

ISOTONIC REGRESSION IN GENERAL DIMENSIONS

BY QIYANG HAN^{*}, TENGGAO WANG[†], SABYASACHI
CHATTERJEE^{‡,§} AND RICHARD J. SAMWORTH[†]

University of Washington^{}, University of Cambridge[†], University of
Chicago[‡] and University of Illinois at Urbana-Champaign[§]*

We study the least squares regression function estimator over the class of real-valued functions on $[0, 1]^d$ that are increasing in each coordinate. For uniformly bounded signals and with a fixed, cubic lattice design, we establish that the estimator achieves the minimax rate of order $n^{-\min\{2/(d+2), 1/d\}}$ in the empirical L_2 loss, up to poly-logarithmic factors. Further, we prove a sharp oracle inequality, which reveals in particular that when the true regression function is piecewise constant on k hyperrectangles, the least squares estimator enjoys a faster, adaptive rate of convergence of $(k/n)^{\min(1, 2/d)}$, again up to poly-logarithmic factors. Previous results are confined to the case $d \leq 2$. Finally, we establish corresponding bounds (which are new even in the case $d = 2$) in the more challenging random design setting. There are two surprising features of these results: first, they demonstrate that it is possible for a global empirical risk minimisation procedure to be rate optimal up to poly-logarithmic factors even when the corresponding entropy integral for the function class diverges rapidly; second, they indicate that the adaptation rate for shape-constrained estimators can be strictly worse than the parametric rate.

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 (Schell and Singh, 1997), 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 (Dykstra and Robertson, 1982). 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 (Mani et al., 2008; Roth, Lipshitz and Andrews, 2009; Luss, Rosset and Shahar, 2012). In these latter con-

AMS 2000 subject classifications: 62G05, 62G08, 62C20

Keywords and phrases: isotonic regression, block increasing functions, adaptation, least squares, sharp oracle inequality, statistical dimension

texts, 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 (Shao et al., 2008; Goldstein, 2009; Eichler et al., 2010). Alternative simplifying interaction structures have also been explored, including those based on products (Elena and Lenski, 1997), logarithms (Sanjuan and Elena, 2006) and minima (Tong et al., 2001), but the form of genetic interaction between factors is not always clear and may vary between phenotypes (Mani et al., 2008; Luss, Rosset and Shahar, 2012).

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

$$\mathcal{F}_d := \left\{ f : [0, 1]^d \rightarrow \mathbb{R}, f(x_1, \dots, x_d) \leq f(x'_1, \dots, x'_d) \right. \\ \left. \text{when } x_j \leq x'_j \text{ for } j = 1, \dots, d \right\}.$$

In this paper, we suppose that we observe data $(X_1, Y_1), \dots, (X_n, Y_n)$, with $n \geq 2$, satisfying

$$(1) \quad Y_i = f_0(X_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where $f_0 : [0, 1]^d \rightarrow \mathbb{R}$ is Borel measurable, $\epsilon_1, \dots, \epsilon_n$ are independent $N(0, 1)$ noise, and the covariates X_1, \dots, X_n , which take values in the set $[0, 1]^d$, can either be fixed or random (independent of $\epsilon_1, \dots, \epsilon_n$). 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

$$(2) \quad R_n(\hat{f}_n, f_0) := \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \{\hat{f}_n(X_i) - f_0(X_i)\}^2 \right].$$

Note that this loss function only considers the errors made at the design points X_1, \dots, X_n , and these points naturally induce a directed acyclic graph $G_X = (V(G_X), E(G_X))$ with $V(G_X) = \{1, \dots, n\}$ and $E(G_X) = \{(i, i') : (X_i)_j \leq (X_{i'})_j \forall j = 1, \dots, d\}$. 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 $\hat{\theta}_n = ((\hat{\theta}_n)_i)_{i=1}^n := (f_n(X_i))_{i=1}^n$ 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., [Dykstra \(1983\)](#); [Kyng, Rao and Sachdeva \(2015\)](#); [Stout \(2015\)](#)).

In the special case where $d = 1$, model (1) reduces to the univariate isotonic regression problem that has a long history (e.g. [Brunk, 1955](#); [van Eeden, 1958](#); [Barlow *et al.*, 1972](#); [van de Geer, 1990, 1993](#); [Donoho, 1991](#); [Birgé and Massart, 1993](#); [Meyer and Woodroffe, 2000](#); [Durot, 2007, 2008](#); [Yang and Barber, 2017](#)). See [Groeneboom and Jongbloed \(2014\)](#) 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 (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 (1) with fixed design points. When $(\theta_0)_1 \leq \dots \leq (\theta_0)_n$ (i.e. $X_1 \leq \dots \leq X_n$), [Zhang \(2002\)](#) proved that for $d = 1$ there exists a universal constant $C > 0$ such that

$$R_n(\hat{\theta}_n, \theta_0) \leq C \left\{ \left(\frac{(\theta_0)_n - (\theta_0)_1}{n} \right)^{2/3} + \frac{\log n}{n} \right\},$$

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 [Zhang \(2002\)](#) in the context of univariate isotonic regression, followed by [Chatterjee, Guntuboyina and Sen \(2015\)](#) and most recently [Bellec \(2018\)](#), who proved that

$$(3) \quad R_n(\hat{\theta}_n, \theta_0) \leq \inf_{\theta \in \mathcal{M}(G_X)} \left\{ \frac{\|\theta - \theta_0\|_2^2}{n} + \frac{k(\theta)}{n} \log \left(\frac{en}{k(\theta)} \right) \right\},$$

where $k(\theta)$ is the number of constant pieces in the isotonic vector θ . The inequality (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 (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 ([Kim, Guntuboyina and Samworth, 2018](#)), and the least squares estimator in unimodal regression ([Chatterjee and Lafferty, 2017](#)).

Much less is known about the rate of convergence of the least squares estimator in the model (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 [Chatterjee, Guntuboyina and Sen \(2018\)](#), which deals with the fixed, lattice design case when $d = 2$. For a general dimension d , and for $n_1, \dots, n_d \in \mathbb{N}$, we define this lattice by $\mathbb{L}_{d,n_1,\dots,n_d} := \prod_{j=1}^d \{1/n_j, 2/n_j, \dots, 1\}$; when $n_1 = \dots = n_d = n^{1/d}$ for some $n \in \mathbb{N}$, we also write $\mathbb{L}_{d,n} := \mathbb{L}_{d,n_1,\dots,n_d}$ as shorthand. When $\{X_1, \dots, X_n\} = \mathbb{L}_{2,n_1,n_2}$, [Chatterjee, Guntuboyina and Sen \(2018\)](#) showed that there exists a universal constant $C > 0$ such that

$$R_n(\hat{\theta}_n, \theta_0) \leq C \left\{ \frac{((\theta_0)_{n_1,n_2} - (\theta_0)_{1,1}) \log^4 n}{n^{1/2}} + \frac{\log^8 n}{n} \right\},$$

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

$$(4) \quad R_n(\hat{\theta}_n, \theta_0) \leq \inf_{\theta \in \mathcal{M}(\mathbb{L}_{2,n_1,n_2})} \left(\frac{\|\theta - \theta_0\|_2^2}{n} + \frac{Ck(\theta) \log^8 n}{n} \right),$$

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. [Bacchetti \(1989\)](#); [Morton-Jones et al. \(2000\)](#); [Mammen and Yu \(2007\)](#); [Chen and Samworth \(2016\)](#)). In the simplest setting, these works investigate the regression problem (1), where the signal f_0 belongs to

$$\mathcal{F}_d^{\text{add}} := \left\{ f \in \mathcal{F}_d : f(x_1, \dots, x_d) = \sum_{j=1}^d f_j(x_j), f_j \in \mathcal{F}_1, \|f_j\|_\infty \leq 1 \right\}.$$

The additive structure greatly reduces the complexity of the class; indeed, it can be shown that the least squares estimator over $\mathcal{F}_d^{\text{add}}$ attains the univariate risk $n^{-2/3}$, up to multiplicative constants depending on d (e.g. [van de Geer, 2000](#), Theorem 9.1).

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

$$(5) \quad \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}), \|\theta_0\|_\infty \leq 1} R_n(\hat{\theta}_n, \theta_0) \leq C n^{-1/d} \log^4 n,$$

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 $\theta_0 \in \mathbb{R}^{\mathbb{L}_{d,n}}$,

$$(6) \quad R_n(\hat{\theta}_n, \theta_0) \leq \inf_{\theta \in \mathcal{M}(\mathbb{L}_{d,n})} \left\{ \frac{\|\theta - \theta_0\|_2^2}{n} + C \left(\frac{k(\theta)}{n} \right)^{2/d} \log^8 \left(\frac{en}{k(\theta)} \right) \right\},$$

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 (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 (8) below for a formal definition. A key step in the proof of (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 1. The adaptation rate in (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 2.

We further demonstrate that analogues of the worst-case bounds and oracle inequalities (5) and (6), with slightly different poly-logarithmic exponents, remain valid for random design points X_1, \dots, 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 in Section 3) that arise in handling the relevant empirical processes.

In addition to the risk $R_n(\hat{f}_n, f_0)$ in (2), for random designs we also study the natural population squared risk

$$R(\hat{f}_n, f_0) := \mathbb{E} \|\hat{f}_n - f_0\|_{L_2(P)}^2 = \mathbb{E} [\{\hat{f}_n(X) - f_0(X)\}^2],$$

where $(X, Y) \stackrel{d}{=} (X_1, Y_1)$ and is independent of $(X_1, Y_1), \dots, (X_n, Y_n)$. We note that the quantity $\mathbb{E} [\{Y - \hat{f}_n(X)\}^2]$, 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 [Birgé and Massart \(1993\)](#) have spawned a heuristic belief that one should not use global empirical risk minimisation procedures¹ when the entropy integral for the corresponding function class diverges (e.g. [van de Geer \(2000, pp. 121–122\)](#), [Rakhlin, Sridharan and Tsybakov \(2017\)](#)). 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 1](#).
2. Sharp adaptive behaviour for shape-constrained estimators has previously only been shown when the adaptive rate is nearly parametric (see, e.g., [Guntuboyina and Sen \(2015\)](#); [Chatterjee, Guntuboyina and Sen \(2015\)](#); [Bellec \(2018\)](#); [Kim, Guntuboyina and Samworth \(2018\)](#)). 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 2](#), we state the main results for the fixed lattice design model. [Section 3](#) describes corresponding results in the random design case. Proofs of all main theoretical results are contained in [Sections 4](#) and [5](#), whereas proofs of ancillary results are deferred until [Appendix B](#) in the online supplement ([Han et al., 2018](#)).

1.1. Notation. For a real-valued measurable function f defined on a probability space $(\mathcal{X}, \mathcal{A}, P)$ and for $p \in [1, \infty)$, we let $\|f\|_{L_p(P)} := (P|f|^p)^{1/p}$

¹The term ‘global’ refers here to procedures that involve minimisation over the entire function class, as opposed to only over a sieve; cf. [van de Geer \(2000\)](#).

denote the usual $L_p(P)$ -norm, and write $\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|$. Moreover, for any Borel measurable $\mathcal{R} \subseteq \mathcal{X}$, we write $\|f\|_{L_p(P; \mathcal{R})} := (\int_{\mathcal{R}} |f|^p dP)^{1/p}$. For $r \geq 0$, we write $B_p(r, P) := \{f : \mathcal{X} \rightarrow \mathbb{R}, \|f\|_{L_p(P)} \leq r\}$ and $B_\infty(r) := \{f : \mathcal{X} \rightarrow \mathbb{R}, \|f\|_\infty \leq r\}$. 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 $j = 1, \dots, d$.

For $\varepsilon > 0$, the ε -covering number of a (semi-)normed space $(\mathcal{F}, \|\cdot\|)$, denoted $N(\varepsilon, \mathcal{F}, \|\cdot\|)$, is the smallest number of closed ε -balls whose union covers \mathcal{F} . The ε -bracketing number, denoted $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$, is the smallest number of ε -brackets, of the form $[l, u] := \{f \in \mathcal{F} : l \leq f \leq u\}$, such that $\|u - l\| \leq \varepsilon$, and whose union covers \mathcal{F} . The *metric/bracketing entropy* is the logarithm of the covering/bracketing number.

Throughout the article $\epsilon_1, \dots, \epsilon_n$ and $\{\epsilon_w : w \in \mathbb{L}_{d, n_1, \dots, n_d}\}$ denote independent standard normal random variables and ξ_1, \dots, ξ_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 $d_{\text{TV}}(P, Q) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$ for their total variation distance, and $d_{\text{KL}}^2(P, Q) := \int_{\mathcal{X}} \log \frac{dP}{dQ} dP$ for their Kullback–Leibler divergence.

We use c, C to denote generic universal positive constants and use c_x, C_x to denote generic positive constants that depend only on x . Exact numeric values of these constants may change from line to line unless otherwise specified. Also, $a \lesssim_x b$ 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 universal constant $C > 0$). Finally, we define $\log_+(x) := \log(x \vee e)$.

2. Fixed lattice design. In this section, we focus on the model (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 [Chatterjee, Guntuboyina and Sen \(2018\)](#). Our main interest lies in the cases $d \geq 3$.

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 $\hat{\theta}_n = \hat{\theta}_n(Y_1, \dots, Y_n)$ of $\theta_0 \in \mathcal{M}(\mathbb{L}_{d, n})$.

THEOREM 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 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 1 is matched, up to poly-logarithmic factors, by the following minimax lower bound.

PROPOSITION 1. *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 $\tilde{\theta}_n = \tilde{\theta}_n(Y_1, \dots, Y_n)$ 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, \dots, i_L) , where $(i_j, i_{j+1}) \in E$ for each $j = 1, \dots, L-1$; an *antichain* in G of cardinality L is a subset $\{i_1, \dots, i_L\}$ of V such that for each distinct $j, j' \in \{1, \dots, L\}$ there is no chain containing both i_j and $i_{j'}$. A key observation in the proof of Proposition 1 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 1 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 1 and Proposition 1, 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 1, 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 Kim and Samworth (2016) 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 $i_j \in \{2, \dots, n_1\}$. 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 1 and Proposition 1.

COROLLARY 1. *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. Gao and Wellner (2007, Theorem 1.1) proved that when $d \geq 3$ and Q denotes Lebesgue measure on $[0, 1]^d$,

$$(7) \quad \log N(\varepsilon, \mathcal{F}_d \cap B_\infty(1), \|\cdot\|_{L_2(Q)}) \asymp_d \varepsilon^{-2(d-1)}.$$

In particular, for $d \geq 3$, the classes $\mathcal{F}_d \cap B_\infty(1)$ are massive in the sense that the entropy integral $\int_\delta^1 \log^{1/2} N(\varepsilon, \mathcal{F}_d \cap B_\infty(1), \|\cdot\|_{L_2(Q)}) d\varepsilon$ 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.

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 Chatterjee, Guntuboyina and Sen (2018)). Our main result is the sharp oracle inequality in Theorem 2 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 $j = 1, \dots, d$. 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, \dots, A_K are disjoint two-dimensional sheets and the restricted vector $\theta_{A_\ell \cap \mathbb{L}_{d,n}}$ is constant for each $\ell = 1, \dots, K$.

THEOREM 2. *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 2 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 $K(\theta_0) = n^{(d-2)/d} \rightarrow \infty$ as $n \rightarrow \infty$. The following corollary of Theorem 2 gives an alternative (weaker) form of oracle inequality that offers easier comparison to lower dimensional results given in (3) and (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 $\mathcal{R}_1, \dots, \mathcal{R}_k$ 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 3. *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 2 and 3 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 3 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 3 cannot be improved.

