On the Minimal Error of Empirical Risk Minimization

Gil Kur
MIT

Alexander Rakhlin
MIT

Abstract

We study the minimal error of the Empirical Risk Minimization (ERM) procedure in the task of regression, both in the random and the fixed design settings. Our sharp lower bounds shed light on the possibility (or impossibility) of adapting to simplicity of the model generating the data. In the fixed design setting, we show that the error is governed by the global complexity of the entire class. In contrast, in random design, ERM may only adapt to simpler models if the local neighborhoods around the regression function are nearly as complex as the class itself, a somewhat counter-intuitive conclusion. We provide sharp lower bounds for performance of ERM for both Donsker and non-Donsker classes. We also discuss our results through the lens of recent studies on interpolation in overparameterized models.

1 Introduction

An increasing number of machine learning applications employ flexible overparameterized models to fit the training data. Theoretical analysis of such 'overfitted' solutions has been a recent focus of the learning community. It is conjectured that the use of large overparameterized neural networks makes the loss landscape amenable to optimization through local search methods, such as stochastic gradient descent. It is also hypothesized that implicit regularization, arising from the choice of the optimization algorithm and the neural network architecture, mitigates the large complexity and ensures that the 'overfitted' solutions generalize.

Suppose a 'simple' class $\mathcal{H}$ of models captures the relationship between the covariates X and the response variable Y. Inspired by the use of overparameterized models, we may take a much larger class $\mathcal{F} \supset \mathcal{H}$ for computational or other purposes (such as lack of explicit description of $\mathcal{H}$) and minimize training loss over this larger class. It is natural to ask whether the learning procedure can *adapt* to the fact that data comes from a simple model $f_0 \in \mathcal{H}$, in the sense that the prediction error depends on the statistical complexity of $\mathcal{H}$ rather than $\mathcal{F}$. We do have positive examples of this type: the least squares solution (that is, empirical risk minimization with respect to square loss) over the class of all convex functions $\mathcal{F}$ on a convex compact subset of $\mathbb{R}^d$ (with $d \le 4$) automatically enjoys the faster "parametric" rate $\tilde{O}(k/n)$ of convergence to the true regression function $f_0 \in \mathcal{H}$ if f_0 is a piece-wise linear convex function with k pieces. This rate should be contrasted with the slow non-parametric rate $\Theta(n^{-4/(d+4)})$ when the true regression function is 'complex' and cannot be approximated well by a piece-wise linear convex function.

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

This paper studies the fundamental limits of adaptivitiy of empirical risk minimization (ERM) in the setting of nonparametric regression (or, prediction with square loss and a well-specified model), in both random and fixed design. In contrast with the standard minimax approach to lower bounds, which may hide the true performance of ERM on simple models, we focus on lower bounds that hold for *any* (rather than the worst-case) regression function in a given class. In the fixed design setting, we show that—informally speaking—for rich classes $\mathcal{F}$, dependence on the global statistical complexity of the class is unavoidable, as it controls the error of ERM for any true regression function f_0, no matter how 'simple' it is. In contrast, in the random design case, the situation is more subtle. Somewhat counter-intuitively, we show that for rich classes $\mathcal{F}$, adaptation to the simplicity of f_0 may only be possible if the local neighborhood of f_0 in $\mathcal{F}$ is

nearly as rich as the class $\mathcal{F}$. This finding can be viewed through the lens of recent results on interpolation (Belkin et al., 2019, 2018; Bartlett et al., 2020; Liang et al., 2020b). In these papers, the solutions can be seen as 'simple-plus-spiky' (Wyner et al., 2017) with spikes responsible for fitting the training data without affecting the error with respect to the population. Since in these models there are enough degrees of freedom to fit any noisy data, the effective function classes have rich local neighborhoods. In such cases, it is still possible that 'overfitting' to the training data does not result it large out-of-sample error. Conversely, we show that—again, informally speaking—if f_0 is embedded in a local neighborhood in $\mathcal{F}$ with low complexity, the empirical minimizer will necessarily be attracted to a solution far away from f_0 with respect to the out-of-sample loss. This finding initially appeared counter-intuitive to the authors.

2 Formal Model

We now present the formal model. Let $\mathcal{F}$ be a *convex* class of real-valued functions on some domain $\mathcal{X}$. We aim to recover $f_0 \in \mathcal{F}$ based on n samples $Y_i = f_0(X_i) + \xi_i$, $i = 1, \ldots, n$, under the assumption $f_0 \in \mathcal{F}$ and $\xi_1, \ldots, \xi_n \overset{i.i.d.}{\sim} N(0,1)$. In the random design setting, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$, where $\mathbb{P}$ is some unknown distribution on $\mathcal{X}$, while in the fixed design setting $X_1, \ldots, X_n$ are some fixed points in $\mathcal{X}$.

The Least Squares Estimator, or ERM with respect to square loss, is defined as

$$\hat{f}_n = \Psi \left(\underset{f \in \mathcal{F}}{\operatorname{argmin}} \ \sum_{i=1}^{n} (Y_i - f(X_i))^2 \right), \tag{1}$$

where Ψ is a function that selects a particular solution in the set of possible minimizers (for example, a minimal norm solution).

One of the most important questions regarding ERM is its statistical performance as compared to other estimators, defined as maps from $\{(X_i, Y_i)\}_{i=1}^{n}$ to $\mathcal{F}$ (or to $\mathbb{R}^{\mathcal{X}}$ for *improper* methods). While there are multiple ways of measuring the statistical performance, perhaps the most popular is the *minimax risk* (Tsybakov, 2003), defined in the random design case for any estimator $\bar{f}_n$ as

$$\mathcal{R}(\bar{f}_n, \mathcal{F}, \mathbb{P}) := \sup_{f_0 \in \mathcal{F}} \mathbb{E}_{x,\xi} \int (\bar{f}_n((X_1, Y_1), \ldots, (X_n, Y_n)) - f_0)^2 d\mathbb{P},$$

where $\mathbb{E}_{x,\xi}$ denotes expectation over the training data and the integral represents the expected out-of-sample performance with respect to $\mathbb{P}$. One can also write this measure of performance as excess square loss

$$\sup_{f_0 \in \mathcal{F}} \mathbb{E}_{x,\xi} \mathbb{E}_{(X,Y)} (\bar{f}_n(X) - Y)^2 - \mathbb{E}_{(X,Y)} (f_0(X) - Y)^2.$$

We say that the ERM $\hat{f}_n$ is *minimax optimal*, if for for all $n \geq 0$,

$$\mathcal{R}(\hat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim \inf_{\bar{f}_n} \mathcal{R}(\bar{f}_n, \mathcal{F}, \mathbb{P}),$$

where $\lesssim$ denotes less or equal up to a constant that only depends on $\mathbb{P}, \mathcal{F}$. The quantity $\inf_{\bar{f}_n} \mathcal{R}(\bar{f}_n, \mathcal{F}, \mathbb{P})$ is known as the *minimax rate* for $(\mathcal{F}, \mathbb{P})$. In the fixed design setting, the risk measure is defined in an analogous way, except that instead of drawing n i.i.d. points from $\mathbb{P}$, we consider a sequence of measures that are supported uniformly on n points.

Clearly, the definitions of the risk and the minimax optimality measure "the worst case scenario" of a given estimator, and may hide the true statistical performance of the ERM in real-life applications (cf. (Bellec, 2017)). For example, as mentioned in the introduction, if f_0 is known to belong to a smaller class $\mathcal{H}$, the relevant quantity is

$$\mathcal{R}_{\mathcal{H}}(\hat{f}_n, \mathcal{F}, \mathbb{P}) := \sup_{f_0 \in \mathcal{H}} \mathbb{E} \int (\hat{f}_n - f_0)^2 d\mathbb{P},$$

which may be significantly smaller than the minimax risk. We remark that the ERM $\hat{f}_n$ is still defined over $\mathcal{F}$, due to computational or other considerations. As an example, consider linear regression in $\mathbb{R}^d$ when the true coefficient vector is sparse, i.e. supported on $k \ll d$ coordinates. Then, due to computational considerations, it is standard to replace the original problem of minimizing square loss over sparse vectors in $\mathbb{R}^d$ by minimization over a larger ℓ_1 ball in $\mathbb{R}^d$ (the Lasso procedure).

The second example was already briefly mentioned in the introduction, and we expand on it here. Let $\mathcal{F}_d$ be the family of convex 1-Lipschitz functions on $\mathcal{X} = [0,1]^d$, and let $\mathbb{P} = \mathrm{Unif}(\mathcal{X})$. The subset $\mathcal{H}_d$ (of 'simple functions') is the set of 1-Lipschitz k-affine piece-wise linear functions with $k = \Theta(1)$. It is well known that ERM over $\mathcal{H}_d$ is NP-hard since the problem is highly non-convex; moreover, even estimating the number of pieces is computationally hard (cf. the recent paper Ghosh et al. (2019) for more details). In contrast, ERM over $\mathcal{F}_d$ can be efficiently computed (Ghosh et al., 2019). While the minimax rate for $(\mathcal{F}_d, \mathbb{P}_d)$ is $\Theta(n^{-4/(d+4)})$ (Dudley, 1999; Bronshtein, 1976), it was proved recently in (Kur et al., 2020b) that the risk of ERM is $\tilde{\Theta}_d(\max\{n^{-2/d}, n^{-4/(d+4)}\})$, which is minimax-suboptimal when $d \geq 5$. Furthermore, it was shown in (Han and Wellner, 2016; Feng et al., 2018) that

$$\mathcal{R}_{\mathcal{H}_d}(\hat{f}_n, \mathcal{F}_d, \mathbb{P}_d) \lesssim \tilde{O}(\max\{n^{-4/d}, n^{-1}\}), \tag{2}$$

which is *significantly* smaller than both the risk of ERM and the minimax rate. When the ERM (or MLE) satisfies such improved bounds, we say that it exhibits *adaptation* (cf. (Feng et al., 2018; Kim et al., 2018; Samworth, 2018; Han et al., 2019; Kur et al., 2020b)).

In this paper we answer the two following questions: Does there exist a *uniform* lower bound on the *minimal error*

$$\inf_{f_0 \in \mathcal{F}} \mathbb{E}_{x,\xi} \int (\hat{f}_n - f_0)^2 d\mathbb{Q}$$

of ERM $\hat{f}_n$, where $\mathbb{Q}$ is either fixed or random design measure? Does the richness of the entire class $\mathcal{F}$ affect the minimal error, or is there a more refined notion of complexity that governs its behavior?

