Empirical Entropy, Minimax Regret and Minimax Risk

Alexander Rakhlin Karthik Sridharan Alexandre B. Tsybakov

February 22, 2013

Abstract

We consider the random design regression with square loss. We propose a method that aggregates empirical minimizers (ERM) over appropriately chosen random subsets and reduces to ERM in the extreme case, and we establish exact oracle inequalities for its risk. We show that, under the ϵ^{-p} growth of the empirical ϵ-entropy, the excess risk of the proposed method attains the rate $n^{-\frac{2}{2+p}}$ for $p \in (0,2]$ and $n^{-1/p}$ for $p > 2$. We provide lower bounds to show that these rates are optimal. Furthermore, for $p \in (0,2]$, the excess risk rate matches the behavior of the minimax risk of function estimation in regression problems under the well-specified model. This yields a surprising conclusion that the rates of statistical estimation in well-specified models (minimax risk) and in misspecified models (minimax regret) are equivalent in the regime $p \in (0,2]$. In other words, for $p \in (0,2]$ the problem of statistical learning enjoys the same minim! ax rate as the problem of statistical estimation. Our oracle inequalities also imply the $\log(n)/n$ rates for Vapnik-Chervonenkis type classes without the typical convexity assumption on the class; we show that these rates are optimal. Finally, for a slightly modified method, we derive a bound on the excess risk of s-sparse convex aggregation improving that of Lounici [30] and we show that it yields the optimal rate.

1 Introduction

Let $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be an i.i.d. sample from distribution P_{XY} of a pair of random variables (X, Y), $X \in \mathcal{X}$, $Y \in \mathcal{Y}$ where $\mathcal{X}$ is any set and $\mathcal{Y}$ is a subset of $\mathbb{R}$. We consider the problem of prediction of Y given X. For any function $f : \mathcal{X} \to \mathcal{Y}$ called the predictor, we define the prediction risk under squared loss:

$$L(f) = \mathbb{E}_{XY}[(f(X) - Y)^2]$$

where $\mathbb{E}_{XY}$ is the expectation with respect to P_{XY}. Let now $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$ and assume that the aim is to mimic the best predictor in this class. This means that we want to find an estimator $\hat{f}$ based the sample D_n and having a small excess risk

$$L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f) \tag{1}$$

in expectation or with high probability. The minimizer of $L(f)$ over all measurable functions is the regression function $\eta(x) = \mathbb{E}_{XY}[Y|X = x]$ and it is straightforward to see that for the expected

excess risk we have

$$\mathbb{E}L(\hat{f}) - \inf_{f\in\mathcal{F}} L(f) = \mathbb{E}\left\|\hat{f}-\eta\right\|^2 - \inf_{f\in\mathcal{F}}\|f-\eta\|^2 \tag{2}$$

where $\mathbb{E}$ is the generic expectation sign, $\|f\|^2 = \int f^2(x)P_X(dx)$, and P_X denotes the marginal distribution of X. The left-hand side of (2) has been studied within Statistical Learning Theory characterizing the error of "agnostic learning" [44], [11], [23], while the object on the right-hand side has been the topic of oracle inequalities in nonparametric statistics [32], [39], and in the literature on aggregation [40], [36]. Upper bounds on the right-hand side of (2) are called *exact* oracle inequalities, which refers to constant 1 in front of the infimum over $\mathcal{F}$. However, some of the key results in the literature were only obtained with a constant greater than 1, i.e., they yield upper bounds for the difference

$$\mathbb{E}\left\|\hat{f}-\eta\right\|^2 - C\inf_{f\in\mathcal{F}}\|f-\eta\|^2 \tag{3}$$

with $C > 1$ and not for the excess risk. In this paper, we obtain exact oracle inequalities, which allows us to consider the excess risk formulation of the problem as described above.

In what follows we assume that $\mathcal{Y} = [0,1]$. For results in expectation, the extension to unbounded $\mathcal{Y}$ with some condition on the tails of the distribution is straightforward. For high probability statements, more care has to be taken, and the requirements on the tail behavior are more stringent. To avoid this extra level of complication, we assume boundedness.