PROPOSITION 2. *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 $n^{-1} \log n$ (Bellec, 2018), and that a poly-logarithmic penalty relative to the parametric rate is unavoidable for the least squares estimator. Moreover, we see from Proposition 2 that for $d \geq 3$, although the least squares estimator achieves a faster rate of convergence than the worst-case bound in Theorem 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 (3) and (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 by²

$$(8) \quad \delta(C) := \int_{\mathbb{R}^n} \|\Pi_C(x)\|_2^2 (2\pi)^{-n/2} e^{-\|x\|_2^2/2} dx,$$

where Π_C is the projection onto the set C (Amelunxen et al., 2014). The proofs of Theorem 3 and Proposition 2 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 1).

TABLE 1
Bounds* for $\delta(\mathcal{M}(\mathbb{L}_{d,n}))$.

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.

† Amelunxen et al. (2014)

‡ Chatterjee, Guntuboyina and Sen (2018)

The following corollary of Theorem 2 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 $r = 0, 1, \dots, d$, 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 $\mathcal{J} \subseteq \{1, \dots, d\}$, of cardinality r , such that $(\theta_0)_{(x_1, \dots, x_d)} = (\theta_0)_{(x'_1, \dots, x'_d)}$ whenever $x_j = x'_j$ for all $j \in \mathcal{J}$.

COROLLARY 2. For $d \geq 2$, there exists constant $C_d > 0$, depending only on d , such that

$$\sup_{\theta_0 \in \mathcal{M}_r(\mathbb{L}_{d,n}) \cap B_\infty(1)} R_n(\hat{\theta}_n, \theta_0) \leq C_d \begin{cases} n^{-2/d} \log^8 n & \text{if } r \leq d - 2 \\ n^{-4/(3d)} \log^{16/3} n & \text{if } r = d - 1 \\ n^{-1/d} \log^4 n & \text{if } r = d. \end{cases}$$

²Our reason for defining the statistical dimension via an integral rather than as $\mathbb{E}\|\Pi_C(\epsilon)\|_2^2$ is because, in the random design setting, the cone C is itself random, and in that case $\delta(C)$ is a random quantity.

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 2 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 2, Corollary 2 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 2 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 Chatterjee, Guntuboyina and Sen (2018, Theorem 2.4). In this sense, the adaptation of the isotonic least squares estimator in general dimensions is essentially a two-dimensional phenomenon.

3. Random design. In this section, we consider the setting where the design points X_1, \dots, 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, \dots, X_n , for definiteness, we extend \hat{f}_n to $[0, 1]^d$ by defining $\hat{f}_n(x) := \min(\{\hat{f}_n(X_i) : 1 \leq i \leq n, X_i \succeq x\} \cup \{\max_i \hat{f}_n(X_i)\})$. If we let $\mathbb{P}_n := n^{-1} \sum_{i=1}^n \delta_{X_i}$, then we can consider the empirical and population risks $R_n(\hat{f}_n, f_0) = \mathbb{E} \|\hat{f}_n - f_0\|_{L_2(\mathbb{P}_n)}^2$ and $R(\hat{f}_n, f_0) = \mathbb{E} \|\hat{f}_n - f_0\|_{L_2(P)}^2$.

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}_ℓ is a hyperrectangle and f is a constant function when restricted to each \mathcal{R}_ℓ . Let $\gamma_2 := 9/2$ and $\gamma_d := (d^2 + d + 1)/2$ for $d \geq 3$.

THEOREM 4. *Let $d \geq 2$. There exists $C_{d,m_0,M_0} > 0$, depending only on d, m_0 and M_0 , such that*

$$\sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \max\{R(\hat{f}_n, f_0), R_n(\hat{f}_n, f_0)\} \leq C_{d,m_0,M_0} n^{-1/d} \log^{\gamma_d} n.$$

THEOREM 5. Fix $d \geq 2$, and a Borel measurable function $f_0 : [0, 1]^d \rightarrow \mathbb{R}$. There exists $C_{d, m_0, M_0} > 0$, depending only on d, m_0 and M_0 , such that

$$R_n(\hat{f}_n, f_0) \leq \inf_{k \in \mathbb{N}} \left\{ \inf_{f \in \mathcal{F}_d^{(k)}} \|f - f_0\|_{L_2(P)}^2 + C_{d, m_0, M_0} \left(\frac{k}{n}\right)^{2/d} \log_+^{2\gamma_d} \left(\frac{n}{k}\right) \right\}.$$

On the other hand, if we also have $\|f_0\|_\infty \leq 1$, then there exists a universal constant $C > 0$ such that

$$R(\hat{f}_n, f_0) \leq \inf_{k \in \mathbb{N}} \left\{ C \log n \inf_{f \in \mathcal{F}_d^{(k)}} \|f - f_0\|_{L_2(P)}^2 + C_{d, m_0, M_0} \left(\frac{k}{n}\right)^{2/d} \log^{2\gamma_d} n \right\}.$$

To the best of our knowledge, the bound in $L_2(\mathbb{P}_n)$ risk in Theorem 5 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 $\mathbb{E}\|f - f_0\|_{L_2(\mathbb{P}_n)}^2 = \|f - f_0\|_{L_2(P)}^2$ 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 4 and 5 are considerably more involved than those of the corresponding Theorems 1 and 2 in Section 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. (7)). As rates obtained from the entropy integral (e.g. van de Geer, 2000, 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. Yang and Barron, 1999, 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 4 and 5. 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 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,*

$$\inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \min\{R(\tilde{f}_n, f_0), R_n(\tilde{f}_n, f_0)\} \geq c_{d,m_0,M_0} n^{-1/d},$$

where the infimum is taken over all measurable functions \tilde{f}_n of the data $(X_1, Y_1), \dots, (X_n, Y_n)$.

We can also provide lower bounds on the adaptation rate risks for the least squares estimator when f_0 is constant.

PROPOSITION 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 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 $\{1, \dots, n\}$, a famous object of study in probability and combinatorics. See Romik (2014) and references therein.

4. Proofs of results in Section 2. Throughout this section, $\epsilon = (\epsilon_w)_{w \in \mathbb{L}_{d,n_1, \dots, n_d}}$ 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. Chatterjee (2014). 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 $T \subseteq \mathbb{R}^{\mathbb{L}_{d,n_1, \dots, n_d}}$ is defined as $w_T := \mathbb{E} \sup_{\theta \in T} \langle \epsilon, \theta \rangle$. Thus the result in the following proposition constitutes a key ingredient in analysing the risk of the least squares estimator.

PROPOSITION 5. *There exists a universal constant $C > 0$ such that for $d \geq 2$ and every $1 \leq n_1 \leq \dots \leq n_d$ with $\prod_{j=1}^d n_j = n$, we have*

$$\frac{\sqrt{2/\pi}}{(d-1)^{d-1}} n_1^{d-1} n^{-1/2} \leq \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n_1,\dots,n_d}) \cap B_2(1)} \langle \epsilon, \theta \rangle \leq C \sqrt{\frac{n}{n_{d-1}n_d}} \log^4 n.$$

We remark that in the case $n_1 = \dots = n_d = n^{1/d}$, 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 $n_1 \leq \dots \leq n_d$ 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 $\{n_1, \dots, n_d\}$ (considered here as a multiset).

PROOF. We first prove the lower bound. Consider $W := \{w \in \mathbb{L}_{d,n_1,\dots,n_d} : \sum_{j=1}^d n_j w_j = n_1\}$, $W^+ := \{w \in \mathbb{L}_{d,n_1,\dots,n_d} : \sum_{j=1}^d n_j w_j > n_1\}$ and $W^- := \{w \in \mathbb{L}_{d,n_1,\dots,n_d} : \sum_{j=1}^d n_j w_j < n_1\}$. For each realisation of the Gaussian random vector $\epsilon = (\epsilon_w)_{w \in \mathbb{L}_{d,n_1,\dots,n_d}}$, we define $\theta(\epsilon) = (\theta_w(\epsilon))_{w \in \mathbb{L}_{d,n_1,\dots,n_d}} \in \mathcal{M}(\mathbb{L}_{d,n_1,\dots,n_d})$ by

$$\theta_w := \begin{cases} 1 & \text{if } w \in W^+ \\ \text{sgn}(\epsilon_w) & \text{if } w \in W \\ -1 & \text{if } w \in W^-. \end{cases}$$

Since $\|\theta(\epsilon)\|_2^2 = n$, it follows that

$$\begin{aligned} \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n_1,\dots,n_d}) \cap B_2(1)} \langle \epsilon, \theta \rangle &\geq \mathbb{E} \left\langle \epsilon, \frac{\theta(\epsilon)}{\|\theta(\epsilon)\|_2} \right\rangle \\ &= \frac{1}{n^{1/2}} \mathbb{E} \left(\sum_{w \in W^+} \epsilon_w - \sum_{w \in W^-} \epsilon_w + \sum_{w \in W} |\epsilon_w| \right) = \frac{\sqrt{2/\pi}}{n^{1/2}} |W|. \end{aligned}$$

The proof of the lower bound is now completed by noting that

$$(9) \quad |W| = \binom{n_1-1}{d-1} \geq \left(\frac{n_1-1}{d-1} \right)^{d-1}.$$

We next prove the upper bound. For $j = 1, \dots, d-2$ and for $x_j \in \{1/n_j, 2/n_j, \dots, 1\}$, we define $A_{x_1, \dots, x_{d-2}} := \{w = (w_1, \dots, w_d) \in \mathbb{L}_{d,n_1, \dots, n_d} : (w_1, \dots, w_{d-2}) = (x_1, \dots, x_{d-2})\}$. Each $A_{x_1, \dots, x_{d-2}}$ can be viewed as a directed acyclic graph with graph structure inherited from $\mathbb{L}_{d,n_1, \dots, n_d}$. Since monotonicity is preserved on subgraphs, we have that $\mathcal{M}(\mathbb{L}_{d,n_1, \dots, n_d}) \subseteq$

$\bigoplus_{x_1, \dots, x_{d-2}} \mathcal{M}(A_{x_1, \dots, x_{d-2}})$. Hence, by the Cauchy–Schwarz inequality and [Amelunxen et al. \(2014, Proposition 3.1\(5, 9, 10\)\)](#), we obtain that

$$\begin{aligned} \left(\mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d, n_1, \dots, n_d}) \cap B_2(1)} \langle \epsilon, \theta \rangle \right)^2 &\leq \mathbb{E} \left\{ \left(\sup_{\theta \in \mathcal{M}(\mathbb{L}_{d, n_1, \dots, n_d}) \cap B_2(1)} \langle \epsilon, \theta \rangle \right)^2 \right\} \\ &= \delta(\mathcal{M}(\mathbb{L}_{d, n_1, \dots, n_d})) \leq \sum_{x_1, \dots, x_{d-2}} \delta(\mathcal{M}(A_{x_1, \dots, x_{d-2}})) \\ &= \delta(\mathcal{M}(\mathbb{L}_{2, n_{d-1}, n_d})) \prod_{j=1}^{d-2} n_j \lesssim \frac{n}{n_{d-1} n_d} \log_+^8(n_{d-1} n_d), \end{aligned}$$

as desired. Here, the final inequality follows from [Chatterjee, Guntuboyina and Sen \(2018, Theorem 2.1\)](#) by setting $\theta^* = 0$ (in their notation) and observing that $\delta(\mathcal{M}(\mathbb{L}_{2, n_{d-1}, n_d})) = n_{d-1} n_d R_n(\hat{\theta}_n, 0) \lesssim \log_+^8(n_{d-1} n_d)$. \square

PROOF OF THEOREM 1. Fix $\theta_0 \in \mathcal{M}(\mathbb{L}_{d, n}) \cap B_\infty(1)$. We have by [Chatterjee \(2014, Theorem 1.1\)](#) that the function

$$t \mapsto \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d, n}), \|\theta - \theta_0\| \leq t} \langle \epsilon, \theta - \theta_0 \rangle - t^2/2$$

is strictly concave on $[0, \infty)$ with a unique maximum at, say, $t_0 \geq 0$. We note that $t_0 \leq t_*$ for any t_* satisfying

$$(10) \quad \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d, n}), \|\theta - \theta_0\| \leq t_*} \langle \epsilon, \theta - \theta_0 \rangle \leq \frac{t_*^2}{2}.$$

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

$$\begin{aligned} \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d, n}), \|\theta - \theta_0\|_2 \leq t_*} \langle \epsilon, \theta - \theta_0 \rangle &= \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d, n}), \|\theta - \bar{\theta}_0 \mathbf{1}_n\|_2 \leq t_*} \langle \epsilon, \theta - \bar{\theta}_0 \mathbf{1}_n \rangle \\ &\leq \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d, n}), \|\theta - \bar{\theta}_0 \mathbf{1}_n\|_2 \leq t_* + n^{1/2}} \langle \epsilon, \theta - \bar{\theta}_0 \mathbf{1}_n \rangle \\ &= \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d, n}) \cap B_2(t_* + n^{1/2})} \langle \epsilon, \theta \rangle = \{t_* + n^{1/2}\} w_{\mathcal{M}(\mathbb{L}_{d, n}) \cap B_2(1)}, \end{aligned}$$

where we recall that $w_{\mathcal{M}(\mathbb{L}_{d, n}) \cap B_2(1)} = \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d, n}) \cap B_2(1)} \langle \epsilon, \theta \rangle$. Therefore, to satisfy (10), it suffices to choose

$$(11) \quad \begin{aligned} t_* &= w_{\mathcal{M}(\mathbb{L}_{d, n}) \cap B_2(1)} + \left\{ w_{\mathcal{M}(\mathbb{L}_{d, n}) \cap B_2(1)}^2 + 2n^{1/2} w_{\mathcal{M}(\mathbb{L}_{d, n}) \cap B_2(1)} \right\}^{1/2} \\ &\lesssim \max \left\{ w_{\mathcal{M}(\mathbb{L}_{d, n}) \cap B_2(1)}, n^{1/4} w_{\mathcal{M}(\mathbb{L}_{d, n}) \cap B_2(1)}^{1/2} \right\}. \end{aligned}$$

Consequently, by [Chatterjee \(2014, Corollary 1.2\)](#) and [Proposition 5](#), 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. \square

The following proposition is the main ingredient of the proof of the min-max lower bound in [Proposition 1](#). 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 $Y_v = \theta_v + \epsilon_v$, where $\theta = (\theta_v)_{v \in V(G)} \in \mathcal{M}(G)$ and $\epsilon = (\epsilon_v)_{v \in V(G)}$ is a vector of independent $N(0, 1)$ random variables.

