d-1} n^{-1/2} \leq \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d}) \cap B_2(1)} \langle \xi, \theta \rangle \leq C \sqrt{\frac{n}{n_{d-1} n_d}} \log^4 n.$$

We remark that in the case $n_1 = \cdots = n_d = n^{1/d}$, we have $n_1^{d-1} n^{-1/2} = \sqrt{\frac{n}{n_{d-1} n_d}} = n^{1/2-1/d}$. Also, from the symmetry of the problem, we see that the restriction that $n_1 \leq \cdots \leq n_d$ is not essential. In the general case, for the lower bound, n_1 should be replaced with $\min_j n_j$, while in the upper bound, $n_{d-1} n_d$ should be replaced with the product of the two largest elements of $\{n_1, \ldots, n_d\}$ (considered here as a multiset).

Proof. We first prove the lower bound. Consider $W := \{w \in \mathbb{L}_{d,n_1,\ldots,n_d} : \sum_{j=1}^d n_j w_j = n_1\}$, $W^+ := \{w \in \mathbb{L}_{d,n_1,\ldots,n_d} : \sum_{j=1}^d n_j w_j > n_1\}$ and $W^- := \{w \in \mathbb{L}_{d,n_1,\ldots,n_d} : \sum_{j=1}^d n_j w_j < n_1\}$. For each realisation of the Gaussian random vector $\xi = (\xi_w)_{w \in \mathbb{L}_{d,n_1,\ldots,n_d}}$, we define $\theta(\xi) = (\theta_w(\xi))_{w \in \mathbb{L}_{d,n_1,\ldots,n_d}} \in \mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d})$ by

$$\theta_w := \begin{cases} 1 & \text{if } w \in W^+ \\ \operatorname{sgn}(\xi_w) & \text{if } w \in W \\ -1 & \text{if } w \in W^-. \end{cases}$$

Since $\|\theta(\xi)\|_2^2 = n$, it follows that

$$\begin{aligned} \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d}) \cap B_2(1)} \langle \xi, \theta \rangle &\geq \mathbb{E} \left\langle \xi, \frac{\theta(\xi)}{\|\theta(\xi)\|_2} \right\rangle \\ &= \frac{1}{n^{1/2}} \mathbb{E} \left(\sum_{w \in W^+} \xi_w - \sum_{w \in W^-} \xi_w + \sum_{w \in W} |\xi_w| \right) = \frac{\sqrt{2/\pi}}{n^{1/2}} |W|. \end{aligned}$$

The proof of the lower bound is now completed by noting that

$$|W| = \binom{n_1 - 1}{d-1} \geq \left(\frac{n_1 - 1}{d-1} \right)^{d-1}. \tag{5.4.1}$$

We next prove the upper bound. For $j = 1, \ldots, d-2$ and for $x_j \in \{1/n_j, 2/n_j, \ldots, 1\}$, we define $A_{x_1,\ldots,x_{d-2}} := \{w = (w_1, \ldots, w_d) \in \mathbb{L}_{d,n_1,\ldots,n_d} : (w_1, \ldots, w_{d-2}) = (x_1, \ldots, x_{d-2})\}$.

Each $A_{x_1,\ldots,x_{d-2}}$ can be viewed as a directed acyclic graph with graph structure inherited from $\mathbb{L}_{d,n_1,\ldots,n_d}$. Since monotonicity is preserved on subgraphs, we have that $\mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d}) \subseteq \bigoplus_{x_1,\ldots,x_{d-2}} \mathcal{M}(A_{x_1,\ldots,x_{d-2}})$. Hence, by the Cauchy–Schwarz inequality and [6] Proposition 3.1 (5,9,10), we obtain that

$$\begin{aligned}\biggl(\mathbb{E}\sup_{\theta\in\mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d})\cap B_2(1)}\langle\xi,\theta\rangle\biggr)^2 &\le \mathbb{E}\biggl\{\biggl(\sup_{\theta\in\mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d})\cap B_2(1)}\langle\xi,\theta\rangle\biggr)^2\biggr\}\\ &= \delta\bigl(\mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d})\bigr) \le \sum_{x_1,\ldots,x_{d-2}}\delta\bigl(\mathcal{M}(A_{x_1,\ldots,x_{d-2}})\bigr)\\ &= \delta\bigl(\mathcal{M}(\mathbb{L}_{2,n_{d-1},n_d})\bigr)\prod_{j=1}^{d-2}n_j \lesssim \frac{n}{n_{d-1}n_d}\log_+^8(n_{d-1}n_d),\end{aligned}$$

as desired. Here, the final inequality follows from [34] Theorem 2.1 by setting $\boldsymbol{\theta}^* = 0$ (in their notation) and observing that $\delta\bigl(\mathcal{M}(\mathbb{L}_{2,n_{d-1},n_d})\bigr) = n_{d-1}n_d R_n(\hat{\theta}_n, 0) \lesssim \log_+^8(n_{d-1}n_d)$. □

Proof of Theorem 5.2.1. Fix $\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)$. We have by [36] Theorem 1.1 that the function

$$t \mapsto \mathbb{E}\sup_{\theta\in\mathcal{M}(\mathbb{L}_{d,n}),\|\theta-\theta_0\|\le t}\langle\xi,\theta-\theta_0\rangle - t^2/2$$

is strictly concave on $[0,\infty)$ with a unique maximum at, say, $t_0 \ge 0$. We note that $t_0 \le t_*$ for any t_* satisfying

$$\mathbb{E}\sup_{\theta\in\mathcal{M}(\mathbb{L}_{d,n}),\|\theta-\theta_0\|\le t_*}\langle\xi,\theta-\theta_0\rangle \le \frac{t_*^2}{2}. \tag{5.4.2}$$

For a vector $\theta = (\theta_x)_{x\in\mathbb{L}_{d,n}}$, define $\bar{\theta} := n^{-1}\sum_{x\in\mathbb{L}_{d,n}}\theta_x$ and write $\mathbf{1}_n \in \mathbb{R}^{\mathbb{L}_{d,n}}$ for the all-one vector. Then

$$\begin{aligned}\mathbb{E}\sup_{\theta\in\mathcal{M}(\mathbb{L}_{d,n}),\|\theta-\theta_0\|_2\le t_*}\langle\xi,\theta-\theta_0\rangle &= \mathbb{E}\sup_{\theta\in\mathcal{M}(\mathbb{L}_{d,n}),\|\theta-\theta_0\|_2\le t_*}\langle\xi,\theta-\bar{\theta}_0\mathbf{1}_n\rangle\\ &\le \mathbb{E}\sup_{\theta\in\mathcal{M}(\mathbb{L}_{d,n}),\|\theta-\bar{\theta}_0\mathbf{1}_n\|_2\le t_*+n^{1/2}}\langle\xi,\theta-\bar{\theta}_0\mathbf{1}_n\rangle\\ &= \mathbb{E}\sup_{\theta\in\mathcal{M}(\mathbb{L}_{d,n})\cap B_2(t_*+n^{1/2})}\langle\xi,\theta\rangle = \bigl\{t_*+n^{1/2}\bigr\}w_{\mathcal{M}(\mathbb{L}_{d,n})\cap B_2(1)},\end{aligned}$$

where we recall that $w_{\mathcal{M}(\mathbb{L}_{d,n})\cap B_2(1)} = \mathbb{E}\sup_{\theta\in\mathcal{M}(\mathbb{L}_{d,n})\cap B_2(1)}\langle\xi,\theta\rangle$. Therefore, to satisfy (5.4.2), it suffices to choose

$$\begin{aligned} t_* &= w_{\mathcal{M}(\mathbb{L}_{d,n})\cap B_2(1)} + \big\{w^2_{\mathcal{M}(\mathbb{L}_{d,n})\cap B_2(1)} + 2n^{1/2}w_{\mathcal{M}(\mathbb{L}_{d,n})\cap B_2(1)}\big\}^{1/2} \\ &\lesssim \max\big\{w_{\mathcal{M}(\mathbb{L}_{d,n})\cap B_2(1)}, n^{1/4}w^{1/2}_{\mathcal{M}(\mathbb{L}_{d,n})\cap B_2(1)}\big\}. \end{aligned} \tag{5.4.3}$$

Consequently, by [36] Corollary 1.2 and Proposition 5.4.1, we have that

$$R_n(\hat{\theta}_n,\theta_0) \lesssim n^{-1}\max(1,t_0^2) \lesssim n^{-1}t_*^2 \lesssim n^{-1/d}\log^4 n,$$

which completes the proof. □

The following proposition is the main ingredient of the proof of the minimax lower bound in Proposition 5.2.2. It exhibits a combinatorial obstacle, namely the existence of a large antichain, that prevents any estimator from achieving a faster rate of convergence. We state the result in the more general and natural setting of least squares isotonic regression on directed acyclic graphs. Recall that the isotonic regression problem on a directed acyclic graph $G = (V(G), E(G))$ is of the form $Y_v = \theta_v + \xi_v$, where $\theta = (\theta_v)_{v\in V(G)} \in \mathcal{M}(G)$ and $\xi = (\xi_v)_{v\in V(G)}$ is a vector of independent $N(0,1)$ random variables.

Proposition 5.4.2. *If $G = (V(G), E(G))$ is a directed acyclic graph with $|V(G)| = n$ and $W \subseteq V(G)$ is an antichain of G, then*

$$\inf_{\tilde{\theta}_n}\sup_{\theta_0\in\mathcal{M}(G)\cap B_\infty(1)} R_n(\tilde{\theta}_n,\theta_0) \geq \frac{4|W|}{27n},$$

where the infimum is taken over all measurable functions $\tilde{\theta}_n$ of $\{Y_v : v \in V(G)\}$.

Proof. Let W_0 be a maximal antichain of G containing W. If $v \notin W_0$, then by the maximality of W_0, there exists $u_0 \in W_0$ such that either $u_0 \leq v$ or $u_0 \geq v$. Suppose without loss of generality that it is the former. Then $v \not\leq u$ for any $u \in W_0$, because otherwise we would have $u_0 \leq u$, contradicting the fact that W_0 is an antichain. It follows that we can write $V(G) = W_0^+ \sqcup W_0 \sqcup W_0^-$, where for all $v \in W_0^+$, $u \in W_0$, we have $u \not\geq v$, and similarly for all $v \in W_0^-$, $u \in W_0$, we have $v \not\geq u$.

For $\tau = (\tau_w) \in \{0,1\}^{W_0} =: T$, we define $\theta^\tau = (\theta^\tau_v) \in \mathcal{M}(G) \cap B_\infty(1)$ by

$$\theta^\tau_v = \begin{cases} -1 & \text{if } v \in W_0^- \\ \rho(2\tau_v - 1) & \text{if } v \in W_0 \\ 1 & \text{if } v \in W_0^+, \end{cases}$$

where $\rho \in (0,1)$ is a constant to be chosen later. Let P_τ denote the distribution of $\{Y_v : v \in V(G)\}$ when the isotonic signal is θ^τ. Then, for $\tau, \tau' \in T$, by Pinsker's inequality [e.g. [125] page 62], we have

$$d^2_{\mathrm{TV}}(P_\tau, P_{\tau'}) \leq \frac{1}{2} d^2_{\mathrm{KL}}(P_\tau, P_{\tau'}) = \frac{1}{4}\|\theta^\tau - \theta^{\tau'}\|_2^2 = \rho^2 \|\tau - \tau'\|_0.$$

Thus, setting $\rho = 2/3$, by Assouad's Lemma [cf. [171] Lemma 2], we have that

$$\begin{aligned} \inf_{\tilde{\theta}_n} \sup_{\theta_0 \in \mathcal{M}(G) \cap B_\infty(1)} R_n(\tilde{\theta}_n, \theta_0) &\geq \inf_{\tilde{\theta}_n} \sup_{\tau \in T} R_n(\tilde{\theta}_n, \theta^\tau) \\ &\geq \frac{\rho^2 |W_0|}{n}(1-\rho) \geq \frac{4|W|}{27n}, \end{aligned}$$

as desired. □

Proof of Proposition 5.2.2. Recall that $n_1 = n^{1/d}$. We note that the set

$$W := \left\{ v = (v_1, \ldots, v_d)^\top \in \mathbb{L}_{d,n} : \sum_{j=1}^d v_j = 1 \right\}$$

is an antichain in $\mathbb{L}_{d,n}$ of cardinality $\binom{n_1 - 1}{d-1} \geq \left(\frac{n_1 - 1}{d-1}\right)^{d-1}$. The desired result therefore follows from Proposition 5.4.2. □

Proof of Corollary 5.2.3. For $Q = P_n$, the result is an immediate consequence of Theorem 5.2.1 and Proposition 5.2.2, together with the facts that

$$\inf_{\tilde{\theta}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} R_n(\tilde{\theta}_n, \theta_0) = \inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E}\|\tilde{f}_n - f_0\|^2_{L_2(P_n)}$$

and

$$\sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} R_n(\hat{\theta}_n, \theta_0) = \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E}\|\hat{f}_n - f_0\|^2_{L_2(P_n)}.$$

Now suppose that Q is Lebesgue measure on $[0,1]^d$. For any $f : [0,1]^d \to \mathbb{R}$, we may define $\theta(f) := f\big|_{\mathbb{L}_{d,n}}$. On the other hand, for any $\theta : \mathbb{L}_{d,n} \to \mathbb{R}$, we can also define $f(\theta) : [0,1]^d \to \mathbb{R}$ by

$$f(\theta)(x_1, \ldots, x_d) := \theta(n_1^{-1}\lfloor n_1 x_1 \rfloor, \ldots, n_1^{-1}\lfloor n_1 x_d \rfloor).$$

We first prove the upper bound by observing from Lemma 5.7.1 and Theorem 5.2.1 that

$$\begin{aligned}
&\sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E}\|\hat{f}_n - f_0\|_{L_2(Q)}^2 \\
&\qquad \le 2 \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \big\{ n^{-1}\mathbb{E}\|\theta(\hat{f}_n) - \theta(f_0)\|_2^2 + \|f_0 - f(\theta(f_0))\|_{L_2(Q)}^2 \big\} \\
&\qquad \le 2 \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} \frac{1}{n}\mathbb{E}\|\hat{\theta}_n - \theta_0\|_2^2 + 8dn^{-1/d} \le C_d n^{-1/d} \log^4 n,
\end{aligned}$$

as desired. Then by convexity of $\mathcal{H}$ and Proposition 5.2.2, we have

$$\begin{aligned}
\inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E}\|\tilde{f}_n - f_0\|_{L_2(Q)}^2 &\ge \inf_{\tilde{f}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} \mathbb{E}\|\tilde{f}_n - f(\theta_0)\|_{L_2(Q)}^2 \\
&= \inf_{\tilde{f}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} \mathbb{E}\|f(\theta(\tilde{f}_n)) - f(\theta_0)\|_{L_2(Q)}^2 \\
&= \inf_{\tilde{\theta}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} \frac{1}{n}\mathbb{E}\|\tilde{\theta}_n - \theta_0\|_2^2 \ge c_d n^{-1/d},
\end{aligned}$$

which completes the proof. □

Proof of Theorem 5.2.4. Recall that the tangent cone at a point x in a closed, convex set K is defined as $T(x, K) := \{t(y - x) : y \in K, t \ge 0\}$. By [18] Proposition 2.1 (see also [34] Lemma 4.1), we have

$$R_n(\hat{\theta}_n, \theta_0) \le \frac{1}{n} \inf_{\theta \in \mathcal{M}(\mathbb{L}_{d,n})} \Big\{ \|\theta - \theta_0\|_2^2 + \delta\big(T(\theta, \mathcal{M}(\mathbb{L}_{d,n}))\big) \Big\}. \tag{5.4.4}$$

For a fixed $\theta \in \mathcal{M}(\mathbb{L}_{d,n})$ such that $K(\theta) = K$, let $\mathbb{L}_{d,n} = \sqcup_{\ell=1}^K A_\ell$ be the partition of $\mathbb{L}_{d,n}$ into two-dimensional sheets A_ℓ such that θ is constant on each A_ℓ. Define $m_\ell := |A_\ell|$. Then any $u \in T(\theta, \mathcal{M}(\mathbb{L}_{d,n}))$ must be isotonic when restricted to each of the two-dimensional sheets; in other words

$$T(\theta, \mathcal{M}(\mathbb{L}_{d,n})) \subseteq \bigoplus_{\ell=1}^K T(0, \mathcal{M}(A_\ell)) = \bigoplus_{\ell=1}^K \mathcal{M}(A_\ell).$$

By [6] Proposition 3.1(9,10), we have

$$\delta\big(T(\theta, \mathcal{M}(\mathbb{L}_{d,n}))\big) \le \delta\bigg(\bigoplus_{\ell=1}^{K} \mathcal{M}(A_\ell)\bigg) = \sum_{\ell=1}^{K} \delta\big(\mathcal{M}(A_\ell)\big). \tag{5.4.5}$$

By a consequence of the Gaussian Poincaré inequality [cf. [22] page 73] and Proposition 5.4.1, we have

$$\delta\big(\mathcal{M}(A_\ell)\big) \le \bigg(\mathbb{E} \sup_{\theta \in \mathcal{M}(A_\ell) \cap B_2(1)} \langle \xi_{A_\ell}, \theta \rangle\bigg)^2 + 1 \lesssim \log_+^8 m_\ell. \tag{5.4.6}$$