3 Main Results

We start with definitions. For n points $\mathbf{x}_n := \{x_1, \ldots, x_n\}$ in $\mathcal{X}$ and $\mathcal{G} \subseteq \mathcal{F}$, we define the Gaussian averages of $\mathcal{G}$ as

$$\widehat{\mathcal{W}}(\mathcal{G}) := \mathbb{E}_\xi \sup_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i f(x_i), \quad \mathcal{W}(\mathcal{G}) := \mathbb{E}\widehat{\mathcal{W}}(\mathcal{G}).$$

For a measure $\mathbb{Q}$ on $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$ we denote by $\|f\|_\mathbb{Q}$ the $L_2(\mathbb{Q})$ norm of f. Finally, for any $\mathbb{Q}$, $f \in \mathcal{F}$, and $r \geq 0$ we denote by $B_\mathbb{Q}(f, r) := \{g \in \mathcal{F} : \|g - f\|_\mathbb{Q} \leq r\}$, the intersection of the $L_2(\mathbb{Q})$ ball around f and the class $\mathcal{F}$.

3.1 Fixed Design

We now state our sharp lower bound for the fixed design error, for simplicity of exposition under the assumption of uniform boundedness of $\mathcal{F}$ (the general statement is given below in Lemma 3.1).

Corollary 3.1. Let $\mathbb{P}_n$ be the empirical measure on some n points in $\mathcal{X}$, and assume $\mathcal{F} \subseteq [-1, 1]^\mathcal{X}$ is convex. Then the minimal error of ERM over $\mathcal{F}$ satisfies

$$\inf_{f_0 \in \mathcal{F}} \mathbb{E}_\xi \int (\hat{f}_n - f_0)^2 d\mathbb{P}_n \geq 64^{-1}(\widehat{\mathcal{W}}(\mathcal{F}) - Cn^{-1})^2,$$

where C is some positive absolute constant.

When $\mathcal{F}$ is uniformly bounded (say, by 1), a classical result in non-parametric statistics (van de Geer, 2000) and our theorem imply that

$$64^{-1}(\widehat{\mathcal{W}}(\mathcal{F}) - Cn^{-1})^2 \leq \inf_{f_0 \in \mathcal{F}} \mathbb{E}_\xi \int (\hat{f}_n - f_0)^2 d\mathbb{P}_n \leq \underbrace{\sup_{f_0 \in \mathcal{F}} \mathbb{E}_\xi \int (\hat{f}_n - f_0)^2 d\mathbb{P}_n}_{=\mathcal{R}(\hat{f}_n, \mathcal{F}, \mathbb{P}_n)} \leq 2\widehat{\mathcal{W}}(\mathcal{F}).$$

Moreover, both of these bounds are tight, in the sense that they can be attained on certain families of functions, up to constants (cf. Birgé et al. (1998); Han et al. (2019)). Therefore, we conclude that in the *fixed design* case, both the minimax risk and the minimal error of the ERM depend on the entire Gaussian complexity of $\mathcal{F}$ (when it is convex and uniformly bounded). In particular, for the case of convex regression, Corollary 3.1 recovers the rate in (2) (up to logarithmic factors) for the fixed design case, since with high probability the global complexity $\widehat{\mathcal{W}}(\mathcal{F})$ is of the order $\max\{n^{-2/d}, n^{-1/2}\}$.

3.2   Random Design

We now turn to the random design setting, which is significantly more subtle. Before stating the result, we describe a direct proof strategy that fails. This approach would attempt to pass from the fixed design lower bound to the random design lower bound by relating the population and empirical norms $\|\cdot\|_\mathbb{P}$ and $\|\cdot\|_{\mathbb{P}_n}$, uniformly over the class. A statement of this type (which may be called "upper isometry," in contrast with "lower isometry" studied, for instance, in Mendelson (2014)) could be derived under additional assumptions on the geometry of $(\mathcal{F}, \mathbb{P})$, such as a small-ball condition (Mendelson, 2014), Kolchinskii-Pollard entropy (Rakhlin et al., 2017), or an ϵ-covering with respect to the *sup*-norm van de Geer (2000). To the best of our knowledge, such upper-isometry statements can at best read

$$\|f - g\|_\mathbb{P}^2 \geq \frac{1}{2}\|f - g\|_{\mathbb{P}_n}^2 - C \cdot \mathcal{W}(\mathcal{F})^2 \quad \forall f, g \in \mathcal{F},$$

where $C \geq 1$. Since $\mathcal{W}(\mathcal{F})^2$ is larger than the lower bound on the fixed design error, this technique does not appear to work.

Moreover, a uniform lower bound of order $\mathcal{W}(\mathcal{F})^2$ in random design cannot be true in general. For instance, it was shown in a string of recent works (Liang et al., 2020a; Belkin et al., 2019; Bartlett et al., 2020; Tsigler and Bartlett, 2020) that it is possible to completely interpolate $Y_1, \ldots, Y_n$ (i.e. achieve zero empirical error) and still have a small generalization error (of order n^{-c}, for some $c \in (0, 1)$), and even be minimax optimal (with an appropriate function Ψ in Eq. (1)) . In these examples, because of the ability to interpolate any data, we know that $\mathcal{W}(B_n(f_0, 1)) = \Theta(1)$; therefore, the lower bound in the fixed design case cannot be always true in random design.

The last paragraph motivates the need to consider additional properties of the model $\mathcal{F}$ and the underlying distribution $\mathbb{P}$. With the interpolation examples in mind, we might hope that the relation between the global complexity of the class and complexity of local neighborhoods around the regression function f_0 may play a role in determining rates of convergence of ERM. To this end, for every n and $f_0 \in \mathcal{F}$, we define the following notion of complexity:

$$t_{n,\mathbb{P}}(f_0, \mathcal{F}) := \max\{t \in \mathbb{R}^+ : \mathcal{W}(B_\mathbb{P}(f_0, t)) \leq l_\xi \mathcal{W}(B_n(f_0, 1))\}, \tag{3}$$

where $l_\xi \in (0, 1)$ is a small absolute constant that will be chosen in the proofs. We remark that under the additional assumption of $\mathcal{F}$ being uniformly bounded by 1, we have that $\mathcal{W}(B_n(f_0, 1)) \leq \frac{1}{2}\mathcal{W}(B_n(f_0, 2)) = \frac{1}{2}\mathcal{W}(\mathcal{F})$, and thus we can replace the term on the right-hand side of (3) with global Gaussian averages $\mathcal{W}(\mathcal{F})$.

The quantity $t_{n,\mathbb{P}}(f_0, \mathcal{F})$ is the maximal radius of the population ball around f_0 that has Gaussian complexity comparable to that of the entire class (in the uniformly bounded case), up to some absolute constant, or to a ball of constant radius within the class. As we show next, this local richness is necessary in order to avoid the rate being dominated by the global complexity of $\mathcal{F}$. In the aforementioned interpolation examples we have both $t_{n,\mathbb{P}}(f_0, \mathcal{F}) = O(n^{-c})$ and $\mathcal{W}(B_n(f_0, 1)) = \Theta(1)$. The last two relations must be true for any $f_0 \in \mathcal{F}$ for which ERM attains perfect fit to data, and yet a small generalization error of order n^{-c}.

We now state the main result of this paper for the random design setting, under the additional assumption of $\mathcal{F}$ being uniformly bounded. Remarkably, $t_{n,\mathbb{P}}(f_0, \mathcal{F})$ is the only additional quantity that we need to consider for a general uniform lower bound on a general family $\mathcal{F}$. Specifically, we prove the following:

Theorem 3.1. *Let $\mathcal{F}$ be a convex class of functions[1] uniformly bounded by one. Then for large enough n, the minimal error of ERM over $\mathcal{F}$ is lower bounded as*

$$\inf_{f_0 \in \mathcal{F}} \frac{\mathbb{E}_{x,\xi} \int (\hat{f}_n - f_0)^2 d\mathbb{P}}{\min\{\mathcal{W}(\mathcal{F})^2, t_{n,\mathbb{P}}(f_0, \mathcal{F})^2\}} \geq c,$$

where $c \in (0,1)$.

Remark 1. *Notably, Theorem 3.1 holds under only convexity and uniform boundedness assumptions on the class $\mathcal{F}$. Furthermore, one can easily design a convex uniformly bounded family and an $f_0 \in \mathcal{F}$ such that the ERM attains an error of order $t_{n,\mathbb{P}}(f_0, \mathcal{F})^2 \ll \mathcal{W}(\mathcal{F})^2$ for all n that are large enough (for completeness see Section B.1). Therefore, under no additional assumption on $\mathcal{F}$, the above lower bound is sharp up to absolute constants.*

An almost immediate corollary of this theorem is the following key insight on the behavior of the ERM procedure in the random design setting:

Corollary 3.2. Let $\mathcal{F}$ be convex and uniformly bounded by 1. For any $f_0 \in \mathcal{F}$ such that

$$\underbrace{\mathbb{E}_{x,\xi} \int (\hat{f}_n - f_0)^2 d\mathbb{P}}_{:=E^2(f_0)} \ll \mathcal{W}(\mathcal{F})^2,$$

there *must* exists some constant $t(f_0) \leq c_1 \cdot E(f_0)$ such that

$$\mathcal{W}(B_{\mathbb{P}}(f_0, t(f_0))) = \Theta(\mathcal{W}(\mathcal{F})),$$

where $c_1 \in (0,1)$ is some absolute constant.

Informally speaking, if ERM learns some $f_0 \in \mathcal{F}$ at a rate faster than $\mathcal{W}(\mathcal{F})^2$, then the local complexity of a population ball centered at f_0 with a very small radius must be as rich as the entire complexity of $\mathcal{F}$. A more prescriptive recipe for guaranteeing such fast rates is an interesting direction of further work.

3.3   Donsker and non-Donsker Classes

The lower bounds stated thus far assumed little about the geometry of the class $\mathcal{F}$ beyond convexity and global and local Gaussian averages. Under additional assumptions on the behavior of entropy $\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbb{P})$ (defined as the logarithm of the smallest number of balls with respect to $L_2(\mathbb{P})$ of radius ϵ sufficient to cover $\mathcal{F}$) or entropy with bracketing $\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \mathbb{P})$ (defined as the logarithm of the smallest number of brackets $l_i, u_i \in \mathcal{F}$ such that $l_i \leq u_i$, $\|l_i - u_i\|_{\mathbb{P}} \leq \epsilon$ and $\mathcal{F}$ is contained in the union of the brackets), we can provide specific upper bounds on the Gaussian averages via chaining and other techniques. In particular, we say that a convex uniformly bounded $\mathcal{F}$ is $\mathbb{P}$-Donsker if $\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \mathbb{P}) \sim \epsilon^{-\alpha}$ with $\alpha \in (0, 2)$ or if $\mathcal{F}$ is parametric with $\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \mathbb{P}) \sim v \log(1/\epsilon)$ for some 'dimension' v. In seminal works of (Birgé and Massart, 1993) it was shown that for any $\mathbb{P}$-Donsker class, the ERM is minimax optimal, i.e.

$$\mathcal{R}(\hat{f}_n, \mathcal{F}, \mathbb{P}) \sim \inf_{\bar{f}_n} \mathcal{R}(\bar{f}_n, \mathcal{F}, \mathbb{P}) \sim n^{-\frac{2}{2+\alpha}}.$$

Note that for $\alpha \in (0, 2)$ we have that $\mathcal{W}(\mathcal{F}) \sim n^{-1/2}$.