PROPOSITION 6. *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 $u \not\leq v$, 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 $\theta^\tau = (\theta_v^\tau) \in \mathcal{M}(G) \cap B_\infty(1)$ by

$$\theta_v^\tau = \begin{cases} -1 & \text{if } v \in W_0^- \\ \rho(2\tau_v - 1) & \text{if } v \in W_0 \\ 1 & \text{if } v \in W_0^+, \end{cases}$$

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. [Pollard, 2002](#), p. 62), we have

$$d_{\text{TV}}^2(P_\tau, P_{\tau'}) \leq \frac{1}{2} d_{\text{KL}}^2(P_\tau, P_{\tau'}) = \frac{1}{4} \|\theta^\tau - \theta^{\tau'}\|_2^2 = \rho^2 \|\tau - \tau'\|_0.$$

Thus, setting $\rho = 2/3$, by Assouad's Lemma (cf. Yu, 1997, 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. \square

PROOF OF PROPOSITION 1. Recall that $n_1 = n^{1/d}$. We note that the set

$$W := \left\{ v = (v_1, \dots, v_d)^\top \in \mathbb{L}_{d,n} : \sum_{j=1}^d v_j = 1 \right\}$$

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 6. \square

PROOF OF COROLLARY 1. For $Q = P_n$, the result is an immediate consequence of Theorem 1 and Proposition 1, together with the facts that

$$\inf_{\tilde{\theta}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} R_n(\tilde{\theta}_n, \theta_0) = \inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E} \|\tilde{f}_n - f_0\|_{L_2(P_n)}^2$$

and

$$\sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} R_n(\hat{\theta}_n, \theta_0) = \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E} \|\hat{f}_n - f_0\|_{L_2(P_n)}^2.$$

Now suppose that Q is Lebesgue measure on $[0, 1]^d$. For any $f : [0, 1]^d \rightarrow \mathbb{R}$, we may define $\theta(f) := f|_{\mathbb{L}_{d,n}}$. On the other hand, for any $\theta : \mathbb{L}_{d,n} \rightarrow \mathbb{R}$, we can also define $f(\theta) : [0, 1]^d \rightarrow \mathbb{R}$ by

$$f(\theta)(x_1, \dots, x_d) := \theta(n_1^{-1} \lfloor n_1 x_1 \rfloor, \dots, n_1^{-1} \lfloor n_1 x_d \rfloor).$$

We first prove the upper bound by observing from Lemma 1 and Theorem 1 that

$$\begin{aligned} &\sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E} \|\hat{f}_n - f_0\|_{L_2(Q)}^2 \\ &\leq 2 \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \left\{ n^{-1} \mathbb{E} \|\theta(\hat{f}_n) - \theta(f_0)\|_2^2 + \|f_0 - f(\theta(f_0))\|_{L_2(Q)}^2 \right\} \\ &\leq 2 \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} \frac{1}{n} \mathbb{E} \|\hat{\theta}_n - \theta_0\|_2^2 + 8dn^{-1/d} \leq Cdn^{-1/d} \log^4 n, \end{aligned}$$

as desired. Then by convexity of \mathcal{H} and Proposition 1, we have

$$\begin{aligned} \inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} \mathbb{E} \|\tilde{f}_n - f_0\|_{L_2(Q)}^2 &\geq \inf_{\tilde{f}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} \mathbb{E} \|\tilde{f}_n - f(\theta_0)\|_{L_2(Q)}^2 \\ &= \inf_{\tilde{f}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} \mathbb{E} \|f(\theta(\tilde{f}_n)) - f(\theta_0)\|_{L_2(Q)}^2 \\ &= \inf_{\tilde{\theta}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} \frac{1}{n} \mathbb{E} \|\tilde{\theta}_n - \theta_0\|_2^2 \geq c_d n^{-1/d}, \end{aligned}$$

which completes the proof. \square

PROOF OF THEOREM 2. Recall that the tangent cone at a point x in a closed, convex set K is defined as $T(x, K) := \{t(y - x) : y \in K, t \geq 0\}$. By Bellec (2018, Proposition 2.1) (see also Chatterjee, Guntuboyina and Sen (2018, Lemma 4.1)), we have

$$(12) \quad R_n(\hat{\theta}_n, \theta_0) \leq \frac{1}{n} \inf_{\theta \in \mathcal{M}(\mathbb{L}_{d,n})} \left\{ \|\theta - \theta_0\|_2^2 + \delta(T(\theta, \mathcal{M}(\mathbb{L}_{d,n}))) \right\}.$$

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 Amelunxen et al. (2014, Proposition 3.1(9, 10)), we have

$$(13) \quad \delta(T(\theta, \mathcal{M}(\mathbb{L}_{d,n}))) \leq \delta\left(\bigoplus_{\ell=1}^K \mathcal{M}(A_\ell)\right) = \sum_{\ell=1}^K \delta(\mathcal{M}(A_\ell)).$$

By a consequence of the Gaussian Poincaré inequality (cf. Boucheron, Lugosi and Massart, 2013, p. 73) and Proposition 5, we have

$$(14) \quad \delta(\mathcal{M}(A_\ell)) \leq \left(\mathbb{E} \sup_{\theta \in \mathcal{M}(A_\ell) \cap B_2(1)} \langle \epsilon_{A_\ell}, \theta \rangle \right)^2 + 1 \lesssim \log_+^8 m_\ell.$$

Thus, by (13), (14) and Lemma 2 applied to $x \mapsto \log_+^8 x$, we have

$$\delta(T(\theta, \mathcal{M}(\mathbb{L}_{d,n}))) \lesssim \sum_{\ell=1}^K \log_+^8 m_\ell \lesssim K \log_+^8 \left(\frac{n}{K} \right),$$

which together with (12) proves the desired result. \square

PROOF OF THEOREM 3. 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}_ℓ . 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 $d \geq 2$), we obtain

$$(15) \quad K(\theta) \leq \sum_{\ell=1}^k |\mathcal{R}_\ell \cap \mathbb{L}_{d,n}|^{1-2/d} \leq k \left(\frac{n}{k}\right)^{1-2/d}.$$

This, combined with the oracle inequality in Theorem 2, gives the desired result. \square

PROOF OF PROPOSITION 2. 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

$$(16) \quad R_n(\hat{\theta}_n, 0) = \frac{1}{n} \delta(\mathcal{M}(\mathbb{L}_{d,n})) \geq \frac{1}{n} \left(\mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_2(1)} \langle \epsilon, \theta \rangle \right)^2,$$

which, together with Proposition 5, establishes the desired lower bound when $d \geq 3$. For the $d = 2$ case, by Sudakov minorisation for Gaussian processes (e.g. Pisier, 1999, Theorem 5.6 and the remark following it) and Lemma 3, there exists a universal constant $\varepsilon_0 > 0$ such that

$$\mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1)} \langle \epsilon, \theta \rangle \gtrsim \varepsilon_0 \log^{1/2} N(\varepsilon_0, \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1), \|\cdot\|_2) \gtrsim \log n.$$

This, together with (16), establishes the desired conclusion when $d = 2$. \square

PROOF OF COROLLARY 2. 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 $x_3, \dots, x_d \in \{1/n_1, 2/n_1, \dots, 1\}$, we define $A_{x_3, \dots, x_d} := \{(x_1, \dots, x_d) : x_1, x_2 \in [0, 1]\}$. When $r \leq d - 2$, we have that θ_0 is constant on each $A_{x_3, \dots, x_d} \cap \mathbb{L}_{d,n}$. Hence, by Theorem 2,

$$R_n(\hat{\theta}_n, \theta_0) \lesssim \frac{K(\theta_0) \log_+^8(n/K(\theta_0))}{n} \lesssim n^{-2/d} \log^8 n.$$

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 $A_{x_3, \dots, x_d} \cap \mathbb{L}_{d,n} = \sqcup_{\ell=-m}^m A_{x_3, \dots, x_d}^{(\ell)}$, where

$$A_{x_3, \dots, x_d}^{(\ell)} := A_{x_3, \dots, x_d} \cap \left\{ v \in \mathbb{L}_{d,n} : \frac{\ell-1}{m} < (\theta_0)_v \leq \frac{\ell}{m} \right\}.$$

Let $\theta^{(m)} \in \mathcal{M}(\mathbb{L}_{d,n})$ be the vector that takes the constant value ℓ/m on $A_{x_3, \dots, x_d}^{(\ell)}$ for each $\ell = -m, \dots, m$. Then setting $m \asymp n^{2/(3d)} \log^{-8/3} n$, we have by Theorem 2 that

$$\begin{aligned} R_n(\hat{\theta}_n, \theta_0) &\lesssim \frac{\|\theta^{(m)} - \theta_0\|_2^2}{n} + \frac{K(\theta^{(m)}) \log_+^8(n/K(\theta^{(m)}))}{n} \\ &\leq \frac{1}{m^2} + \frac{m}{n^{2/d}} \log^8 n \lesssim n^{-4/(3d)} \log^{16/3} n. \end{aligned}$$

as desired.

Finally, the $r = d$ case is covered in Theorem 1. \square

5. Proof of results in Section 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 dense³ 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.

5.1. Preparatory results. We first state a few intermediate results that will be used in the proofs of Theorems 4 and 5. The proofs of propositions in this subsection are contained Section A 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 (18) and (19). For $f_0 \in \mathcal{F}_d$, $r, a > 0$, define

$$(17) \quad \mathcal{G}(f_0, r, a) := \{f \in \mathcal{F}_d : f - f_0 \in B_2(r, P) \cap B_\infty(a)\}.$$

PROPOSITION 7. *Suppose that $f_0 \in \mathcal{F}_d \cap B_\infty(1)$ and that for each $n \geq 2$ there exist both a function $\phi_n : [0, \infty) \rightarrow [0, \infty)$ 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 $r \geq$*

³Here ‘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 $\|f_m - f\|_\infty \rightarrow 0$. This can be done by defining, e.g., $f_m(x) := m^{-1} \lceil mf(x) \rceil$.

r_n the map $r \mapsto \phi_n(r)/r$ is non-increasing and

$$(18) \quad \mathbb{E} \sup_{f \in \mathcal{G}(f_0, r, 6 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \epsilon_i \{f(X_i) - f_0(X_i)\} \right| \leq K \phi_n(r),$$

$$(19) \quad \mathbb{E} \sup_{f \in \mathcal{G}(f_0, r, 6 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i \{f(X_i) - f_0(X_i)\}^2 \right| \leq K \phi_n(r),$$

for some $K \geq 1$ that does not depend on r and n . Then, there exist universal constants $C, C' > 0$ such that for all $r \geq C' K r_n$, we have

$$\mathbb{P}(\{\|\hat{f}_n - f_0\|_{L_2(P)} \geq r\} \cap \{\|\hat{f}_n - f_0\|_\infty \leq 6 \log^{1/2} n\}) \leq C \exp\left(-\frac{nr^2}{C \log n}\right).$$

Consequently,

$$\mathbb{E}\{\|\hat{f}_n - f_0\|_{L_2(P)}^2 \mathbb{1}_{\{\|\hat{f}_n - f_0\|_\infty \leq 6 \log^{1/2} n\}}\} \lesssim K^2 r_n^2.$$

By means of Lemmas 5 and 6, the control of the empirical processes (18) and (19) in turn reduces to the study of the symmetrised local empirical process

$$(20) \quad \mathbb{E} \sup_{f \in \mathcal{G}(0, r, 1)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \right|,$$

for a suitable $L_2(P)$ radius r . To obtain a sharp bound on the empirical process in (20), 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 $\ell = 1, \dots, n_1$, and decompose $\sum_{i=1}^n \xi_i f(X_i)$ 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 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 9). This leads to the following conclusion:

PROPOSITION 8. *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*

$$\mathbb{E} \sup_{f \in \mathcal{G}(0, r, 1)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \right| \leq C_{d, m_0, M_0} r n^{1/2-1/d} \log^{\gamma_{d-1/2}} n.$$

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

$$\mathbb{E} \sup_{f \in \mathcal{G}(0,r,1)} \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \geq c_{d,m_0} r n^{1/2-1/d}.$$

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 9. 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

$$\mathbb{E} \|\tilde{f}_n\|_{L_2(\mathbb{P}_n)}^2 \leq C_{d,m_0,M_0} \left\{ n^{-2/d} \log^{2\gamma_d} n + \mathbb{E} \|\tilde{f}_n\|_{L_2(P)}^2 \right\}.$$

Propositions 7, 8 and 9 allow us to control the risk of the least squares estimator when the true signal $f_0 = 0$.

PROPOSITION 10. 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

$$\max\{R(\hat{f}_n, 0), R_n(\hat{f}_n, 0)\} \leq C_{d,m_0,M_0} n^{-2/d} \log^{2\gamma_d} n.$$

5.2. *Proofs of Theorems 4 and 5 and Propositions 3 and 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 4 IN $L_2(P)$ LOSS. Recall the definition of the function class $\mathcal{G}(f_0, r, a)$ in (17). Let $r_n := n^{-1/(2d)} \log^{\gamma_d/2} n$. For any $r, a > 0$, by the triangle inequality, Lemma 5 and Proposition 8, we have that for $r \geq r_n$,

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{G}(f_0, r, 4 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \epsilon_i \{f(X_i) - f_0(X_i)\} \right| \\ \leq \mathbb{E} \sup_{f \in \mathcal{G}(0, r+1, 6 \log^{1/2} n)} \left| \frac{2 \log^{1/2} n}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \right| + 1 \\ \lesssim_{d,m_0,M_0} (r+1) n^{1/2-1/d} \log^{\gamma_d} n \lesssim n^{1/2} r r_n. \end{aligned}$$

Similarly, by Lemma 6 and Proposition 8, we have that for $r \geq r_n$,

$$\begin{aligned} & \mathbb{E} \sup_{f \in \mathcal{G}(f_0, r, 4 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i \{f(X_i) - f_0(X_i)\}^2 \right| \\ & \lesssim \mathbb{E} \sup_{f \in \mathcal{G}(0, r+1, 6 \log^{1/2} n)} \left| \frac{\log^{1/2} n}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \right| + \log^{1/2} n \lesssim_{d, m_0, M_0} n^{1/2} r r_n. \end{aligned}$$

Thus, conditions (18) and (19) in Proposition 7 are satisfied with $\phi_n(r) = n^{1/2} r r_n$ and $1 \leq K \lesssim_{d, m_0, M_0} 1$. Let $\Omega_0 := \{\|\hat{f}_n - f_0\|_\infty \leq 6 \log^{1/2} n\}$. It follows from Proposition 7 and Lemma 10 that

$$\begin{aligned} R(\hat{f}_n, f_0) &= \mathbb{E}\{\|\hat{f}_n - f_0\|_{L_2(P)}^2 \mathbb{1}_{\Omega_0}\} + \mathbb{E}\{\|\hat{f}_n - f_0\|_{L_2(P)}^2 \mathbb{1}_{\Omega_0^c}\} \\ &\lesssim_{d, m_0, M_0} r_n^2 + n^{-1} \lesssim n^{-1/d} \log^{\gamma_d} n, \end{aligned}$$

as desired. \square

PROOF OF THEOREM 4 IN $L_2(\mathbb{P}_n)$ LOSS. Since the argument used in the proof of Theorem 1, up to (11), does not depend on the design, we deduce from Chatterjee (2014, Corollary 1.2), Amelunxen et al. (2014, Proposition 3.1(5)) and the Cauchy–Schwarz inequality that

$$(21) \quad R_n(\hat{f}_n, f_0) \lesssim \frac{1}{n} \mathbb{E} \max\{1, \delta(\mathcal{M}(G_X)), n^{1/2} \delta(\mathcal{M}(G_X))^{1/2}\}.$$

On the other hand, by Proposition 10, we have

$$(22) \quad \mathbb{E} \delta(\mathcal{M}(G_X)) \lesssim_{d, m_0, M_0} n^{1-2/d} \log^{2\gamma_d} n.$$

We obtain the desired result by combining (21) and (22). \square

PROOF OF THEOREM 5 IN $L_2(\mathbb{P}_n)$ LOSS. For any $f \in \mathcal{F}_d$, we can define a random vector $\theta_{f, X} := (f(X_1), \dots, f(X_n))^\top$. By Bellec (2018, Proposition 2.1), we have

$$\begin{aligned} (23) \quad R_n(\hat{f}_n, f_0) &\leq \frac{1}{n} \mathbb{E} \left[\inf_{f \in \mathcal{F}_d} \left\{ \|\theta_{f, X} - \theta_{f_0, X}\|_2^2 + \delta(T(\theta_{f, X}, \mathcal{M}(G_X))) \right\} \right] \\ &\leq \frac{1}{n} \inf_{k \in \mathbb{N}} \inf_{f \in \mathcal{F}_d^{(k)}} \left\{ \mathbb{E} \|\theta_{f, X} - \theta_{f_0, X}\|_2^2 + \mathbb{E} \delta(T(\theta_{f, X}, \mathcal{M}(G_X))) \right\}. \end{aligned}$$