Thus, by (5.4.5), (5.4.6) and Lemma 5.7.2 applied to $x \mapsto \log_+^8 x$, we have

$$\delta\big(T(\theta, \mathcal{M}(\mathbb{L}_{d,n}))\big) \lesssim \sum_{\ell=1}^{K} \log_+^8 m_\ell \lesssim K \log_+^8\bigg(\frac{n}{K}\bigg),$$

which together with (5.4.4) proves the desired result. □

Proof of Theorem 5.2.5. For a fixed $\theta \in \mathcal{M}^{(k)}(\mathbb{L}_{d,n})$, let $\mathbb{L}_{d,n} \subseteq \sqcup_{\ell=1}^{k} \mathcal{R}_\ell$ be a covering of $\mathbb{L}_{d,n}$ by disjoint hyperrectangles such that θ is constant on each hyperrectangle $\mathcal{R}_\ell$. Suppose $\mathcal{R}_\ell \cap \mathbb{L}_{d,n}$ has side lengths $m_1, \dots, m_d$ (so $|\mathcal{R}_\ell \cap \mathbb{L}_{d,n}| = \prod_{j=1}^{d} m_j$). Then it can be covered by the union of $\frac{|\mathcal{R}_\ell|}{m_j m_{j'}}$ parallel two-dimensional sheets, where m_j and $m_{j'}$ are the largest two elements of the multiset $\{m_1, \dots, m_d\}$. By Jensen's inequality (noting that $x \mapsto x^{1-2/d}$ is concave when $d \ge 2$), we obtain

$$K(\theta) \le \sum_{\ell=1}^{k} |\mathcal{R}_\ell \cap \mathbb{L}_{d,n}|^{1-2/d} \le k\bigg(\frac{n}{k}\bigg)^{1-2/d}. \tag{5.4.7}$$

This, combined with the oracle inequality in Theorem 5.2.4, gives the desired result. □

Proof of Proposition 5.2.6. Since the convex cone $\mathcal{M}(\mathbb{L}_{d,n})$ is invariant under translation by any $\theta_0 \in \mathcal{M}^{(1)}(\mathbb{L}_{d,n})$, we may assume without loss of generality that $\theta_0 = 0$. By the Cauchy–Schwarz inequality, we have

$$R_n(\hat{\theta}_n, 0) = \frac{1}{n}\delta\big(\mathcal{M}(\mathbb{L}_{d,n})\big) \ge \frac{1}{n}\bigg(\mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_2(1)} \langle \xi, \theta \rangle\bigg)^2, \tag{5.4.8}$$

which, together with Proposition 5.4.1, establishes the desired lower bound when $d \geq 3$. For the $d = 2$ case, by Sudakov minorisation for Gaussian processes [e.g. [123] Theorem5.6 and the remark following it] and Lemma 5.7.3, there exists a universal constant $\varepsilon_0 > 0$ such that

$$\mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1)} \langle \xi, \theta \rangle \gtrsim \varepsilon_0 \log^{1/2} N\big(\varepsilon_0, \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1), \|\cdot\|_2\big) \gtrsim \log n.$$

This, together with (5.4.8), establishes the desired conclusion when $d = 2$. □

Proof of Corollary 5.2.7. Without loss of generality, we may assume that $\theta_0 \in \mathcal{M}_r(\mathbb{L}_{d,n})$ is a function of the final r variables. For $x_3, \ldots, x_d \in \{1/n_1, 2/n_1, \ldots, 1\}$, we define $A_{x_3,\ldots,x_d} := \big\{(x_1, \ldots, x_d) : x_1, x_2 \in [0,1]\big\}$. When $r \leq d-2$, we have that θ_0 is constant on each $A_{x_3,\ldots,x_d} \cap \mathbb{L}_{d,n}$. Hence, by Theorem 5.2.4,

$$R_n(\hat{\theta}_n, \theta_0) \lesssim \frac{K(\theta_0) \log^8_+\big(n/K(\theta_0)\big)}{n} \lesssim n^{-2/d} \log^8 n.$$

Now suppose that $\theta_0 \in \mathcal{M}_{d-1}(\mathbb{L}_{d,n})$. Let m be a positive integer to be chosen later. Then $A_{x_3,\ldots,x_d} \cap \mathbb{L}_{d,n} = \sqcup_{\ell=-m}^{m} A^{(\ell)}_{x_3,\ldots,x_d}$, where

$$A^{(\ell)}_{x_3,\ldots,x_d} := A_{x_3,\ldots,x_d} \cap \bigg\{v \in \mathbb{L}_{d,n} : \frac{\ell-1}{m} < (\theta_0)_v \leq \frac{\ell}{m}\bigg\}.$$

Let $\theta^{(m)} \in \mathcal{M}(\mathbb{L}_{d,n})$ be the vector that takes the constant value ℓ/m on $A^{(\ell)}_{x_3,\ldots,x_d}$ for each $\ell = -m, \ldots, m$. Then setting $m \asymp n^{2/(3d)} \log^{-8/3} n$, we have by Theorem 5.2.4 that

$$\begin{aligned} R_n(\hat{\theta}_n, \theta_0) &\lesssim \frac{\|\theta^{(m)} - \theta_0\|_2^2}{n} + \frac{K(\theta^{(m)}) \log^8_+\big(n/K(\theta^{(m)})\big)}{n} \\ &\leq \frac{1}{m^2} + \frac{m}{n^{2/d}} \log^8 n \lesssim n^{-4/(3d)} \log^{16/3} n. \end{aligned}$$

as desired.

Finally, the $r = d$ case is covered in Theorem 5.2.1. □

5.5 *Proof of results in Section 5.3*

From now on we write $\mathbb{G}_n := n^{1/2}(\mathbb{P}_n - P)$. Recall that $\gamma_2 = 9/2$ and $\gamma_d = (d^2 + d + 1)/2$ for $d \geq 3$.

In our empirical process theory arguments, we frequently need to consider suprema over subsets of $\mathcal{F}_d$. In order to avoid measurability digressions, and since our least squares estimator $\hat{f}_n$ is defined to be lower semi-continuous, we always assume implicitly that such suprema are in fact taken over the intersection of the relevant subset of $\mathcal{F}_d$ with $\mathcal{L}$, the class of real-valued lower semi-continuous functions on $[0,1]^d$. Then $\mathcal{F}_d' := \{f \in \mathcal{F}_d \cap \mathcal{L} : f|_{(\mathbb{Q}\cap[0,1])^d} \subseteq \mathbb{Q}\}$ is a countable, uniformly dense[3] subset of $\mathcal{F}_d \cap \mathcal{L}$ so that, for example, $\sup_{f\in\mathcal{F}_d\cap\mathcal{L}} \mathbb{G}_n f = \sup_{f\in\mathcal{F}_d'} \mathbb{G}_n f$, which ensures measurability.

5.5.1 Preparatory results

We first state a few intermediate results that will be used in the proofs of Theorems 5.3.1 and 5.3.2. The proofs of propositions in this subsection are contained Section 5.6 in the online supplementary material.

The following proposition controls the tail probability of $\|\hat{f}_n - f_0\|_{L_2(P)}$ on the event $\{\|\hat{f}_n - f_0\|_\infty \leq 6\log^{1/2} n\}$ by two multiplier empirical processes (5.5.2) and (5.5.3). For $f_0 \in \mathcal{F}_d$, $r, a > 0$, define

$$\mathcal{G}(f_0, r, a) := \big\{f \in \mathcal{F}_d : f - f_0 \in B_2(r, P) \cap B_\infty(a)\big\}. \tag{5.5.1}$$

Proposition 5.5.1. *Suppose that $f_0 \in \mathcal{F}_d \cap B_\infty(1)$ and that for each $n \geq 2$ there exist both a function $\phi_n : [0,\infty) \to [0,\infty)$ and a sequence $r_n \geq n^{-1/2}\log^{1/2} n$ such that $\phi_n(r_n) \leq n^{1/2} r_n^2$. Moreover, assume that for all $r \geq r_n$ the map $r \mapsto \phi_n(r)/r$ is non-increasing and*

$$\mathbb{E} \sup_{f\in\mathcal{G}(f_0,r,6\log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i\{f(X_i) - f_0(X_i)\} \right| \leq K\phi_n(r), \tag{5.5.2}$$

$$\mathbb{E} \sup_{f\in\mathcal{G}(f_0,r,6\log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \varepsilon_i\{f(X_i) - f_0(X_i)\}^2 \right| \leq K\phi_n(r), \tag{5.5.3}$$

for some $K \geq 1$ that does not depend on r and n. Then, there exist universal constants

[3]Here 'uniformly dense' means that for any $f \in \mathcal{F}_d \cap \mathcal{L}$, we can find a sequence (f_m) in $\mathcal{F}_d'$ such that $\|f_m - f\|_\infty \to 0$. This can be done by defining, e.g., $f_m(x) := m^{-1}\lceil m f(x)\rceil$.

$C, C' > 0$ *such that for all* $r \geq C' K r_n$, *we have*

$$\mathbb{P}\big(\big\{\|\hat{f}_n - f_0\|_{L_2(P)} \geq r\big\} \cap \big\{\|\hat{f}_n - f_0\|_\infty \leq 6\log^{1/2} n\big\}\big) \leq C\exp\bigg(-\frac{nr^2}{C\log n}\bigg).$$

Consequently,

$$\mathbb{E}\big\{\|\hat{f}_n - f_0\|^2_{L_2(P)} \mathbb{1}_{\{\|\hat{f}_n - f_0\|_\infty \leq 6\log^{1/2} n\}}\big\} \lesssim K^2 r_n^2.$$

By means of Lemmas 5.7.5 and 5.7.6, the control of the empirical processes (5.5.2) and (5.5.3) in turn reduces to the study of the symmetrised local empirical process

$$\mathbb{E}\sup_{f \in \mathcal{G}(0,r,1)} \bigg|\frac{1}{n^{1/2}} \sum_{i=1}^n \varepsilon_i f(X_i)\bigg|, \tag{5.5.4}$$

for a suitable $L_2(P)$ radius r. To obtain a sharp bound on the empirical process in (5.5.4), which constitutes the main technical challenge of the proof, we slice $[0,1]^d$ into strips of the form $[0,1]^{d-1} \times [\frac{\ell-1}{n_1}, \frac{\ell}{n_1}]$, for $\ell = 1, \ldots, n_1$, and decompose $\sum_{i=1}^n \varepsilon_i f(X_i)$ into sums of smaller empirical processes over these strips. Each of these smaller empirical processes is then controlled via a bracketing entropy chaining argument (Lemma 5.7.7). The advantage of this decomposition is that the block monotonicity permits good control of the $L_2(P)$ norm of the envelope function in each strip (Lemma 5.7.9). This leads to the following conclusion:

Proposition 5.5.2. *Let* $d \geq 2$. *There exists* $C_{d,m_0,M_0} > 0$, *depending only on* d, m_0 *and* M_0, *such that if* $r \geq n^{-1/2}(\log_+ \log n)^2$ *when* $d = 2$ *and* $r \geq n^{-(1-2/d)} \log^{\gamma_d - 1/2} n$ *when* $d \geq 3$, *then*

$$\mathbb{E}\sup_{f \in \mathcal{G}(0,r,1)} \bigg|\frac{1}{n^{1/2}} \sum_{i=1}^n \varepsilon_i f(X_i)\bigg| \leq C_{d,m_0,M_0} r n^{1/2-1/d} \log^{\gamma_d - 1/2} n.$$

On the other hand, there exists $c_{d,m_0} > 0$, *depending only on* d *and* m_0, *such that if* $r \leq 1$, *then*

$$\mathbb{E}\sup_{f \in \mathcal{G}(0,r,1)} \frac{1}{n^{1/2}} \sum_{i=1}^n \varepsilon_i f(X_i) \geq c_{d,m_0} r n^{1/2-1/d}.$$

Our next proposition controls the discrepancy between the $L_2(P)$ and $L_2(\mathbb{P}_n)$ risks for the truncated estimator, $\tilde{f}_n := \hat{f}_n \mathbb{1}_{\{\|\hat{f}_n\|_\infty \leq 6\log^{1/2} n\}}$, when the true signal $f_0 = 0$.

Proposition 5.5.3. *Fix $d \geq 2$ and suppose that $f_0 = 0$. There exists $C_{d,m_0,M_0} > 0$, depending only on d, m_0 and M_0, such that*

$$\mathbb{E}\big\|\tilde{f}_n\big\|_{L_2(\mathbb{P}_n)}^2 \leq C_{d,m_0,M_0}\Big\{n^{-2/d}\log^{2\gamma_d} n + \mathbb{E}\big\|\tilde{f}_n\big\|_{L_2(P)}^2\Big\}.$$

Propositions 5.5.1, 5.5.2 and 5.5.3 allow us to control the risk of the least squares estimator when the true signal $f_0 = 0$.

Proposition 5.5.4. *Let $d \geq 2$. There exists a constant $C_{d,m_0,M_0} > 0$, depending only on d, m_0 and M_0, such that*

$$\max\big\{R(\hat{f}_n, 0), R_n(\hat{f}_n, 0)\big\} \leq C_{d,m_0,M_0} n^{-2/d}\log^{2\gamma_d} n.$$

5.5.2 Proofs of Theorems 5.3.1 and 5.3.2 and Propositions 5.3.3 and 5.3.4

The risk bounds in $L_2(P)$ loss and $L_2(\mathbb{P}_n)$ loss are proved with different arguments and hence presented separately below.

Proof of Theorem 5.3.1 in $L_2(P)$ loss. Recall the definition of the function class $\mathcal{G}(f_0, r, a)$ in (5.5.1). Let $r_n := n^{-1/(2d)}\log^{\gamma_d/2} n$. For any $r, a > 0$, by the triangle inequality, Lemma 5.7.5 and Proposition 5.5.2, we have that for $r \geq r_n$,

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{G}(f_0,r,4\log^{1/2} n)}\left|\frac{1}{n^{1/2}}\sum_{i=1}^n \xi_i\{f(X_i) - f_0(X_i)\}\right| \\
&\qquad\leq \mathbb{E}\sup_{f\in\mathcal{G}(0,r+1,6\log^{1/2} n)}\left|\frac{2\log^{1/2} n}{n^{1/2}}\sum_{i=1}^n \varepsilon_i f(X_i)\right| + 1 \\
&\qquad\lesssim_{d,m_0,M_0} (r+1)n^{1/2-1/d}\log^{\gamma_d} n \lesssim n^{1/2} r r_n.
\end{aligned}$$

Similarly, by Lemma 5.7.6 and Proposition 5.5.2, we have that for $r \geq r_n$,

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{G}(f_0,r,4\log^{1/2} n)}\left|\frac{1}{n^{1/2}}\sum_{i=1}^n \varepsilon_i\{f(X_i) - f_0(X_i)\}^2\right| \\
&\lesssim \mathbb{E}\sup_{f\in\mathcal{G}(0,r+1,6\log^{1/2} n)}\left|\frac{\log^{1/2} n}{n^{1/2}}\sum_{i=1}^n \varepsilon_i f(X_i)\right| + \log^{1/2} n \lesssim_{d,m_0,M_0} n^{1/2} r r_n.
\end{aligned}$$

Thus, conditions (5.5.2) and (5.5.3) in Proposition 5.5.1 are satisfied with $\phi_n(r) = n^{1/2} r r_n$ and $1 \le K \lesssim_{d,m_0,M_0} 1$. Let $\Omega_0 := \{\|\hat{f}_n - f_0\|_\infty \le 6 \log^{1/2} n\}$. It follows from Proposition 5.5.1 and Lemma 5.7.10 that

$$\begin{aligned} R(\hat{f}_n, f_0) &= \mathbb{E}\big\{\|\hat{f}_n - f_0\|_{L_2(P)}^2 \mathbb{1}_{\Omega_0}\big\} + \mathbb{E}\big\{\|\hat{f}_n - f_0\|_{L_2(P)}^2 \mathbb{1}_{\Omega_0^c}\big\} \\ &\lesssim_{d,m_0,M_0} r_n^2 + n^{-1} \lesssim n^{-1/d} \log^{\gamma_d} n, \end{aligned}$$

as desired. □

Proof of Theorem 5.3.1 in $L_2(\mathbb{P}_n)$ loss. Since the argument used in the proof of Theorem 5.2.1, up to (5.4.3), does not depend on the design, we deduce from [36] Corollary 1.2, [6] Proposition 3.1(5) and the Cauchy–Schwarz inequality that

$$R_n(\hat{f}_n, f_0) \lesssim \frac{1}{n} \mathbb{E} \max\big\{1, \delta(\mathcal{M}(G_X)), n^{1/2} \delta(\mathcal{M}(G_X))^{1/2}\big\}. \tag{5.5.5}$$

On the other hand, by Proposition 5.5.4, we have

$$\mathbb{E}\, \delta\big(\mathcal{M}(G_X)\big) \lesssim_{d,m_0,M_0} n^{1-2/d} \log^{2\gamma_d} n. \tag{5.5.6}$$

We obtain the desired result by combining (5.5.5) and (5.5.6). □

Proof of Theorem 5.3.2 in $L_2(\mathbb{P}_n)$ loss. For any $f \in \mathcal{F}_d$, we can define a random vector $\theta_{f,X} := (f(X_1), \ldots, f(X_n))^\top$. By[18] Proposition 2.1, we have

$$\begin{aligned} R_n(\hat{f}_n, f_0) &\le \frac{1}{n} \mathbb{E}\left[\inf_{f \in \mathcal{F}_d} \Big\{\|\theta_{f,X} - \theta_{f_0,X}\|_2^2 + \delta\big(T(\theta_{f,X}, \mathcal{M}(G_X))\big)\Big\}\right] \\ &\le \frac{1}{n} \inf_{k \in \mathbb{N}} \inf_{f \in \mathcal{F}_d^{(k)}} \Big\{\mathbb{E}\|\theta_{f,X} - \theta_{f_0,X}\|_2^2 + \mathbb{E}\, \delta\big(T(\theta_{f,X}, \mathcal{M}(G_X))\big)\Big\}. \end{aligned} \tag{5.5.7}$$

Now, for a fixed $f \in \mathcal{F}_d^{(k)}$, let $\mathcal{R}_1, \ldots, \mathcal{R}_k$ be the corresponding hyperrectangles such that f is constant when restricted to each $\mathcal{R}_\ell$. Define $\mathcal{X}_\ell := \mathcal{R}_\ell \cap \{X_1, \ldots, X_n\}$ and $N_\ell := |\mathcal{X}_\ell|$. Then for fixed $X_1, \ldots, X_n$, we have $T(\theta_{f,X}, \mathcal{M}(G_X)) \subseteq \bigoplus_{\ell=1}^k T\big(0, \mathcal{M}(G_{\mathcal{X}_\ell})\big) = \bigoplus_{\ell=1}^k \mathcal{M}(G_{\mathcal{X}_\ell})$.