The next result shows that without further assumptions we cannot learn *any* function in a convex uniformly bounded $\mathbb{P}$-Donsker class faster than a parametric rate.

---

[1]We assume that $\mathcal{F}$ is non-degenerate and contains at least two functions such that $\|f_1 - f_2\|_{\mathbb{P}} \geq 1/2$.

Corollary 3.3. Let $\mathcal{F}$ be a convex uniformly bounded $\mathbb{P}$-Donsker class, and let $X_1, \ldots, X_n \sim \mathbb{P}$. Then

$$n^{-1} \lesssim \inf_{f_0 \in \mathcal{F}} \mathbb{E} \int (\hat{f}_n - f_0)^2 d\mathbb{P}_n \sim \inf_{f_0 \in \mathcal{F}} \mathbb{E} \int (\hat{f}_n - f_0)^2 d\mathbb{P}$$

This lower bound is sharp, namely there are classical $\mathbb{P}$-Donsker classes, such as the convex regression example mentioned in the introduction and Section 2, where ERM can attain a parametric rate (up to logarithmic factors) when optimizing over all convex Lipschitz functions, but only for $d \leq 4$ which puts us in the Donsker regime.

For non-Donsker classes, i.e when $\alpha > 2$, the ERM procedure may not be optimal. One can show that

$$n^{-\frac{2}{2+\alpha}} \lesssim \mathcal{R}(\hat{f}_n, \mathcal{F}, \mathbb{P}) \lesssim n^{-\frac{1}{\alpha}}$$

and both of these bounds can be tight, up to logarithmic factors. Furthermore, one can show that

$$n^{-\frac{2}{2+\alpha}} \lesssim \mathcal{W}(\mathcal{F}) \lesssim n^{-\frac{1}{\alpha}}$$

and, again, both of these can be tight. Our next corollary shows that in this regime, the fixed-design error is at least of the order $\mathcal{W}(\mathcal{F})^2$, i.e. it is impossible to learn at a parametric rate in the non-Donsker regime.

Corollary 3.4. Let $\mathcal{F}$ be a convex uniformly bounded non-$\mathbb{P}$-Donsker class, and let $X_1, \ldots, X_n \sim \mathbb{P}$. Then the following holds:

$$n^{-\frac{4}{2+\alpha}} \lesssim \mathcal{W}(\mathcal{F})^2 \lesssim \inf_{f_0 \in \mathcal{F}} \mathbb{E} \int (\hat{f}_n - f_0)^2 d\mathbb{P}_n$$

The proof of these two corollaries appears in the appendix.

Remark 2. *Due to the geometry of general non-Donsker classes, in random design case the same lower bound may not hold. However, in all the examples in the literature (Han and Wellner, 2016; Feng et al., 2018; Kim et al., 2018; Han et al., 2019; Kur et al., 2020b) that study the adaptivity of ERM in non-Donsker families (such as convex functions when $d \geq 5$, isotonic functions when $d \geq 3$), the term of $t_{n,\mathbb{P}}(f_0, \mathcal{F})$ of Theorem 3.1 is significantly larger than $\mathcal{W}(\mathcal{F})$. As a consequence, one may use Theorem 3.1 to show that the bound in Eq. (2) is tight up to logarithmic factors.*

3.4 General Lower Bound for Fixed Design

In this section, we state the general lower bound for fixed design. In comparison to its consequence, Corollary 3.1, the version below captures complexity of local neighborhoods around regression functions that are close to f_0. Note that this lemma holds for any convex family (and not necessarily bounded).

Lemma 3.1. Let $\mathcal{F}$ be a convex family of functions and and let $x_1, \ldots, x_n \in \mathcal{X}$ be some n points, and let $\mathbb{P}_n := n^{-1} \sum_{i=1}^n \delta_{x_i}$. For all $f_0 \in \mathcal{F}$ define

$$r(f_0) := \operatorname*{argmax}_{r \geq 0} \widehat{\mathcal{W}}(B_n(f_0, r)) - \frac{r^2}{2} \tag{4}$$

and

$$L_x(f_0) := \max_{g \in B_n(f_0, 1), t \geq 0} \frac{\widehat{\mathcal{W}}(B_n(g, t)) - \widehat{\mathcal{W}}(B_n(f_0, r(f_0))) - Cn^{-1}}{\|g - f_0\|_{\mathbb{P}_n} + t}$$

where $C \in (1, \infty)$ is some absolute constant. Then the following lower bound holds:

$$\mathbb{E}_\xi \int (\hat{f}_n - f_0)^2 d\mathbb{P}_n \geq \max \{ \frac{(\widehat{\mathcal{W}}(B_n(f_0, 1)) - Cn^{-1})^2}{4}, L_x(f_0)^2 \}.$$

Note that Corollary 3.1 follows almost immediately from the last lemma. To see this, the convexity of $\mathcal{F}$, and the uniform bounded by 1 assumption imply that

$$\mathcal{W}(\mathcal{F}) = \widehat{\mathcal{W}}(B_n(f_0, 2)) \leq 2\widehat{\mathcal{W}}(B_n(f_0, 1)).$$

Remark 3. *The second term in our lower bound may be significantly larger than $\widehat{\mathcal{W}}(\mathcal{F})^2$. For example, the second term may be equal to $\widehat{\mathcal{W}}(\mathcal{F})$ in several non-Donsker families that appear in (Birgé and Massart, 1993; Kur et al., 2020a; Birgé, 2006). We also remark that constant $\frac{1}{4}$ is tight (up to $o_n(1)$).*

The rest of this paper is devoted to proofs. While the fixed design lower bound follows a rather simple argument, the corresponding lower bound in the random design case is more subtle. In particular, we employ a particular version of Talagrand's inequality that, in our particular regime, provides control on certain empirical processes, while the more commonly used versions (including Bousquet's inequality) result in vacuous estimates.

4   Proof of Lemma 3.1

Notation   Throughout this section, $c, c_1, c_2 \in (0, 1)$ and $C, C_1, C_2 \in (1, \infty)$ are some absolute constants that may change from to line to line. Also S_1, s_1, S_2, s_2 are absolute constants, but we use this notation to emphasize that we have some freedom to control their size. We also use the notation $C(c_1, C_2)$ to mean that the constant depends on c_1, C_2.

To recap, we assume that $\mathcal{F}$ is a convex family of functions, $Y_i = f_0(x_i) + \xi_i$, where $\xi_i \sim N(0, 1)$ i.i.d., $x_1, \ldots, x_n \in \mathcal{X}$, and $f_0 \in \mathcal{F}$. We write $\|f\|_n = \|f\|_{\mathbb{P}_n}$ and $\langle f, g \rangle_n = \int f g \, d\mathbb{P}_n$. With slight abuse of notation, we write $\langle \boldsymbol{\xi}, f \rangle_n = \frac{1}{n} \sum_{i=1}^n \xi_i f(x_i)$ for $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_n)$. We also abbreviate $B_n(f_0, t) := B_{\mathbb{P}_n}(f_0, t)$ to be the $L_2(\mathbb{P}_n)$ ball with respect to empirical measure $\mathbb{P}_n$.

Recall the definition of $r(f_0)$ in (4). The following lemma that was proven in (Chatterjee, 2014):

Lemma 4.1. [(Chatterjee, 2014, Thm 1.1)] The following holds under the above assumptions:

$$\Pr\left(|\|\hat{f}_n - f_0\|_n - r(f_0)| \geq t\right) \leq \begin{cases} 3\exp(-\frac{nt^2}{64}) & t \geq r(f_0) \\ 3\exp(-\frac{nt^4}{64r(f_0)^2}) & 0 \leq t \leq r(f_0) \end{cases} \tag{5}$$

Moreover, for each $t \geq 0$ the following holds

$$\Pr\left(|\langle \hat{f}_n - f_0, \boldsymbol{\xi} \rangle_n - \widehat{\mathcal{W}}(B_n(f_0, r(f_0)))| \geq t \cdot r(f_0)\right) \leq \begin{cases} 3\exp(-\frac{nt^2}{64}) & t \geq r(f_0) \\ 3\exp(-\frac{nt^4}{64r(f_0)^2}) & 0 \leq t \leq r(f_0) \end{cases} \tag{6}$$

Also, we state a simple corollary that follows from this lemma (cf. (Boucheron et al., 2013),(Chatterjee, 2014, Thm 1.2))

Corollary 4.1. The following two bounds hold

$$\mathbb{E}\left|\langle \hat{f}_n - f_0, \boldsymbol{\xi} \rangle_n - \widehat{\mathcal{W}}(B_n(f_0, r(f_0)))\right| \leq C_1 \max\{r(f_0)^{3/2} n^{-1/4}, n^{-1}\}.$$

and

$$\mathbb{E}\left|\|\hat{f}_n - f_0\|_n^2 - r(f_0)^2\right| \leq C_2 \max\{r(f_0)^{3/2} n^{-1/4}, n^{-1}\}.$$

Proof of Lemma 3.1. For brevity, denote $\hat{r} := r(f_0)$, where $r(f_0)$ is defined in Lemma 4.1. Define

$$g_\xi := \underset{h \in B_n(g, t)}{\operatorname{argmax}} \langle h - g, \boldsymbol{\xi} \rangle_n.$$

Optimality of $\hat{f}_n$ and convexity of $\mathcal{F}$ imply that $\langle \nabla_f \|f - \boldsymbol{y}\|_n^2|_{f=\hat{f}_n}, g - \hat{f}_n \rangle_n \geq 0$ for any $g \in \mathcal{F}$. In particular, for $g = g_\xi$ this implies

$$0 \geq \mathbb{E}\langle \boldsymbol{\xi} + f_0 - \hat{f}_n, g_\xi - \hat{f}_n \rangle_n,$$

where the expectation is over $\boldsymbol{\xi}$, conditionally on $x_1, \ldots, x_n$. For any $g \in \mathcal{F}$, we may write the right-hand side as

$$\mathbb{E}\langle \boldsymbol{\xi} + f_0 - \hat{f}_n, g_\xi - g + g - f_0 + f_0 - \hat{f}_n \rangle_n$$
$$= \widehat{\mathcal{W}}(B_n(g,t)) - \mathbb{E}\left[\langle \boldsymbol{\xi}, \hat{f}_n - f_0 \rangle_n + \langle \hat{f}_n - f_0, g_\xi - g \rangle_n + \langle \hat{f}_n - f_0, g - f_0 \rangle_n \right] + \mathbb{E}\|\hat{f}_n - f_0\|_n^2$$

where we used the definition of g_ξ and the fact that $\mathbb{E}\langle \boldsymbol{\xi}, g - f_0 \rangle_n = 0$. Using Corollary 4.1, we obtain a further lower bound of

$$\widehat{\mathcal{W}}(B_n(g,t)) - \widehat{\mathcal{W}}(B_n(f_0,\hat{r})) - \mathbb{E}\left[\langle \hat{f}_n - f_0, g_\xi - g \rangle_n + \langle \hat{f}_n - f_0, g - f_0 \rangle_n \right]$$
$$+ \hat{r}^2 - C\hat{r}^{3/2}n^{-1/4} - Cn^{-1}$$
$$\geq \widehat{\mathcal{W}}(B_n(g,t)) - \widehat{\mathcal{W}}(B_n(f_0,\hat{r})) - \mathbb{E}\left[\langle \hat{f}_n - f_0, g_\xi - g \rangle_n + \langle \hat{f}_n - f_0, g - f_0 \rangle_n \right] \qquad (7)$$
$$+ \hat{r}^2/2 - C_1 n^{-1}.$$