From the minimax point of view, the object studied in statistical learning theory can be written as the *minimax regret*

$$V_n(\mathcal{F}) = \inf_{\hat{f}}\sup_{P_{XY}\in\mathcal{P}}\left\{\mathbb{E}L(\hat{f}) - \inf_{f\in\mathcal{F}} L(f)\right\} \tag{4}$$

where $\mathcal{P}$ is the set of all distributions on $\mathcal{X}\times\mathcal{Y}$ and $\inf_{\hat{f}}$ denotes the infimum over all estimators. We observe that the study of this object leads to a *distribution-free* theory, as no model is assumed. Instead, the goal is to achieve predictive performance competitive with a reference class $\mathcal{F}$. In view of (2), an equivalent way to write $V_n(\mathcal{F})$ is

$$V_n(\mathcal{F}) = \inf_{\hat{f}}\sup_{P_{XY}\in\mathcal{P}}\left\{\mathbb{E}\left\|\hat{f}-\eta\right\|^2 - \inf_{f\in\mathcal{F}}\|f-\eta\|^2\right\} \tag{5}$$

The expression in curly brackets in (5) can be viewed as a "distance" between the estimator $\hat{f}$ and the regression function η (which might lie outside of the specified set of models $\mathcal{F}$) defined through a comparison to the best possible performance within this set of models. Thus, the minimax regret can be interpreted as a measure of performance of estimators for misspecified models. The study of $V_n(\mathcal{F})$ will be further referred to as the misspecified models setting.

A special instance of the minimax regret has been studied in the context aggregation of estimators, with the aim to characterize optimal rates of aggregation, cf., e.g., [40, 36]. There, $\mathcal{F}$ is a subclass of the linear span of M given functions $f_1,\dots,f_M$, for example, their convex hull or sparse linear (convex) hull. Functions $f_1,\dots,f_M$ are interpreted as some initial estimators of the regression

function η based on another sample from the distribution of (X, Y). This sample is supposed to be independent from D_n and is considered as frozen when dealing with the minimax regret. The aim of aggregation is to construct an estimator $\hat{f}$, called the aggregate, that mimics the best linear combination of $f_1, \dots, f_M$ with coefficients of the combination lying in a given set in $\mathbb{R}^M$. Our results below apply to this setting as well and we will provide their consequences for some important examples of aggregation.

In the nonparametric regression setting, it is typically assumed that the model is well-specified specified, i.e., we have $Y_i = f(X_i) + \xi_i$ where the random errors ξ_i satisfy $\mathbb{E}(\xi_i | X_i) = 0$ and f belongs to a given functional class $\mathcal{F}$. Then $f = \eta$ and the infimum on the right-hand side of (2) is zero. The value of reference characterizing the best estimation in this problem is the *minimax risk*

$$W_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\hat{f} - f\|^2 \tag{6}$$

where $\mathbb{E}_f$ is the expectation w.r.t. the distribution of the sample D_n when $\mathbb{E}(Y|X) = f(X)$ for a fixed marginal distribution P_X and a fixed conditional distribution of $\xi = Y - f(X)$ given X. It is not difficult to see that

$$W_n(\mathcal{F}) \le V_n(\mathcal{F}),$$

yet the minimax risk and the minimax regret are quite different and it is not clear whether the two quantities can be of the same order for particular $\mathcal{F}$. The main message of this paper is to show that this is indeed the case, under an assumption on the behavior of the empirical entropy of $\mathcal{F}$ satisfied in many interesting examples. We also show that this assumption is tight in the sense that the minimax regret and the minimax risk can have different rates of convergence when it is violated.

Observe a certain duality between $W_n(\mathcal{F})$ and $V_n(\mathcal{F})$. In the former, the assumption about the reality is placed on the way data are generated. In the latter, no such assumption is made, yet the assumption is placed in the term that is being subtracted off. As we describe in Section 5, the study of these two quantities represents two parallel developments: the former has been a subject mostly studied within nonparametric statistics, while the second – within statistical learning theory. We aim to bring out a connection between these two objects.

The paper is organized as follows. In Section 3 we present our estimation procedure and the upper bounds on its risk. These include the main oracle inequality in Theorem 1 and its consequences given in Theorems 2-4. In Section 5, we compare the results to those in the literature. Section 6 is devoted to proving Theorems 2-4. The main part of the proof of Theorem 1 is in Section 8, with some technical results further postponed to Section 10. Lower bounds are proved in Section 9.