Therefore, by [6] Proposition 3.1(9,10) and (5.5.6), we have that

$$\begin{aligned}\mathbb{E}\,\delta\big(T(\theta_{f,X},\mathcal{M}(G_X))\big) &= \mathbb{E}\Big[\mathbb{E}\Big\{\delta\big(T(\theta_{f,X},\mathcal{M}(G_X))\big)\;\Big|\;N_1,\dots,N_k\Big\}\Big]\\ &\le \mathbb{E}\Big[\sum_{\ell:N_\ell\ge 1}\mathbb{E}\Big\{\delta\big(\mathcal{M}(G_{\mathcal{X}_\ell})\big)\;\Big|\;N_\ell\Big\}\Big] \lesssim_{d,m_0,M_0} \mathbb{E}\Big\{\sum_{\ell:N_\ell\ge 1} N_\ell^{1-2/d}\log_+^{2\gamma_d} N_\ell\Big\}\\ &\lesssim_d n(k/n)^{2/d}\log_+^{2\gamma_d}(n/k), \end{aligned} \tag{5.5.8}$$

where the final bound follows from applying Lemma 5.7.2 to the function $x \mapsto x^{1-2/d}\log_+^{2\gamma_d} x$. We complete the proof by substituting (5.5.8) into (5.5.7) and observing that

$$\frac{1}{n}\inf_{f\in\mathcal{F}_d^{(k)}}\mathbb{E}\|\theta_{f,X}-\theta_{f_0,X}\|_2^2 = \inf_{f\in\mathcal{F}_d^{(k)}}\mathbb{E}\|f-f_0\|_{L_2(\mathbb{P}_n)}^2 = \inf_{f\in\mathcal{F}_d^{(k)}}\|f-f_0\|_{L_2(P)}^2,$$

as desired. □

Proof of Theorem 5.3.2 in $L_2(P)$ loss. Fix $k\in\mathbb{N}$, $f_k\in\mathcal{F}_d^{(k)}\cap B_\infty(1)$ and let $\mathcal{R}_1,\dots,\mathcal{R}_k$ be the corresponding hyperrectangles such that f_k is constant when restricted to each $\mathcal{R}_\ell$. Define $N_\ell := |\{X_1,\dots,X_n\}\cap\mathcal{R}_\ell|$.

We let $\mathbb{P}_{f_0}$ and $\mathbb{P}_{f_k}$ denote the probability with respect to the data generating mechanisms $Y_i = f_0(X_i)+\xi_i$ and $Y_i = f_k(X_i)+\xi_i$ respectively, and write $\mathbb{E}_{f_0}$ and $\mathbb{E}_{f_k}$ for the respective expectations. For any $t\ge 0$, write $\Omega_t' := \big\{\|\hat{f}_n-f_0\|_{L_2(P)} > \|f_k-f_0\|_{L_2(P)}+t\big\}\cap\big\{\|\hat{f}_n-f_0\|_\infty \le 3\log^{1/2} n\big\}$. We have that

$$\begin{aligned}\mathbb{P}_{f_0}(\Omega_t') &\le \mathbb{P}_{f_0}\big(\{\|\hat{f}_n-f_k\|_{L_2(P)}>t\}\cap\{\|\hat{f}_n-f_k\|_\infty\le 6\log^{1/2} n\}\big)\\ &= \mathbb{E}_{f_k}\Big\{e^{-\frac{n}{2}\|f_k-f_0\|_{L_2(\mathbb{P}_n)}^2-\sum_{i=1}^n\xi_i(f_k-f_0)(X_i)}\mathbb{1}_{\{\hat{f}_n-f_k\in B_2(t,P)^{\mathrm{c}}\cap B_\infty(6\log^{1/2} n)\}}\Big\}\\ &\le \mathbb{P}_{f_k}\big\{\hat{f}_n-f_k\in B_2(t,P)^{\mathrm{c}}\cap B_\infty(6\log^{1/2} n)\big\}^{1/2}\big\{\mathbb{E}e^{n\|f_k-f_0\|_{L_2(\mathbb{P}_n)}^2}\big\}^{1/2}, \end{aligned}\tag{5.5.9}$$

where the equality follows from a change of measure (the Radon–Nikodym theorem), and the final step uses the Cauchy–Schwarz inequality. We control the two factors on the right-hand side separately. For the second factor, since $\|f_k-f_0\|_\infty\le 2$, we have by Lemma 5.7.12 that

$$\mathbb{E}e^{n\|f_k-f_0\|_{L_2(\mathbb{P}_n)}^2} \le e^{14n\|f_k-f_0\|_{L_2(P)}^2}. \tag{5.5.10}$$

For the first factor, for all $r \geq (k/n)^{1/d} \log^{\gamma_d} n =: r_{n,k}$, we have that

$$\begin{aligned}
&\mathbb{E}_{f_k} \sup_{f \in \mathcal{G}(f_k, r, 1)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \\
&\leq \mathbb{E}_{f_k} \sum_{\ell : N_\ell \geq 1} \frac{N_\ell^{1/2}}{n^{1/2}} \mathbb{E}_{f_k} \left\{ \sup_{\substack{f \in \mathcal{F}_d, \|f - f_k\|_\infty \leq 1 \\ \|f - f_k\|_{L_2(P; \mathcal{R}_\ell)} \leq r}} \left| \frac{1}{N_\ell^{1/2}} \sum_{i : X_i \in \mathcal{R}_\ell} \varepsilon_i f(X_i) \right| \,\middle|\, N_\ell \right\} \\
&\lesssim_{d, m_0, M_0} \frac{r \log^{\gamma_d - 1/2} n}{n^{1/2}} \mathbb{E}_{f_k} \sum_{\ell : N_\ell \geq 1} N_\ell^{1 - 1/d} \\
&\lesssim r n^{1/2} \left(\frac{k}{n} \right)^{1/d} \log^{\gamma_d - 1/2} n,
\end{aligned}$$

where the penultimate inequality follows from Proposition 5.5.2 and the final step uses Jensen's inequality. Using the above bound together with Lemmas 5.7.5 and 5.7.6 as in the proof of Theorem 5.3.1, we see that (5.5.2) and (5.5.3) (with f_0 replaced with f_k there) are satisfied with $1 \leq K \lesssim_{d, m_0, M_0} 1$ and $\phi_n(r) = n^{1/2} r_{n,k} r$, so by Proposition 5.5.1, there exist universal constants $C, C' > 1$ such that for $t \geq C' K r_{n,k}$,

$$\mathbb{P}_{f_k} \big\{ \hat{f}_n - f_k \in B_2(t, P)^{\mathrm{c}} \cap B_\infty(6 \log^{1/2} n) \big\} \leq C e^{-nt^2/(C \log n)}. \tag{5.5.11}$$

Substituting (5.5.11) and (5.5.10) into (5.5.9) and writing $t_0 := (28 C \log n)^{1/2} \|f_k - f_0\|_{L_2(P)}$, we have for all $t \geq t_0 + C' K r_{n,k}$ that

$$\mathbb{P}_{f_0}(\Omega_t') \lesssim e^{7n \|f_k - f_0\|_{L_2(P)}^2 - nt^2/(2C \log n)} \leq e^{-nt^2/(4C \log n)}.$$

Combining the above probability bound with Lemma 5.7.10, we obtain that

$$\begin{aligned}
R(\hat{f}_n, f_0) &\lesssim \mathbb{E}_{f_0} \big\{ \|\hat{f}_n - f_0\|_{L_2(P)}^2 \mathbb{1}_{\{\|\hat{f}_n - f_0\|_\infty \leq 3 \log^{1/2} n\}} \big\} + \frac{1}{n} \\
&\lesssim \|f_k - f_0\|_{L_2(P)}^2 \log n + K^2 r_{n,k}^2 + \int_{t_0/2 + C' K r_{n,k}}^{\infty} (t + t_0) \mathbb{P}_{f_0}(\Omega_t') \, \mathrm{d}t \\
&\lesssim \|f_k - f_0\|_{L_2(P)}^2 \log n + K^2 r_{n,k}^2 \\
&\lesssim \|f_k - f_0\|_{L_2(P)}^2 \log n + C_{d, m_0, M_0} \left(\frac{k}{n} \right)^{2/d} \log^{2\gamma_d} n,
\end{aligned}$$

where $C_{d, m_0, M_0} > 0$ depends only on d, m_0 and M_0. The desired result follows since the above inequality holds for all $k \in \mathbb{N}$ and $f_k \in \mathcal{F}_d^{(k)} \cap B_\infty(1)$, and $\inf_{f \in \mathcal{F}_d^{(k)} \cap B_\infty(1)} \|f - f_0\|_{L_2(P)} = \inf_{f \in \mathcal{F}_d^{(k)}} \|f - f_0\|_{L_2(P)}$. □

Proof of Proposition 5.3.3 in $L_2(P)$ loss. By [57] Theorem 1.1, we have

$$\log N\big(\varepsilon, \mathcal{F}_d \cap B_\infty(1), \|\cdot\|_{L_2(P)}\big) \gtrsim_{m_0,d} \varepsilon^{-2(d-1)}.$$

The desired lower bound in $L_2(P)$ risk then follows from [169] Proposition 1. □

Proof of Proposition 5.3.3 in $L_2(\mathbb{P}_n)$ loss. Without loss of generality, we may assume that $n = n_1^d$ for some $n_1 \in \mathbb{N}$. Let $W := \{w \in \mathbb{L}_{d,n} : \sum_{j=1}^d w_j = 1\}$. For any $w = (w_1, \ldots, w_d)^\top \in W$, we define $\mathcal{C}_w := \prod_{j=1}^d (w_j - 1/n_1, w_j]$. Note that $x = (x_1, \ldots, x_d)^\top \in \cup_{w\in W}\mathcal{C}_w$ if and only if $\lceil n_1 x_1\rceil + \cdots + \lceil n_1 x_d\rceil = n_1$. For any $\tau = (\tau_w) \in \{0,1\}^{|W|} =: T$, we define $f_\tau \in \mathcal{F}_d$ by

$$f_\tau(x) := \begin{cases} 0 & \text{if } \lceil n_1 x_1\rceil + \cdots + \lceil n_1 x_d\rceil \le n_1 - 1 \\ 1 & \text{if } \lceil n_1 x_1\rceil + \cdots + \lceil n_1 x_d\rceil \ge n_1 + 1 \\ \rho\tau_{(\lceil n_1 x_1\rceil, \ldots, \lceil n_1 x_d\rceil)} & \text{if } x \in \cup_{w\in W}\mathcal{C}_w, \end{cases}$$

where $\rho \in [0,1]$ is to be specified later. Moreover, let τ^w be the binary vector differing from τ in only the w coordinate. We write $\mathbb{E}_\tau$ for the expectation over $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $Y_i = f_\tau(X_i) + \xi_i$ for $i = 1, \ldots, n$. We let $\mathbb{E}_X$ be the expectation over $(X_i)_{i=1}^n$ alone and $\mathbb{E}_{Y|X,\tau}$ be the conditional expectation of $(Y_i)_{i=1}^n$ given $(X_i)_{i=1}^n$. Given any estimator $\tilde{f}_n$, we have

$$\begin{aligned}
&\max_{\tau\in T} \mathbb{E}_\tau \big\|\tilde{f}_n - f_\tau\big\|_{L_2(\mathbb{P}_n)}^2 \ge \frac{1}{2^{|W|}} \sum_{w\in W}\sum_{\tau\in T} \mathbb{E}_\tau \int_{\mathcal{C}_w} (\tilde{f}_n - f_\tau)^2 \,\mathrm{d}\mathbb{P}_n \\
&= \frac{1}{2^{|W|+1}} \sum_{w\in W}\sum_{\tau\in T} \bigg\{ \mathbb{E}_\tau \int_{\mathcal{C}_w} (\tilde{f}_n - f_\tau)^2 \,\mathrm{d}\mathbb{P}_n + \mathbb{E}_{\tau^w} \int_{\mathcal{C}_w} (\tilde{f}_n - f_{\tau^w})^2 \,\mathrm{d}\mathbb{P}_n \bigg\} \\
&\ge \frac{1}{2^{|W|+3}} \sum_{w\in W}\sum_{\tau\in T} \mathbb{E}_X \bigg\{ \int_{\mathcal{C}_w} (f_\tau - f_{\tau^w})^2 \,\mathrm{d}\mathbb{P}_n \Big[1 - d_{\mathrm{TV}}\Big(P_{Y|X,\tau}, P_{Y|X,\tau^w}\Big)\Big] \bigg\}, \qquad (5.5.12)
\end{aligned}$$

where $P_{Y|X,\tau}$ (respectively $P_{Y|X,\tau^w}$) is the conditional distribution of $(Y_i)_{i=1}^n$ given $(X_i)_{i=1}^n$ when the true signal is f_τ (respectively f_{τ^w}). The final inequality in the above display follows because for $\Delta := \big(\int_{\mathcal{C}_w} (f_\tau - f_{\tau^w})^2 \,\mathrm{d}\mathbb{P}_n\big)^{1/2}$ and $A := \big\{\int_{\mathcal{C}_w} (\tilde{f}_n - f_\tau)^2 \,\mathrm{d}\mathbb{P}_n \ge \Delta^2/4\big\}$, we have

$$\begin{aligned}
&\mathbb{E}_{Y|X,\tau} \int_{\mathcal{C}_w} (\tilde{f}_n - f_\tau)^2 \,\mathrm{d}\mathbb{P}_n + \mathbb{E}_{Y|X,\tau^w} \int_{\mathcal{C}_w} (\tilde{f}_n - f_{\tau^w})^2 \,\mathrm{d}\mathbb{P}_n \\
&\quad \ge \frac{\Delta^2}{4}\Big\{P_{Y|X,\tau}(A) + P_{Y|X,\tau^w}(A^{\mathrm{c}})\Big\} \ge \frac{\Delta^2}{4}\big\{1 - d_{\mathrm{TV}}\big(P_{Y|X,\tau}, P_{Y|X,\tau^w}\big)\big\}.
\end{aligned}$$

By Pinsker's inequality [cf. [125] page 62], we obtain that

$$d_{\mathrm{TV}}^2\Big(P_{Y|X,\tau}, P_{Y|X,\tau^w}\Big) \leq \frac{1}{2} d_{\mathrm{KL}}^2(P_{Y|X,\tau}, P_{Y|X,\tau^w}) = \frac{n}{4}\|f_\tau - f_{\tau^w}\|_{L_2(\mathbb{P}_n)}^2. \tag{5.5.13}$$

Writing $N_w := \sum_{i=1}^n \mathbb{1}_{\{X_i \in \mathcal{C}_w\}}$, we have $N_w \sim \mathrm{Bin}(n, P(\mathcal{C}_w))$, so $\mathbb{E}_X N_w \geq m_0$ and $\mathbb{E}_X N_w^{3/2} \leq (\mathbb{E}_X N_w^2 \, \mathbb{E}_X N_w)^{1/2} \leq 2^{1/2} M_0^{3/2}$. Thus, together with (5.5.13), we have

$$\begin{aligned}
&\mathbb{E}_X\bigg\{\int_{\mathcal{C}_w} (f_\tau - f_{\tau^w})^2 \,\mathrm{d}\mathbb{P}_n\Big[1 - d_{\mathrm{TV}}\Big(P_{Y|X,\tau}, P_{Y|X,\tau^w}\Big)\Big]\bigg\} \\
&\qquad\qquad \geq \mathbb{E}_X\bigg\{\|f_\tau - f_{\tau^w}\|_{L_2(\mathbb{P}_n)}^2\bigg(1 - \frac{n^{1/2}}{2}\|f_\tau - f_{\tau^w}\|_{L_2(\mathbb{P}_n)}\bigg)\bigg\} \\
&\qquad\qquad = \frac{\rho^2}{n}\mathbb{E}_X N_w - \frac{\rho^3}{2n}\mathbb{E}_X N_w^{3/2} \geq \frac{\rho^2}{n}\bigg(m_0 - \frac{\rho}{2^{1/2}} M_0^{3/2}\bigg).
\end{aligned} \tag{5.5.14}$$

Substituting (5.5.14) into (5.5.12), we obtain that for $\rho = 2^{3/2} m_0/(3M_0^{3/2})$,

$$\max_{\tau \in T} \mathbb{E}_\tau \big\|\tilde{f}_n - f_\tau\big\|_{L_2(\mathbb{P}_n)}^2 \geq \frac{|W|}{27n}\frac{m_0^3}{M_0^3} \geq c_{d,m_0,M_0} n^{-1/d},$$

where the final inequality follows from a counting argument as in (5.4.1). This completes the proof. □

Proof of Proposition 5.3.4 in $L_2(P)$ loss. Case $d = 2$. First note that, by translation invariance, $R(\hat{f}_n, f_0)$ is constant for $f_0 \in \mathcal{F}_d^{(1)}$. We then observe that, given any estimator $\tilde{f}_n = \tilde{f}_n(X_1, Y_1, \ldots, X_n, Y_n)$ of $f_0 \in \mathcal{F}_d^{(1)}$, we can construct a new estimator $\tilde{f}_n'$ by setting $\tilde{f}_n'(x) := P\tilde{f}_n$ for all $x \in [0,1]^d$. Then

$$R(\tilde{f}_n, f_0) = R(\tilde{f}_n', f_0) + \int_{[0,1]^d} (\tilde{f}_n - \tilde{f}_n')^2 \,\mathrm{d}P \geq R(\tilde{f}_n', f_0),$$

so in seeking to minimise $\sup_{f \in \mathcal{F}_d^{(1)}} R(\tilde{f}_n, f)$, we may restrict attention to estimators that are constant on $[0,1]^d$. It follows that for any $f_0 \in \mathcal{F}_d^{(1)}$,

$$R(\tilde{f}_n, f_0) = \sup_{f \in \mathcal{F}_d^{(1)}} R(\hat{f}_n, f) \geq \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}_d^{(1)}} R(\tilde{f}_n, f) = \inf_{\tilde{\mu}_n} \sup_{\mu \in \mathbb{R}} \mathbb{E}\{(\tilde{\mu}_n - \mu)^2\} \gtrsim \frac{1}{n},$$

where the second infimum is taken over all estimators $\tilde{\mu}_n = \tilde{\mu}_n(Y_1, \ldots, Y_n)$ of $\mu = f_0(0)$.