To verify the last inequality, observe that $\hat{r}^2/2 \geq C_1 \hat{r}^{3/2} n^{-1/4}$ when $\hat{r} \geq C_2 n^{-1/2}$ for C_2 that is large enough; on the other hand, if $\hat{r} \leq C_2 n^{-1/2}$, the Cn^{-1} term is dominant for C large enough. Since $\langle \hat{f}_n - f_0, g - f_0 \rangle_n \leq \|\hat{f}_n - f_0\|_n \|g - f_0\|_n$ and $\langle \hat{f}_n - f_0, g_\xi - g \rangle_n \leq t \cdot \|\hat{f}_n - f_0\|_n$, we conclude that

$$0 \geq \widehat{\mathcal{W}}(B_n(g,t)) - \widehat{\mathcal{W}}(B_n(f_0,\hat{r})) - \mathbb{E}[\|\hat{f}_n - f_0\|_{\mathbb{P}_n}](\|g - f_0\|_{\mathbb{P}_n} + t) - Cn^{-1}. \qquad (8)$$

By re-arranging the terms and using Jensen's inequality, we have

$$\mathbb{E}\|\hat{f}_n - f_0\|_n^2 \geq \left(\mathbb{E}\|\hat{f}_n - f_0\|_n\right)^2 \geq \left(\frac{\widehat{\mathcal{W}}(B_n(g,t)) - \widehat{\mathcal{W}}(B_n(f_0,\hat{r})) - Cn^{-1}}{t + \|f_0 - g\|_n}\right)_+^2$$

where $(a)_+^2 = \max\{a, 0\}^2$. Since $L_x(f_0)$ in the statement of the Lemma is non-negative, the lower bound of $L_x(f_0)^2$ follows.

Now, for the first part of the lower bound, we have to consider two cases. The first one is when $\widehat{\mathcal{W}}(B_n(f_0,\hat{r})) \geq 2^{-1}\widehat{\mathcal{W}}(B_n(f_0,1)) + \hat{r}^2/2$, and we have

$$\mathbb{E}\|\hat{f}_n - f_0\|_n \geq \mathbb{E}\|\boldsymbol{\xi}\|_n \cdot \mathbb{E}\|\hat{f}_n - f_0\|_n \geq \mathbb{E}\langle \boldsymbol{\xi}, \hat{f}_n - f_0 \rangle_n$$
$$\geq 2^{-1}\widehat{\mathcal{W}}(B_n(f_0,1)) + \hat{r}^2/2 - C\hat{r}^{3/2}n^{-1/4} \geq 2^{-1}\widehat{\mathcal{W}}(B_n(f_0,1)) - C_1 n^{-1},$$

where we used Cauchy-Schwartz inequality and Corollary 4.1. In the other case, we use Eq. (8) with $g = f_0$ and $t = 1$:

$$\mathbb{E}\|\hat{f}_n - f_0\|_n \geq \widehat{\mathcal{W}}(B_n(f_0,1)) - \widehat{\mathcal{W}}(B_n(f_0,\hat{r})) - Cn^{-1} \geq 2^{-1}\widehat{\mathcal{W}}(B_n(f_0,1)) - Cn^{-1},$$

concluding the proof. $\qquad \square$

5   Proof of Theorem 3.1

Throughout the proof of Theorem 3.1, $\mathbb{P}_n$ denotes the random empirical measure of $X_1, \ldots, X_n$. Denote by $\hat{r} := \text{argmax}\, \widehat{\mathcal{W}}(B_n(f_0,r)) - \frac{r^2}{2}$, with the hat emphasizing the dependence on $\mathbf{x}_n = (X_1, \ldots, X_n)$. We adopt the notation $\|\cdot\|_n, \langle\cdot,\cdot\rangle_n, B_n$ in the previous section for the norm and the inner product with respect to $\mathbb{P}_n$, and the $L_2(\mathbb{P}_n)$ ball. Recall that we assumed that $\mathcal{F}$ is not degenerate: $\mathcal{W}(\mathcal{F}) \geq c/\sqrt{n}$, for some $c \in (0,1)$.[2]

---

[2]See the proof of Lemma A.3 for further details

Proof of Theorem 3.1. Denote

$$t_* := \min\{t_{n,\mathbb{P}}(f_0, \mathcal{F}), s_1 \sqrt{\mathcal{W}(\mathcal{F})}, s_2 \mathcal{W}(\mathcal{F})\} \tag{9}$$

where $s_1, s_2 \in (0,1)$ are small enough absolute constants that will be defined in the proof, and $t_{n,\mathbb{P}}(f_0, \mathcal{F})$ is defined in Eq. (3).

Denote by $\mathcal{M}$ the maximal separated set with respect to $L_2(\mathbb{P})$ at scale $6\sqrt{\mathcal{W}(\mathcal{F})}$, and let

$$M = \mathcal{M}(6\sqrt{\mathcal{W}(\mathcal{F})}, \mathcal{F}, \mathbb{P}) \tag{10}$$

denote its size.

For a constant $K_1 \in (1, \infty)$, let $\mathcal{E}_1$ denote the high-probability event that is defined by the intersection of the events of Lemma A.4 and Lemma A.2:

$$\mathcal{E}_1 := \left\{ \mathbf{x}_n : \sup_{f,g \in \mathcal{F}} \left| \|f - g\|_n^2 - \|f - g\|_{\mathbb{P}}^2 \right| \leq 10\mathcal{W}(\mathcal{F}), \right.$$
$$\left. \sup_{h \in B_{\mathbb{P}}(f_0, t_*), g \in \mathcal{M}} |\langle (g - f_0), (h - f_0) \rangle_n - \mathbb{E}[(g - f_0)(h - f_0)]| \leq (8K_1)^{-1} \mathcal{W}(\mathcal{F}) \right\}. \tag{11}$$

Further, define the events

$$\mathcal{E}_2 = \left\{ \mathbf{x}_n : K_1^{-1} \mathcal{W}(\mathcal{F}) \leq \widehat{\mathcal{W}}(\mathcal{F}) \leq K_1 \mathcal{W}(\mathcal{F}) + C n^{-1/2} \right\},$$
$$\mathcal{E}_3 = \left\{ \mathbf{x}_n : \widehat{\mathcal{W}}(B_{\mathbb{P}}(f_0, t_*)) \leq K_1 \mathcal{W}(B_{\mathbb{P}}(f_0, t_*)) + C_1 \mathcal{W}(\mathcal{F})^{1/2} n^{-1/2} \right\}, \tag{12}$$
$$\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3.$$

Lemma A.3, proved in the appendix, shows that the event $\mathcal{E}$ holds with probability of at least 0.9. Note that under the event $\mathcal{E}_1$, $\mathcal{M}$ is also a $2\sqrt{\mathcal{W}(\mathcal{F})}$ separated set with respect to the random empirical measure $\mathbb{P}_n$. Hence, we may apply Sudakov's minoration (Lemma A.5) with $\epsilon = 2\sqrt{\mathcal{W}(\mathcal{F})}$ and empirical measure $\mathbb{P}_n$ defined on any $\mathbf{x}_n \in \mathcal{E}$:

$$c_1 \sqrt{4\mathcal{W}(\mathcal{F}) \cdot \frac{\log M}{n}} \leq \widehat{\mathcal{W}}(\mathcal{F}) \leq K_1 \mathcal{W}(\mathcal{F}) + C n^{-1/2} \leq C_1 K_1 \mathcal{W}(\mathcal{F}), \tag{13}$$

where in the last inequality we used the assumption that $\mathcal{W}(\mathcal{F}) \geq c \cdot n^{-1/2}$, and $C_1 \geq 0$ is defined to be large enough to satisfy the last inequality. Hence, the last equation implies that

$$M \leq \exp(C_2 K_1^2 n \mathcal{W}(\mathcal{F})). \tag{14}$$

First, recall the definition of $t_{n,\mathbb{P}}(f_0, \mathcal{F})$ where in Lemma A.2 (that appears in the supplementary) we set $l_\xi = (256K_1)^{-3}$. Recall Eq. (9), where in Lemma A.2 we set $s_1 = c(K, K_1, C_2)$, and the three constants K, K_1, C_2 follow from Sudakov's minoration lemma, Talagrand's inequality, and Adamzcak's bound. We define $s_2 := 16^{-1}(K_1)^{-1}$.