2 Notation

Set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For $S = \{z_1, \dots, z_n\} \in \mathcal{Z}^n$ and a class $\mathcal{G}$ of real-valued functions on $\mathcal{Z}$, consider the Rademacher average on $\mathcal{G}$:

$$\hat{\mathfrak{R}}_n(\mathcal{G}, S) = \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(z_i) \right]$$

where $\mathbb{E}_\sigma$ denotes the expectation with respect to the joint distribution of i.i.d. random variables $\sigma_1, \ldots, \sigma_n$ taking values 1 and -1 with probabilities $1/2$. Let

$$\mathfrak{R}_n(\mathcal{G}) = \sup_{S \in \mathcal{Z}^n} \hat{\mathfrak{R}}_n(\mathcal{G}, S).$$

For any $\epsilon > 0, 1 \le p < \infty$, $S = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$, we will denote by $\mathcal{N}_p(\mathcal{F}, \epsilon, S)$ the empirical ϵ-covering number of the class $\mathcal{F}$ with respect to the L_p pseudonorm

$$\left(\frac{1}{n} \sum_{i=1}^{n} |f(z_i)|^p \right)^{1/p},$$

and by $\mathcal{N}_\infty(\mathcal{F}, \epsilon)$ the ϵ-covering number of the class $\mathcal{F}$ with respect to the supremum norm.

Given $r > 0$, we denote by $\mathcal{G}[r, S]$ the set of functions in $\mathcal{G}$ with empirical average at most r on S:

$$\mathcal{G}[r, S] = \left\{ g \in \mathcal{G} : \frac{1}{n} \sum_{i=1}^{n} g(z_i) \le r \right\}.$$

We write $\ell \circ f$ for the function $(x, y) \mapsto (f(x) - y)^2$ and $\ell \circ \mathcal{F}$ for the class of functions $\{\ell \circ f : f \in \mathcal{F}\}$. Thus,

$$(\ell \circ \mathcal{F})[r, S] = \left\{ \ell \circ f : f \in \mathcal{F}, \quad \frac{1}{n} \sum_{i=1}^{n} (\ell \circ f)(x_i, y_i) \le r \right\}$$

for $S = \{z_1, \ldots, z_n\}$ with $z_i = (x_i, y_i)$. The minimum risk on the class of functions $\mathcal{F}$ is denoted by

$$L^* = \inf_{f \in \mathcal{F}} L(f).$$

The set $\{1, \ldots, N\}$ is denoted by $[N]$. Let $\lceil x \rceil$ denote the minimal integer strictly greater than x, and $|\mathcal{F}|$ the cardinality of $\mathcal{F}$. Notation C will be used for absolute positive constants that can vary on different occasions.

3 Main Results

In this section we introduce the estimator studied along the paper, state the main oracle inequality for its risk and provide corollaries for the minimax risk and minimax regret. The estimation procedure comprises three steps. The first step is to construct a random ϵ-net on $\mathcal{F}$ with respect to the empirical ℓ_2 metric and to form the induced partition of $\mathcal{F}$. The second step is to compute empirical risk minimizers (in our case, the least squares estimators) over each cell of this random partition. Finally, the third step is to aggregate these minimizers using a suitable aggregation procedure. If the radius ϵ of the initial net is taken to be large enough, the method reduces to a single empirical risk minimization (ERM) procedure over the class $\mathcal{F}$. While such an ERM procedure is known (or in some cases suspected) to be suboptimal, the proposed method enjoys optimal rates.

To ease the notation, assume that we have a sample D_{3n} of size $3n$ and we divide it into three parts: $D_{3n} = S \cup S' \cup S''$, where the subsamples S, S', S'' are each of size n. Fix $\epsilon > 0$. Let

$$d_S(f,g) = \sqrt{\frac{1}{n}\sum_{(x,y)\in S}(f(x)-g(x))^2}$$

be the empirical ℓ_2 pseudometric associated with the subsample S of cardinality n, and