Case $d \geq 3$. It suffices to only consider the case when $f_0 = 0$. For $i = 1, \ldots, n$, let $\tilde{\xi}_i := \xi_i \mathbb{1}_{\{|\xi_i| \leq 2\log^{1/2} n\}}$ and for $r, b \geq 0$, define

$$E_n(r,b) := \sup_{f \in \mathcal{F}_d \cap B_2(r,P) \cap B_\infty(b)} \frac{1}{n} \sum_{i=1}^n \big\{ 2\tilde{\xi}_i f(X_i) - f^2(X_i) + \|f\|_{L_2(P)}^2 \big\}.$$

Observe that for $r \geq n^{-1/2} \log n$, $b \in [0, 6\log^{1/2} n]$ and any $f \in \mathcal{F}_d \cap B_2(r,P) \cap B_\infty(b)$, we have

$$\begin{aligned} \mathrm{Var}\big\{ 2\tilde{\xi}_1 f(X_1) - f^2(X_1) \big\} &\leq r^2(8 + 2b^2) \lesssim r^2 \log n, \\ \|2\tilde{\xi}_1 f - f^2\|_\infty &\leq 4b \log^{1/2} n + b^2 \lesssim \log n. \end{aligned}$$

It follows by Talagrand's concentration inequality [145] in the form given by [106] Theorem 3, that for each $r \geq n^{-1/2} \log n$ and $b \in [0, 6\log^{1/2} n]$, there is a universal constant $C_0 > 0$ and an event $\Omega_{r,b}$, with probability at least $1 - n^{-1}$, such that on $\Omega_{r,b}$,

$$\frac{1}{2} \mathbb{E} E_n(r,b) - C_0 r^2 \leq E_n(r,b) \leq 2\mathbb{E} E_n(r,b) + C_0 r^2. \tag{5.5.15}$$

Let $F_n(r) := E_n(r, 6\log^{1/2} n) - r^2$ and choose

$$\tilde{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}_d \cap B_\infty(6\log^{1/2} n)} \sum_{i=1}^n \{\tilde{\xi}_i - f(X_i)\}^2$$

such that $\tilde{f}_n = \hat{f}_n$ on the event $\Omega_0 := \{\|\hat{f}_n\|_\infty \leq 6\log^{1/2} n\} \cap \cap_{i=1}^n \{|\xi_i| \leq 2\log^{1/2} n\}$. Then for any $r \geq 0$, we have

$$\begin{aligned} F_n(r) &\leq \sup_{f \in \mathcal{F}_d \cap B_2(r,P) \cap B_\infty(6\log^{1/2} n)} \frac{1}{n} \sum_{i=1}^n \big\{ 2\tilde{\xi}_i f(X_i) - f^2(X_i) \big\} \\ &\leq \frac{1}{n} \sum_{i=1}^n \big\{ 2\tilde{\xi}_i \tilde{f}_n(X_i) - \tilde{f}_n^2(X_i) \big\} = F_n(\|\tilde{f}_n\|_{L_2(P)}). \end{aligned}$$

In other words, $\|\tilde{f}_n\|_{L_2(P)} \in \mathrm{argmax}_{r \geq 0} F_n(r)$.

If we can find $0 < r_1 < r_2$ such that

$$E_n(r_1, 6\log^{1/2} n) < F_n(r_2), \tag{5.5.16}$$

then for all $r \in [0, r_1]$, we have $F_n(r) \leq E_n(r_1, 6\log^{1/2} n) < F_n(r_2)$. This means that r_1 is a lower bound for $\operatorname{argmax}_{r\geq 0} F_n(r)$ and therefore

$$\|\hat{f}_n\|^2_{L_2(P)} \geq r_1^2 \mathbb{1}_{\Omega_0}. \tag{5.5.17}$$

It remains to choose suitable r_1 and r_2 that satisfy (5.5.16).

By (5.5.15), the symmetrisation inequality [162] Lemma 2.3.1, Lemmas 5.7.5 and 5.7.6 and Proposition 5.5.2, we have that for $r_1 \geq n^{-1/2}\log n$ and on $\Omega_{r_1, 6\log^{1/2} n}$,

$$\begin{aligned}
&E_n(r_1, 6\log^{1/2} n) \\
&\quad\leq 2\mathbb{E} \sup_{f\in\mathcal{F}_d\cap B_2(r_1,P)\cap B_\infty(6\log^{1/2} n)} \left\{\frac{2}{n}\sum_{i=1}^n \tilde{\xi}_i f(X_i) - \frac{1}{n^{1/2}}\mathbb{G}_n f^2\right\} + C_0 r_1^2 \\
&\quad\leq 104\log^{1/2} n\, \mathbb{E} \sup_{f\in\mathcal{F}_d\cap B_2(r_1,P)\cap B_\infty(6\log^{1/2} n)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| + C_0 r_1^2 \\
&\quad\leq C_{d,m_0,M_0} r_1 n^{-1/d}\log^{\gamma_d} n + C_0 r_1^2,
\end{aligned}$$

for some $C_{d,m_0,M_0} > 0$ depending only on d, m_0 and M_0. Similarly, for $r_2 \in [n^{-1/2}\log n, 1]$, $b \in [r_2, 6\log^{1/2} n]$ and on $\Omega_{r_2,b}$,

$$\begin{aligned}
F_n(r_2) &= E_n(r_2, 6\log^{1/2} n) - r_2^2 \\
&\geq \frac{1}{2}\mathbb{E} \sup_{f\in\mathcal{F}_d\cap B_2(r_2,P)\cap B_\infty(b)} \left\{\frac{2}{n}\sum_{i=1}^n \tilde{\xi}_i f(X_i) - \frac{1}{n^{1/2}}\mathbb{G}_n f^2\right\} - (C_0+1) r_2^2 \\
&\geq (\mathbb{E}|\tilde{\xi}_1| - 4b)\mathbb{E} \sup_{f\in\mathcal{F}_d\cap B_2(r_2,P)\cap B_\infty(b)} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) - (C_0+1) r_2^2 \\
&\geq (1/2 - 4b) c_{d,m_0} r_2 n^{-1/d} - (C_0+1) r_2^2,
\end{aligned}$$

for some $c_{d,m_0} > 0$ depending only on d and m_0. Hence, when $d \geq 3$, we can choose $b = 1/10$, $r_2 = (2C_0+2)^{-1}(1/2 - 4b) c_{d,m_0} n^{-1/d}$ and $r_1 = c'_{d,m_0,M_0} n^{-1/d}\log^{-\gamma_d} n$, where $c'_{d,m_0,M_0} > 0$ is chosen such that

$$C_{d,m_0,M_0} r_1 n^{-1/d}\log^{\gamma_d} n + C_0 r_1^2 < \frac{1}{2}\left(\frac{1}{2} - 4b\right) c_{d,m_0} r_2 n^{-1/d}.$$

We then see that for all n larger than some integer depending on d, m_0, M_0 only, (5.5.16) is satisfied. We therefore conclude from (5.5.17), Lemma 5.7.10 and the fact that $\mathbb{P}(|\xi_1| > 2\log^{1/2} n) \leq n^{-2}$ that

$$R(\hat{f}_n, 0) \geq \mathbb{E}\big\{\|\hat{f}_n\|^2_{L_2(P)} \mathbb{1}_{\Omega_0 \cap \Omega_{r_1, 6\log^{1/2} n} \cap \Omega_{r_2, b}}\big\} \gtrsim_{d, m_0, M_0} n^{-2/d} \log^{-2\gamma_d} n,$$

as desired. □

Proof of Proposition 5.3.4 in $L_2(\mathbb{P}_n)$ loss. Due to translation invariance we only need to establish the claim for $f_0 = 0$. By Lemma 5.7.4, there is an event $\mathcal{E}$ with probability at least $1 - e^{-ed^{-1}(M_0 n)^{1/d}\log(M_0 n)}$ on which the data points $X_1, \ldots, X_n$ contain an antichain W_X of cardinality at least $n^{1-1/d}/(2eM_0^{1/d})$. Write $W_X^+ := \{X_i : \exists w \in W_X, X_i \succ w\}$ and $W_X^- := \{X_i : \exists w \in W_X, X_i \prec w\}$. For each realisation of the n-dimensional Gaussian random vector ξ, we define $\theta_X = \theta_X(\xi) = ((\theta_X)_w)$ by

$$(\theta_X)_w := \begin{cases} 1 & \text{if } w \in W_X^+ \\ \operatorname{sgn}(\xi_w) & \text{if } w \in W_X \\ -1 & \text{if } w \in W_X^-, \end{cases}$$

so $\theta_X \in \mathcal{M}(G_X)$. By [36] Theorem 1.1, for $f_0 = 0$, we have that

$$n^{1/2}\big\|\hat{f}_n\big\|_{L_2(\mathbb{P}_n)} = \operatorname*{argmax}_{t \geq 0}\bigg(\sup_{\theta \in \mathcal{M}(G_X) \cap B_2(t)} \langle \xi, \theta\rangle - \frac{t^2}{2}\bigg) = \sup_{\theta \in \mathcal{M}(G_X) \cap B_2(1)} \langle \xi, \theta\rangle.$$

Hence

$$\begin{aligned} \mathbb{E}\big\|\hat{f}_n\big\|_{L_2(\mathbb{P}_n)} &= \frac{1}{n^{1/2}} \mathbb{E} \sup_{\theta \in \mathcal{M}(G_X) \cap B_2(1)} \langle \xi, \theta\rangle \geq \frac{1}{n^{1/2}} \mathbb{E}\bigg(\bigg\langle \xi, \frac{\theta_X(\xi)}{\|\theta_X(\xi)\|_2}\bigg\rangle \mathbb{1}_{\mathcal{E}}\bigg) \\ &= \frac{1}{n} \mathbb{E}\bigg(\sum_{i: X_i \in W_X^+} \xi_i \mathbb{1}_{\mathcal{E}} - \sum_{i: X_i \in W_X^-} \xi_i \mathbb{1}_{\mathcal{E}} + \sum_{i: X_i \in W_X} |\xi_i| \mathbb{1}_{\mathcal{E}}\bigg). \end{aligned} \tag{5.5.18}$$

The first two terms in the bracket are seen to be zero by computing the expectation conditionally on $X_1, \ldots, X_n$. For the third term, we have that

$$\begin{aligned} \mathbb{E}\bigg(\sum_{i: X_i \in W_X} |\xi_i| \mathbb{1}_{\mathcal{E}}\bigg) &= \mathbb{E} \sum_{i: X_i \in W_X} \mathbb{E}\big(|\xi_i| \mathbb{1}_{\mathcal{E}} \,\big|\, X_1, \ldots, X_n\big) \\ &\geq (2/\pi)^{1/2} \mathbb{E}\big(|W_X| \mathbb{1}_{\mathcal{E}}\big) \gtrsim_{d, M_0} n^{1-1/d}. \end{aligned} \tag{5.5.19}$$

By (5.5.18), (5.5.19) and the Cauchy–Schwarz inequality, we have that

$$\mathbb{E}\big\|\hat{f}_n\big\|_{L_2(\mathbb{P}_n)}^2 \geq \big\{\mathbb{E}\big\|\hat{f}_n\big\|_{L_2(\mathbb{P}_n)}\big\}^2 \gtrsim_{d,M_0} n^{-2/d},$$

as desired. □

5.6 Proofs of Preparatory Propositions

Proof of Proposition 5.5.1. For any $f : [0,1]^d \to \mathbb{R}$, define $\mathbb{M}_n f := 2\sum_{i=1}^n \xi_i\{f(X_i) - f_0(X_i)\} - \sum_{i=1}^n \{f(X_i) - f_0(X_i)\}^2$ and $Mf := \mathbb{E}\mathbb{M}_n f = -n\|f - f_0\|_{L_2(P)}^2$. By the definition of $\hat{f}_n$, we have that $\sum_{i=1}^n (\hat{f}_n(X_i) - f_0(X_i) - \xi_i)^2 \leq \sum_{i=1}^n \xi_i^2$, which implies that $\mathbb{M}_n \hat{f}_n \geq 0$. We therefore have that for any $r > 0$,

$$\begin{aligned}
&\mathbb{P}\big(\{\|\hat{f}_n - f_0\|_{L_2(P)} \geq r\} \cap \{\|\hat{f}_n - f_0\|_\infty \leq 6\log^{1/2} n\}\big) \\
&\leq \sum_{\ell=1}^\infty \mathbb{P}\bigg(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6\log^{1/2} n) \setminus \mathcal{G}(f_0, 2^{\ell-1} r, 6\log^{1/2} n)} (\mathbb{M}_n - M)f \geq n 2^{2\ell-2} r^2\bigg) \\
&\leq \sum_{\ell=1}^\infty \mathbb{P}\bigg(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6\log^{1/2} n)} \bigg|\frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i (f - f_0)(X_i)\bigg| \geq 2^{2\ell-4} n^{1/2} r^2\bigg) \\
&\qquad + \sum_{\ell=1}^\infty \mathbb{P}\bigg(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6\log^{1/2} n)} \big|\mathbb{G}_n (f - f_0)^2\big| \geq 2^{2\ell-3} n^{1/2} r^2\bigg). \qquad (5.6.1)
\end{aligned}$$

By a moment inequality for empirical processes [61] Proposition 3.1 and (5.5.2) in the main text, we have for all $p \geq 1$ that

$$\begin{aligned}
&\mathbb{E}\bigg[\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6\log^{1/2} n)} \bigg|\frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i \{f(X_i) - f_0(X_i)\}\bigg|^p\bigg]^{1/p} \\
&\qquad \lesssim K\phi_n(2^\ell r) + 2^\ell r p^{1/2} + n^{-1/2} p \log n. \qquad (5.6.2)
\end{aligned}$$

For any $C' > 0$ and $r \geq C' K r_n$, we have $\phi_n(2^\ell r) \leq 2^\ell (r/r_n)\phi_n(r_n) \leq 2^\ell n^{1/2} r_n r \leq (C'K)^{-1} 2^\ell n^{1/2} r^2$. It therefore follows from (5.6.2) and Lemma 5.7.11 that there exist universal constants $C, C' > 0$ such that for all $\ell \in \mathbb{N}$ and $r \geq C' K r_n$,

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6\log^{1/2} n)} \bigg|\frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i \{f(X_i) - f_0(X_i)\}\bigg| \geq 2^{2\ell-4} n^{1/2} r^2\bigg) \\
&\qquad \leq C \exp\bigg(-\frac{2^{2\ell} n r^2}{C \log n}\bigg). \qquad (5.6.3)
\end{aligned}$$

Similarly, by a symmetrisation inequality (cf. [162] Lemma 2.3.1), (5.5.3) in the main text and the same argument as above, and by increasing C, C' if necessary, we have that for all $\ell \in \mathbb{N}$ and $r \geq C'Kr_n$,

$$\mathbb{P}\bigg(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6\log^{1/2} n)} \Big|\mathbb{G}_n(f - f_0)^2\Big| \geq 2^{2\ell-3} n^{1/2} r^2\bigg) \leq C \exp\bigg(-\frac{2^{2\ell} n r^2}{C \log n}\bigg). \tag{5.6.4}$$

Substituting (5.6.3) and (5.6.4) into (5.6.1), we obtain that for all $r \geq C'Kr_n$,

$$\begin{aligned}\mathbb{P}\big(\big\{\|\hat{f}_n - f_0\|_{L_2(P)} \geq r\big\} \cap \big\{\|\hat{f}_n - f_0\|_\infty \leq 6\log^{1/2} n\big\}\big) \\ \lesssim \sum_{\ell=1}^{\infty} \exp\bigg(-\frac{2^{2\ell} n r^2}{C \log n}\bigg) \lesssim \exp\bigg(-\frac{n r^2}{C \log n}\bigg).\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbb{E}\big(\|\hat{f}_n - f_0\|_{L_2(P)}^2 \mathbb{1}_{\{\|\hat{f}_n - f_0\|_\infty \leq 6\log^{1/2} n\}}\big) \\ &\quad = \int_0^\infty 2t \mathbb{P}\big(\big\{\big\|\hat{f}_n - f_0\big\|_{L_2(P)} \geq t\big\} \cap \big\{\|\hat{f}_n - f_0\|_\infty \leq 6\log^{1/2} n\big\}\big)\, \mathrm{d}t \\ &\quad \lesssim K^2 r_n^2 + \int_{C'Kr_n}^\infty 2t \exp\bigg(-\frac{t^2}{C r_n^2}\bigg)\, \mathrm{d}t \lesssim K^2 r_n^2,\end{aligned}$$

as desired, where we have used $r_n^2 \geq n^{-1} \log n$ in the penultimate inequality. □

Proof of Proposition 5.5.2. [Upper bound] It is convenient here to work with the class of block decreasing functions $\mathcal{F}_{d,\downarrow} := \{f : [0,1]^d \to \mathbb{R} : -f \in \mathcal{F}_d\}$ instead. We write $\mathcal{F}_d^+ := \{f \in \mathcal{F}_d : f \geq 0\}$ and $\mathcal{F}_{d,\downarrow}^+ := \{f \in \mathcal{F}_{d,\downarrow} : f \geq 0\}$. By replacing f with $-f$ and decomposing any function f into its positive and negative parts, it suffices to prove the result with $\mathcal{G}_\downarrow^+(0,r,1) := \mathcal{F}_{d,\downarrow}^+ \cap B_2(r,P) \cap B_\infty(1)$ in place of $\mathcal{G}(0,r,1)$. Since $\mathcal{G}_\downarrow^+(0,r,1) = \mathcal{G}_\downarrow^+(0,1,1)$ for $r \geq 1$, we may also assume without loss of generality that $r \leq 1$. We handle the cases $d = 2$ and $d \geq 3$ separately.