Define the event

$$\mathcal{A} = \left\{ (\boldsymbol{\xi}, \mathbf{x}_n) : \hat{f}_n \in B_{\mathbb{P}}(f_0, t_*) \right\} \tag{15}$$

and, for any $\mathbf{x}_n$, define the conditional event

$$\mathcal{A}(\mathbf{x}_n) = \left\{ \boldsymbol{\xi} : \hat{f}_n \in B_{\mathbb{P}}(f_0, t_*) \right\}. \tag{16}$$

Assume by the way of contradiction that $\Pr_{x,\xi}(\mathcal{A}) > 0.5$. Then, using the average principle (Fubini) and the fact $\Pr(\mathcal{E}) \geq 0.9$, we can find an event $\mathcal{E}_4 \subseteq \mathcal{E}$ that has a probability of at least 0.4 (when n is large enough) such that

$$\forall \mathbf{x}_n \in \mathcal{E}_4 \quad \Pr_\xi(\mathcal{A}(\mathbf{x}_n)) \geq 0.5.$$

9

Our first step is to prove that, for all $\mathbf{x}_n \in \mathcal{E}_4$,

$$K_1 \mathcal{W}(B_{\mathbb{P}}(f_0, t_*)) + l_\xi \mathcal{W}(\mathcal{F}) \geq \widehat{\mathcal{W}}(B_n(f_0, \widehat{r})). \tag{17}$$

First, recall that $t_* \leq s_1 \sqrt{\mathcal{W}(\mathcal{F})}$ and therefore under the event $\mathcal{E}_1$, we have

$$\sup_{h \in B_{\mathbb{P}}(f_0, t_*)} \|h - f_0\|_n^2 \leq 11 \cdot \mathcal{W}(\mathcal{F}). \tag{18}$$

Now, for each $\mathbf{x}_n \in \mathcal{E}_4 \subseteq \mathcal{E}$, the map $\boldsymbol{\xi} \mapsto \sup_{h \in B_{\mathbb{P}}(f_0, t_*)} \langle \boldsymbol{\xi}, h - f_0 \rangle_n$ is Lipschitz with constant at most

$$\sup_{h \in B_{\mathbb{P}}(f_0, t_*)} n^{-1/2} \|h - f_0\|_n \leq C_3 \sqrt{\mathcal{W}(\mathcal{F}) n^{-1}}$$

by (18), and thus by Lipschitz concentration (Lemma A.8), conditionally on $\mathbf{x}_n$,

$$\Pr_{\xi} \left(\left| \sup_{h \in B_{\mathbb{P}}(f_0, t_*)} \langle \boldsymbol{\xi}, h - f_0 \rangle_n - \widehat{\mathcal{W}}(B_{\mathbb{P}}(f_0, t_*)) \right| \geq \epsilon \right) \leq 2 \exp(-Cn\mathcal{W}(\mathcal{F})^{-1} \epsilon^2)$$

for some absolute constant C. By setting $\epsilon = C_5 (n^{-1} \mathcal{W}(\mathcal{F}))^{1/2}$ in the last equation, we may define the event

$$\mathcal{A}_1(\mathbf{x}_n) = \left\{ \boldsymbol{\xi} : \left| \sup_{h \in B_{\mathbb{P}}(f_0, t_*)} \langle \boldsymbol{\xi}, h - f_0 \rangle_n - \widehat{\mathcal{W}}(B_{\mathbb{P}}(f_0, t_*)) \right| \leq C_5 \sqrt{\mathcal{W}(\mathcal{F}) n^{-1}} \right\} \cap \mathcal{A}(\mathbf{x}_n).$$

that holds with probability of at least 0.25 (over ξ) for any $\mathbf{x}_n \in \mathcal{E}_4$.

Before defining the next event, observe that

$$\widehat{r} = \operatorname*{argmax}_{r \geq 0} \widehat{\mathcal{W}}(B_n(f_0, r)) - r^2/2 \leq 2\sqrt{\widehat{\mathcal{W}}(\mathcal{F})},$$

according to Lemma 4.1 and the fact that for $r_n := 2\sqrt{\widehat{\mathcal{W}}(\mathcal{F})}$ we have $\widehat{\mathcal{W}}(B_n(f_0, r_n)) - r_n^2/2 \leq 0$. As we already argued in (13), for any $\mathbf{x}_n \in \mathcal{E}_2$ we have that $\widehat{\mathcal{W}}(\mathcal{F}) \leq C_1 K_1 \mathcal{W}(\mathcal{F})$ for some absolute constant C_1, and thus

$$\forall \mathbf{x}_n \in \mathcal{E}_2, \quad \widehat{r} \leq C\sqrt{K_1 \mathcal{W}(\mathcal{F})}. \tag{19}$$

Now, from Eq. (6) in Lemma 4.1, for C_8 large enough, the event

$$\left\{ \boldsymbol{\xi} : |\langle \boldsymbol{\xi}, \widehat{f}_n - f_0 \rangle_n - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r}))| \leq C_8(n^{-1/4} \widehat{r}^{3/2} + n^{-1}) \right\}$$

holds with probability of at least 0.5, and thus, in view of (19), for all $\mathbf{x}_n \in \mathcal{E}_4$, the event

$$\mathcal{A}_2(\mathbf{x}_n) = \left\{ \boldsymbol{\xi} : |\langle \boldsymbol{\xi}, \widehat{f}_n - f_0 \rangle_n - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r}))| \leq C_6 K_1 \mathcal{W}(\mathcal{F})^{3/4} n^{-1/4} \right\} \cap \mathcal{A}_1(\mathbf{x}_n)$$

that holds with probability of at least 0.1 over ξ.

We are now ready to prove Eq. (17), using the fact that $\mathcal{A}_2(\mathbf{x}_n)$ is not empty for each $\mathbf{x}_n \in \mathcal{E}_4$. To this end, fix $\mathbf{x}_n \in \mathcal{E}_4$ and $\boldsymbol{\xi} \in \mathcal{A}_2(\mathbf{x}_n)$. First, by definition of $\mathcal{E}_3$, we have

$$K_1 \mathcal{W}(B_{\mathbb{P}}(f_0, t_*)) \geq \widehat{\mathcal{W}}(B_{\mathbb{P}}(f_0, t_*)) - C\sqrt{\mathcal{W}(\mathcal{F}) n^{-1}}$$

which can be further lower bounded, by definition of $\mathcal{A}_1(\mathbf{x}_n)$, by

$$\sup_{h \in B_{\mathbb{P}}(f_0, t_*)} \langle \boldsymbol{\xi}, h - f_0 \rangle_n - C\sqrt{\mathcal{W}(\mathcal{F}) n^{-1}}.$$

Since $\boldsymbol{\xi} \in \mathcal{A}_2(\mathbf{x}_n) \subseteq \mathcal{A}(\mathbf{x}_n)$, the above expression is further lower bounded by

$$\langle \boldsymbol{\xi}, \hat{f}_n - f_0 \rangle_n - C\sqrt{\mathcal{W}(\mathcal{F})n^{-1}}$$

which, under the assumption of $\boldsymbol{\xi} \in \mathcal{A}_2(\mathbf{x}_n)$, is lower bounded by

$$\widehat{\mathcal{W}}(B_n(f_0, \hat{r})) - C_2\sqrt{\mathcal{W}(\mathcal{F})n^{-1}} - C_6 K_1 \mathcal{W}(\mathcal{F})^{3/4} n^{-1/4}.$$

When n is large enough, the above estimate is lower bounded by

$$\widehat{\mathcal{W}}(B_n(f_0, \hat{r})) - l_\xi \mathcal{W}(\mathcal{F}).$$

To see this, observe that under the assumption of $\mathcal{W}(\mathcal{F}) \geq c/\sqrt{n}$, both $\sqrt{\mathcal{W}(\mathcal{F})n^{-1}} = o_n(\mathcal{W}(\mathcal{F}))$ and $\mathcal{W}(\mathcal{F})^{3/4}n^{-1/4} = o_n(\mathcal{W}(\mathcal{F}))$. Therefore, we proved Eq. (17) holds, namely that

$$K_1 \mathcal{W}(B_\mathbb{P}(f_0, t_*)) + l_\xi \mathcal{W}(\mathcal{F}) \geq \widehat{\mathcal{W}}(B_n(f_0, \hat{r}))$$

for all $\mathbf{x}_n \in \mathcal{E}_4$. Using the definition of $l_\xi = (256K_1)^{-3}$, we have

$$\mathcal{W}(B_\mathbb{P}(f_0, t_*)) \leq l_\xi \mathcal{W}(\mathcal{F}) \leq 128^{-3} K_1^{-3} \mathcal{W}(\mathcal{F}),$$

and thus for any $\mathbf{x}_n \in \mathcal{E}_4$,

$$8^{-1} K_1^{-1} \mathcal{W}(\mathcal{F}) \geq \widehat{\mathcal{W}}(B_n(f_0, \hat{r})). \tag{20}$$

By Lemma A.1 and (20), for any $\mathbf{x}_n \in \mathcal{E}_4$,

$$0 \geq (2^{-1} K_1^{-1} - 8^{-1} K_1^{-1}) \mathcal{W}(\mathcal{F}) - \mathbb{E}_\xi \max_{g \in \mathcal{M}} \langle g - f_0, \hat{f}_n - f_0 \rangle_n - 16\sqrt{\mathcal{W}(\mathcal{F})}\hat{r},$$

and since $\mathbf{x}_n \in \mathcal{E}_4 \subseteq \mathcal{E}_1$, we also have

$$\begin{aligned}
0 \geq &(2^{-1} K_1^{-1} - 8^{-1} K_1^{-1} - 8^{-1} K_1^{-1}) \mathcal{W}(\mathcal{F}) - \sup_{h \in B_\mathbb{P}(f_0, t_*), g \in \mathcal{M}} \int (g - f_0)(h - f_0) d\mathbb{P} - 16\sqrt{\mathcal{W}(\mathcal{F})}\hat{r} \\
\geq &(4K_1)^{-1} \mathcal{W}(\mathcal{F}) - \max_{g \in \mathcal{M}} \|g - f_0\| t_* - 16\sqrt{\mathcal{W}(\mathcal{F})}\hat{r} \\
\geq &(4K_1)^{-1} \mathcal{W}(\mathcal{F}) - 2t_* - 16\sqrt{\mathcal{W}(\mathcal{F})}\hat{r} \tag{21}
\end{aligned}$$

where we used the Cauchy-Schwartz inequality, the fact that $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, and the definition of $l_\xi = (256K_1)^{-3}$.

If $16\sqrt{\mathcal{W}(\mathcal{F})}\hat{r} < (8K_1)^{-1} \mathcal{W}(\mathcal{F})$, then the last equation implies that

$$s_2 \mathcal{W}(\mathcal{F}) = (16K_1)^{-1} \mathcal{W}(\mathcal{F}) < t_*.$$

However, this inequality contradicts the definition of t_*, and thus cannot hold for any $\mathbf{x}_n \in \mathcal{E}_4$. In the other case, we assume that $16\sqrt{\mathcal{W}(\mathcal{F})}\hat{r} \geq (8K_1)^{-1} \mathcal{W}(\mathcal{F})$, or equivalently, $\hat{r} \geq (128K_1)^{-1}\sqrt{\mathcal{W}(\mathcal{F})}$. Now, from Lemma 4.1 one can see that the maximizing value $\hat{r}$ ensures

$$\widehat{\mathcal{W}}(B_n(f_0, \hat{r})) - 2^{-1} \hat{r}^2 > 0$$

and hence

$$\widehat{\mathcal{W}}(B_n(f_0, \hat{r})) > 2^{-1} (128K_1)^{-2} \mathcal{W}(\mathcal{F}).$$

Therefore, under the event $\mathcal{E}_4$ and by Eq. (17)

$$2K_1 l_\xi \mathcal{W}(\mathcal{F}) \geq K_1 \mathcal{W}(B_\mathbb{P}(f_0, t_*)) + l_\xi \mathcal{W}(\mathcal{F}) \geq \widehat{\mathcal{W}}(B_n(f_0, \hat{r})) > 2^{-1} (128K_1)^{-2} \mathcal{W}(\mathcal{F}).$$

Once again, we have a contradiction for any $\mathbf{x}_n \in \mathcal{E}_4$, since we assumed that $l_\xi = (256K_1)^{-3}$.