$$N = \mathcal{N}_2(\mathcal{F}, \epsilon, S).$$

Let $\hat{c}_1, \dots, \hat{c}_N$ be an ϵ-net on $\mathcal{F}$ with respect to $d_S(\cdot,\cdot)$. We assume without loss of generality that it is *proper*, i.e., $\hat{c}_i \in \mathcal{F}$ for $i = 1, \dots, N$. Let $\hat{\mathcal{F}}_1^S, \dots, \hat{\mathcal{F}}_N^S$ be the following partition of $\mathcal{F}$ induced by $\hat{c}_i$'s:

$$\hat{\mathcal{F}}_i^S = \hat{\mathcal{F}}_i^S(\epsilon) = \left\{ f \in \mathcal{F} \ : \ i \in \operatorname*{argmin}_{j=1,\dots,N} d_S(f, \hat{c}_j) \right\}$$

with ties broken in an arbitrary way. Now, for each $\hat{\mathcal{F}}_i^S$, define the least squares estimators over the subsets $\hat{\mathcal{F}}_i^S$ with respect to the second subsample S':

$$\hat{f}_i^{S,S'} \in \operatorname*{argmin}_{f\in\hat{\mathcal{F}}_i^S} \ \frac{1}{n}\sum_{(x,y)\in S'}(f(x)-y)^2.$$

Finally, at the third step we use the subsample S'' to aggregate the estimators $\left\{\hat{f}_1^{S,S'}, \dots, \hat{f}_N^{S,S'}\right\}$. We call a function $\tilde{f}(x, D_{3n})$ with values in $\mathcal{Y}$ a *sharp MS-aggregate*[1] if it has the following property.

There exists a constant $C > 0$ such that, for any $\delta > 0$,

$$L\left(\tilde{f}\right) \le \min_{i=1,\dots,N} L\left(\hat{f}_i^{S,S'}\right) + C\frac{\log(N/\delta)}{n} \tag{7}$$

with probability at least $1-\delta$ over the sample S'', conditionally on $S \cup S'$.

Note that, in (7), the subsamples S, S' are fixed, so that the estimators $\hat{f}_i^{S,S'} \triangleq g_i$ can be considered as fixed (non-random) functions, and $\tilde{f}$ as a function of S'' only. There exist several examples of sharp MS-aggregates of fixed functions $g_1, \dots, g_N$ [2, 28]. They are realized as mixtures:

$$\tilde{f} = \sum_{i=1}^N \theta_i g_i = \sum_{i=1}^N \theta_i \hat{f}_i^{S,S'},$$

where θ_i are random weights measurable with respect to S''.

The next theorem contains the main oracle inequality for the aggregate $\tilde{f}$ constructed by this three-step procedure. To state the result, we will need some definitions. Consider the class of functions $\mathcal{G} = \{(f-g)^2 : f, g \in \mathcal{F}\}$. Let $\phi_n : [0,\infty) \mapsto \mathbb{R}$ be a function that satisfies

$$\sup_{S\in\mathcal{Z}^n} \hat{\mathfrak{R}}_n\Big(\mathcal{G}[r,S], S\Big) \le \phi_n(r) \tag{8}$$

[1]Here, MS-aggregate is an abbreviation for *model selection type aggregate*. The word *sharp* indicates that (7) is an oracle inequality with leading constant 1.

for all $r > 0$ and assume that ϕ_n is non-negative, non-decreasing, and $\phi_n(r)/\sqrt{r}$ is non-increasing. Let $r^* = r^*(\mathcal{G})$ denote an upper bound on the largest solution of $\phi_n(r) = r$. Define

$$\beta = \frac{\log(\mathcal{N}_2(\mathcal{F}, \epsilon, S)/\delta) + \log\log n}{n}, \quad \text{and} \quad \gamma = \sqrt{\epsilon^2 + r^* + \beta}.$$

Theorem 1. *Let $0 \le f \le 1$ for all $f \in \mathcal{F}$, and let $\tilde{f}$ be a sharp MS-aggregate defined by the above three-stage procedure. Fix $\epsilon > 0$. Then there exists an absolute constant C such that for any $\delta > 0$, with probability at least $1 - 3\delta$,*

$$L(\tilde{f}) \le \inf_{f \in \mathcal{F}} L(f) + C(\beta + \Xi(n, \epsilon, S')), \tag{9}$$

where

$$\Xi(n, \epsilon, S') = \min\Bigg\{ \gamma\sqrt{r^*} + \frac{1}{\sqrt{n}} \int_0^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, S')} d\rho \ , \tag{10}$$

$$\sqrt{L^*}\sqrt{\log^3(n)\mathfrak{R}_n^2(\mathcal{F}) + \beta} \Bigg\}. \tag{11}$$