Case $d = 2$. We apply Lemma 5.7.7 with $\eta = r/(2n)$ and Lemma 5.7.8 to obtain

$$\begin{aligned}&\mathbb{E} \sup_{f \in \mathcal{F}_{2,\downarrow}^+ \cap B_2(r,P) \cap B_\infty(1)} \bigg|\frac{1}{n^{1/2}} \sum_{i=1}^n \varepsilon_i f(X_i)\bigg| \\ &\quad \lesssim_{d, m_0, M_0} n^{1/2}\eta + \log^3 n \int_\eta^r \frac{r}{\varepsilon}\, \mathrm{d}\varepsilon + \frac{(\log^4 n)(\log\log n)^2}{n^{1/2}} \lesssim r \log^4 n,\end{aligned}$$

as desired.

Case $d \geq 3$. We assume without loss of generality that $n = n_1^d$ for some $n_1 \in \mathbb{N}$. We define strips $I_\ell := [0,1]^{d-1} \times [\frac{\ell-1}{n_1}, \frac{\ell}{n_1}]$ for $\ell = 1, \ldots, n_1$, so that $[0,1]^d = \cup_{\ell=1}^{n_1} I_\ell$. Our strategy is to analyse the expected supremum of the symmetrised empirical process when restricted to each strip. To this end, define $S_\ell := \{X_1, \ldots, X_n\} \cap I_\ell$ and $N_\ell := |S_\ell|$, and let $\Omega_0 := \{m_0 n^{1-1/d}/2 \leq \min_\ell N_\ell \leq \max_\ell N_\ell \leq 2M_0 n^{1-1/d}\}$. Then by Hoeffding's inequality,

$$\mathbb{P}(\Omega_0^c) \leq \sum_{\ell=1}^{n_1} \mathbb{P}\biggl(\Big|N_\ell - \mathbb{E}N_\ell\Big| > \frac{m_0 n}{2n_1}\biggr) \leq 2n_1 \exp\bigl(-m_0^2 n^{1-2/d}/8\bigr).$$

Hence we have

$$\begin{aligned}\mathbb{E} \sup_{f \in \mathcal{F}_{d,\downarrow}^+ \cap B_2(r,P) \cap B_\infty(1)} &\Big|\frac{1}{n^{1/2}} \sum_{i=1}^n \varepsilon_i f(X_i)\Big| \\ &\leq \mathbb{E}\biggl(\sum_{\ell: N_\ell \geq 1} \frac{N_\ell^{1/2}}{n^{1/2}} E_\ell \, \mathbb{1}_{\Omega_0}\biggr) + C\exp\bigl(-m_0^2 n^{1-2/d}/16\bigr), \end{aligned} \tag{5.6.5}$$

where

$$E_\ell := \mathbb{E}\biggl\{\sup_{f \in \mathcal{F}_{d,\downarrow}^+ \cap B_2(r,P) \cap B_\infty(1)} \Big|\frac{1}{N_\ell^{1/2}} \sum_{i: X_i \in S_\ell} \varepsilon_i f(X_i)\Big| \;\Big|\; N_1, \ldots, N_{n_1}\biggr\}.$$

By Lemma 5.7.9, for any $f \in \mathcal{F}_{d,\downarrow}^+ \cap B_2(r,P) \cap B_\infty(1)$ and $\ell \in \{1, \ldots, n_1\}$, we have $\int_{I_\ell} f^2 \, \mathrm{d}P \leq 7(M_0/m_0)\ell^{-1} r^2 \log^d n =: r_{n,\ell}^2$. Consequently, we have by Lemma 5.7.7 that for any $\eta \in [0, r_{n,\ell}/3)$,

$$E_\ell \lesssim N_\ell^{1/2}\eta + \int_\eta^{r_{n,\ell}} H_{[],\ell}^{1/2}(\varepsilon) \, \mathrm{d}\varepsilon + \frac{H_{[],\ell}(r_{n,\ell})}{N_\ell^{1/2}}, \tag{5.6.6}$$

where $H_{[],\ell}(\varepsilon) := \log N_{[]}\bigl(\varepsilon, \mathcal{F}_{d,\downarrow}^+(I_\ell) \cap B_2(r_{n,\ell}, P; I_\ell) \cap B_\infty(1; I_\ell), \|\cdot\|_{L_2(P;I_\ell)}\bigr)$. Here, the set $\mathcal{F}_{d,\downarrow}^+(I_\ell)$ is the class of non-negative functions on I_ℓ that are block decreasing, $B_\infty(1; I_\ell)$ is the class of functions on I_ℓ that are bounded by 1 and $B_2(r_{n,\ell}, P; I_\ell)$ is the class of measurable functions f on I_ℓ with $\|f\|_{L_2(P;I_\ell)} \leq r_{n,\ell}$. Note that any $g \in \mathcal{F}_{d,\downarrow}^+(I_\ell) \cap B_2(r_{n,\ell}, P; I_\ell) \cap B_\infty(1; I_\ell)$ can be rescaled into a function $f_g \in \mathcal{F}_{d,\downarrow}^+ \cap B_2\bigl(n_1^{1/2}(M_0/m_0)^{1/2} r_{n,\ell}, P\bigr) \cap B_\infty(1)$ via the invertible map $f_g(x_1, \ldots, x_{d-1}, x_d) := g(x_1, \ldots, x_{d-1}, (x_d + \ell - 1)/n_1)$. Moreover, we have

$\int_{[0,1]^d}(f_g - f_{g'})^2 \,\mathrm{d}P \geq n_1(m_0/M_0)\int_{I_\ell}(g-g')^2\,\mathrm{d}P$. Thus, by Lemma 5.7.8, for $\varepsilon \in [\eta, r_{n,\ell}]$,

$$\begin{aligned} H_{[],\ell}(\varepsilon) &\leq \log N_{[]}\big(n^{1/(2d)}(m_0/M_0)^{1/2}\varepsilon, \\ &\qquad \mathcal{F}^+_{d,\downarrow} \cap B_2\big(n^{1/(2d)}(M_0/m_0)^{1/2} r_{n,\ell}, P\big) \cap B_\infty(1), \|\cdot\|_{L_2(P)}\big) \\ &\lesssim_{d,m_0,M_0} \left(\frac{r_{n,\ell}}{\varepsilon}\right)^{2(d-1)} \log_+^{d^2}(1/\varepsilon). \end{aligned}$$

Substituting the above bound into (5.6.6), and choosing $\eta = n^{-1/(2d)} r_{n,\ell}$, we obtain

$$\begin{aligned} E_\ell &\lesssim_{d,m_0,M_0} N_\ell^{1/2}\eta + \log^{d^2/2} n \int_\eta^{r_{n,\ell}} \left(\frac{r_{n,\ell}}{\varepsilon}\right)^{d-1} \mathrm{d}\varepsilon + \frac{\log^{d^2} n}{N_\ell^{1/2}} \\ &\lesssim N_\ell^{1/2}\eta + \frac{r_{n,\ell}^{d-1}\log^{d^2/2} n}{\eta^{d-2}} + \frac{\log^{d^2} n}{N_\ell^{1/2}}. \end{aligned}$$

Hence

$$\begin{aligned} E_\ell \mathbb{1}_{\Omega_0} &\lesssim_{d,m_0,M_0} r_{n,\ell}\, n^{1/2-1/d}\log^{d^2/2} n + n^{-1/2+1/(2d)}\log^{d^2} n \\ &\lesssim_{m_0,M_0} r_{n,\ell}\, n^{1/2-1/d}\log^{d^2/2} n, \end{aligned} \tag{5.6.7}$$

where in the final inequality we used the conditions that $d \geq 3$ and $r \geq n^{-(1-2/d)}\log^{(d^2-d)/2} n$. Combining (5.6.5) and (5.6.7), we have that

$$\begin{aligned} &\mathbb{E} \sup_{f \in \mathcal{F}^+_{d,\downarrow} \cap B_2(r,P) \cap B_\infty(1)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \\ &\lesssim_{d,m_0,M_0} r n^{1/2-3/(2d)} \log^{(d^2+d)/2} n \sum_{\ell=1}^{n_1} \ell^{-1/2} \lesssim r n^{1/2-1/d}\log^{(d^2+d)/2} n, \end{aligned}$$

which completes the proof.

[Lower bound] Assume without of loss of generality that $n = n_1^d$ for some $n_1 \in \mathbb{N}$. For a multi-index $w = (w_1, \ldots, w_d) \in \mathbb{L}_{d,n}$, let $L_w := \prod_{j=1}^d (w_j - 1/n_1, w_j]$ and $N_w := |\{X_1, \ldots, X_n\} \cap L_w|$. We define $W := \{(w_1, \ldots, w_d) : \sum_{j=1}^d w_j = 1\}$ to be indices of a mutually incomparable collection of cubelets and define $\tilde{W} := \{w \in W : N_w \geq 1\}$ to be the (random) set of indices of cubelets in this collection that contain at least one design point. For each $w \in \tilde{W}$, associate $i_w := \min\{i : X_i \in L_w\}$. For each realisation of the Rademacher

random variables $\varepsilon = (\varepsilon_i)_{i=1}^n$ and design points $X = \{X_i\}_{i=1}^n$, define $f_{\varepsilon,X} : [0,1]^d \to [-1,1]$ to be the function such that

$$f_{\varepsilon,X}(x) := \begin{cases} r\,\varepsilon_{i_w} & \text{if } x \in L_w,\ w \in \tilde{W} \\ r & \text{if } x \in L_w \text{ with } \sum_{j=1}^d w_j > n_1 \\ -r & \text{otherwise.} \end{cases}$$

For $r \leq 1$, we have $f_{\varepsilon,X} \in \mathcal{F}_d \cap B_2(r,P) \cap B_\infty(1)$. Therefore,

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r,P) \cap B_\infty(1)} \sum_{i=1}^n \varepsilon_i f(X_i) &\geq \mathbb{E} \sum_{i=1}^n \varepsilon_i f_{\varepsilon,X}(X_i) \\ &\geq \mathbb{E}\Big[\mathbb{E}\Big\{\sum_{i=1}^n \varepsilon_i f_{\varepsilon,X}(X_i) \;\Big|\; X_1,\ldots,X_n, \{\varepsilon_{i_w} : w \in \tilde{W}\}\Big\}\Big] \\ &= \mathbb{E} \sum_{w \in \tilde{W}} \varepsilon_{i_w} f_{\varepsilon,X}(X_{i_w}) = r\mathbb{E}|\tilde{W}|. \end{aligned}$$

The desired lower bound follows since $\mathbb{E}|\tilde{W}| \geq \{1-(1-m_0/n)^n\}|W| \geq (1-e^{-m_0})|W| \gtrsim_{d,m_0} n^{1-1/d}$, where the final bound follows as in the proof of Proposition 5.4.1. □

Proof of Proposition 5.5.3. Let $r_n := n^{-1/d}\log^{\gamma_d} n$. We write

$$\mathbb{E}\big\|\tilde{f}_n\big\|_{L_2(\mathbb{P}_n)}^2 = \mathbb{E}\big\{\big\|\tilde{f}_n\big\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} \leq r_n\}}\big\} + \mathbb{E}\big\{\big\|\tilde{f}_n\big\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} > r_n\}}\big\} \tag{5.6.8}$$

and control the two terms on the right hand side of (5.6.8) separately. For the first term, we have

$$\begin{aligned} \mathbb{E}\big\{\big\|\tilde{f}_n\big\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} \leq r_n\}}\big\} &\leq \mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_n,P) \cap B_\infty(6\log^{1/2} n)} \frac{1}{n}\sum_{i=1}^n f^2(X_i) \\ &\lesssim r_n^2 + \frac{1}{n}\mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_n,P) \cap B_\infty(6\log^{1/2} n)} \Big|\sum_{i=1}^n \varepsilon_i f^2(X_i)\Big| \\ &\lesssim r_n^2 + \frac{\log^{1/2} n}{n}\mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_n,P) \cap B_\infty(6\log^{1/2} n)} \Big|\sum_{i=1}^n \varepsilon_i f(X_i)\Big| \\ &\lesssim_{d,m_0,M_0} r_n^2 + r_n n^{-1/d}\log^{\gamma_d} n \lesssim r_n^2, \end{aligned} \tag{5.6.9}$$

where the second line uses the symmetrisation inequality [cf. [162] Lemma 2.3.1], the third inequality follows from Lemma 5.7.6 and the penultimate inequality follows from Proposition 5.5.2. For the second term on the right-hand side of (5.6.8), we first claim that there exists $C'_{d,m_0,M_0} > 0$, depending only on d, m_0 and M_0, such that

$$\mathbb{P}(\mathcal{E}^c) \leq \frac{2}{n^2}, \tag{5.6.10}$$

where

$$\mathcal{E} := \left\{ \sup_{f \in \mathcal{F}_d \cap B_2(r_n,P)^c \cap B_\infty(6\log^{1/2} n)} \left| \frac{\mathbb{P}_n f^2}{Pf^2} - 1 \right| \leq C'_{d,m_0,M_0} \right\}.$$

To see this, we adopt a peeling argument as follows. Let $\mathcal{F}_{d,\ell} := \{f \in \mathcal{F}_d \cap B_\infty(6\log^{1/2} n) : 2^{\ell-1} r_n^2 < Pf^2 \leq 2^\ell r_n^2\}$ and let m be the largest integer such that $2^m r_n^2 < 32 \log n$ (so that $m \asymp \log n$). We have that

$$\sup_{\substack{f \in \mathcal{F}_d \cap B_\infty(6\log^{1/2} n) \\ \|f\|_{L_2(P)} > r_n}} \left| \frac{\mathbb{P}_n f^2}{Pf^2} - 1 \right| \leq \frac{2}{n^{1/2}} \max_{\ell=1,\ldots,m} \Big\{ (2^\ell r_n^2)^{-1} \sup_{f \in \mathcal{F}_{d,\ell}} |\mathbb{G}_n f^2| \Big\}.$$

By Talagrand's concentration inequality for empirical processes [145] in the form given by [106] Theorem 3, applied to the class $\{f^2 : f \in \mathcal{F}_{d,\ell}\}$, we have that for any $s_\ell > 0$,

$$\mathbb{P}\Big\{ \sup_{f \in \mathcal{F}_{d,\ell}} |\mathbb{G}_n f^2| > 2\mathbb{E} \sup_{f \in \mathcal{F}_{d,\ell}} |\mathbb{G}_n f^2| + 12\sqrt{2}\big(2^\ell s_\ell \log n\big)^{1/2} r_n + \frac{1242 s_\ell \log n}{n^{1/2}} \Big\} \leq e^{-s_\ell}.$$

Here we have used the fact that $\sup_{f \in \mathcal{F}_{d,\ell}} \mathrm{Var}_P f^2 \leq \sup_{f \in \mathcal{F}_{d,\ell}} Pf^2 \|f\|_\infty^2 \leq 36 \cdot 2^\ell r_n^2 \log n$. Further, by the symmetrisation inequality again, Lemma 5.7.6 and Proposition 5.5.2, we have that

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_{d,\ell}} |\mathbb{G}_n f^2| &\leq \frac{2}{n^{1/2}} \mathbb{E} \sup_{f \in \mathcal{F}_{d,\ell}} \left| \sum_{i=1}^n \varepsilon_i f^2(X_i) \right| \leq \frac{48 \log^{1/2} n}{n^{1/2}} \mathbb{E} \sup_{f \in \mathcal{F}_{d,\ell}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \\ &\lesssim_{d,m_0,M_0} 2^{\ell/2} r_n n^{1/2-1/d} \log^{\gamma_d} n. \end{aligned}$$

By a union bound, we have that with probability at least $1 - \sum_{\ell=1}^m e^{-s_\ell}$,

$$\sup_{f \in \mathcal{F}_d \cap B_2(r_n, P)^c \cap B_\infty(6 \log^{1/2} n)} \left| \frac{\mathbb{P}_n f^2}{P f^2} - 1 \right| \lesssim_{d, m_0, M_0} \max_{\ell = 1, \ldots, m} \left\{ \frac{n^{1/2 - 1/d} \log^{\gamma_d} n + s_\ell^{1/2} \log^{1/2} n}{2^{\ell/2} n^{1/2} r_n} + \frac{s_\ell \log n}{2^\ell n r_n^2} \right\}.$$

By choosing $s_\ell := 2^\ell \log n$, we see that on an event of probability at least $1 - \sum_{\ell=1}^m e^{-s_\ell} \geq 1 - \sum_{\ell=1}^\infty n^{-\ell-1} \geq 1 - 2n^{-2}$, we have

$$\sup_{f \in \mathcal{F}_d \cap B_2(r_n, P)^c \cap B_\infty(6 \log^{1/2} n)} \left| \frac{\mathbb{P}_n f^2}{P f^2} - 1 \right| \lesssim_{d, m_0, M_0} 1,$$

which verifies (5.6.10). Thus

$$\begin{aligned} \mathbb{E}\big\{ \big\| \tilde{f}_n \big\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} > r_n\}} \big\} &\leq \mathbb{E}\big\{ \big\| \tilde{f}_n \big\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} > r_n\}} \mathbb{1}_{\mathcal{E}} \big\} + \frac{72 \log n}{n^2} \\ &\leq (C'_{d, m_0, M_0} + 1) \mathbb{E} \big\| \tilde{f}_n \big\|_{L_2(P)}^2 + \frac{72 \log n}{n^2}. \end{aligned} \tag{5.6.11}$$

Combining (5.6.8), (5.6.9) and (5.6.11), we obtain

$$\mathbb{E} \big\| \tilde{f}_n \big\|_{L_2(\mathbb{P}_n)}^2 \lesssim_{d, m_0, M_0} r_n^2 + \mathbb{E} \big\| \tilde{f}_n \big\|_{L_2(P)}^2,$$

as desired. □

Proof of Proposition 5.5.4. Let $r_n := n^{-1/d} \log^{\gamma_d} n$ and observe that by Lemma 5.7.5 and Proposition 5.5.2, we have that for $r \geq r_n$,

$$\mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r, P) \cap B_\infty(6 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \right| \lesssim_{d, m_0, M_0} r n^{1/2 - 1/d} \log^{\gamma_d} n.$$

On the other hand, by Lemma 5.7.6 and Proposition 5.5.2, for $r \geq r_n$,

$$\mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r, P) \cap B_\infty(6 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \varepsilon_i f^2(X_i) \right| \lesssim_{d, m_0, M_0} r n^{1/2 - 1/d} \log^{\gamma_d} n.$$

It follows that the conditions of Proposition 5.5.1 are satisfied for this choice of r_n with $\phi_n(r) := r n^{1/2 - 1/d} \log^{\gamma_d} n$ and $K \lesssim_{d, m_0, M_0} 1$. By Lemma 5.7.10, Propositions 5.5.3 and 5.5.1,

we have that

$$R_n(\hat{f}_n, 0) \leq \mathbb{E}\|\tilde{f}_n\|^2_{L_2(\mathbb{P}_n)} + n^{-1}$$
$$\lesssim_{d,m_0,M_0} n^{-2/d}\log^{2\gamma_d} n + \mathbb{E}\|\tilde{f}_n\|^2_{L_2(P)} \lesssim_{d,m_0,M_0} n^{-2/d}\log^{2\gamma_d} n,$$

as desired. □

5.7 Ancillary lemmas

We collect here various ancillary results used in the proofs.