11

Therefore, we showed that Eq. (21), cannot hold under the event $\mathcal{E}_4$, i.e. the set $\mathcal{E}_4$ is empty. This contradicts our earlier conclusion that $\Pr(\mathcal{E}_4) \geq 0.4$, which was made under the assumption that event $\mathcal{A}$ has probability at least 0.5. Hence, we conclude that $\Pr(\mathcal{A}) \leq 0.5$, or, equivalently, with probability at least 0.5, $\hat{f}_n \notin B_{\mathbb{P}}(f_0, t_*)$. Therefore, we must have that

$$\mathbb{E} \int (\hat{f}_n - f_0)^2 d\mathbb{P} \geq \frac{t_*^2}{2} = \frac{1}{2} \min\{t_{n,\mathbb{P}}(f_0, \mathcal{F})^2, s_1^2 \mathcal{W}(\mathcal{F}), s_2^2 \mathcal{W}(\mathcal{F})^2\} \geq c_1 \cdot \min\{t_{n,\mathbb{P}}(f_0, \mathcal{F})^2, \mathcal{W}(\mathcal{F})^2\}.$$

where in the last inequality, we used the fact that

$$\mathcal{W}(\mathcal{F}) \leq \mathcal{W}([-1,1]^{\mathcal{X}}) \leq \mathbb{E}|\xi| \leq \sqrt{\mathbb{E}\xi^2} = 1.$$

The theorem follows. $\square$

Acknowledgements

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

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

Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *arXiv preprint arXiv:1806.05161*, 2018.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

Pierre C Bellec. Optimistic lower bounds for convex regularized least-squares. *arXiv preprint arXiv:1703.01332*, 2017.

Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'IHP Probabilités et statistiques*, volume 42, pages 273–325, 2006.

Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150, 1993.

Lucien Birgé, Pascal Massart, et al. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

EM Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17(3):393–398, 1976.

Sourav Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, 2014.

Richard M Dudley. *Uniform central limit theorems*. Number 63. Cambridge university press, 1999.

Oliver Y Feng, Adityanand Guntuboyina, Arlene KH Kim, and Richard J Samworth. Adaptation in multivariate log-concave density estimation. *arXiv preprint arXiv:1812.11634*, 2018.

Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran. Max-affine regression: Provable, tractable, and near-optimal statistical estimation. *arXiv preprint arXiv:1906.09255*, 2019.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Number 40. Cambridge University Press, 2016.

Qiyang Han and Jon A Wellner. Multivariate convex regression: global risk bounds and adaptation. *arXiv preprint arXiv:1601.06844*, 2016.

Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, Richard J Samworth, et al. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019.

Arlene KH Kim, Adityanand Guntuboyina, Richard J Samworth, et al. Adaptation in log-concave density estimation. *The Annals of Statistics*, 46(5):2279–2306, 2018.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Gil Kur, Fuchang Gao, Adityanand Guntuboyina, and Bodhisattva Sen. Convex regression in multidimensions: Suboptimality of least squares estimators. *arXiv preprint arXiv:2006.02044*, 2020a.

Gil Kur, Alexander Rakhlin, and Adityanand Guntuboyina. On suboptimality of least squares with application to estimation of convex bodies. *arXiv preprint arXiv:2006.04046*, 2020b.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020a.

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020b.

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

Gilles Pisier. Some applications of the metric entropy condition to harmonic analysis. In *Banach Spaces, Harmonic Analysis, and Probability Theory*, pages 123–154. Springer, 1983.

Alexander Rakhlin, Karthik Sridharan, Alexandre B Tsybakov, et al. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

Richard J Samworth. Recent progress in log-concave density estimation. *Statistical Science*, 33(4):493–509, 2018.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2003.

Sara A van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017.

A   Lemmas

Lemma A.1. Under the event $\mathcal{E}$ in (12), and for n that is large enough, the following holds:

$$0 \geq 2^{-1} K_1^{-1} \mathcal{W}(\mathcal{F}) - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r})) - \mathbb{E}_\xi \max_{g \in \mathcal{M}} \langle g - f_0, \widehat{f}_n - f_0 \rangle_n - 16 \sqrt{\mathcal{W}(\mathcal{F})} \widehat{r}, \tag{22}$$

where K_1 is defined in Eq. (12), and set $\mathcal{M}$ is defined in Eq. (10).

Lemma A.2. Let $X_1, \dots, X_n \underset{i.i.d}{\sim} \mathbb{P}$, then the following holds with probability of at least $1 - K \exp(-n \mathcal{W}(\mathcal{F}))$

$$\sup_{g \in \mathcal{M}, h \in B_\mathbb{P}(f_0, t_*)} \left| \langle h - f_0, g - f_0 \rangle_n - \int_\mathcal{X} (h - f_0)(g - f_0) d\mathbb{P} \right| \leq (8K_1)^{-1} \mathcal{W}(\mathcal{F}).$$

where $\mathcal{M}$ is defined in Eq. (10), t_* is defined in Eq. (9), and K_1, K are defined in Eq. (12), Lemma A.6.

Lemma A.3. The event $\mathcal{E}$ defined in Eq. (12) holds with probability of at least 0.9.

A.1   Auxiliary Lemmas

Lemma A.4. [(Koltchinskii, 2011, pgs. 25-26)] Let $\mathcal{F} \subseteq [-1, 1]^\mathcal{X}$ be family of functions. Then with probability of at least $1 - 2 \exp(-c_1 n \mathcal{W}(\mathcal{F}))$,

$$\forall f, g \in \mathcal{F} \quad \left| \|f - g\|_n^2 - \|f - g\|_\mathbb{P}^2 \right| \leq 10 \mathcal{W}(\mathcal{F}),$$

and

$$\|\widehat{f}_n - f_0\|_n^2 \leq 10 \mathcal{W}(\mathcal{F}).$$

Lemma A.5 (Sudakov's minoration lemma). Let $\mathcal{H} \subset [-1, 1]^\mathcal{X}$. There exists a constant c_1 such that for any $\mathbb{P}_n$,

$$c_1 \sup_{\epsilon \geq 0} \epsilon \sqrt{\frac{\log \mathcal{M}(\epsilon, \mathcal{H}, \mathbb{P}_n)}{n}} \leq \widehat{\mathcal{W}}(\mathcal{H}).$$

where $\mathcal{M}(\epsilon, \mathcal{H}, \mathbb{P}_n)$ denotes the size of the largest ϵ-separated set in $\mathcal{H}$ with respect to $L_2(\mathbb{P}_n)$.

The next two lemmas appear in (Koltchinskii, 2011, pgs. 24-25), (Adamczak, 2008).

Lemma A.6 (Talagrand's inequality). Let $X_1, \dots, X_n \underset{i.i.d.}{\sim} \mathbb{P}$, and $\mathcal{H} \subseteq [-U, U]^\mathcal{X}$ be a family of functions. Let $Z = \sup_{f \in \mathcal{H}} |n^{-1} \sum_{i=1}^n f(X_i) - \mathbb{E}[f]|$. Then there exists an absolute constant $K \geq 0$ such that for any $s \geq 0$

$$\Pr(|Z - \mathbb{E}Z| \geq s) \leq K \exp\left(-K^{-1} U^{-1} \log(1 + \frac{sU}{V^2}) ns\right),$$

where $V^2 = \sup_{f \in \mathcal{H}} \int f^2 d\mathbb{P}$.

Lemma A.7 (Adamczak's inequality). Let $\mathcal{G}$ be a centred family of functions supported on $\mathcal{D}$, and $\mathbb{Q}$ be some distribution on $\mathcal{D}$. Let $Z = \sup_{g \in \mathcal{G}} |n^{-1} \sum_{i=1}^n g(X_i)|$. Assume that there exists an envelope function G such that $|g(x)| \leq G(x)$ for all $g \in \mathcal{G}, x \in \mathcal{D}$. Then, the following holds for all $t \geq 0$

$$K_2^{-1} \mathbb{E}Z - V\sqrt{\frac{t}{n}} - \frac{\|\max_{1 \leq i \leq n} f'(X_i)\|_{\psi_1} t}{n} \leq Z \leq K_2 \mathbb{E}Z + V\sqrt{\frac{t}{n}} + \frac{\|\max_{1 \leq i \leq n} f(X_i)\|_{\psi_1} t}{n},$$

where $V^2 := \sup_{g \in \mathcal{G}} \int g^2 d\mathbb{Q}$, and $K_2 \in (1, \infty)$ is some universal constant, and ψ_1 is the Orlicz norm.

Lemma A.8 (Lipschitz Concentration). Let $\xi_1, \dots, \xi_n \underset{i.i.d}{\sim} N(0, 1)$, and $f : \mathbb{R}^n \to \mathbb{R}$ be a L-Lipschitz function with respect to $\| \cdot \|_2$. Then, for all $\epsilon > 0$,

$$\Pr(|f - \mathbb{E}[f]| \geq \epsilon) \leq \exp(-c\epsilon^2 L^{-2}).$$

B Proofs

Proof of Lemma A.1. We invoke the lower bound of Eq. (7) with $g = f_0$ and $t = 2$, implying

$$0 \geq \widehat{\mathcal{W}}(B_n(f_0, 2)) - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r})) - \mathbb{E}\langle \widehat{f}_n - f_0, g_\xi - f_0 \rangle_n - Cn^{-1}$$
$$= \widehat{\mathcal{W}}(\mathcal{F}) - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r})) - \mathbb{E}\langle \widehat{f}_n - f_0, g_\xi - \Pi(g_\xi) + \Pi(g_\xi) - f_0 \rangle_n - Cn^{-1}, \qquad (23)$$

where $\Pi(g_\xi) := \operatorname{argmin}_{g \in \mathcal{M}} \|g_\xi - g\|_{\mathbb{P}_n}$, and the equality follows for the fact that for $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ we have $B_n(f_0, 2) = \mathcal{F}$.