Now, for a fixed $f \in \mathcal{F}_d^{(k)}$, let $\mathcal{R}_1, \dots, \mathcal{R}_k$ be the corresponding hyperrectangles such that f is constant when restricted to each \mathcal{R}_ℓ . Define $\mathcal{X}_\ell := \mathcal{R}_\ell \cap \{X_1, \dots, X_n\}$ and $N_\ell := |\mathcal{X}_\ell|$. Then for fixed X_1, \dots, X_n , we have

$T(\theta_{f,X}, \mathcal{M}(G_X)) \subseteq \bigoplus_{\ell=1}^k T(0, \mathcal{M}(G_{\mathcal{X}_\ell})) = \bigoplus_{\ell=1}^k \mathcal{M}(G_{\mathcal{X}_\ell})$. Therefore, by [Amelunxen et al. \(2014, Proposition 3.1\(9, 10\)\)](#) and [\(22\)](#), we have that

$$\begin{aligned} \mathbb{E} \delta(T(\theta_{f,X}, \mathcal{M}(G_X))) &= \mathbb{E} \left[\mathbb{E} \left\{ \delta(T(\theta_{f,X}, \mathcal{M}(G_X))) \mid N_1, \dots, N_k \right\} \right] \\ &\leq \mathbb{E} \left[\sum_{\ell: N_\ell \geq 1} \mathbb{E} \left\{ \delta(\mathcal{M}(G_{\mathcal{X}_\ell})) \mid N_\ell \right\} \right] \lesssim_{d, m_0, M_0} \mathbb{E} \left\{ \sum_{\ell: N_\ell \geq 1} N_\ell^{1-2/d} \log_+^{2\gamma_d} N_\ell \right\} \\ (24) \quad &\lesssim_d n(k/n)^{2/d} \log_+^{2\gamma_d}(n/k), \end{aligned}$$

where the final bound follows from applying [Lemma 2](#) to the function $x \mapsto x^{1-2/d} \log_+^{2\gamma_d} x$. We complete the proof by substituting [\(24\)](#) into [\(23\)](#) 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. \square

PROOF OF THEOREM 5 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}_ℓ . 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 $Y_i = f_0(X_i) + \epsilon_i$ and $Y_i = f_k(X_i) + \epsilon_i$ respectively, and write \mathbb{E}_{f_0} and \mathbb{E}_{f_k} for the respective expectations. For any $t \geq 0$, write $\Omega'_t := \{\|\hat{f}_n - f_0\|_{L_2(P)} > \|f_k - f_0\|_{L_2(P)} + t\} \cap \{\|\hat{f}_n - f_0\|_\infty \leq 3 \log^{1/2} n\}$. We have that

$$\begin{aligned} \mathbb{P}_{f_0}(\Omega'_t) &\leq \mathbb{P}_{f_0}(\{\|\hat{f}_n - f_k\|_{L_2(P)} > t\} \cap \{\|\hat{f}_n - f_k\|_\infty \leq 6 \log^{1/2} n\}) \\ &= \mathbb{E}_{f_k} \left\{ e^{-\frac{n}{2} \|f_k - f_0\|_{L_2(\mathbb{P}_n)}^2 - \sum_{i=1}^n \epsilon_i (f_k - f_0)(X_i)} \mathbb{1}_{\{\hat{f}_n - f_k \in B_2(t, P)^c \cap B_\infty(6 \log^{1/2} n)\}} \right\} \\ (25) \quad &\leq \mathbb{P}_{f_k} \left\{ \hat{f}_n - f_k \in B_2(t, P)^c \cap B_\infty(6 \log^{1/2} n) \right\}^{1/2} \left\{ \mathbb{E} e^{n \|f_k - f_0\|_{L_2(\mathbb{P}_n)}^2} \right\}^{1/2}, \end{aligned}$$

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 $\|f_k - f_0\|_\infty \leq 2$, we have by [Lemma 12](#) that

$$(26) \quad \mathbb{E} e^{n \|f_k - f_0\|_{L_2(\mathbb{P}_n)}^2} \leq e^{14n \|f_k - f_0\|_{L_2(P)}^2}.$$

For the first factor, for all $r \geq (k/n)^{1/d} \log^{\gamma_d} n =: r_{n,k}$, we have that

$$\begin{aligned}
& \mathbb{E}_{f_k} \sup_{f \in \mathcal{G}(f_k, r, 1)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \right| \\
& \leq \mathbb{E}_{f_k} \sum_{\ell: N_\ell \geq 1} \frac{N_\ell^{1/2}}{n^{1/2}} \mathbb{E}_{f_k} \left\{ \sup_{\substack{f \in \mathcal{F}_d, \|f - f_k\|_\infty \leq 1 \\ \|f - f_k\|_{L_2(P; \mathcal{R}_\ell)} \leq r}} \left| \frac{1}{N_\ell^{1/2}} \sum_{i: X_i \in \mathcal{R}_\ell} \xi_i f(X_i) \right| \middle| N_\ell \right\} \\
& \lesssim_{d, m_0, M_0} \frac{r \log^{\gamma_d - 1/2} n}{n^{1/2}} \mathbb{E}_{f_k} \sum_{\ell: N_\ell \geq 1} N_\ell^{1-1/d} \\
& \lesssim r n^{1/2} \left(\frac{k}{n} \right)^{1/d} \log^{\gamma_d - 1/2} n,
\end{aligned}$$

where the penultimate inequality follows from Proposition 8 and the final step uses Jensen's inequality. Using the above bound together with Lemmas 5 and 6 as in the proof of Theorem 4, we see that (18) and (19) (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 7, there exist universal constants $C, C' > 1$ such that for $t \geq C' K r_{n,k}$,

$$(27) \quad \mathbb{P}_{f_k} \{ \hat{f}_n - f_k \in B_2(t, P)^c \cap B_\infty(6 \log^{1/2} n) \} \leq C e^{-nt^2/(C \log n)}.$$

Substituting (27) and (26) into (25) and writing $t_0 := (28C \log n)^{1/2} \|f_k - f_0\|_{L_2(P)}$, 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 10, we obtain that

$$\begin{aligned}
R(\hat{f}_n, f_0) & \lesssim \mathbb{E}_{f_0} \left\{ \| \hat{f}_n - f_0 \|_{L_2(P)}^2 \mathbb{1}_{\{ \| \hat{f}_n - f_0 \|_\infty \leq 3 \log^{1/2} n \}} \right\} + \frac{1}{n} \\
& \lesssim \|f_k - f_0\|_{L_2(P)}^2 \log n + K^2 r_{n,k}^2 + \int_{t_0/2 + C' K r_{n,k}}^\infty (t + t_0) \mathbb{P}_{f_0}(\Omega'_t) dt \\
& \lesssim \|f_k - f_0\|_{L_2(P)}^2 \log n + K^2 r_{n,k}^2 \\
& \lesssim \|f_k - f_0\|_{L_2(P)}^2 \log n + C_{d, m_0, M_0} \left(\frac{k}{n} \right)^{2/d} \log^{2\gamma_d} n,
\end{aligned}$$

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 $\inf_{f \in \mathcal{F}_d^{(k)} \cap B_\infty(1)} \|f - f_0\|_{L_2(P)} = \inf_{f \in \mathcal{F}_d^{(k)}} \|f - f_0\|_{L_2(P)}$. \square

PROOF OF PROPOSITION 3 IN $L_2(P)$ LOSS. By Gao and Wellner (2007, Theorem 1.1), we have

$$\log N(\varepsilon, \mathcal{F}_d \cap B_\infty(1), \|\cdot\|_{L_2(P)}) \gtrsim_{m_0, d} \varepsilon^{-2(d-1)}.$$

The desired lower bound in $L_2(P)$ risk then follows from Yang and Barron (1999, Proposition 1). \square

PROOF OF PROPOSITION 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 $w = (w_1, \dots, w_d)^\top \in W$, we define $\mathcal{C}_w := \prod_{j=1}^d (w_j - 1/n_1, w_j]$. Note that $x = (x_1, \dots, x_d)^\top \in \cup_{w \in W} \mathcal{C}_w$ if and only if $\lceil n_1 x_1 \rceil + \dots + \lceil n_1 x_d \rceil = n_1$. For any $\tau = (\tau_w) \in \{0, 1\}^{|W|} =: T$, we define $f_\tau \in \mathcal{F}_d$ by

$$f_\tau(x) := \begin{cases} 0 & \text{if } \lceil n_1 x_1 \rceil + \dots + \lceil n_1 x_d \rceil \leq n_1 - 1 \\ 1 & \text{if } \lceil n_1 x_1 \rceil + \dots + \lceil n_1 x_d \rceil \geq n_1 + 1 \\ \rho \tau_{(\lceil n_1 x_1 \rceil, \dots, \lceil n_1 x_d \rceil)} & \text{if } x \in \cup_{w \in W} \mathcal{C}_w, \end{cases}$$

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}_τ for the expectation over $(X_1, Y_1), \dots, (X_n, Y_n)$, where $Y_i = f_\tau(X_i) + \epsilon_i$ for $i = 1, \dots, n$. 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

$$\begin{aligned} \max_{\tau \in T} \mathbb{E}_\tau \|\tilde{f}_n - f_\tau\|_{L_2(\mathbb{P}_n)}^2 &\geq \frac{1}{2^{|W|}} \sum_{w \in W} \sum_{\tau \in T} \mathbb{E}_\tau \int_{\mathcal{C}_w} (\tilde{f}_n - f_\tau)^2 d\mathbb{P}_n \\ &= \frac{1}{2^{|W|+1}} \sum_{w \in W} \sum_{\tau \in T} \left\{ \mathbb{E}_\tau \int_{\mathcal{C}_w} (\tilde{f}_n - f_\tau)^2 d\mathbb{P}_n + \mathbb{E}_{\tau^w} \int_{\mathcal{C}_w} (\tilde{f}_n - f_{\tau^w})^2 d\mathbb{P}_n \right\} \\ (28) \quad &\geq \frac{1}{2^{|W|+3}} \sum_{w \in W} \sum_{\tau \in T} \mathbb{E}_X \left\{ \int_{\mathcal{C}_w} (f_\tau - f_{\tau^w})^2 d\mathbb{P}_n \left[1 - d_{\text{TV}}(P_{Y|X, \tau}, P_{Y|X, \tau^w}) \right] \right\}, \end{aligned}$$

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 $\Delta := (\int_{\mathcal{C}_w} (f_\tau - f_{\tau^w})^2 d\mathbb{P}_n)^{1/2}$ and $A := \{\int_{\mathcal{C}_w} (\tilde{f}_n - f_\tau)^2 d\mathbb{P}_n \geq \Delta^2/4\}$, we have

$$\begin{aligned} &\mathbb{E}_{Y|X, \tau} \int_{\mathcal{C}_w} (\tilde{f}_n - f_\tau)^2 d\mathbb{P}_n + \mathbb{E}_{Y|X, \tau^w} \int_{\mathcal{C}_w} (\tilde{f}_n - f_{\tau^w})^2 d\mathbb{P}_n \\ &\geq \frac{\Delta^2}{4} \{P_{Y|X, \tau}(A) + P_{Y|X, \tau^w}(A^c)\} \geq \frac{\Delta^2}{4} \{1 - d_{\text{TV}}(P_{Y|X, \tau}, P_{Y|X, \tau^w})\}. \end{aligned}$$

By Pinsker's inequality (cf. [Pollard, 2002](#), p. 62), we obtain that

$$(29) \quad d_{\text{TV}}^2(P_{Y|X,\tau}, P_{Y|X,\tau^w}) \leq \frac{1}{2} d_{\text{KL}}^2(P_{Y|X,\tau}, P_{Y|X,\tau^w}) = \frac{n}{4} \|f_\tau - f_{\tau^w}\|_{L_2(\mathbb{P}_n)}^2.$$

Writing $N_w := \sum_{i=1}^n \mathbb{1}_{\{X_i \in \mathcal{C}_w\}}$, we have $N_w \sim \text{Bin}(n, P(\mathcal{C}_w))$, 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 (29), we have

$$(30) \quad \begin{aligned} \mathbb{E}_X \left\{ \int_{\mathcal{C}_w} (f_\tau - f_{\tau^w})^2 d\mathbb{P}_n \left[1 - d_{\text{TV}}(P_{Y|X,\tau}, P_{Y|X,\tau^w}) \right] \right\} \\ \geq \mathbb{E}_X \left\{ \|f_\tau - f_{\tau^w}\|_{L_2(\mathbb{P}_n)}^2 \left(1 - \frac{n^{1/2}}{2} \|f_\tau - f_{\tau^w}\|_{L_2(\mathbb{P}_n)} \right) \right\} \\ = \frac{\rho^2}{n} \mathbb{E}_X N_w - \frac{\rho^3}{2n} \mathbb{E}_X N_w^{3/2} \geq \frac{\rho^2}{n} \left(m_0 - \frac{\rho}{2^{1/2}} M_0^{3/2} \right). \end{aligned}$$

Substituting (30) into (28), we obtain that for $\rho = 2^{3/2} m_0 / (3M_0^{3/2})$,

$$\max_{\tau \in \mathcal{I}} \mathbb{E}_\tau \| \tilde{f}_n - f_\tau \|_{L_2(\mathbb{P}_n)}^2 \geq \frac{|W|}{27n} \frac{m_0^3}{M_0^3} \geq c_{d,m_0,M_0} n^{-1/d},$$

where the final inequality follows from a counting argument as in (9). This completes the proof. \square

PROOF OF PROPOSITION 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 $\tilde{f}_n = \tilde{f}_n(X_1, Y_1, \dots, X_n, Y_n)$ 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

$$R(\tilde{f}_n, f_0) = R(\tilde{f}'_n, f_0) + \int_{[0,1]^d} (\tilde{f}_n - \tilde{f}'_n)^2 dP \geq R(\tilde{f}'_n, f_0),$$

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 $f_0 \in \mathcal{F}_d^{(1)}$,

$$R(\tilde{f}_n, f_0) = \sup_{f \in \mathcal{F}_d^{(1)}} R(\hat{f}_n, f) \geq \inf_{\tilde{f}_n} \sup_{f \in \mathcal{F}_d^{(1)}} R(\tilde{f}_n, f) = \inf_{\tilde{\mu}_n} \sup_{\mu \in \mathbb{R}} \mathbb{E}\{(\tilde{\mu}_n - \mu)^2\} \gtrsim \frac{1}{n},$$

where the second infimum is taken over all estimators $\tilde{\mu}_n = \tilde{\mu}_n(Y_1, \dots, Y_n)$ of $\mu = f_0(0)$.

Case $d \geq 3$. It suffices to only consider the case when $f_0 = 0$. For $i = 1, \dots, n$, let $\tilde{\epsilon}_i := \epsilon_i \mathbb{1}_{\{|\epsilon_i| \leq 2 \log^{1/2} n\}}$ and for $r, b \geq 0$, define

$$E_n(r, b) := \sup_{f \in \mathcal{F}_d \cap B_2(r, P) \cap B_\infty(b)} \frac{1}{n} \sum_{i=1}^n \{2\tilde{\epsilon}_i f(X_i) - f^2(X_i) + \|f\|_{L_2(P)}^2\}.$$

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

$$\begin{aligned} \text{Var}\{2\tilde{\epsilon}_1 f(X_1) - f^2(X_1)\} &\leq r^2(8 + 2b^2) \lesssim r^2 \log n, \\ \|2\tilde{\epsilon}_1 f - f^2\|_\infty &\leq 4b \log^{1/2} n + b^2 \lesssim \log n. \end{aligned}$$

It follows by Talagrand's concentration inequality (Talagrand, 1996) in the form given by (Massart, 2000, 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 $\Omega_{r,b}$,

$$(31) \quad \frac{1}{2} \mathbb{E} E_n(r, b) - C_0 r^2 \leq E_n(r, b) \leq 2 \mathbb{E} E_n(r, b) + C_0 r^2.$$

Let $F_n(r) := E_n(r, 6 \log^{1/2} n) - r^2$ and choose

$$\tilde{f}_n \in \underset{f \in \mathcal{F}_d \cap B_\infty(6 \log^{1/2} n)}{\text{argmin}} \sum_{i=1}^n \{\tilde{\epsilon}_i - f(X_i)\}^2$$

such that $\tilde{f}_n = \hat{f}_n$ on the event $\Omega_0 := \{\|\hat{f}_n\|_\infty \leq 6 \log^{1/2} n\} \cap \cap_{i=1}^n \{|\epsilon_i| \leq 2 \log^{1/2} n\}$. Then for any $r \geq 0$, we have

$$\begin{aligned} F_n(r) &\leq \sup_{f \in \mathcal{F}_d \cap B_2(r, P) \cap B_\infty(6 \log^{1/2} n)} \frac{1}{n} \sum_{i=1}^n \{2\tilde{\epsilon}_i f(X_i) - f^2(X_i)\} \\ &\leq \frac{1}{n} \sum_{i=1}^n \{2\tilde{\epsilon}_i \tilde{f}_n(X_i) - \tilde{f}_n^2(X_i)\} = F_n(\|\tilde{f}_n\|_{L_2(P)}). \end{aligned}$$

In other words, $\|\tilde{f}_n\|_{L_2(P)} \in \text{argmax}_{r \geq 0} F_n(r)$.