The proof of Corollary 5.2.3 in the main document requires the following lemma on Riemann approximation of block increasing functions.

Lemma 5.7.1. *Suppose $n_1 = n^{1/d}$ is a positive integer. For any $f \in \mathcal{F}_d$, define $f_L(x_1, \ldots, x_d) := f\big(n_1^{-1}\lfloor n_1 x_1\rfloor, \ldots, n_1^{-1}\lfloor n_1 x_d\rfloor\big)$ and $f_U(x_1, \ldots, x_d) := f\big(n_1^{-1}\lceil n_1 x_1\rceil, \ldots, n_1^{-1}\lceil n_1 x_d\rceil\big)$. Then*

$$\int_{[0,1]^d} (f_U - f_L)^2 \leq 4dn^{-1/d}\|f\|^2_\infty.$$

Proof. For $x = (x_1, \ldots, x_d)^\top$ and $x' = (x'_1, \ldots, x'_d)^\top$ in $\mathbb{L}_{d,n}$, we say x and x' are equivalent if and only if $x_j - x_1 = x'_j - x'_1$ for $j = 1, \ldots, d$. Let $\mathbb{L}_{d,n} = \bigsqcup_{r=1}^N P_r$ be the partition of $\mathbb{L}_{d,n}$ into equivalence classes. Since each P_r has non-empty intersection with a different element of the set $\{(x_1, \ldots, x_d) \in \mathbb{L}_{d,n} : \max_j x_j = 1\}$, we must have $N \leq dn^{1-1/d}$. Therefore, we have

$$\int_{[0,1]^d} (f_U - f_L)^2 = \sum_{r=1}^N \int_{P_r + n_1^{-1}(-1,0]^d} (f_U - f_L)^2$$
$$\leq \frac{2}{n}\|f\|_\infty \sum_{r=1}^N \sum_{x=(x_1,\ldots,x_d)^\top \in P_r} \big\{f(x_1, \ldots, x_d) - f\big(x_1 - n_1^{-1}, \ldots, x_d - n_1^{-1}\big)\big\}$$
$$\leq \frac{2N}{n}\|f\|_\infty\big(f(1, \ldots, 1) - f(0, \ldots, 0)\big) \leq 4dn^{-1/d}\|f\|^2_\infty,$$

as desired. □

The following is a simple generalisation of Jensen's inequality.

Lemma 5.7.2. *Suppose $h : [0,\infty) \to (0,\infty)$ is a non-decreasing function satisfying the following:*

(i) *There exists $x_0 \geq 0$ such that h is concave on $[x_0,\infty)$.*

(ii) *There exists some $x_1 > x_0$ such that $h(x_1) - x_1 h'_+(x_1) \geq h(x_0)$, where h'_+ is the right derivative of h.*

Then there exists $C_h > 0$, depending only on h, such that for any nonnegative random variable X with $\mathbb{E}X < \infty$, we have

$$\mathbb{E}h(X) \leq C_h h(\mathbb{E}X).$$

Proof. Define $H : [0,\infty) \to [h(0),\infty)$ by

$$H(x) := \begin{cases} h(x_1) - x_1 h'_+(x_1) + x h'_+(x_1) & \text{if } x \in [0, x_1) \\ h(x) & \text{if } x \in [x_1, \infty). \end{cases}$$

Then H is a concave majorant of h. Moreover, we have $H \leq (h(x_1)/h(0))h$. Hence, by Jensen's inequality, we have

$$\mathbb{E}h(X) \leq \mathbb{E}H(X) \leq H(\mathbb{E}X) \leq \frac{h(x_1)}{h(0)} h(\mathbb{E}X),$$

as desired. □

We need the following lower bound on the metric entropy of $\mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1)$ for the proof of Proposition 5.2.6.

Lemma 5.7.3. *There exist universal constants $c > 0$ and $\varepsilon_0 > 0$ such that*

$$\log N\big(\varepsilon_0, \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1), \|\cdot\|_2\big) \geq c \log^2 n.$$

Proof. It suffices to prove the equivalent result that there exist universal constants $c, \varepsilon_0 > 0$ such that the packing number $D\big(\varepsilon_0, \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1), \|\cdot\|_2\big)$ (i.e. the maximum number of disjoint open Euclidean balls of radius ε_0 that can be fitted into $\mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1)$) is at least

$\exp(c\log^2 n)$. Without loss of generality, we may also assume that $n_1 := n^{1/2} = 2^\ell - 1$ for some $\ell \in \mathbb{N}$, so that $\ell \asymp \log n$. Now, for $r = 1, \ldots, \ell$, let $I_r := n_1^{-1}\{2^{r-1}, \ldots, 2^r - 1\}$ and consider the set

$$\bar{\mathcal{M}} := \left\{\theta \in \mathbb{R}^{\mathbb{L}_{2,n}} : \theta_{I_r \times I_s} \in \left\{\frac{-\mathbf{1}_{I_r \times I_s}}{\sqrt{2^{r+s+1}}\log n}, \frac{-\mathbf{1}_{I_r \times I_s}}{\sqrt{2^{r+s}}\log n}\right\}\right\} \subseteq \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1),$$

where $\mathbf{1}_{I_r \times I_s}$ denotes the all-one vector on $I_r \times I_s$. Define a bijection $\psi : \bar{\mathcal{M}} \to \{0,1\}^{\ell^2}$ by

$$\psi(\theta) := \left(\mathbb{1}_{\left\{\theta_{I_r \times I_s} = -\mathbf{1}_{I_r \times I_s}/\sqrt{2^{r+s+1}}\log n\right\}}\right)_{r,s=1}^{\ell}.$$

Then, for $\theta, \theta' \in \bar{\mathcal{M}}$,

$$\|\theta - \theta'\|_2^2 = \frac{d_{\mathrm{H}}(\psi(\theta), \psi(\theta'))}{\log^2 n}\frac{1}{4}\left(1 - \frac{1}{2^{1/2}}\right)^2,$$

where $d_{\mathrm{H}}(\cdot,\cdot)$ denotes the Hamming distance. On the other hand, by the Gilbert–Varshamov inequality [e.g. [107] Lemma 4.7], there exists a subset $\mathcal{I} \subseteq \{0,1\}^{\ell^2}$ such that $|\mathcal{I}| \geq \exp(\ell^2/8)$ and $d_{\mathrm{H}}(v, v') \geq \ell^2/4$ for any distinct $v, v' \in \mathcal{I}$. Then the set $\psi^{-1}(\mathcal{I}) \subseteq \bar{\mathcal{M}}$ has cardinality at least $\exp(\ell^2/8) \geq \exp(\log^2 n/32)$, and each pair of distinct elements have squared ℓ_2 distance at least $\varepsilon_0 := \frac{\ell^2/4}{\log^2 n}\frac{1}{4}(1 - \frac{1}{2^{1/2}})^2 \gtrsim 1$, as desired. □

Lemma 5.7.4 below gives a lower bound on the size of the maximal antichain (with respect to the natural partial ordering on $\mathbb{R}^d$) among independent and identically distributed $X_1, \ldots, X_n$.

Lemma 5.7.4. *Let $d \geq 2$. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$, where P is a distribution on $[0,1]^d$ with Lebesgue density bounded above by $M_0 \in [1, \infty)$. Then with probability at least $1 - e^{-ed^{-1}(M_0 n)^{1/d}\log(M_0 n)}$, there is an antichain in G_X with cardinality at least $n^{1-1/d}/(2eM_0^{1/d})$.*

Proof. By Dilworth's Theorem [44], for each realisation of the directed acyclic graph G_X, there exists a covering of $V(G_X)$ by chains $\mathcal{C}_1, \ldots, \mathcal{C}_M$, where M denotes the cardinality of a maximum antichain of G_X. Thus, it suffices to show that with the given probability, the

maximum chain length of G_X is at most $k := \lceil e(M_0 n)^{1/d} \rceil \leq 2e(M_0 n)^{1/d}$. By a union bound, we have that

$$\begin{aligned}\mathbb{P}(\exists \text{ a chain of length } k \text{ in } G_X) &\leq \frac{n!}{(n-k)!}\mathbb{P}(X_1 \preceq \cdots \preceq X_k) \\ &\leq \binom{n}{k}(k!)^{-(d-1)}M_0^k \leq \left(\frac{en}{k}\right)^k \left(\frac{k}{e}\right)^{-k(d-1)} M_0^k \\ &\leq (M_0 n)^{-k/d} \leq e^{-ed^{-1}(M_0 n)^{1/d}\log(M_0 n)},\end{aligned}$$

as desired. □

The following two lemmas control the empirical processes in (5.5.2) and (5.5.3) in the main text by the symmetrised empirical process in (5.5.4) in the main text.

Lemma 5.7.5. *Let $n \geq 2$, and suppose that $X_1, \ldots, X_n, \tilde{\xi}_1, \ldots, \tilde{\xi}_n$ are independent, with $X_1, \ldots, X_n$ identically distributed on $\mathcal{X}$ and $\tilde{\xi}_1, \ldots, \tilde{\xi}_n$ identically distributed, with $|\tilde{\xi}_1|$ stochastically dominated by $|\xi_1|$. Then for any countable class $\mathcal{F}$ of measurable, real-valued functions defined on $\mathcal{X}$, we have*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \tilde{\xi}_i f(X_i)\right| \leq 2\log^{1/2} n\, \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|.$$

Proof. Let $\alpha_0 := 0$, and for $k = 1, \ldots, n$, let $\alpha_k := \mathbb{E}|\tilde{\xi}_{(k)}|$, where $|\tilde{\xi}_{(1)}| \leq \cdots \leq |\tilde{\xi}_{(n)}|$ are the order statistics of $\{|\tilde{\xi}_1|, \ldots, |\tilde{\xi}_n|\}$, so that $\alpha_n \leq (2\log n)^{1/2}$. Observe that for any $k = 1, \ldots, n$,

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^k \varepsilon_i f(X_i)\right| &= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^k \varepsilon_i f(X_i) + \mathbb{E}\sum_{i=k+1}^n \varepsilon_i f(X_i)\right| \\ &\leq \mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{E}\left\{\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \,\middle|\, X_1, \ldots, X_k, \varepsilon_1, \ldots, \varepsilon_k\right\} \\ &\leq \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|. \qquad (5.7.1)\end{aligned}$$

We deduce from [76] Proposition 1 and (5.7.1) that

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \tilde{\xi}_i f(X_i)\right| &\leq 2^{1/2}\sum_{k=1}^n(\alpha_{n+1-k} - \alpha_{n-k})\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^k \varepsilon_i f(X_i)\right| \\ &\leq 2^{1/2}\alpha_n \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|,\end{aligned}$$

as required. □

Lemma 5.7.6. *Let $X_1, \ldots, X_n$ be random variables taking values in $\mathcal{X}$ and $\mathcal{F}$ be a countable class of measurable functions $f : \mathcal{X} \to [-1, 1]$. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f^2(X_i) \right| \leq 4 \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|.$$

Proof. By [93] Theorem 4.12, applied to $\phi_i(y) = y^2/2$ for $i = 1, \ldots, n$ (note that $y \mapsto y^2/2$ is a contraction on $[0, 1]$), we have

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f^2(X_i) \right| &= \mathbb{E} \left\{ \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f^2(X_i) \right| \;\middle|\; X_1, \ldots, X_n \right\} \\ &\leq 4 \mathbb{E} \left\{ \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \;\middle|\; X_1, \ldots, X_n \right\} = 4 \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|, \end{aligned}$$

as required. □

The following is a local maximal inequality for empirical processes under bracketing entropy conditions. This result is well known for $\eta = 0$ in the literature, but we provide a proof for the general case $\eta \geq 0$ for the convenience of the reader.

Lemma 5.7.7. *Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$ on $\mathcal{X}$ with empirical distribution $\mathbb{P}_n$, and, for some $r > 0$, let $\mathcal{G} \subseteq B_2(r, P) \cap B_\infty(1)$ be a countable class of measurable functions. Then for any $\eta \in [0, r/3)$, we have*

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{G}} |\mathbb{G}_n f| \lesssim n^{1/2} \eta + \int_\eta^r \log_+^{1/2} N_{[]}(\varepsilon, \mathcal{G}, \| \cdot \|_{L_2(P)}) \, \mathrm{d}\varepsilon \\ + \frac{1}{n^{1/2}} \log_+ N_{[]}(r, \mathcal{G}, \| \cdot \|_{L_2(P)}). \end{aligned}$$

The above inequality also holds if we replace $\mathbb{G}_n f$ with the symmetrised empirical process $n^{-1/2} \sum_{i=1}^n \varepsilon_i f(X_i)$.

Proof. Writing $N_r := N_{[]}(r, \mathcal{G}, \| \cdot \|_{L_2(P)})$, there exists $\{(f_\ell^L, f_\ell^U) : \ell = 1, \ldots, N_r\}$ that form an r-bracketing set for $\mathcal{G}$ in the $L_2(P)$ norm. Letting $\mathcal{G}_1 := \{f \in \mathcal{G} : f_1^L \leq f \leq f_1^U\}$ and

$\mathcal{G}_\ell := \{f \in \mathcal{G} : f_\ell^L \le f \le f_\ell^U\} \setminus \cup_{j=1}^{\ell-1} \mathcal{G}_j$ for $\ell = 2, \dots, N_r$, we see that $\{\mathcal{G}_\ell\}_{\ell=1}^{N_r}$ is a partition of $\mathcal{G}$ such that the $L_2(P)$-diameter of each $\mathcal{G}_\ell$ is at most r. It follows by [162] Lemma 2.14.3 that for any choice of $f_\ell \in \mathcal{G}_\ell$, we have that

$$\mathbb{E} \sup_{f \in \mathcal{G}} |\mathbb{G}_n f| \lesssim n^{1/2}\eta + \int_\eta^r \log_+^{1/2} N_{[]}(\varepsilon, \mathcal{G}, \|\cdot\|_{L_2(P)}) \, \mathrm{d}\varepsilon \\ + \mathbb{E} \max_{\ell=1,\dots,N_r} |\mathbb{G}_n f_\ell| + \mathbb{E} \max_{\ell=1,\dots,N_r} \Big| \mathbb{G}_n \Big(\sup_{f \in \mathcal{G}_\ell} |f - f_\ell| \Big) \Big|. \tag{5.7.2}$$

The third and fourth terms of (5.7.2) can be controlled by Bernstein's inequality (in the form of (2.5.5) in [162]):

$$\mathbb{E} \max_{\ell=1,\dots,N_r} |\mathbb{G}_n f_\ell| \vee \mathbb{E} \max_{\ell=1,\dots,N_r} \Big| \mathbb{G}_n \Big(\sup_{f \in \mathcal{G}_\ell} |f - f_\ell| \Big) \Big| \lesssim \frac{\log_+ N_r}{n^{1/2}} + r \log_+^{1/2} N_r.$$

Since $\eta < r/3$, the last term $r \log_+^{1/2} N_r$ in the above display can be assimilated into the entropy integral in (5.7.2), which establishes the claim for $\mathbb{E} \sup_{f \in \mathcal{G}} |\mathbb{G}_n f|$.