Now, recall that $\mathcal{M}$ is a maximal $6\sqrt{\mathcal{W}(\mathcal{F})}$-separated set with respect to $L_2(\mathbb{P})$, and therefore also a $12\sqrt{\mathcal{W}(\mathcal{F})}$-net with respect to $L_2(\mathbb{P})$. Therefore, under the event $\mathcal{E}$ it is also a $16\sqrt{\mathcal{W}(\mathcal{F})}$-net with respect to $L_2(\mathbb{P}_n)$, and, in particular, $\|\Pi(g_\xi) - g_\xi\|_{\mathbb{P}_n} \leq 16\sqrt{\mathcal{W}(\mathcal{F})}$. Hence, we can rewrite (23) as

$$\mathbb{E}_\xi \max_{g \in \mathcal{M}} \langle \widehat{f}_n - f_0, g - f_0 \rangle_n$$
$$\geq \widehat{\mathcal{W}}(\mathcal{F}) - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r})) - \mathbb{E}_\xi \langle \widehat{f}_n - f_0, g_\xi - \Pi(g_\xi) \rangle_n - Cn^{-1}$$
$$\geq K_1^{-1}\mathcal{W}(\mathcal{F}) - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r})) - 16\sqrt{\mathcal{W}(\mathcal{F})}\mathbb{E}_\xi \|f - \widehat{f}_n\|_{\mathbb{P}_n} - Cn^{-1}$$

Now, we proceed by using the first part of Corollary 4.1 and the assumption of lying in $\mathcal{E}$. The last expression is lower-bounded by

$$K_1^{-1}\mathcal{W}(\mathcal{F}) - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r})) - 16\sqrt{\mathcal{W}(\mathcal{F})}(\widehat{r} + C\widehat{r}^{1/2}n^{-1/4}). \qquad (24)$$

According to (19), under the event $\mathcal{E}$, we have

$$\widehat{r} \leq C_3\sqrt{K_1\mathcal{W}(\mathcal{F})}$$

for some constant C_3. Thus the expression in Eq. (24) is further lower-bounded by

$$K_1^{-1}\mathcal{W}(\mathcal{F}) - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r})) - 16\sqrt{\mathcal{W}(\mathcal{F})}\widehat{r} - C_4\sqrt{K_1}\mathcal{W}(\mathcal{F})^{3/4}n^{-1/4} - Cn^{-1}$$
$$\geq (2K_1)^{-1}\mathcal{W}(\mathcal{F}) - \widehat{\mathcal{W}}(B_n(f_0, \widehat{r})) - 16\sqrt{\mathcal{W}(\mathcal{F})}\widehat{r}$$

where the last inequality holds when n is large enough. To see this, recall that $\mathcal{W}(\mathcal{F}) \geq c/\sqrt{n}$ and under this assumption both $n^{-1} = o_n(\mathcal{W}(\mathcal{F}))$ and $\mathcal{W}(\mathcal{F})^{3/4}n^{-1/4} = o_n(\mathcal{W}(\mathcal{F}))$ hold. Therefore, the lemma follows. $\qquad \square$

Proof of Lemma A.2. First, denote by $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{H}} := \sup_{h \in \mathcal{H}} |n^{-1}\sum_{i=1}^n h(X_i) - \mathbb{E}[h]|$, and for each $g_i \in \mathcal{M}$, define $\mathcal{G}_i = \{(h - f_0)(g_i - f_0) : h \in B_{\mathbb{P}}(f_0, t_*)\}$. By Talagrand's inequality (Lemma A.6), the following holds for and $u \geq 0$

$$\Pr\left(|\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i}| \geq u\right) \leq K\exp\left(-nK^{-1}\log(1 + \frac{4^{-1}us_1^{-2}}{\mathcal{W}(\mathcal{F})})u\right)$$

where we used the fact that $V^2 \leq \sup_{h \in \mathcal{F}} \|g - f_0\|_\infty^2 t_*^2 \leq 4s_1^2\mathcal{W}(\mathcal{F})$. Now, we set $u = (16K_1)^{-1}\mathcal{W}(\mathcal{F})$ in the last equation

$$\Pr\left(|\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i}| \geq (16K_1)^{-1}\mathcal{W}(\mathcal{F})\right)$$
$$\leq K\exp\left(-n(16K \cdot K_1)^{-1}\mathcal{W}(\mathcal{F})\log(1 + 4^{-1}(16K_1)^{-1}s_1^{-2})\right).$$

Next, we aim to take a union bound over $\mathcal{M}$, and recall that $\log M \leq C_2 K_1^2 n\mathcal{W}(\mathcal{F}) \leq C(K_1)n\mathcal{W}(\mathcal{F})$, for some absolute constant that does not depend on s_1. Therefore, we may choose

$$s_1 := c(K, K_1, C_2) \qquad (25)$$

where $c(K, K_1, C_2)$ is a constant that satisfies the following:

$$\Pr\left(|\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i}| \geq (16K_1)^{-1}\mathcal{W}(\mathcal{F})\right) \leq K\exp(-2C_2K_1n\mathcal{W}(\mathcal{F})).$$

Therefore, we have

$$\Pr\left(\exists 1 \leq i \leq M : |\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i}| \geq (16K_1)^{-1}\mathcal{W}(\mathcal{F})\right) \leq MK\exp(-2C_2K_1 n\mathcal{W}(\mathcal{F})))$$
$$\leq K\exp(-C_2K_1 n\mathcal{W}(\mathcal{F}))$$
$$\leq K\exp(-n\mathcal{W}(\mathcal{F})).$$

We conclude that with probability of at least $1 - K\exp(-n\mathcal{W}(\mathcal{F}))$ the following holds for $\mathcal{G} := \{(h - f_0)(g - f_0) : g \in \mathcal{M}, h \in B_\mathbb{P}(f_0, t_*)\}$:

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}} \leq \max_{1 \leq i \leq M} \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} + (16K_1)^{-1}\mathcal{W}(\mathcal{F}). \tag{26}$$

The lemma will follow as soon as we show that

$$\max_{1 \leq i \leq M} \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} \leq (16K_1)^{-1}\mathcal{W}(\mathcal{F}).$$

In order to prove the last inequality, we first apply the symmetrization lemma (cf. (Koltchinskii, 2011, p. 20)) and majorize the resulting Rademacher averages by a constant multiple of the Gaussian averages

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} \leq 8\mathcal{W}(\mathcal{G}_i) \tag{27}$$

where we used the fact that $0 \in \mathcal{G}_i$.

Next, since $\|g_i - f_0\|_\infty \leq 2$, a standard argument (e.g. (Giné and Nickl, 2016, Theorem 3.1.17)) gives

$$\mathbb{E}_\xi \sup_{h \in B_\mathbb{P}(f_0, t_*)} n^{-1} \sum_{k=1}^n (h - f_0)(g_i - f_0)(X_k)\xi_k \leq 2\mathbb{E}_\xi \sup_{h \in B_\mathbb{P}(f_0, t_*)} n^{-1} \sum_{k=1}^n (h - f_0)(X_i)\xi_k.$$

Then, by taking expectation over $X_1, \ldots, X_n$ over the last equation and by Eq. (27), we conclude

$$\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{G}_i} \leq 16\mathcal{W}(B_\mathbb{P}(f_0, t_*)) \leq 16l_\xi \mathcal{W}(\mathcal{F}) \leq (16K_1)^{-1}\mathcal{W}(\mathcal{F}),$$

where we set $l_\xi = (256K_1)^{-3}$. Then, by Eq. (26) and the last equation, the claim follows. $\square$

Proof of Lemma A.3. It is enough to show that $\mathcal{E}_2$, $\mathcal{E}_3$ hold with probability of at least 0.99 for n large enough. First, we prove this claim for $\mathcal{E}_2$.

We aim to apply Adamczak bound for concentration of the suprema of unbounded empirical processes (Lemma A.7). For this purpose, define the family of functions $\mathcal{G} := \{yf(x), y \in \mathbb{R}, f \in \mathcal{F} - f_0\}$, and the distribution $\mathbb{Q} = \mathbb{P} \otimes N(0, 1)$. Note that $\mathcal{F} \subseteq [-1, 1]^\mathcal{X}$ and, ξ is Gaussian. Therefore, by Pisier's inequality (cf. Pisier (1983),(Adamczak, 2008, Eq. 13)), we have

$$\|\max_{1 \leq i \leq n} |\xi_i f(X_i)|\|_{\psi_1} \leq C\log(n)\max_{1 \leq i \leq n}\|\xi_i f(X_i)\|_{\psi_1} \leq C_2\log(n).$$

By Adamczak's bound (Lemma A.7),

$$K_2^{-1}\mathbb{E}_{x,\xi} \sup_{f \in \mathcal{F} - f_0} |\frac{1}{n}\sum_{i=1}^n f(X_i)\xi_i| - \frac{10}{\sqrt{n}} - \frac{C\log(n)}{n}$$
$$\leq \sup_{f \in \mathcal{F} - f_0} |\frac{1}{n}\sum_{i=1}^n f(X_i)\xi_i| \leq K_2\mathbb{E}_{x,\xi} \sup_{f \in \mathcal{F} - f_0} |\frac{1}{n}\sum_{i=1}^n f(X_i)\xi_i| + \frac{10}{\sqrt{n}} + \frac{C\log(n)}{n}, \tag{28}$$

with probability of at least 0.99 both $X_1, \ldots, X_n$ and $\boldsymbol{\xi}$.

Now, using the average principle, for n large enough, we can find an event $\mathcal{E}_7$ (that depends only on $X_1, \ldots, X_n$) that holds with probability 0.98, such that for any fixed $\mathbf{x}_n \in \mathcal{E}_7$, there exists an event $\mathcal{A}_3(\mathbf{x}_n)$ of probability at least 0.98 (over $\boldsymbol{\xi}$) such that Eq. (28) holds. For each $\mathbf{x}_n \in \mathcal{E}_7$, Lemma A.8 (with Lipschitz

16

constant $\sup_{f\in\mathcal{F}}\|f-f_0\|_n \leq 2$) implies that the middle term in (28) is, with high probability, within $Cn^{-1/2}$ from its expectation (with respect to $\boldsymbol{\xi}$). Therefore, we have for all $\mathbf{x}_n \in \mathcal{E}_7$:

$$K_2^{-1}\mathbb{E}_{x,\xi} \sup_{f\in\mathcal{F}-f_0} |n^{-1}\sum_{i=1}^n f(X_i)\xi_i| - \frac{C}{\sqrt{n}}$$

$$\leq \mathbb{E}_\xi \sup_{f\in\mathcal{F}-f_0} |n^{-1}\sum_{i=1}^n f(X_i)\xi_i| \leq K_2\mathbb{E}_{x,\xi} \sup_{f\in\mathcal{F}-f_0} |n^{-1}\sum_{i=1}^n f(X_i)\xi_i| + \frac{C}{\sqrt{n}}.$$

Finally, since $0 \in \mathcal{F} - f_0$, we have

$$\mathbb{E}_\xi \sup_{f\in\mathcal{F}-f_0} n^{-1}\sum_{i=1}^n f(X_i)\xi_i \leq \mathbb{E}_\xi \sup_{f\in\mathcal{F}-f_0} |n^{-1}\sum_{i=1}^n f(X_i)\xi_i| \leq 2\mathbb{E}_\xi \sup_{f\in\mathcal{F}-f_0} n^{-1}\sum_{i=1}^n f(X_i)\xi_i.$$

Hence, the last two equations imply that when $\mathcal{W}(\mathcal{F}) \geq C_1 n^{-1/2}$, for C_1 that is large enough, the claim follows for $\mathcal{E}_2$. To handle the remaining case of $\mathcal{W}(\mathcal{F}) \leq C_1 n^{-1/2}$, recall that we assumed that our class is not degenerate (i.e it has two functions that are $\|f_1 - f_2\|_{\mathbb{P}} \geq 0.5$. Then, it is easy to see that with probability of 0.99 it holds that

$$\widehat{\mathcal{W}}(\mathcal{F} - f_0) \geq \mathcal{W}(\{0, f_2 - f_0, f_1 - f_0\}) \geq \mathbb{E}\max\{n^{-0.5}g, 0\} \geq c \cdot n^{-1/2} \geq c \cdot C_1^{-1}\mathcal{W}(\mathcal{F}),$$

where $g \sim N(0, 1/4)$. Therefore, for some $K_1^{-1} = c(K_2, c)$, the claim follows for $\mathcal{E}_2$.