If we can find $0 < r_1 < r_2$ such that

$$(32) \quad E_n(r_1, 6 \log^{1/2} n) < F_n(r_2),$$

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 $\text{argmax}_{r \geq 0} F_n(r)$ and therefore

$$(33) \quad \|\hat{f}_n\|_{L_2(P)}^2 \geq r_1^2 \mathbb{1}_{\Omega_0}.$$

It remains to choose suitable r_1 and r_2 that satisfy (32).

By (31), the symmetrisation inequality (van der Vaart and Wellner, 1996, Lemma 2.3.1), Lemmas 5 and 6 and Proposition 8, we have that for $r_1 \geq n^{-1/2} \log n$ and on $\Omega_{r_1, 6 \log^{1/2} n}$,

$$\begin{aligned} & E_n(r_1, 6 \log^{1/2} n) \\ & \leq 2\mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_1, P) \cap B_\infty(6 \log^{1/2} n)} \left\{ \frac{2}{n} \sum_{i=1}^n \tilde{\epsilon}_i f(X_i) - \frac{1}{n^{1/2}} \mathbb{G}_n f^2 \right\} + C_0 r_1^2 \\ & \leq 104 \log^{1/2} n \mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_1, P) \cap B_\infty(6 \log^{1/2} n)} \left| \frac{1}{n} \sum_{i=1}^n \xi_i f(X_i) \right| + C_0 r_1^2 \\ & \leq C_{d, m_0, M_0} r_1 n^{-1/d} \log^{\gamma_d} n + C_0 r_1^2, \end{aligned}$$

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 $\Omega_{r_2, b}$,

$$\begin{aligned} F_n(r_2) &= E_n(r_2, 6 \log^{1/2} n) - r_2^2 \\ &\geq \frac{1}{2} \mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_2, P) \cap B_\infty(b)} \left\{ \frac{2}{n} \sum_{i=1}^n \tilde{\epsilon}_i f(X_i) - \frac{1}{n^{1/2}} \mathbb{G}_n f^2 \right\} - (C_0 + 1) r_2^2 \\ &\geq (\mathbb{E} |\tilde{\epsilon}_1| - 4b) \mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_2, P) \cap B_\infty(b)} \frac{1}{n} \sum_{i=1}^n \xi_i f(X_i) - (C_0 + 1) r_2^2 \\ &\geq (1/2 - 4b) c_{d, m_0} r_2 n^{-1/d} - (C_0 + 1) r_2^2, \end{aligned}$$

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, (32) is satisfied. We therefore conclude from (33), Lemma 10 and the fact that $\mathbb{P}(|\epsilon_1| > 2 \log^{1/2} n) \leq n^{-2}$ that

$$R(\hat{f}_n, 0) \geq \mathbb{E} \left\{ \|\hat{f}_n\|_{L_2(P)}^2 \mathbb{1}_{\Omega_0 \cap \Omega_{r_1, 6 \log^{1/2} n} \cap \Omega_{r_2, b}} \right\} \gtrsim_{d, m_0, M_0} n^{-2/d} \log^{-2\gamma_d} n,$$

as desired. \square

PROOF OF PROPOSITION 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 4, there

is an event \mathcal{E} with probability at least $1 - e^{-ed^{-1}(M_0n)^{1/d} \log(M_0n)}$ on which the data points X_1, \dots, 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 ϵ , we define $\theta_X = \theta_X(\epsilon) = ((\theta_X)_w)$ by

$$(\theta_X)_w := \begin{cases} 1 & \text{if } w \in W_X^+ \\ \text{sgn}(\epsilon_w) & \text{if } w \in W_X \\ -1 & \text{if } w \in W_X^-, \end{cases}$$

so $\theta_X \in \mathcal{M}(G_X)$. By Chatterjee (2014, Theorem 1.1), for $f_0 = 0$, we have that

$$n^{1/2} \|\hat{f}_n\|_{L_2(\mathbb{P}_n)} = \operatorname{argmax}_{t \geq 0} \left(\sup_{\theta \in \mathcal{M}(G_X) \cap B_2(t)} \langle \epsilon, \theta \rangle - \frac{t^2}{2} \right) = \sup_{\theta \in \mathcal{M}(G_X) \cap B_2(1)} \langle \epsilon, \theta \rangle.$$

Hence

$$\begin{aligned} \mathbb{E} \|\hat{f}_n\|_{L_2(\mathbb{P}_n)} &= \frac{1}{n^{1/2}} \mathbb{E} \sup_{\theta \in \mathcal{M}(G_X) \cap B_2(1)} \langle \epsilon, \theta \rangle \geq \frac{1}{n^{1/2}} \mathbb{E} \left(\left\langle \epsilon, \frac{\theta_X(\epsilon)}{\|\theta_X(\epsilon)\|_2} \right\rangle \mathbb{1}_{\mathcal{E}} \right) \\ (34) \quad &= \frac{1}{n} \mathbb{E} \left(\sum_{i: X_i \in W_X^+} \epsilon_i \mathbb{1}_{\mathcal{E}} - \sum_{i: X_i \in W_X^-} \epsilon_i \mathbb{1}_{\mathcal{E}} + \sum_{i: X_i \in W_X} |\epsilon_i| \mathbb{1}_{\mathcal{E}} \right). \end{aligned}$$

The first two terms in the bracket are seen to be zero by computing the expectation conditionally on X_1, \dots, X_n . For the third term, we have that

$$\begin{aligned} \mathbb{E} \left(\sum_{i: X_i \in W_X} |\epsilon_i| \mathbb{1}_{\mathcal{E}} \right) &= \mathbb{E} \sum_{i: X_i \in W_X} \mathbb{E}(|\epsilon_i| \mathbb{1}_{\mathcal{E}} \mid X_1, \dots, X_n) \\ (35) \quad &\geq (2/\pi)^{1/2} \mathbb{E}(|W_X| \mathbb{1}_{\mathcal{E}}) \gtrsim_{d, M_0} n^{1-1/d}. \end{aligned}$$

By (34), (35) and the Cauchy–Schwarz inequality, we have that

$$\mathbb{E} \|\hat{f}_n\|_{L_2(\mathbb{P}_n)}^2 \geq \left\{ \mathbb{E} \|\hat{f}_n\|_{L_2(\mathbb{P}_n)} \right\}^2 \gtrsim_{d, M_0} n^{-2/d},$$

as desired. \square

Acknowledgements. The research of the first author is supported in part by NSF Grant DMS-1566514. The research of the second and fourth authors is supported by EPSRC fellowship EP/J017213/1 and a grant from the Leverhulme Trust RG81761. We thank the anonymous reviewers for their helpful and constructive comments on an earlier draft, which led to

significant improvements. We also thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme ‘Statistical Scalability’ when work on this paper was undertaken. This work was supported by Engineering and Physical Sciences Research Council grant numbers EP/K032208/1 and EP/R014604/1.

SUPPLEMENTARY MATERIAL

Supplement A: Supplementary material to ‘Isotonic regression in general dimensions’

(). Auxiliary results.

References.

- Amelunxen, D., Lotz, M., McCoy, M. B. and Tropp, J. A. (2014) Living on the edge: phase transition in convex programs with random data. *Inf. Inference*, **3**, 224–294.
- Bacchetti, P. (1989) Additive isotonic models. *J. Amer. Statist. Assoc.*, **84**, 289–294.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972) *Statistical Inference under Order Restrictions*. Wiley, New York.
- Bellec, P. C. (2018) Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.*, **46**, 745–780.
- Birgé, L. and Massart, P. (1993) Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, **97**, 113–150.
- Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities*. Oxford University Press, Oxford.
- Brunk, H. D. (1955) Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, **26**, 607–616.
- Chatterjee, S. (2014) A new perspective on least squares under convex constraint. *Ann. Statist.*, **42**, 2340–2381.
- Chatterjee, S., Guntuboyina, A. and Sen, B. (2015) On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, **43**, 1774–1800.
- Chatterjee, S., Guntuboyina, A. and Sen, B. (2018) On matrix estimation under monotonicity constraints. *Bernoulli*, **24**, 1072–1100.
- Chatterjee, S. and Lafferty, J. (2017) Adaptive risk bounds in unimodal regression. *arXiv preprint*, arxiv:1512.02956v5.
- Chen, Y. and Samworth, R. J. (2016) Generalized additive and index models with shape constraints. *J. Roy. Statist. Soc., Ser. B*, **78**, 729–754.
- Donoho, D. (1991) Gelfand n -widths and the method of least squares. *Technical Report*, University of California.
- Durot, C. (2007) On the L_p -error of monotonicity constrained estimators. *Ann. Statist.*, **35**, 1080–1104.
- Durot, C. (2008) Monotone nonparametric regression with random design. *Math. Methods Statist.*, **17**, 327–341.
- Dykstra, R. L. and Robertson, T. (1982) An algorithm for isotonic regression for two or more independent variables. *Ann. Statist.*, **10**, 708–716.
- Dykstra, R. L. (1983) An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.*, **78**, 837–842.
- Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. and Nadeau, J. H.

- (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.
- Elena, S. F. and Lenski R. E. (1997) Test of synergistic interactions among deleterious mutations in bacteria. *Nature*, **390**, 395–398.
- Gao, F. and Wellner, J. A. (2007) Entropy estimate for high-dimensional monotonic functions. *J. Mult. Anal.*, **98**, 1751–1764.
- Goldstein, D. B. (2009) Common genetic variation and human traits. *N. Engl. J. Med.*, **360**, 1696–1698.
- Groeneboom, P. and Jongbloed, G. (2014) *Nonparametric Estimation under Shape Constraints*. Cambridge University Press, Cambridge.
- Guntuboyina, A. and Sen, B. (2015) Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields*, **163**, 379–411.
- Han, Q., Wang, T., Chatterjee, S. and Samworth, R. J. (2018) Supplementary material to ‘Isotonic regression in general dimensions’. *Online supplement*.
- Kim, A. K. H., Guntuboyina, A. and Samworth, R. J. (2018) Adaptation in log-concave density estimation. *Ann. Statist.*, to appear.
- Kim, A. K. H. and Samworth, R. J. (2017) Global rates of convergence in log-concave density estimation. *Ann. Statist.*, **44**, 2756–2779.
- Kyng, R., Rao, A. and Sachdeva, S. (2015) Fast, provable algorithms for isotonic regression in all ℓ_p -norms. In *Advances in Neural Information Processing Systems*, pp. 2719–2727.
- Luss, R., Rosset, S. and Shahar, M. (2012) Efficient regularized isotonic regression with application to gene-gene interaction search. *Ann. Appl. Statist.*, **6**, 253–283.
- Mammen, E. and Yu, K. (2007) Additive isotonic regression. In Cator, E. A. et al. (Eds.), *Asymptotics: Particles, Processes and Inverse Problems*, 179–195. Institute of Mathematical Statistics, Beachwood.
- Mani, R., Onge, R. P. S., Hartman, J. L., Giaever, G. and Roth, F. P. (2008) Defining genetic interaction. *Proc. Nat. Acad. Sci. USA*, **105**, 3461–3466.
- Massart, P. (2000) About the constant in Talagrand’s concentration inequalities for empirical processes. *Ann. Probab.*, **28**, 863–884.
- Meyer, M. and Woodroffe, M. (2000) On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, **28**, 1083–1104.
- Morton-Jones, T., Diggle, P., Parker, L., Dickinson, H. O. and Binks, K. (2000) Additive isotonic regression models in epidemiology. *Statist. Med.*, **19**, 849–859.
- Pisier, G. (1999) *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press, Cambridge.
- Pollard, D. (2002) *A User’s Guide to Measure Theoretic Probability*. Cambridge University Press, Cambridge.
- Rakhlin, A., Sridharan, K. and Tsybakov, A. B. (2017) Empirical entropy, minimax regret and minimax risk. *Bernoulli*, **23**, 789–824.
- Romik, D. (2014) *The Surprising Mathematics of Longest Increasing Subsequences*. Cambridge University Press, Cambridge.
- Roth, F. P., Lipshitz, H. D. and Andrews, B. J. (2009) Q&A: epistasis. *J. Biol.*, **8**, 35.
- Sanjuan, R. and Elena, S. F. (2006) Epistasis correlates to genomic complexity. *Proc. Natl. Acad. Sci. USA*, **103**, 14402–14405.
- Shao, H., Burrage, L. C., Sinasac, D. S., Hill, A. E., Ernest, S. R., O’Brien, W., Courtland, H.-W., Jepsen, K. J., Kirby, A., Kulbokas, E. J., Daly, M. J., Broman, K. W., Lander, E. S. and Nadeau, J. H. (2008) Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. *Proc. Natl. Acad. Sci. USA*, **105**, 19910–19914.
- Schell, M. J. and Singh, B. (1997) The reduced monotonic regression method. *J. Amer. Statist. Assoc.*, **92**, 128–135.

- Stout, Q. F. (2015) Isotonic regression for multiple independent variables. *Algorithmica*, **71**, 450–470.
- Talagrand, M. (1996) New concentration inequalities in product spaces. *Invent. Math.*, **126**, 505–563.
- Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Pagé, N., Robinson, M., Raghbizadeh, S., Hogue, C. W. V., Bussey, H., Andrews, B., Tyers, M. and Boone, C. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.
- van de Geer, S. A. (1990) Estimating a regression function. *Ann. Statist.* **18**, 907–924.
- van de Geer, S. A. (1993) Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14–44.
- van de Geer, S. A. (2000) *Applications of Empirical Process Theory*. Cambridge University Press, Cambridge.
- van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. Springer, New York.
- van Eeden, C. (1958) *Testing and estimating ordered parameters of probability distributions*. Ph.D. thesis, University of Amsterdam.
- Yang, F. and Barber, R. F. (2017) Uniform convergence of isotonic regression. *arXiv preprint*, arxiv:1706.01852.
- Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, **27**, 1564–1599.
- Yu, B. (1997) Assouad, Fano and Le Cam. In Pollard, D., Torgersen, E. and Yang G. L. (Eds.) *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, 423–435. Springer, New York.
- Zhang, C.-H. (2002) Risk bounds in isotonic regression. *Ann. Statist.*, **30**, 528–555.