We now study the symmetrised empirical process. For $f \in \mathcal{G}$, we define $e \otimes f : \{-1, 1\} \times \mathcal{X} \to \mathbb{R}$ by $(e \otimes f)(t, x) := t f(x)$, and apply the previous result to the function class $e \otimes \mathcal{G} := \{e \otimes f : f \in \mathcal{G}\} \subseteq B_2(r, P_\varepsilon \otimes P) \cap B_\infty(1)$, where P_ε denotes the Rademacher distribution on $\{-1, 1\}$. Here the randomness is induced by the independently and identically distributed pairs $(\varepsilon_i, X_i)_{i=1}^n$. For any $f \in \mathcal{G}$ and any ε-bracket $[\underline{f}, \bar{f}]$ containing f, we have that $[e_+ \otimes \underline{f} - e_- \otimes \bar{f}, e_+ \otimes \bar{f} - e_- \otimes \underline{f}]$ is an ε-bracket for $e \otimes f$ in the $L_2(P_\varepsilon \otimes P)$ metric, where $e_+(t) := \max\{e(t), 0\} = \max(t, 0)$ and $e_-(t) := \max(-t, 0)$. It follows that for every $\xi > 0$,

$$N_{[]}(\varepsilon, e \otimes \mathcal{G}, L_2(P_\varepsilon \otimes P)) \le N_{[]}(\varepsilon, \mathcal{G}, L_2(P)),$$

which proves the claim for the symmetrised empirical process. □

In the next two lemmas, we assume, as in the main text, that P is a distribution on $[0, 1]^d$ with Lebesgue density bounded above and below by $M_0 \in [1, \infty)$ and $m_0 \in (0, 1]$ respectively. As in the proof of Proposition 5.5.2, let $\mathcal{F}_{d,\downarrow}^+ = \{f : -f \in \mathcal{F}_d, f \ge 0\}$. The following result is used to control the bracketing entropy terms that appear in Lemma 5.7.7 when we apply it in the proof of Proposition 5.5.2.

Lemma 5.7.8. *There exists a constant $C_d > 0$, depending only on d, such that for any $r, \xi > 0$,*

$$\log N_{[]}\big(\varepsilon, \mathcal{F}^+_{d,\downarrow} \cap B_2(r,P) \cap B_\infty(1), \|\cdot\|_{L_2(P)}\big) \leq C_d \begin{cases} (r/\varepsilon)^2 \frac{M_0}{m_0} \log^2(\frac{M_0}{m_0}) \log^4_+(1/\varepsilon) \log^2_+\big(\frac{r \log_+(1/\varepsilon)}{\varepsilon}\big) & \text{if } d = 2, \\ (r/\varepsilon)^{2(d-1)} (\frac{M_0}{m_0})^{d-1} \log^{d^2}_+(1/\varepsilon) & \text{if } d \geq 3. \end{cases}$$

Proof. We first claim that for any $\eta \in (0, 1/4]$,

$$\log N_{[]}\big(\varepsilon, \mathcal{F}^+_{d,\downarrow} \cap B_2(r,P), \|\cdot\|_{L_2(P;[\eta,1]^d)}\big) \lesssim_d \begin{cases} (\frac{r}{\varepsilon})^2 \frac{M_0}{m_0} \log^2(\frac{M_0}{m_0}) \log^4(1/\eta) \log^2_+\big(\frac{r \log(1/\eta)}{\varepsilon}\big) & \text{if } d = 2, \\ (\frac{r}{\varepsilon})^{2(d-1)} (\frac{M_0}{m_0})^{d-1} \log^{d^2}(1/\eta) & \text{if } d \geq 3. \end{cases} \tag{5.7.3}$$

By the cone property of $\mathcal{F}^+_{d,\downarrow}$, it suffices to establish the above claim when $r = 1$. We denote by $\mathrm{vol}(S)$ the d-dimensional Lebesgue measure of a measurable set $S \subseteq [0,1]^d$. By [57] Theorem 1.1 and a scaling argument, we have for any $\delta, M > 0$ and any hyperrectangle $A \subseteq [0,1]^d$ that

$$\log N_{[]}\big(\delta, \mathcal{F}^+_{d,\downarrow} \cap B_\infty(M), \|\cdot\|_{L_2(P;A)}\big) \lesssim_d \begin{cases} (\gamma/\delta)^2 \log^2_+(\gamma/\delta) & \text{if } d = 2, \\ (\gamma/\delta)^{2(d-1)} & \text{if } d \geq 3, \end{cases} \tag{5.7.4}$$

where $\gamma := M_0^{1/2} M \mathrm{vol}^{1/2}(A)$. We define $m := \lfloor \log_2(1/\eta) \rfloor$ and set $I_\ell := [2^\ell \eta, 2^{\ell+1}\eta] \cap [0,1]$ for each $\ell = 0, \ldots, m$. Then for $\ell_1, \ldots, \ell_d \in \{0, \ldots, m\}$, any $f \in \mathcal{F}^+_{d,\downarrow} \cap B_2(1,P)$ is uniformly bounded by $\big\{m_0 \prod_{j=1}^d (2^{\ell_j}\eta)\big\}^{-1/2}$ on the hyperrectangle $\prod_{j=1}^d I_{\ell_j}$. Then by (5.7.4) we see that for any $\delta > 0$,

$$\log N_{[]}\big(\delta, \mathcal{F}^+_{d,\downarrow} \cap B_2(1,P), \|\cdot\|_{L_2(P;\prod_{j=1}^d I_{\ell_j})}\big) \lesssim_d \begin{cases} \delta^{-2}(M_0/m_0) \log^2(\frac{M_0}{m_0}) \log^2_+(1/\delta) & \text{if } d = 2, \\ \delta^{-2(d-1)} (M_0/m_0)^{d-1} & \text{if } d \geq 3, \end{cases}$$

where we have used the fact that $\log_+(ax) \le 2\log_+(a)\log_+(x)$ for any $a, x > 0$. Note that these bounds do not depend on η, since the dependence of M and $\mathrm{vol}(A)$ on η is such that it cancels in the expression for γ. Global brackets for $\mathcal{F}^+_{d,\downarrow} \cap B_2(1)$ on $[\eta,1]^d$ can then be constructed by taking all possible combinations of local brackets on $I_{\ell_1} \times \cdots \times I_{\ell_d}$ for $\ell_1, \dots, \ell_d \in \{0, \dots, m\}$. Overall, for any $\varepsilon > 0$, setting $\delta = (m+1)^{-d/2}\varepsilon$ establishes the claim (5.7.3) in the case $r = 1$.

We conclude that if we fix any $\varepsilon > 0$, take $\eta = \varepsilon^2/(4d) \wedge 1/4$ and take a single bracket consisting of the constant functions 0 and 1 on $[0,1]^d \setminus [\eta,1]^d$, we have

$$\begin{aligned}
&\log N_{[]}\big(\varepsilon, \mathcal{F}^+_{d,\downarrow} \cap B_2(r,P) \cap B_\infty(1), \|\cdot\|_{L_2(P)}\big) \\
&\qquad \le \log N_{[]}\big(\varepsilon/2, \mathcal{F}^+_{d,\downarrow} \cap B_2(r,P), \|\cdot\|_{L_2(P;[\eta,1]^d)}\big) \\
&\qquad \lesssim_d \begin{cases} (r/\varepsilon)^2 \frac{M_0}{m_0} \log^2(\frac{M_0}{m_0}) \log_+^4(1/\varepsilon) \log_+^2\big(\frac{r\log_+(1/\varepsilon)}{\varepsilon}\big) & \text{if } d = 2, \\ (r/\varepsilon)^{2(d-1)} (\frac{M_0}{m_0})^{d-1} \log_+^{d^2}(1/\varepsilon) & \text{if } d \ge 3, \end{cases}
\end{aligned}$$

completing the proof. □

For $0 < r < 1$, let F_r be the envelope function of $\mathcal{F}^+_{d,\downarrow} \cap B_2(r,P) \cap B_\infty(1)$. The lemma below controls the $L_2(P)$ norm of F_r when restricted to strips of the form $I_\ell := [0,1]^{d-1} \times [\frac{\ell-1}{n_1}, \frac{\ell}{n_1}]$ for $\ell = 1, \dots, n_1$.

Lemma 5.7.9. *For any $r \in (0,1]$ and $\ell = 1, \dots, n_1$, we have*

$$\int_{I_\ell} F_r^2 \, \mathrm{d}P \le \frac{7 M_0 r^2 \log_+^d(1/r^2)}{m_0 \ell}.$$

Proof. By monotonicity and the $L_2(P)$ and L_∞ constraints, we have $F_r^2(x_1, \dots, x_d) \le \frac{r^2}{m_0 x_1 \cdots x_d} \wedge 1$. We first claim that for any $d \in \mathbb{N}$,

$$\int_{[0,1]^d} \bigg(\frac{t}{x_1 \cdots x_d} \wedge 1 \bigg) \mathrm{d}x_1 \cdots \mathrm{d}x_d \le 5t \log_+^d(1/t).$$

To see this, we define $S_d := \big\{(x_1, \dots, x_d) : \prod_{j=1}^d x_j \ge t\big\}$ and set $a_d := \int_{S_d} \frac{t}{x_1 \cdots x_d} \mathrm{d}x_1 \cdots \mathrm{d}x_d$ and $b_d := \int_{S_d} \mathrm{d}x_1 \cdots \mathrm{d}x_d$. By integrating out the last coordinate, we obtain the following

relation

$$b_d = \int_{S_{d-1}} \left(1 - \frac{t}{x_1 \cdots x_{d-1}}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} = b_{d-1} - a_{d-1}. \tag{5.7.5}$$

On the other hand, we have by direct computation that

$$\begin{aligned} a_d &= \int_t^1 \cdots \int_{\frac{t}{x_1 \cdots x_{d-1}}}^1 \frac{t}{x_1 \cdots x_d} \, \mathrm{d}x_d \cdots \mathrm{d}x_1 \\ &\leq a_{d-1} \log(1/t) \leq \cdots \leq a_1 \log^{d-1}(1/t) = t \log^d(1/t). \end{aligned} \tag{5.7.6}$$

Combining (5.7.5) and (5.7.6), we have

$$\begin{aligned} \int_{[0,1]^d} \left(\frac{t}{x_1 \cdots x_d} \wedge 1\right) \mathrm{d}x_1 \cdots \mathrm{d}x_d &= a_d + 1 - b_d \\ &\leq \min\{a_d + 1, a_d + a_{d-1} + \cdots + a_1 + 1 - b_1\} \\ &\leq \min\left\{t \log^d(1/t) + 1, \frac{t \log^{d+1}(1/t)}{\log(1/t) - 1}\right\} \leq 5t \log_+^d(1/t), \end{aligned}$$

as claimed, where the final inequality follows by considering the cases $t \in [1/e, 1]$, $t \in [1/4, 1/e)$ and $t \in [0, 1/4)$ separately. Consequently, for $\ell = 2, \ldots, n_1$, we have that

$$\begin{aligned} \int_{I_\ell} F_r^2 \, \mathrm{d}P &\leq \frac{M_0}{m_0} \int_{(\ell-1)/n_1}^{\ell/n_1} \int_{[0,1]^{d-1}} \left(\frac{r^2/x_d}{x_1 \cdots x_{d-1}} \wedge 1\right) \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} \mathrm{d}x_d \\ &\leq \frac{M_0}{m_0} \int_{(\ell-1)/n_1}^{\ell/n_1} 5(r^2/x_d) \log_+^{d-1}(x_d/r^2) \, \mathrm{d}x_d \\ &\leq \frac{M_0}{m_0} 5r^2 \log_+^{d-1}(1/r^2) \log\bigl(\ell/(\ell-1)\bigr) \leq \frac{7M_0 r^2 \log_+^{d-1}(1/r^2)}{m_0 \ell}, \end{aligned}$$

as desired. For the remaining case $\ell = 1$, we have

$$\int_{I_1} F_r^2 \, \mathrm{d}P \leq M_0 \int_{[0,1]^d} F_r^2 \, \mathrm{d}x_1 \cdots \mathrm{d}x_d \leq \frac{5M_0}{m_0} r^2 \log_+^d(1/r^2),$$

which is also of the correct form. □

Lemma 5.7.10. *For any Borel measurable $f_0 : [0,1]^d \to [-1,1]$ and any $a > 2$, we have $\mathbb{P}(\|\hat{f}_n - f_0\|_\infty > a) \leq n e^{-(a-2)^2/2}$. Consequently,*

$$\mathbb{E}\Bigl(\|\hat{f}_n - f_0\|_\infty^2 \mathbb{1}_{\{\|\hat{f}_n - f_0\|_\infty > a\}}\Bigr) \leq n\bigl(a^2 + 2 + 2\sqrt{2\pi}\bigr) e^{-(a-2)^2/2}.$$

Proof. Recall that we say $U \subseteq \mathbb{R}^d$ is an *upper set* if whenever $x \in U$ and $x \preceq y$, we have $y \in U$; we say, $L \subseteq \mathbb{R}^d$ is a *lower set* if $-L$ is an upper set. We write $\mathcal{U}$ and $\mathcal{L}$ respectively for the collections of upper and lower sets in $[0,1]^d$. The least squares estimator $\hat{f}_n$ over $\mathcal{F}_d$ then has a well-known min-max representation [129] Theorem 1.4.4:

$$\hat{f}_n(X_i) = \min_{L \in \mathcal{L}, L \ni X_i} \max_{U \in \mathcal{U}, U \ni X_i} \overline{Y_{L \cap U}},$$

where $\overline{Y_{L\cap U}}$ denotes the average value of the elements of $\{Y_i : X_i \in L \cap U\}$. Thus we have

$$\|\hat{f}_n\|_\infty = \max_{1 \le i \le n} |\hat{f}_n(X_i)| \le \max_{1 \le i \le n} |Y_i|.$$

Since $Y_i = f_0(X_i) + \xi_i$ and $\|f_0\|_\infty \le 1$, we have by a union bound that

$$\mathbb{P}\big(\|\hat{f}_n - f_0\|_\infty \ge t\big) \le n\mathbb{P}(|\xi_1| \ge t - 2).$$

The first claim follows using the fact that $\mathbb{P}(\xi_1 \ge t) \le \frac{1}{2}e^{-t^2/2}$ for any $t \ge 0$. Moreover, for any $a > 2$,

$$\begin{aligned}
\mathbb{E}\Big(\|\hat{f}_n - f_0\|_\infty^2 \mathbb{1}_{\{\|\hat{f}_n - f_0\|_\infty > a\}}\Big) &= \int_0^\infty 2t\, \mathbb{P}\big(\|\hat{f}_n - f_0\|_\infty \ge \max\{a, t\}\big)\, \mathrm{d}t \\
&\le na^2\mathbb{P}(|\xi_1| \ge a - 2) + n\int_a^\infty 2t\, \mathbb{P}(|\xi_1| \ge t - 2)\, \mathrm{d}t \\
&\le n\big(a^2 + 2 + 2\sqrt{2\pi}\big)e^{-(a-2)^2/2},
\end{aligned}$$

as desired. □

Lemma 5.7.11. *If Y is a non-negative random variable such that $(\mathbb{E}Y^p)^{1/p} \le A_1 p + A_2 p^{1/2} + A_3$ for all $p \in [1, \infty)$ and some $A_1, A_2 > 0$, $A_3 \ge 0$, then for every $t \ge 0$,*

$$\mathbb{P}(Y \ge t + eA_3) \le e \exp\left(-\min\left\{\frac{t}{2eA_1}, \frac{t^2}{4e^2A_2^2}\right\}\right).$$

Proof. Let $s := \min\{t/(2eA_1), t^2/(2eA_2)^2\}$. For values of t such that $s \ge 1$, we have by Markov's inequality that

$$\mathbb{P}(Y \ge t + eA_3) \le \left(\frac{A_1 s + A_2 s^{1/2} + A_3}{t + eA_3}\right)^s \le e^{-s} \le e^{1-s}.$$

For values of t such that $s < 1$, we trivially have $\mathbb{P}(Y \ge t + eA_3) \le \mathbb{P}(Y \ge t) \le e^{1-s}$, as desired. □

Lemma 5.7.12. *Let X be a non-negative random variable satisfying $X \leq b$ almost surely. Then*

$$\mathbb{E}e^X \leq \exp\left\{\frac{e^b - 1}{b}\mathbb{E}X\right\}.$$

Proof. We have

$$\mathbb{E}e^X = \sum_{r=0}^{\infty} \frac{\mathbb{E}(X^r)}{r!} \leq 1 + \sum_{r=1}^{\infty} \frac{b^{r-1}\mathbb{E}X}{r!} = 1 + \frac{\mathbb{E}X}{b}(e^b - 1) \leq \exp\left\{\frac{e^b - 1}{b}\mathbb{E}X\right\},$$

as required. □

BIBLIOGRAPHY

[1] Kenneth S. Alexander. Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.*, 12(4):1041–1067, 1984.

[2] Kenneth S Alexander. The non-existence of a universal multiplier moment for the central limit theorem. In *Probability in Banach Spaces V*, pages 15–16. Springer, 1985.

[3] Kenneth S. Alexander. Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 475–493. Wadsworth, Belmont, CA, 1985.

[4] Kenneth S. Alexander. The central limit theorem for weighted empirical processes indexed by sets. *J. Multivariate Anal.*, 22(2):313–339, 1987.

[5] Kenneth S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probab. Theory Related Fields*, 75(3):379–423, 1987.

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

[7] Niels T. Andersen, Evarist Giné, and Joel Zinn. The central limit theorem for empirical processes under local conditions: the case of Radon infinitely divisible limits without Gaussian component. *Trans. Amer. Math. Soc.*, 308(2):603–635, 1988.

[8] Ery Arias-Castro, David L. Donoho, and Xiaoming Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory*, 51(7):2402–2425, 2005.

[9] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011.

[10] Peter Bacchetti. Additive isotonic models. *J. Amer. Statist. Assoc.*, 84(405):289–294, 1989.