Next, we handle $\mathcal{E}_3$. By using the definition of $B_{\mathbb{P}}(f_0, t_*)$, and similar considerations that led to Eq. (28), we have

$$\sup_{f\in B_{\mathbb{P}}(f_0,t_*)-f_0} |n^{-1}\sum_{i=1}^n f(X_i)\xi_i| \leq K_2\mathbb{E}_{x,\xi} \sup_{B_{\mathbb{P}}(f_0,t_*)-f_0} |n^{-1}\sum_{i=1}^n f(X_i)\xi_i| + \frac{10t_*}{\sqrt{n}} + \frac{C\log(n)}{n}, \qquad (29)$$

with probability of at least 0.99 over both $X_1, \ldots, X_n$ and $\boldsymbol{\xi}$.

As above, for n large enough, we can find an event $\mathcal{E}_8 \subseteq \mathcal{E}_1$ (where $\mathcal{E}_1$ is defined in Eq. (11)) of probability at least 0.98 (over $X_1, \ldots, X_n$), such that for any $\mathbf{x}_n \in \mathcal{E}_8$, there exists an event $\mathcal{A}_4(\mathbf{x}_n)$ of probability at least 0.98 (over $\boldsymbol{\xi}$) such that (29) holds. Then, similarly to the case of $\mathcal{E}_2$, we will employ Lipschitz concentration for the middle term in (29), for each $\mathbf{x}_n \in \mathcal{E}_8$. To estimate the Lipschitz constant, recall that under $\mathcal{E}_1$ (more precisely, under the event of Lemma A.4), we also have that

$$\|f - f_0\|_n^2 \leq s_1^2\mathcal{W}(\mathcal{F}) + 10\mathcal{W}(\mathcal{F}) \leq 11\mathcal{W}(\mathcal{F})$$

for all $f \in B_{\mathbb{P}}(f_0, t_*)$, under the choice t_* in (9). Then, using the fact that $\mathcal{A}_4(\mathbf{x}_n)$ holds with probability of at least 0.98, and Lemma A.8 with Lipschitz constant $\sup_{f\in B_{\mathbb{P}}(f_0,t_*)} \|f - f_0\|_{\mathbb{P}_n} \leq \sqrt{11\mathcal{W}(\mathcal{F})}$, imply that for each $\mathbf{x}_n \in \mathcal{E}_8$, the middle term in (29) is within an additive factor of $C_1\sqrt{\mathcal{W}(\mathcal{F})}n^{-1/2}$ from its expectation over $\boldsymbol{\xi}$. Namely, we have for all $\mathbf{x}_n \in \mathcal{E}_8$:

$$\mathbb{E}_\xi \sup_{f\in\mathcal{F}-f_0} |n^{-1}\sum_{i=1}^n f(X_i)\xi_i| \leq K_2\mathbb{E}_{x,\xi} \sup_{f\in\mathcal{F}-f_0} |n^{-1}\sum_{i=1}^n f(X_i)\xi_i| + \frac{C\sqrt{\mathcal{W}(\mathcal{F})}}{\sqrt{n}}.$$

where we used the fact that $t_* \leq s_1\sqrt{\mathcal{W}(\mathcal{F})}$. The claim for $\mathcal{E}_3$ follows by similar considerations that we used earlier. $\qquad\square$

Proof of Corollary 3.3. For any $\mathbb{P}$-Donsker class we have with probability at least 0.9 (van de Geer, 2000, Chap. 5)

$$\widehat{\mathcal{W}}(\mathcal{F}) \sim \mathcal{W}(\mathcal{F}) \sim n^{-1/2}.$$

Then, by Corollary 3.1, we have that

$$\mathbb{E}\int(\hat{f}_n - f_0)^2 d\mathbb{P}_n \gtrsim n^{-1}.$$

17

In order to prove the second part of the bound, we apply Theorem 3.1,

$$\mathbb{E}\int(\hat{f}_n - f_0)^2 d\mathbb{P} \gtrsim \max\{n^{-1}, t_{n,\mathbb{P}}(f_0, \mathcal{F})^2\}.$$

The corollary will follow if we show that for any $f_0 \in \mathcal{F}$, we have that $t_{n,\mathbb{P}} \gtrsim 1$. To see this, we use (van de Geer, 2000, Thm 5.11) that shows that for all $t \geq 0$, we have

$$\mathcal{W}(B_\mathbb{P}(f_0, t)) \lesssim n^{-1/2}\int_0^t u^{-\alpha/2}du \lesssim t^{\frac{-\alpha+2}{2}}n^{-1/2}.$$

Since $\alpha \in (0, 2)$, the right hand side is decreasing in t, therefore we know that if

$$\mathcal{W}(B_\mathbb{P}(f_0, t_*)) \gtrsim \mathcal{W}(\mathcal{F}) \gtrsim n^{-1/2}$$

then we have $t_* \gtrsim 1$. Hence, $t_{n,\mathbb{P}}(f_0, \mathcal{F}) \gtrsim 1$, and the claim follows. □

Proof of Corollary 3.4. For any non $\mathbb{P}$-Donsker class we have with probability of at least 0.9 (van de Geer, 2000, Chap. 5)

$$n^{-\frac{2}{2+\alpha}} \lesssim \widehat{\mathcal{W}}(\mathcal{F}) \sim \mathcal{W}(\mathcal{F}) \lesssim n^{-\frac{1}{\alpha}}.$$

Then, by Corollary 3.1, we have that

$$\mathbb{E}\int(\hat{f}_n - f_0)^2 d\mathbb{P}_n \gtrsim n^{-\frac{4}{2+\alpha}},$$

and the claim follows. □

B.1   An example to the tightness of Theorem 3.1 (a sketch)

Let $\mathbb{P}$ be the uniform density of $[0, 1]$, and denote by $I(x_i, l_i)$ to be an interval with center x_i and length l_i. For each $m \geq 0$ we define

$$\mathcal{F}_m := \big\{m^{-1/6}\sum_{i=1}^m \epsilon_i \mathbb{1}_{I(x_i, m^{-5/4})} : \forall x_1, \ldots, x_m \text{ s.t. } 1 \leq j \neq k \leq m \ \ I(x_k, m^{-5/4}) \cap I(x_j, m^{-5/4}) = \emptyset,$$

$$\forall(\epsilon_1, \ldots, \epsilon_m) \in \{-1, 1\}^m\big\}.$$

Now, we define $\mathcal{F} := \text{conv}\{0, \{\mathcal{F}_m\}_{m=1}^\infty\}$. Clearly, this family is uniformly bounded by one. Also, we assume that $f_0 = 0$.

Using a classical fact, we have that with probability of at least $1 - n^2$,

$$\max_{1 \leq i \neq j \leq n} |X_j - X_i| \geq c \cdot (n \log n)^{-1},$$

and denote this event by A. Clearly, for each $\mathbf{x}_n \in A$, we can find a function $f_\xi \in \mathcal{F}_n$ (that depends on $\mathbf{x}_n$ as well) such that

$$\langle f_\xi, \boldsymbol{\xi}\rangle_n = n^{-1/6}\cdot n^{-1}\sum_{i=1}^n |\xi_i|. \tag{30}$$

Also, note that under the event A, $\hat{f}_n \notin \{\mathcal{F}_m\}_{m=n+1}^\infty$. Therefore, one can easily show that

$$\widehat{\mathcal{W}}(B_n(f_0, n^{-1/6})) \sim n^{-1/6}.$$

Now, denote by $C(n) := C_1(n\log(n))^{4/5}$ for C_1 that is large enough. Note that any $\mathcal{F}_m$ such that $C(n) \leq m \leq n-1$, we can only place m intervals with length of at most $(c/2)\cdot(n\log(n))^{-1}$. Therefore, under the event A, each of these intervals has at most one point. Hence, we have that

$$\max_{f_m \in \mathcal{F}_m}\langle f_m, \boldsymbol{\xi}\rangle_n = m^{-1/6}n^{-1}\max_{S \in \binom{n}{m}, |S|=m}\sum_{i \in S}|\xi_i|.$$

18

Now, for any fixed $C(n) \le m \le n - 1$, one can easily show by standard concentration inequalities that

$$\mathbb{E} \max_{f_m \in \mathcal{F}_m} \langle f_m, \boldsymbol{\xi} \rangle_n \lesssim m^{-1/6}(m/n) + m^{-1/6}(m/n)\sqrt{\log(n/m)} \lesssim m^{-1/6}(m/n)\sqrt{\log(n/m)}. \qquad (31)$$

In the remaining case of $m \le C(n)$, using some standard arguments, it can be shown that with probability of at least $1 - n^2$ (over $X_1, \ldots, X_n$) the following holds:

$$\mathbb{E}_{\xi} \sup_{f_m \in \mathcal{F}_m} \langle f_m, \boldsymbol{\xi} \rangle_n \sim \mathbb{E}_{x,\xi} \sup_{f_m \in \mathcal{F}_m} \langle f_m, \boldsymbol{\xi} \rangle_n \ll n^{-1/6}. \qquad (32)$$

By using Eqs. (30),(31),(32), one can show that with high probability $\hat{f}_n \in B_n(f_0, Cn^{-1/6})$, for some $C \ge 0$, and also

$$\mathcal{W}(\mathcal{F}) \sim n^{-1/6}.$$

Therefore, one can conclude that

$$\mathbb{E} \int (\hat{f}_n - f_0)^2 d\mathbb{P}_n \sim n^{-\frac{1}{3}} \sim \mathcal{W}(\mathcal{F})^2,$$

and

$$\mathbb{E} \int (\hat{f}_n - f_0)^2 d\mathbb{P} \sim n^{-(\frac{1}{4}+\frac{1}{3})} \ll \mathcal{W}(\mathcal{F})^2 \sim n^{-\frac{1}{3}}.$$

Finally, it is easy to see that $t_{n,\mathbb{P}}(f_0, \mathcal{F}) \gtrsim n^{-(\frac{1}{8}+\frac{1}{6})}$, and therefore, by using the last equation

$$\mathbb{E} \int (\hat{f}_n - f_0)^2 d\mathbb{P} \sim t_{n,\mathbb{P}}(f_0, \mathcal{F})^2,$$

and the claim follows.