DEPARTMENT OF STATISTICS, BOX 354322
 UNIVERSITY OF WASHINGTON
 SEATTLE, WA 98195
 UNITED STATES
 E-MAIL: royhan@uw.edu

STATISTICAL LABORATORY
 WILBERFORCE ROAD
 CAMBRIDGE, CB3 0WB
 UNITED KINGDOM
 E-MAIL: t.wang@statslab.cam.ac.uk
 E-MAIL: r.samworth@statslab.cam.ac.uk

ILLINI HALL
 725 S WRIGHT ST #101
 CHAMPAIGN, IL 61820
 UNITED STATES
 E-MAIL: sc1706@illinois.edu

SUPPLEMENTARY MATERIAL TO ‘ISOTONIC REGRESSION IN GENERAL DIMENSIONS’

BY QIYANG HAN^{*}, TENG YAO WANG[†], SABYASACHI
CHATTERJEE^{‡,§} AND RICHARD J. SAMWORTH[†]

University of Washington^{}, University of Cambridge[†], University of
Chicago[‡] and University of Illinois at Urbana-Champaign[§]*

APPENDIX A: PROOFS OF PREPARATORY PROPOSITIONS

PROOF OF PROPOSITION 7. For any $f : [0, 1]^d \rightarrow \mathbb{R}$, define $\mathbb{M}_n f := 2 \sum_{i=1}^n \epsilon_i \{f(X_i) - f_0(X_i)\} - \sum_{i=1}^n \{f(X_i) - f_0(X_i)\}^2$ 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 $\sum_{i=1}^n (\hat{f}_n(X_i) - f_0(X_i) - \epsilon_i)^2 \leq \sum_{i=1}^n \epsilon_i^2$, which implies that $\mathbb{M}_n \hat{f}_n \geq 0$. We therefore have that for any $r > 0$,

$$\begin{aligned}
 & \mathbb{P}(\{\|\hat{f}_n - f_0\|_{L_2(P)} \geq r\} \cap \{\|\hat{f}_n - f_0\|_\infty \leq 6 \log^{1/2} n\}) \\
 & \leq \sum_{\ell=1}^{\infty} \mathbb{P}\left(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6 \log^{1/2} n) \setminus \mathcal{G}(f_0, 2^{\ell-1} r, 6 \log^{1/2} n)} (\mathbb{M}_n - M)f \geq n 2^{2\ell-2} r^2\right) \\
 & \leq \sum_{\ell=1}^{\infty} \mathbb{P}\left(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \epsilon_i (f - f_0)(X_i) \right| \geq 2^{2\ell-4} n^{1/2} r^2\right) \\
 (1) \quad & + \sum_{\ell=1}^{\infty} \mathbb{P}\left(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6 \log^{1/2} n)} \left| \mathbb{G}_n(f - f_0)^2 \right| \geq 2^{2\ell-3} n^{1/2} r^2\right).
 \end{aligned}$$

By a moment inequality for empirical processes (Giné, Latała and Zinn, 2000, Proposition 3.1) and (18) in the main text, we have for all $p \geq 1$ that

$$\begin{aligned}
 & \mathbb{E} \left[\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \epsilon_i \{f(X_i) - f_0(X_i)\} \right|^p \right]^{1/p} \\
 (2) \quad & \lesssim K \phi_n(2^\ell r) + 2^\ell r p^{1/2} + n^{-1/2} p \log n.
 \end{aligned}$$

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 (2) and Lemma 11 that there exist universal constants $C, C' > 0$ such that for all $\ell \in \mathbb{N}$ and

$$r \geq C' K r_n,$$

$$(3) \quad \mathbb{P}\left(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \epsilon_i \{f(X_i) - f_0(X_i)\} \right| \geq 2^{2\ell-4} n^{1/2} r^2\right) \leq C \exp\left(-\frac{2^{2\ell} n r^2}{C \log n}\right).$$

Similarly, by a symmetrisation inequality (cf. [van der Vaart and Wellner \(1996, Lemma 2.3.1\)](#)), (19) 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 $r \geq C' K r_n$,

$$(4) \quad \mathbb{P}\left(\sup_{f \in \mathcal{G}(f_0, 2^\ell r, 6 \log^{1/2} n)} \left| \mathbb{G}_n(f - f_0)^2 \right| \geq 2^{2\ell-3} n^{1/2} r^2\right) \leq C \exp\left(-\frac{2^{2\ell} n r^2}{C \log n}\right).$$

Substituting (3) and (4) into (1), we obtain that for all $r \geq C' K r_n$,

$$\begin{aligned} & \mathbb{P}(\{\|\hat{f}_n - f_0\|_{L_2(P)} \geq r\} \cap \{\|\hat{f}_n - f_0\|_\infty \leq 6 \log^{1/2} n\}) \\ & \lesssim \sum_{\ell=1}^{\infty} \exp\left(-\frac{2^{2\ell} n r^2}{C \log n}\right) \lesssim \exp\left(-\frac{n r^2}{C \log n}\right). \end{aligned}$$

It follows that

$$\begin{aligned} & \mathbb{E}(\|\hat{f}_n - f_0\|_{L_2(P)}^2 \mathbb{1}_{\{\|\hat{f}_n - f_0\|_\infty \leq 6 \log^{1/2} n\}}) \\ & = \int_0^\infty 2t \mathbb{P}(\{\|\hat{f}_n - f_0\|_{L_2(P)} \geq t\} \cap \{\|\hat{f}_n - f_0\|_\infty \leq 6 \log^{1/2} n\}) dt \\ & \lesssim K^2 r_n^2 + \int_{C' K r_n}^\infty 2t \exp\left(-\frac{t^2}{C r_n^2}\right) dt \lesssim K^2 r_n^2, \end{aligned}$$

as desired, where we have used $r_n^2 \geq n^{-1} \log n$ in the penultimate inequality. \square

PROOF OF PROPOSITION 8. [Upper bound] It is convenient here to work with the class of block decreasing functions $\mathcal{F}_{d,\downarrow} := \{f : [0, 1]^d \rightarrow \mathbb{R} : -f \in \mathcal{F}_d\}$ 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 7 with $\eta = r/(2n)$ and Lemma 8 to obtain

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_{2,\downarrow}^+ \cap B_2(r,P) \cap B_\infty(1)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \right| \\ \lesssim_{d,m_0,M_0} n^{1/2} \eta + \log^3 n \int_\eta^r \frac{r}{\varepsilon} d\varepsilon + \frac{(\log^4 n)(\log \log n)^2}{n^{1/2}} \lesssim r \log^4 n, \end{aligned}$$

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 $\ell = 1, \dots, n_1$, 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 $S_\ell := \{X_1, \dots, X_n\} \cap I_\ell$ 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,

$$\mathbb{P}(\Omega_0^c) \leq \sum_{\ell=1}^{n_1} \mathbb{P}\left(\left|N_\ell - \mathbb{E}N_\ell\right| > \frac{m_0 n}{2n_1}\right) \leq 2n_1 \exp(-m_0^2 n^{1-2/d}/8).$$

Hence we have

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_{d,\downarrow}^+ \cap B_2(r,P) \cap B_\infty(1)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \right| \\ (5) \quad \leq \mathbb{E}\left(\sum_{\ell: N_\ell \geq 1} \frac{N_\ell^{1/2}}{n^{1/2}} E_\ell \mathbb{1}_{\Omega_0}\right) + C \exp(-m_0^2 n^{1-2/d}/16), \end{aligned}$$

where

$$E_\ell := \mathbb{E}\left\{ \sup_{f \in \mathcal{F}_{d,\downarrow}^+ \cap B_2(r,P) \cap B_\infty(1)} \left| \frac{1}{N_\ell^{1/2}} \sum_{i: X_i \in S_\ell} \xi_i f(X_i) \right| \middle| N_1, \dots, N_{n_1} \right\}.$$

By Lemma 9, for any $f \in \mathcal{F}_{d,\downarrow}^+ \cap B_2(r,P) \cap B_\infty(1)$ and $\ell \in \{1, \dots, n_1\}$, we have $\int_{I_\ell} f^2 dP \leq 7(M_0/m_0)\ell^{-1}r^2 \log^d n =: r_{n,\ell}^2$. Consequently, we have by Lemma 7 that for any $\eta \in [0, r_{n,\ell}/3]$,

$$(6) \quad E_\ell \lesssim N_\ell^{1/2} \eta + \int_\eta^{r_{n,\ell}} H_{[\cdot],\ell}^{1/2}(\varepsilon) d\varepsilon + \frac{H_{[\cdot],\ell}(r_{n,\ell})}{N_\ell^{1/2}},$$

where $H_{[\cdot],\ell}(\varepsilon) := \log N_{[\cdot]}(\varepsilon, \mathcal{F}_{d,\downarrow}^+(I_\ell) \cap B_2(r_{n,\ell}, P; I_\ell) \cap B_\infty(1; I_\ell), \|\cdot\|_{L_2(P; I_\ell)})$. 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 $f_g \in \mathcal{F}_{d,\downarrow}^+ \cap B_2(n_1^{1/2}(M_0/m_0)^{1/2}r_{n,\ell}, P) \cap B_\infty(1)$ via the invertible map $f_g(x_1, \dots, x_{d-1}, x_d) := g(x_1, \dots, x_{d-1}, (x_d + \ell - 1)/n_1)$. Moreover, we have $\int_{[0,1]^d} (f_g - f_{g'})^2 dP \geq n_1(m_0/M_0) \int_{I_\ell} (g - g')^2 dP$. Thus, by Lemma 8, for $\varepsilon \in [\eta, r_{n,\ell}]$,

$$\begin{aligned} H_{[\cdot], \ell}(\varepsilon) &\leq \log N_{[\cdot]}(n^{1/(2d)}(m_0/M_0)^{1/2}\varepsilon, \\ &\quad \mathcal{F}_{d,\downarrow}^+ \cap B_2(n^{1/(2d)}(M_0/m_0)^{1/2}r_{n,\ell}, P) \cap B_\infty(1), \|\cdot\|_{L_2(P)}) \\ &\lesssim_{d, m_0, M_0} \left(\frac{r_{n,\ell}}{\varepsilon}\right)^{2(d-1)} \log_+^{d^2}(1/\varepsilon). \end{aligned}$$

Substituting the above bound into (6), and choosing $\eta = n^{-1/(2d)}r_{n,\ell}$, we obtain

$$\begin{aligned} E_\ell &\lesssim_{d, m_0, M_0} N_\ell^{1/2}\eta + \log^{d^2/2} n \int_\eta^{r_{n,\ell}} \left(\frac{r_{n,\ell}}{\varepsilon}\right)^{d-1} d\varepsilon + \frac{\log^{d^2} n}{N_\ell^{1/2}} \\ &\lesssim N_\ell^{1/2}\eta + \frac{r_{n,\ell}^{d-1} \log^{d^2/2} n}{\eta^{d-2}} + \frac{\log^{d^2} n}{N_\ell^{1/2}}. \end{aligned}$$

Hence

$$\begin{aligned} E_\ell \mathbb{1}_{\Omega_0} &\lesssim_{d, m_0, M_0} r_{n,\ell} n^{1/2-1/d} \log^{d^2/2} n + n^{-1/2+1/(2d)} \log^{d^2} n \\ (7) \quad &\lesssim_{m_0, M_0} r_{n,\ell} n^{1/2-1/d} \log^{d^2/2} n, \end{aligned}$$

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) and (7), we have that

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_{d,\downarrow}^+ \cap B_2(r, P) \cap B_\infty(1)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i f(X_i) \right| \\ \lesssim_{d, m_0, M_0} r n^{1/2-3/(2d)} \log^{(d^2+d)/2} n \sum_{\ell=1}^{n_1} \ell^{-1/2} \lesssim r n^{1/2-1/d} \log^{(d^2+d)/2} n, \end{aligned}$$

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 $w = (w_1, \dots, w_d) \in \mathbb{L}_{d,n}$, let $L_w := \prod_{j=1}^d (w_j - 1/n_1, w_j]$ and $N_w := |\{X_1, \dots, X_n\} \cap L_w|$. We define $W := \{(w_1, \dots, w_d) :$

$\sum_{j=1}^d w_j = 1$ 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 $\xi = (\xi_i)_{i=1}^n$ and design points $X = \{X_i\}_{i=1}^n$, define $f_{\xi, X} : [0, 1]^d \rightarrow [-1, 1]$ to be the function such that

$$f_{\xi, X}(x) := \begin{cases} r \xi_{i_w} & \text{if } x \in L_w, w \in \tilde{W} \\ r & \text{if } x \in L_w \text{ with } \sum_{j=1}^d w_j > n_1 \\ -r & \text{otherwise.} \end{cases}$$

For $r \leq 1$, we have $f_{\xi, X} \in \mathcal{F}_d \cap B_2(r, P) \cap B_\infty(1)$. Therefore,

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r, P) \cap B_\infty(1)} \sum_{i=1}^n \xi_i f(X_i) &\geq \mathbb{E} \sum_{i=1}^n \xi_i f_{\xi, X}(X_i) \\ &\geq \mathbb{E} \left[\mathbb{E} \left\{ \sum_{i=1}^n \xi_i f_{\xi, X}(X_i) \mid X_1, \dots, X_n, \{\xi_{i_w} : w \in \tilde{W}\} \right\} \right] \\ &= \mathbb{E} \sum_{w \in \tilde{W}} \xi_{i_w} f_{\xi, X}(X_{i_w}) = r \mathbb{E} |\tilde{W}|. \end{aligned}$$

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. \square

PROOF OF PROPOSITION 9. Let $r_n := n^{-1/d} \log^{\gamma_d} n$. We write

$$(8) \quad \mathbb{E} \|\tilde{f}_n\|_{L_2(\mathbb{P}_n)}^2 = \mathbb{E} \left\{ \|\tilde{f}_n\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} \leq r_n\}} \right\} + \mathbb{E} \left\{ \|\tilde{f}_n\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} > r_n\}} \right\}$$

and control the two terms on the right hand side of (8) separately. For the first term, we have

$$\begin{aligned} \mathbb{E} \left\{ \|\tilde{f}_n\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} \leq r_n\}} \right\} &\leq \mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_n, P) \cap B_\infty(6 \log^{1/2} n)} \frac{1}{n} \sum_{i=1}^n f^2(X_i) \\ &\lesssim r_n^2 + \frac{1}{n} \mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_n, P) \cap B_\infty(6 \log^{1/2} n)} \left| \sum_{i=1}^n \xi_i f^2(X_i) \right| \\ &\lesssim r_n^2 + \frac{\log^{1/2} n}{n} \mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r_n, P) \cap B_\infty(6 \log^{1/2} n)} \left| \sum_{i=1}^n \xi_i f(X_i) \right| \\ (9) \quad &\lesssim_{d, m_0, M_0} r_n^2 + r_n n^{-1/d} \log^{\gamma_d} n \lesssim r_n^2, \end{aligned}$$

where the second line uses the symmetrisation inequality (cf. [van der Vaart and Wellner, 1996](#), Lemma 2.3.1), the third inequality follows from Lemma 6 and the penultimate inequality follows from Proposition 8. For the second term on the right-hand side of (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

$$(10) \quad \mathbb{P}(\mathcal{E}^c) \leq \frac{2}{n^2},$$

where

$$\mathcal{E} := \left\{ \sup_{f \in \mathcal{F}_d \cap B_2(r_n, P)^c \cap B_\infty(6 \log^{1/2} n)} \left| \frac{\mathbb{P}_n f^2}{P f^2} - 1 \right| \leq C'_{d,m_0,M_0} \right\}.$$

To see this, we adopt a peeling argument as follows. Let $\mathcal{F}_{d,\ell} := \{f \in \mathcal{F}_d \cap B_\infty(6 \log^{1/2} n) : 2^{\ell-1} r_n^2 < P f^2 \leq 2^\ell r_n^2\}$ 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

$$\sup_{\substack{f \in \mathcal{F}_d \cap B_\infty(6 \log^{1/2} n) \\ \|f\|_{L_2(P)} > r_n}} \left| \frac{\mathbb{P}_n f^2}{P f^2} - 1 \right| \leq \frac{2}{n^{1/2}} \max_{\ell=1, \dots, m} \left\{ (2^\ell r_n^2)^{-1} \sup_{f \in \mathcal{F}_{d,\ell}} |\mathbb{G}_n f^2| \right\}.$$