[11] Z. D. Bai, Jack W. Silverstein, and Y. Q. Yin. A note on the largest eigenvalue of a large-dimensional sample covariance matrix. *J. Multivariate Anal.*, 26(2):166–168, 1988.

[12] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

[13] Gábor Balázs, András György, and Csaba Szepesvári. Near-optimal max-affine estimators for convex regression. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 56–64, 2015.

[14] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517, 2017.

[15] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions. The theory and application of isotonic regression.* John Wiley & Sons, London-New York-Sydney, 1972. Wiley Series in Probability and Mathematical Statistics.

[16] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

[17] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

[18] Pierre C. Bellec. Sharp oracle inequalities for Least Squares estimators in shape restricted regression. *Ann. Statist.*, 46(2):745–780, 2018.

[19] Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and John A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York, 1998. Reprint of the 1993 original.

[20] Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237, 1983.

[21] Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150, 1993.

[22] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[23] Olivier Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 213–247. Birkhäuser, Basel, 2003.

[24] Leif Boysen, Angela Kempe, Volkmar Liebscher, Axel Munk, and Olaf Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.*, 37(1):157–183, 2009.

[25] Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *Ann. Statist.*, 43(6):2507–2536, 2015.

[26] Victor-Emmanuel Brunel. Adaptive estimation of convex polytopes and convex sets from noisy data. *Electron. J. Stat.*, 7:1301–1327, 2013.

[27] H. D. Brunk. Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616, 1955.

[28] Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Trans. Inform. Theory*, 59(11):7711–7717, 2013.

[29] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[30] Bernd Carl. Entropy numbers of embedding maps between Besov spaces with an application to eigenvalue problems. *Proc. Roy. Soc. Edinburgh Sect. A*, 90(1-2):63–70, 1981.

[31] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.

[32] Olivier Catoni. Pac-bayesian bounds for the gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.

[33] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43(4):1774–1800, 2015.

[34] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On matrix estimation under monotonicity constraints. *Bernoulli*, 24(2):1072–1100, 2018.

[35] Sabyasachi Chatterjee and John Lafferty. Adaptive risk bounds in unimodal regression. *arXiv preprint arXiv:1512.02956*, 2015.

[36] Sourav Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, 2014.

[37] Yining Chen and Richard J. Samworth. Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(4):729–754, 2016.

[38] Guang Cheng. Semiparametric additive isotonic regression. *J. Statist. Plann. Inference*, 139(6):1980–1991, 2009.

[39] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013.

[40] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597, 2014.

[41] Víctor H. de la Peña and Evarist Giné. *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

[42] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[43] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 2016.

[44] R. P. Dilworth. A decomposition theorem for partially ordered sets. *Ann. of Math. (2)*, 51:161–166, 1950.

[45] David L Donoho. Gelfand n-widths and the method of least squares. *Preprint*, 1990.

[46] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, 1:290–330, 1967.

[47] R. M. Dudley. Empirical and Poisson processes on classes of sets or functions too large for central limit theorems. *Z. Wahrsch. Verw. Gebiete*, 61(3):355–368, 1982.

[48] R. M. Dudley. *Uniform central limit theorems*, volume 142 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, New York, second edition, 2014.

[49] C. Durot. Monotone nonparametric regression with random design. *Math. Methods Statist.*, 17(4):327–341, 2008.

[50] Cécile Durot. On the $\mathbb{L}_p$-error of monotonicity constrained estimators. *Ann. Statist.*, 35(3):1080–1104, 2007.

[51] Rick Durrett. *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition, 2010.

[52] Richard L. Dykstra. An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.*, 78(384):837–842, 1983.

[53] Richard L. Dykstra and Tim Robertson. An algorithm for isotonic regression for two or more independent variables. *Ann. Statist.*, 10(3):708–716, 1982.

[54] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446, 2010.

[55] Santiago F Elena and Richard E Lenski. Test of synergistic interactions among deleterious mutations in bacteria. *Nature*, 390(6658):395, 1997.

[56] Chao Gao, Fang Han, and Cun-Hui Zhang. Minimax risk bounds for piecewise constant models. *arXiv preprint arXiv:1705.06386*, 2017.

[57] Fuchang Gao and Jon A. Wellner. Entropy estimate for high-dimensional monotonic functions. *J. Multivariate Anal.*, 98(9):1751–1764, 2007.

[58] Xiaoli Gao and Jian Huang. Asymptotic analysis of high-dimensional LAD regression with Lasso. *Statist. Sinica*, 20(4):1485–1506, 2010.

[59] Evarist Giné and Vladimir Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216, 2006.

[60] Evarist Giné, Vladimir Koltchinskii, and Jon A. Wellner. Ratio limit theorems for empirical processes. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 249–278. Birkhäuser, Basel, 2003.

[61] Evarist Giné, Rafał Latała, and Joel Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA, 2000.

[62] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2015.

[63] Evarist Giné and Joel Zinn. Central limit theorems and weak laws of large numbers in certain Banach spaces. *Z. Wahrsch. Verw. Gebiete*, 62(3):323–354, 1983.

[64] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4):929–998, 1984. With discussion.

[65] Evarist Giné and Joel Zinn. Lectures on the central limit theorem for empirical processes. In *Probability and Banach spaces (Zaragoza, 1985)*, volume 1221 of *Lecture Notes in Math.*, pages 50–113. Springer, Berlin, 1986.

[66] Evarist Giné and Joel Zinn. Gaussian characterization of uniform Donsker classes of functions. *Ann. Probab.*, 19(2):758–782, 1991.

[67] David B Goldstein. Common genetic variation and human traits. *New England Journal of Medicine*, 360(17):1696, 2009.

[68] Ulf Grenander. *Abstract inference*. John Wiley & Sons, Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[69] Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*, volume 38 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York, 2014. Estimators, algorithms and asymptotics.

[70] Piet Groeneboom, Geurt Jongbloed, and Jon A. Wellner. Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, 29(6):1653–1698, 2001.

[71] Adityanand Guntuboyina. Optimal rates of convergence for convex set estimation from support functions. *Ann. Statist.*, 40(1):385–411, 2012.

[72] Adityanand Guntuboyina and Bodhisattva Sen. Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields*, 163(1-2):379–411, 2015.

[73] Adityanand Guntuboyina and Bodhisattva Sen. Nonparametric shape-restricted regression. *arXiv preprint arXiv:1709.05707*, 2017.

[74] Qiyang Han. Bayes model selection. *arXiv preprint arXiv:1704.07513*, 2017.

[75] Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J. Samworth. Isotonic regression in general dimensions. *arXiv preprint arXiv:1708.09468*, 2017.

[76] Qiyang Han and Jon A Wellner. A sharp multiplier inequality with applications to heavy-tailed regression problems. *arXiv preprint arXiv:1706.02410*, 2017.

[77] Qiyang Han and Jon A Wellner. Robustness of shape-restricted regression estimators: an envelope perspective. *arXiv preprint arXiv:1805.02542*, 2018.

[78] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London, 1990.

[79] Harold Hotelling. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Ann. Math. Statistics*, 11:271–283, 1940.

[80] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.*, 17:Paper No. 18, 40, 2016.

[81] Emilien Joly, Gábor Lugosi, and Roberto Imbuzeiro Oliveira. On the estimation of the mean of a random vector. *Electron. J. Stat.*, 11(1):440–451, 2017.

[82] Arlene K. H. Kim and Richard J. Samworth. Global rates of convergence in log-concave density estimation. *Ann. Statist.*, 44(6):2756–2779, 2016.

[83] Arlene KH Kim, Adityanand Guntuboyina, and Richard J Samworth. Adaptation in log-concave density estimation. *arXiv preprint arXiv:1609.00861*, 2016.

[84] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[85] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 443–457. Birkhäuser Boston, Boston, MA, 2000.

[86] A. P. Korostelëv and A. B. Tsybakov. Asymptotically minimax image reconstruction problems. In *Topics in nonparametric estimation*, volume 12 of *Adv. Soviet Math.*, pages 45–86. Amer. Math. Soc., Providence, RI, 1992.

[87] A. P. Korostelëv and A. B. Tsybakov. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993.

[88] Rasmus Kyng, Anup Rao, and Sushant Sachdeva. Fast, provable algorithms for isotonic regression in all l_p-norms. In *Advances in Neural Information Processing Systems*, pages 2719–2727, 2015.

[89] Guillaume Lecué. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, 35(4):1698–1721, 2007.

[90] Guillaume Lecué and Shahar Mendelson. Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc. (JEMS)*, 19(3):881–904, 2017.

[91] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.*, 46(2):611–641, 2018.

[92] Michel Ledoux and Michel Talagrand. Conditions d'intégrabilité pour les multiplicateurs dans le TLC banachique. *Ann. Probab.*, 14(3):916–921, 1986.

[93] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[94] Gábor Lugosi and Shahar Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Bernoull (to appear)*, 2017.

[95] Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc. (JEMS) (to appear)*, 2017.

[96] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *Ann. Statist. (to appear)*, 2017.

[97] Ronny Luss, Saharon Rosset, and Moni Shahar. Efficient regularized isotonic regression with application to gene-gene interaction search. *Ann. Appl. Stat.*, 6(1):253–283, 2012.

[98] E. Mammen, O. Linton, and J. Nielsen. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, 27(5):1443–1490, 1999.

[99] E. Mammen and A. B. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.*, 23(2):502–524, 1995.

[100] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[101] Enno Mammen and Kyusang Yu. Additive isotone regression. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 179–195. Inst. Math. Statist., Beachwood, OH, 2007.

[102] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences*, 105(9):3461–3466, 2008.

[103] David M. Mason. The asymptotic distribution of weighted empirical distribution functions. *Stochastic Process. Appl.*, 15(1):99–109, 1983.

[104] David M. Mason, Galen R. Shorack, and Jon A. Wellner. Strong limit theorems for oscillation moduli of the uniform empirical process. *Z. Wahrsch. Verw. Gebiete*, 65(1):83–97, 1983.

[105] P. Massart and E. Rio. A uniform Marcinkiewicz-Zygmund strong law of large numbers for empirical processes. In *Asymptotic methods in probability and statistics (Ottawa, ON, 1997)*, pages 199–211. North-Holland, Amsterdam, 1998.

[106] Pascal Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884, 2000.

[107] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[108] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[109] Shahar Mendelson. Learning without concentration. *J. ACM*, 62(3):Art. 21, 25, 2015.

[110] Shahar Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Process. Appl.*, 126(12):3652–3680, 2016.

[111] Shahar Mendelson. Extending the small-ball method. *arXiv preprint arXiv:1709.00843*, 2017.

[112] Shahar Mendelson. "Local" vs. "global" parameters—breaking the Gaussian complexity barrier. *Ann. Statist.*, 45(5):1835–1862, 2017.

[113] Shahar Mendelson. On aggregation for heavy-tailed classes. *Probab. Theory Related Fields*, 168(3-4):641–674, 2017.

[114] Shahar Mendelson. On multiplier processes under weak moment assumptions. In *Geometric aspects of functional analysis*, volume 2169 of *Lecture Notes in Math.*, pages 301–318. Springer, Cham, 2017.

[115] Mary Meyer and Michael Woodroofe. On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, 28(4):1083–1104, 2000.

[116] Mary C. Meyer. Semi-parametric additive constrained regression. *J. Nonparametr. Stat.*, 25(3):715–730, 2013.

[117] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[118] S. J. Montgomery-Smith. Comparison of sums of independent identically distributed random vectors. *Probab. Math. Statist.*, 14(2):281–285 (1994), 1993.

[119] Tony Morton-Jones, Peter Diggle, Louise Parker, Heather O Dickinson, and Keith Binks. Additive isotonic regression models in epidemiology. *Statistics in Medicine*, 19(6):849–859, 2000.

[120] Fabien Navarro and Adrien Saumard. Slope heuristics and V-fold model selection in heteroscedastic regression using strongly localized bases. *ESAIM Probab. Stat.*, 21:412–451, 2017.

[121] Richard Nickl and Sara van de Geer. Confidence sets in sparse regression. *Ann. Statist.*, 41(6):2852–2876, 2013.

[122] Mina Ossiander. A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.*, 15(3):897–919, 1987.

[123] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[124] David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

[125] David Pollard. *A user's guide to measure theoretic probability*, volume 8 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2002.

[126] Stephen Portnoy and Roger Koenker. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statist. Sci.*, 12(4):279–300, 1997. With comments by Ronald A. Thisted and M. R. Osborne and a rejoinder by the authors.

[127] Alexander Rakhlin, Karthik Sridharan, and Alexandre B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

[128] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011.

[129] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1988.

[130] Dan Romik. *The surprising mathematics of longest increasing subsequences*, volume 4 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, New York, 2015.

[131] Frederick P Roth, Howard D Lipshitz, and Brenda J Andrews. Q& a: epistasis. *Journal of Biology*, 8(4):35, 2009.

[132] Gennady Samorodnitsky and Murad S. Taqqu. *Stable non-Gaussian random processes*. Stochastic Modeling. Chapman & Hall, New York, 1994. Stochastic models with infinite variance.

[133] Rafael Sanjuán and Santiago F Elena. Epistasis correlates to genomic complexity. *Proceedings of the National Academy of Sciences*, 103(39):14402–14405, 2006.

[134] Michael J Schell and Bahadur Singh. The reduced monotonic regression method. *Journal of the American Statistical Association*, 92(437):128–135, 1997.

[135] Haifeng Shao, Lindsay C Burrage, David S Sinasac, Annie E Hill, Sheila R Ernest, William O'Brien, Hayden-William Courtland, Karl J Jepsen, Andrew Kirby, and EJ Kulbokas. Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proceedings of the National Academy of Sciences*, 105(50):19910–19914, 2008.

[136] Galen R. Shorack and Jon A. Wellner. Limit theorems and inequalities for the uniform empirical process indexed by intervals. *Ann. Probab.*, 10(3):639–652, 1982.

[137] Galen R. Shorack and Jon A. Wellner. *Empirical processes with applications to statistics*, volume 59 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009.

[138] Vidyashankar Sivakumar, Arindam Banerjee, and Pradeep K Ravikumar. Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs. In *Advances in neural information processing systems*, pages 2206–2214, 2015.

[139] Charles J. Stone. Additive regression and other nonparametric models. *Ann. Statist.*, 13(2):689–705, 1985.

[140] Quentin F. Stout. Isotonic regression for multiple independent variables. *Algorithmica*, 71(2):450–470, 2015.

[141] Volker Strassen and R. M. Dudley. The central limit theorem and ε-entropy. In *Probability and Information Theory (Proc. Internat. Sympos., McMaster Univ., Hamilton, Ont., 1968)*, pages 224–231. Springer, Berlin, 1969.

[142] Winfried Stute. The oscillation behavior of empirical processes. *Ann. Probab.*, 10(1):86–107, 1982.

[143] Winfried Stute. The oscillation behavior of empirical processes: the multivariate case. *Ann. Probab.*, 12(2):361–379, 1984.

[144] V. N. Sudakov. Gauss and Cauchy measures and ε-entropy. *Dokl. Akad. Nauk SSSR*, 185:51–53, 1969.

[145] Michel Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

[146] Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics*. Springer, Heidelberg, 2014. Modern methods and classical problems.

[147] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[148] Amy Hin Yan Tong, Marie Evangelista, Ainslie B Parsons, Hong Xu, Gary D Bader, Nicholas Pagé, Mark Robinson, Sasan Raghibizadeh, Christopher WV Hogue, and Howard Bussey. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368, 2001.

[149] Hans Tribel′. Interpolation properties of ε-entropy and of diameters. Geometric characteristics of the imbedding of function spaces of Sobolev-Besov type. *Mat. Sb. (N.S.)*, 98(140)(1 (9)):27–41, 157, 1975.

[150] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[151] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[152] Sara van de Geer. A new approach to least-squares estimation, with applications. *Ann. Statist.*, 15(2):587–602, 1987.

[153] Sara van de Geer. Estimating a regression function. *Ann. Statist.*, 18(2):907–924, 1990.

[154] Sara van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1):14–44, 1993.

[155] Sara van de Geer. On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electron. J. Stat.*, 8(1):543–574, 2014.

[156] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

[157] Sara van de Geer and Alan Muro. Penalized least squares estimation in the additive

model with different smoothness for the components. *J. Statist. Plann. Inference*, 162:43–61, 2015.

[158] Sara van de Geer and Martin J. Wainwright. On concentration for (regularized) empirical risk minimization. *Sankhya A*, 79(2):159–200, 2017.

[159] Sara van de Geer and Marten Wegkamp. Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.*, 24(6):2513–2523, 1996.

[160] Sara A. van de Geer. *Applications of Empirical Process Theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[161] Aad van der Vaart and Jon A. Wellner. A local maximal inequality under uniform entropy. *Electron. J. Stat.*, 5:192–203, 2011.

[162] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[163] Constance van Eeden. Testing and estimating ordered parameters of probability distribution. 1958.

[164] Grace Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[165] Jon A. Wellner. Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Z. Wahrsch. Verw. Gebiete*, 45(1):73–88, 1978.

[166] Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, 23(2):339–362, 1995.

[167] Fan Yang and Rina Foygel Barber. Contraction and uniform convergence of isotonic regression. *arXiv preprint arXiv:1706.01852*, 2017.

[168] Yuhong Yang. Nonparametric regression with dependent errors. *Bernoulli*, 7(4):633–655, 2001.

[169] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.

[170] Qi Wei Yao. Tests for change-points with epidemic alternatives. *Biometrika*, 80(1):179–191, 1993.

[171] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.

[172] Cun-Hui Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555, 2002.

VITA

Qiyang Han was born in Shanghai, China, in June 1991. He received a B.Sc. in mathematics from Fudan University in 2013 and a Ph.D. in Statistics from University of Washington in 2018. He will join Rutgers as an assistant professor of Statistics and Biostatistics from fall 2018.