By Talagrand's concentration inequality for empirical processes ([Talagrand, 1996](#)), in the form given by [Massart \(2000, Theorem 3\)](#), applied to the class $\{f^2 : f \in \mathcal{F}_{d,\ell}\}$, we have that for any $s_\ell > 0$,

$$\mathbb{P} \left\{ \sup_{f \in \mathcal{F}_{d,\ell}} |\mathbb{G}_n f^2| > 2\mathbb{E} \sup_{f \in \mathcal{F}_{d,\ell}} |\mathbb{G}_n f^2| + 12\sqrt{2} (2^\ell s_\ell \log n)^{1/2} r_n + \frac{1242 s_\ell \log n}{n^{1/2}} \right\} \leq e^{-s_\ell}.$$

Here we have used the fact that $\sup_{f \in \mathcal{F}_{d,\ell}} \text{Var}_P f^2 \leq \sup_{f \in \mathcal{F}_{d,\ell}} P f^2 \|f\|_\infty^2 \leq 36 \cdot 2^\ell r_n^2 \log n$. Further, by the symmetrisation inequality again, Lemma 6 and Proposition 8, we have that

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_{d,\ell}} |\mathbb{G}_n f^2| &\leq \frac{2}{n^{1/2}} \mathbb{E} \sup_{f \in \mathcal{F}_{d,\ell}} \left| \sum_{i=1}^n \xi_i f^2(X_i) \right| \leq \frac{48 \log^{1/2} n}{n^{1/2}} \mathbb{E} \sup_{f \in \mathcal{F}_{d,\ell}} \left| \sum_{i=1}^n \xi_i f(X_i) \right| \\ &\lesssim_{d,m_0,M_0} 2^{\ell/2} r_n n^{1/2-1/d} \log^{\gamma_d} n. \end{aligned}$$

By a union bound, we have that with probability at least $1 - \sum_{\ell=1}^m e^{-s_\ell}$,

$$\begin{aligned} &\sup_{f \in \mathcal{F}_d \cap B_2(r_n, P)^c \cap B_\infty(6 \log^{1/2} n)} \left| \frac{\mathbb{P}_n f^2}{P f^2} - 1 \right| \\ &\lesssim_{d,m_0,M_0} \max_{\ell=1, \dots, m} \left\{ \frac{n^{1/2-1/d} \log^{\gamma_d} n + s_\ell^{1/2} \log^{1/2} n}{2^{\ell/2} n^{1/2} r_n} + \frac{s_\ell \log n}{2^\ell n r_n^2} \right\}. \end{aligned}$$

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 (10). Thus

$$\begin{aligned} \mathbb{E} \left\{ \|\tilde{f}_n\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} > r_n\}} \right\} &\leq \mathbb{E} \left\{ \|\tilde{f}_n\|_{L_2(\mathbb{P}_n)}^2 \mathbb{1}_{\{\|\hat{f}_n\|_{L_2(P)} > r_n\}} \mathbb{1}_{\mathcal{E}} \right\} + \frac{72 \log n}{n^2} \\ (11) \qquad \qquad \qquad &\leq (C'_{d, m_0, M_0} + 1) \mathbb{E} \|\tilde{f}_n\|_{L_2(P)}^2 + \frac{72 \log n}{n^2}. \end{aligned}$$

Combining (8), (9) and (11), we obtain

$$\mathbb{E} \|\tilde{f}_n\|_{L_2(\mathbb{P}_n)}^2 \lesssim_{d, m_0, M_0} r_n^2 + \mathbb{E} \|\tilde{f}_n\|_{L_2(P)}^2,$$

as desired. \square

PROOF OF PROPOSITION 10. Let $r_n := n^{-1/d} \log^{\gamma_d} n$ and observe that by Lemma 5 and Proposition 8, we have that for $r \geq r_n$,

$$\mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r, P) \cap B_\infty(6 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \epsilon_i f(X_i) \right| \lesssim_{d, m_0, M_0} r n^{1/2-1/d} \log^{\gamma_d} n.$$

On the other hand, by Lemma 6 and Proposition 8, for $r \geq r_n$,

$$\mathbb{E} \sup_{f \in \mathcal{F}_d \cap B_2(r, P) \cap B_\infty(6 \log^{1/2} n)} \left| \frac{1}{n^{1/2}} \sum_{i=1}^n \xi_i f^2(X_i) \right| \lesssim_{d, m_0, M_0} r n^{1/2-1/d} \log^{\gamma_d} n.$$

It follows that the conditions of Proposition 7 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 10, Propositions 9 and 7, we have that

$$\begin{aligned} R_n(\hat{f}_n, 0) &\leq \mathbb{E} \|\tilde{f}_n\|_{L_2(\mathbb{P}_n)}^2 + n^{-1} \\ &\lesssim_{d, m_0, M_0} n^{-2/d} \log^{2\gamma_d} n + \mathbb{E} \|\tilde{f}_n\|_{L_2(P)}^2 \lesssim_{d, m_0, M_0} n^{-2/d} \log^{2\gamma_d} n, \end{aligned}$$

as desired. \square

APPENDIX B: AUXILIARY LEMMAS

We collect here various auxiliary results used in the proofs in the main document ([Han et al., 2018](#)).

The proof of Corollary 1 in the main document requires the following lemma on Riemann approximation of block increasing functions.

LEMMA 1. *Suppose $n_1 = n^{1/d}$ is a positive integer. For any $f \in \mathcal{F}_d$, define $f_L(x_1, \dots, x_d) := f(n_1^{-1}\lfloor n_1 x_1 \rfloor, \dots, n_1^{-1}\lfloor n_1 x_d \rfloor)$ and $f_U(x_1, \dots, x_d) := f(n_1^{-1}\lceil n_1 x_1 \rceil, \dots, n_1^{-1}\lceil n_1 x_d \rceil)$. Then*

$$\int_{[0,1]^d} (f_U - f_L)^2 \leq 4dn^{-1/d} \|f\|_\infty^2.$$

PROOF. For $x = (x_1, \dots, x_d)^\top$ and $x' = (x'_1, \dots, x'_d)^\top$ 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 $j = 1, \dots, d$. 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 $\{(x_1, \dots, x_d) \in \mathbb{L}_{d,n} : \max_j x_j = 1\}$, we must have $N \leq dn^{1-1/d}$. Therefore, we have

$$\begin{aligned} \int_{[0,1]^d} (f_U - f_L)^2 &= \sum_{r=1}^N \int_{P_r + n_1^{-1}(-1,0]^d} (f_U - f_L)^2 \\ &\leq \frac{2}{n} \|f\|_\infty \sum_{r=1}^N \sum_{x=(x_1, \dots, x_d)^\top \in P_r} \{f(x_1, \dots, x_d) - f(x_1 - n_1^{-1}, \dots, x_d - n_1^{-1})\} \\ &\leq \frac{2N}{n} \|f\|_\infty (f(1, \dots, 1) - f(0, \dots, 0)) \leq 4dn^{-1/d} \|f\|_\infty^2, \end{aligned}$$

as desired. □

The following is a simple generalisation of Jensen's inequality.

LEMMA 2. *Suppose $h : [0, \infty) \rightarrow (0, \infty)$ 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 $H : [0, \infty) \rightarrow [h(0), \infty)$ 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. \square

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 2.

LEMMA 3. *There exist universal constants $c > 0$ and $\varepsilon_0 > 0$ such that*

$$\log N(\varepsilon_0, \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1), \|\cdot\|_2) \geq c \log^2 n.$$

PROOF. It suffices to prove the equivalent result that there exist universal constants $c, \varepsilon_0 > 0$ such that the packing number $D(\varepsilon_0, \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1), \|\cdot\|_2)$ (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 $r = 1, \dots, \ell$, let $I_r := n_1^{-1} \{2^{r-1}, \dots, 2^r - 1\}$ and consider the set

$$\begin{aligned} \bar{\mathcal{M}} := \left\{ \theta \in \mathbb{R}^{\mathbb{L}_{2,n}} : \theta_{I_r \times I_s} \in \left\{ \frac{-\mathbf{1}_{I_r \times I_s}}{\sqrt{2^{r+s+1}} \log n}, \frac{-\mathbf{1}_{I_r \times I_s}}{\sqrt{2^{r+s}} \log n} \right\} \right\} \\ \subseteq \mathcal{M}(\mathbb{L}_{2,n}) \cap B_2(1), \end{aligned}$$

where $\mathbf{1}_{I_r \times I_s}$ denotes the all-one vector on $I_r \times I_s$. Define a bijection $\psi : \bar{\mathcal{M}} \rightarrow \{0, 1\}^{\ell^2}$ by

$$\psi(\theta) := \left(\mathbb{1}_{\{\theta_{I_r \times I_s} = -\mathbf{1}_{I_r \times I_s} / \sqrt{2^{r+s+1}} \log n\}} \right)_{r,s=1}^\ell.$$

Then, for $\theta, \theta' \in \bar{\mathcal{M}}$,

$$\|\theta - \theta'\|_2^2 = \frac{d_{\text{H}}(\psi(\theta), \psi(\theta'))}{\log^2 n} \frac{1}{4} \left(1 - \frac{1}{2^{1/2}} \right)^2,$$

where $d_H(\cdot, \cdot)$ denotes the Hamming distance. On the other hand, by the Gilbert–Varshamov inequality (e.g. [Massart, 2007](#), Lemma 4.7), there exists a subset $\mathcal{I} \subseteq \{0, 1\}^{\ell^2}$ such that $|\mathcal{I}| \geq \exp(\ell^2/8)$ and $d_H(v, v') \geq \ell^2/4$ 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. \square

Lemma 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, \dots, X_n .

LEMMA 4. *Let $d \geq 2$. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P$, 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 $1 - e^{-ed^{-1}(M_0n)^{1/d} \log(M_0n)}$, there is an antichain in G_X with cardinality at least $n^{1-1/d}/(2eM_0^{1/d})$.*

PROOF. By Dilworth’s Theorem ([Dilworth, 1950](#)), for each realisation of the directed acyclic graph G_X , there exists a covering of $V(G_X)$ by chains $\mathcal{C}_1, \dots, \mathcal{C}_M$, 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 $k := \lceil e(M_0n)^{1/d} \rceil \leq 2e(M_0n)^{1/d}$. By a union bound, we have that

$$\begin{aligned} \mathbb{P}(\exists \text{ a chain of length } k \text{ in } G_X) &\leq \frac{n!}{(n-k)!} \mathbb{P}(X_1 \preceq \dots \preceq X_k) \\ &\leq \binom{n}{k} (k!)^{-(d-1)} M_0^k \leq \left(\frac{en}{k}\right)^k \left(\frac{k}{e}\right)^{-k(d-1)} M_0^k \\ &\leq (M_0n)^{-k/d} \leq e^{-ed^{-1}(M_0n)^{1/d} \log(M_0n)}, \end{aligned}$$

as desired. \square

The following two lemmas control the empirical processes in (18) and (19) in the main text by the symmetrised empirical process in (20) in the main text.

LEMMA 5. *Let $n \geq 2$, and suppose that $X_1, \dots, X_n, \tilde{\epsilon}_1, \dots, \tilde{\epsilon}_n$ are independent, with X_1, \dots, X_n identically distributed on \mathcal{X} and $\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_n$ identically distributed, with $|\tilde{\epsilon}_1|$ stochastically dominated by $|\epsilon_1|$. Then for any countable class \mathcal{F} of measurable, real-valued functions defined on \mathcal{X} , we have*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \tilde{\epsilon}_i f(X_i) \right| \leq 2 \log^{1/2} n \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \xi_i f(X_i) \right|.$$

PROOF. Let $\alpha_0 := 0$, and for $k = 1, \dots, n$, let $\alpha_k := \mathbb{E}|\tilde{\epsilon}_{(k)}|$, where $|\tilde{\epsilon}_{(1)}| \leq \dots \leq |\tilde{\epsilon}_{(n)}|$ are the order statistics of $\{|\tilde{\epsilon}_1|, \dots, |\tilde{\epsilon}_n|\}$, so that $\alpha_n \leq (2 \log n)^{1/2}$. Observe that for any $k = 1, \dots, n$,

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^k \xi_i f(X_i) \right| &= \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^k \xi_i f(X_i) + \mathbb{E} \sum_{i=k+1}^n \xi_i f(X_i) \right| \\
&\leq \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{E} \left\{ \left| \sum_{i=1}^n \xi_i f(X_i) \right| \middle| X_1, \dots, X_k, \xi_1, \dots, \xi_k \right\} \\
(12) \quad &\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \xi_i f(X_i) \right|.
\end{aligned}$$

We deduce from [Han and Wellner \(2017, Proposition 5\)](#) and (12) that

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \tilde{\epsilon}_i f(X_i) \right| &\leq 2^{1/2} \sum_{k=1}^n (\alpha_{n+1-k} - \alpha_{n-k}) \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^k \xi_i f(X_i) \right| \\
&\leq 2^{1/2} \alpha_n \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \xi_i f(X_i) \right|,
\end{aligned}$$

as required. \square

LEMMA 6. *Let X_1, \dots, X_n be random variables taking values in \mathcal{X} and \mathcal{F} be a countable class of measurable functions $f : \mathcal{X} \rightarrow [-1, 1]$. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \xi_i f^2(X_i) \right| \leq 4 \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \xi_i f(X_i) \right|.$$

PROOF. By [Ledoux and Talagrand \(2013, Theorem 4.12\)](#), applied to $\phi_i(y) = y^2/2$ for $i = 1, \dots, n$ (note that $y \mapsto y^2/2$ is a contraction on $[0, 1]$), we have

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \xi_i f^2(X_i) \right| &= \mathbb{E} \left\{ \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \xi_i f^2(X_i) \right| \middle| X_1, \dots, X_n \right\} \\
&\leq 4 \mathbb{E} \left\{ \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \xi_i f(X_i) \right| \middle| X_1, \dots, X_n \right\} = 4 \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \xi_i f(X_i) \right|,
\end{aligned}$$

as required. \square

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 7. *Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P$ 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*

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{G}} |\mathbb{G}_n f| &\lesssim n^{1/2} \eta + \int_\eta^r \log_+^{1/2} N_{[]}(\varepsilon, \mathcal{G}, \|\cdot\|_{L_2(P)}) \, d\varepsilon \\ &\quad + \frac{1}{n^{1/2}} \log_+ N_{[]} (r, \mathcal{G}, \|\cdot\|_{L_2(P)}). \end{aligned}$$

The above inequality also holds if we replace $\mathbb{G}_n f$ with the symmetrised empirical process $n^{-1/2} \sum_{i=1}^n \xi_i f(X_i)$.

PROOF. Writing $N_r := N_{[]} (r, \mathcal{G}, \|\cdot\|_{L_2(P)})$, there exists $\{(f_\ell^L, f_\ell^U) : \ell = 1, \dots, N_r\}$ 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 $\mathcal{G}_\ell := \{f \in \mathcal{G} : f_\ell^L \leq f \leq f_\ell^U\} \setminus \cup_{j=1}^{\ell-1} \mathcal{G}_j$ 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}_ℓ is at most r . It follows by [van der Vaart and Wellner \(1996, Lemma 2.14.3\)](#) that for any choice of $f_\ell \in \mathcal{G}_\ell$, we have that

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{G}} |\mathbb{G}_n f| &\lesssim n^{1/2} \eta + \int_\eta^r \log_+^{1/2} N_{[]}(\varepsilon, \mathcal{G}, \|\cdot\|_{L_2(P)}) \, d\varepsilon \\ (13) \quad &\quad + \mathbb{E} \max_{\ell=1, \dots, N_r} |\mathbb{G}_n f_\ell| + \mathbb{E} \max_{\ell=1, \dots, N_r} \left| \mathbb{G}_n \left(\sup_{f \in \mathcal{G}_\ell} |f - f_\ell| \right) \right|. \end{aligned}$$

The third and fourth terms of (13) can be controlled by Bernstein's inequality (in the form of (2.5.5) in [van der Vaart and Wellner \(1996\)](#)):

$$\mathbb{E} \max_{\ell=1, \dots, N_r} |\mathbb{G}_n f_\ell| \vee \mathbb{E} \max_{\ell=1, \dots, N_r} \left| \mathbb{G}_n \left(\sup_{f \in \mathcal{G}_\ell} |f - f_\ell| \right) \right| \lesssim \frac{\log_+ N_r}{n^{1/2}} + r \log_+^{1/2} N_r.$$

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 (13), 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 $e \otimes f : \{-1, 1\} \times \mathcal{X} \rightarrow \mathbb{R}$ by $(e \otimes f)(t, x) := tf(x)$, and apply the previous result to the function class $e \otimes \mathcal{G} := \{e \otimes f : f \in \mathcal{G}\} \subseteq B_2(r, P_\xi \otimes P) \cap B_\infty(1)$, where P_ξ denotes the Rademacher distribution on $\{-1, 1\}$. Here the

randomness is induced by the independently and identically distributed pairs $(\xi_i, X_i)_{i=1}^n$. 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 $L_2(P_\xi \otimes P)$ metric, where $e_+(t) := \max\{e(t), 0\} = \max(t, 0)$ and $e_-(t) := \max(-t, 0)$. It follows that for every $\varepsilon > 0$,

$$N_{[\cdot]}(\varepsilon, e \otimes \mathcal{G}, L_2(P_\xi \otimes P)) \leq N_{[\cdot]}(\varepsilon, \mathcal{G}, L_2(P)),$$

which proves the claim for the symmetrised empirical process. \square

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 8, let $\mathcal{F}_{d,\downarrow}^+ = \{f : -f \in \mathcal{F}_d, f \geq 0\}$. The following result is used to control the bracketing entropy terms that appear in Lemma 7 when we apply it in the proof of Proposition 8.

LEMMA 8. *There exists a constant $C_d > 0$, depending only on d , such that for any $r, \varepsilon > 0$,*

$$\begin{aligned} & \log N_{[\cdot]}(\varepsilon, \mathcal{F}_{d,\downarrow}^+ \cap B_2(r, P) \cap B_\infty(1), \|\cdot\|_{L_2(P)}) \\ & \leq C_d \begin{cases} (r/\varepsilon)^2 \frac{M_0}{m_0} \log^2(\frac{M_0}{m_0}) \log^4(1/\varepsilon) \log_+^2(\frac{r \log_+(1/\varepsilon)}{\varepsilon}) & \text{if } d = 2, \\ (r/\varepsilon)^{2(d-1)} (\frac{M_0}{m_0})^{d-1} \log_+^{d^2}(1/\varepsilon) & \text{if } d \geq 3. \end{cases} \end{aligned}$$

PROOF. We first claim that for any $\eta \in (0, 1/4]$,

$$(14) \quad \begin{aligned} & \log N_{[\cdot]}(\varepsilon, \mathcal{F}_{d,\downarrow}^+ \cap B_2(r, P), \|\cdot\|_{L_2(P; [\eta, 1]^d)}) \\ & \lesssim_d \begin{cases} (\frac{r}{\varepsilon})^2 \frac{M_0}{m_0} \log^2(\frac{M_0}{m_0}) \log^4(1/\eta) \log_+^2(\frac{r \log(1/\eta)}{\varepsilon}) & \text{if } d = 2, \\ (\frac{r}{\varepsilon})^{2(d-1)} (\frac{M_0}{m_0})^{d-1} \log^{d^2}(1/\eta) & \text{if } d \geq 3. \end{cases} \end{aligned}$$

By the cone property of $\mathcal{F}_{d,\downarrow}^+$, it suffices to establish the above claim when $r = 1$. We denote by $\text{vol}(S)$ the d -dimensional Lebesgue measure of a measurable set $S \subseteq [0, 1]^d$. By Gao and Wellner (2007, Theorem 1.1) and a scaling argument, we have for any $\delta, M > 0$ and any hyperrectangle $A \subseteq [0, 1]^d$ that

$$(15) \quad \log N_{[\cdot]}(\delta, \mathcal{F}_{d,\downarrow}^+ \cap B_\infty(M), \|\cdot\|_{L_2(P; A)}) \lesssim_d \begin{cases} (\gamma/\delta)^2 \log_+^2(\gamma/\delta) & \text{if } d = 2, \\ (\gamma/\delta)^{2(d-1)} & \text{if } d \geq 3, \end{cases}$$

where $\gamma := M_0^{1/2} M \text{vol}^{1/2}(A)$. 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 $\ell = 0, \dots, m$. Then for $\ell_1, \dots, \ell_d \in \{0, \dots, m\}$,

any $f \in \mathcal{F}_{d,\downarrow}^+ \cap B_2(1, P)$ is uniformly bounded by $\{m_0 \prod_{j=1}^d (2^{\ell_j} \eta)\}^{-1/2}$ on the hyperrectangle $\prod_{j=1}^d I_{\ell_j}$. Then by (15) we see that for any $\delta > 0$,

$$\begin{aligned} \log N_{[\cdot]}(\delta, \mathcal{F}_{d,\downarrow}^+ \cap B_2(1, P), \|\cdot\|_{L_2(P; \prod_{j=1}^d I_{\ell_j})}) \\ \lesssim_d \begin{cases} \delta^{-2} (M_0/m_0) \log^2(\frac{M_0}{m_0}) \log_+^2(1/\delta) & \text{if } d = 2, \\ \delta^{-2(d-1)} (M_0/m_0)^{d-1} & \text{if } d \geq 3, \end{cases} \end{aligned}$$

where we have used the fact that $\log_+(ax) \leq 2 \log_+(a) \log_+(x)$ for any $a, x > 0$. Note that these bounds do not depend on η , since the dependence of M and $\text{vol}(A)$ 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 (14) 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

$$\begin{aligned} \log N_{[\cdot]}(\varepsilon, \mathcal{F}_{d,\downarrow}^+ \cap B_2(r, P) \cap B_\infty(1), \|\cdot\|_{L_2(P)}) \\ \leq \log N_{[\cdot]}(\varepsilon/2, \mathcal{F}_{d,\downarrow}^+ \cap B_2(r, P), \|\cdot\|_{L_2(P; [\eta, 1]^d)}) \\ \lesssim_d \begin{cases} (r/\varepsilon)^2 \frac{M_0}{m_0} \log^2(\frac{M_0}{m_0}) \log_+^4(1/\varepsilon) \log_+^2(\frac{r \log_+(1/\varepsilon)}{\varepsilon}) & \text{if } d = 2, \\ (r/\varepsilon)^{2(d-1)} (\frac{M_0}{m_0})^{d-1} \log_+^{d^2}(1/\varepsilon) & \text{if } d \geq 3, \end{cases} \end{aligned}$$

completing the proof. \square

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 9. *For any $r \in (0, 1]$ and $\ell = 1, \dots, n_1$, we have*

$$\int_{I_\ell} F_r^2 dP \leq \frac{7M_0 r^2 \log_+^d(1/r^2)}{m_0 \ell}.$$

PROOF. By monotonicity and the $L_2(P)$ and L_∞ constraints, we have $F_r^2(x_1, \dots, x_d) \leq \frac{r^2}{m_0 x_1 \cdots x_d} \wedge 1$. We first claim that for any $d \in \mathbb{N}$,

$$\int_{[0, 1]^d} \left(\frac{t}{x_1 \cdots x_d} \wedge 1 \right) dx_1 \cdots dx_d \leq 5t \log_+^d(1/t).$$

To see this, we define $S_d := \{(x_1, \dots, x_d) : \prod_{j=1}^d x_j \geq t\}$ and set $a_d := \int_{S_d} \frac{t}{x_1 \cdots x_d} dx_1 \cdots dx_d$ and $b_d := \int_{S_d} dx_1 \cdots dx_d$. By integrating out the last coordinate, we obtain the following relation

$$(16) \quad b_d = \int_{S_{d-1}} \left(1 - \frac{t}{x_1 \cdots x_{d-1}}\right) dx_1 \cdots dx_{d-1} = b_{d-1} - a_{d-1}.$$

On the other hand, we have by direct computation that

$$(17) \quad \begin{aligned} a_d &= \int_t^1 \cdots \int_{\frac{t}{x_1 \cdots x_{d-1}}}^1 \frac{t}{x_1 \cdots x_d} dx_d \cdots dx_1 \\ &\leq a_{d-1} \log(1/t) \leq \cdots \leq a_1 \log^{d-1}(1/t) = t \log^d(1/t). \end{aligned}$$

Combining (16) and (17), we have

$$\begin{aligned} \int_{[0,1]^d} \left(\frac{t}{x_1 \cdots x_d} \wedge 1\right) dx_1 \cdots dx_d &= a_d + 1 - b_d \\ &\leq \min\{a_d + 1, a_d + a_{d-1} + \cdots + a_1 + 1 - b_1\} \\ &\leq \min\left\{t \log^d(1/t) + 1, \frac{t \log^{d+1}(1/t)}{\log(1/t) - 1}\right\} \leq 5t \log_+^d(1/t), \end{aligned}$$

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 $\ell = 2, \dots, n_1$, we have that

$$\begin{aligned} \int_{I_\ell} F_r^2 dP &\leq \frac{M_0}{m_0} \int_{(\ell-1)/n_1}^{\ell/n_1} \int_{[0,1]^{d-1}} \left(\frac{r^2/x_d}{x_1 \cdots x_{d-1}} \wedge 1\right) dx_1 \cdots dx_{d-1} dx_d \\ &\leq \frac{M_0}{m_0} \int_{(\ell-1)/n_1}^{\ell/n_1} 5(r^2/x_d) \log_+^{d-1}(x_d/r^2) dx_d \\ &\leq \frac{M_0}{m_0} 5r^2 \log_+^{d-1}(1/r^2) \log(\ell/(\ell-1)) \leq \frac{7M_0 r^2 \log_+^{d-1}(1/r^2)}{m_0 \ell}, \end{aligned}$$

as desired. For the remaining case $\ell = 1$, we have

$$\int_{I_1} F_r^2 dP \leq M_0 \int_{[0,1]^d} F_r^2 dx_1 \cdots dx_d \leq \frac{5M_0}{m_0} r^2 \log_+^d(1/r^2),$$

which is also of the correct form. \square

LEMMA 10. *For any Borel measurable $f_0 : [0, 1]^d \rightarrow [-1, 1]$ and any $a > 2$, we have $\mathbb{P}(\|\hat{f}_n - f_0\|_\infty > a) \leq ne^{-(a-2)^2/2}$. Consequently,*

$$\mathbb{E}\left(\|\hat{f}_n - f_0\|_\infty^2 \mathbb{1}_{\{\|\hat{f}_n - f_0\|_\infty > a\}}\right) \leq n(a^2 + 2 + 2\sqrt{2\pi})e^{-(a-2)^2/2}.$$

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 (Robertson, Wright and Dykstra, 1988, 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

$$\|\hat{f}_n\|_\infty = \max_{1 \leq i \leq n} |\hat{f}_n(X_i)| \leq \max_{1 \leq i \leq n} |Y_i|.$$

Since $Y_i = f_0(X_i) + \epsilon_i$ and $\|f_0\|_\infty \leq 1$, we have by a union bound that

$$\mathbb{P}(\|\hat{f}_n - f_0\|_\infty \geq t) \leq n\mathbb{P}(|\epsilon_1| \geq t - 2).$$

The first claim follows using the fact that $\mathbb{P}(\epsilon_1 \geq t) \leq \frac{1}{2}e^{-t^2/2}$ for any $t \geq 0$. Moreover, for any $a > 2$,

$$\begin{aligned} \mathbb{E}\left(\|\hat{f}_n - f_0\|_\infty^2 \mathbb{1}_{\{\|\hat{f}_n - f_0\|_\infty > a\}}\right) &= \int_0^\infty 2t \mathbb{P}(\|\hat{f}_n - f_0\|_\infty \geq \max\{a, t\}) dt \\ &\leq na^2 \mathbb{P}(|\epsilon_1| \geq a - 2) + n \int_a^\infty 2t \mathbb{P}(|\epsilon_1| \geq t - 2) dt \\ &\leq n(a^2 + 2 + 2\sqrt{2\pi})e^{-(a-2)^2/2}, \end{aligned}$$

as desired. \square

LEMMA 11. *If Y is a non-negative random variable such that $(\mathbb{E}Y^p)^{1/p} \leq A_1p + A_2p^{1/2} + A_3$ for all $p \in [1, \infty)$ and some $A_1, A_2 > 0$, $A_3 \geq 0$, then for every $t \geq 0$,*

$$\mathbb{P}(Y \geq t + eA_3) \leq e \exp\left(-\min\left\{\frac{t}{2eA_1}, \frac{t^2}{4e^2A_2^2}\right\}\right).$$

PROOF. Let $s := \min\{t/(2eA_1), t^2/(2eA_2)^2\}$. For values of t such that $s \geq 1$, we have by Markov's inequality that

$$\mathbb{P}(Y \geq t + eA_3) \leq \left(\frac{A_1s + A_2s^{1/2} + A_3}{t + eA_3}\right)^s \leq e^{-s} \leq e^{1-s}.$$

For values of t such that $s < 1$, we trivially have $\mathbb{P}(Y \geq t + eA_3) \leq \mathbb{P}(Y \geq t) \leq e^{1-s}$, as desired. \square

LEMMA 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. \square

REFERENCES

- Dilworth, R. P. (1950) A decomposition theorem for partially ordered sets. *Ann. Math.*, **51**, 161–166.
- Gao, F. and Wellner, J. A. (2007) Entropy estimate for high-dimensional monotonic functions. *J. Mult. Anal.*, **98**, 1751–1764.
- Giné, E., Latała, R and Zinn, J. (2000) Exponential and moment inequalities for U -statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pp. 13–38. Birkhäuser Boston, Boston, MA, 2000.
- Han, Q., Wang, T., Chatterjee, S. and Samworth, R. J. (2018) Isotonic regression in general dimensions. *Ann. Statist.*, to appear.
- Han, Q. and Wellner, J. A. (2017) Convergence rates of least squares regression estimators with heavy-tailed errors. *Ann. Statist.*, to appear. Available at arXiv:1706.02410.
- Ledoux, M. and Talagrand, M. (2013) *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.
- Massart, P. (2000) About the constant in Talagrand’s concentration inequalities for empirical processes. *Ann. Probab.*, **28**, 863–884.
- Massart, P. (2007) *Concentration Inequalities and Model Selection*, Springer, Berlin.
- Robertson, T., Wright, F. T. and Dykstra, R. L. (1988) *Order Restricted Statistical Inference*. John Wiley & Sons, Ltd., Chichester.
- Talagrand, M. (1996) New concentration inequalities in product spaces. *Invent. Math.*, **126**, 505–563.
- van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. Springer, New York.

DEPARTMENT OF STATISTICS, BOX 354322
UNIVERSITY OF WASHINGTON
SEATTLE, WA 98195
UNITED STATES
E-MAIL: royhan@uw.edu

STATISTICAL LABORATORY
WILBERFORCE ROAD
CAMBRIDGE, CB3 0WB
UNITED KINGDOM
E-MAIL: t.wang@statslab.cam.ac.uk
E-MAIL: r.samworth@statslab.cam.ac.uk

ILLINI HALL
725 S WRIGHT ST #101
CHAMPAIGN, IL 61820
UNITED STATES
E-MAIL: sc1706@illinois.edu