Furthermore, if $\mathcal{F}$ is a convex subset of a d-dimensional linear subspace of $L_2(P_X)$ then, with probability at least $1 - 3\delta$, inequality (9) holds with the remainder term

$$\Xi(n, \epsilon, S') = \min\Bigg\{ \frac{d}{n}, \ \gamma\sqrt{r^*} + \frac{1}{\sqrt{n}} \int_0^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, S')} d\rho \ , \tag{12}$$

$$\sqrt{L^*}\sqrt{\log^3(n)\mathfrak{R}_n^2(\mathcal{F}) + \beta} \Bigg\}.$$

Remarks.

1. The term $\Xi(n, \epsilon, S')$ in Theorem 1 is a bound on the rate of convergence of the ERM $\hat{f}_i^{S,S'}$ over the cell $\hat{\mathcal{F}}_i^S$. It is the minimum of three possible rates, the first of which (10) we prove in Section 8 (the main crux of this paper), the second (11) is due to [38], while the third term d/n present only for convex d-dimensional $\mathcal{F}$ is due to [23]. If, in particular instances, there exists a sharper bound for the rate of ERM, as it will be the case in some examples below, one can readily use this bound instead of the expressions for $\Xi(n, \epsilon, S')$ given in Theorem 1.

2. The partitions $\hat{\mathcal{F}}_i^S$ defined above can be viewed as a default option. In some situations, we may better taylor the partitions to the geometry of $\mathcal{F}$. For instance, in the aggregation context (cf. Theorem 4 below), $\mathcal{F}$ is union of convex sets. We choose each convex set as an element of the partition, and use the rate for ERM over individual convex sets instead of the overall rate $\Xi(n, \epsilon, S')$. In this case, the partition is non-random. It is also important to note that in Theorem 1 we can use the localization radius $r^* = r^*(\hat{\mathcal{G}}_i)$ for $\hat{\mathcal{G}}_i = \{(f-g)^2 : f, g \in \hat{\mathcal{F}}_i^S\}$ instead of the larger quantity $r^*(\mathcal{G})$. Inspection of the proof shows that the oracle inequality (9) generalizes to

$$L(\tilde{f}) \le \min_{i=1,\dots,N} \inf_{f \in \hat{\mathcal{F}}_i^S} \big\{ L(f) + C(\beta + \Xi_i(n, \epsilon, S')) \big\}, \tag{13}$$

   where $\Xi_i(n, \epsilon, S')$ is defined in the same way as $\Xi(n, \epsilon, S')$ with the only difference that $r^*(\mathcal{G})$ is replaced by $r^*(\hat{\mathcal{G}}_i)$.

The oracle inequality (9) of Theorem 1 depends on three quantities that should be specified: the empirical entropy numbers $\log \mathcal{N}_2(\mathcal{F}, \cdot, \cdot)$, the optimal localization radius r^* and the Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$. The crucial role in determining the rate belongs to the empirical entropies. We further replace in (9)-(12) these random entropies by their upper bound

$$\mathcal{H}_2(\mathcal{F}, \rho) = \sup_{S \in \mathcal{Z}^n} \log \mathcal{N}_2(\mathcal{F}, \rho, S).$$

The next theorem is a corollary of Theorem 1 in the case of polynomial growth of the empirical entropy. It gives upper bounds on the minimax regret and on the minimax risk derived from (9).

Theorem 2. *Assume that $\mathcal{Y} = [0,1]$ and the empirical entropy satisfies $\mathcal{H}_2(\mathcal{F}, \rho) \le A\rho^{-p}$, $\forall\, \rho > 0$, for some constants $A < \infty$, $p > 0$. Let $\tilde{f}$ be a sharp MS-aggregate defined by the above three-stage procedure with the covering radius $\epsilon > 0$. Then there exist constants $C_p > 0$ depending only on A and p such that the following holds.*

(i) Let $0 \le f \le 1$ for all $f \in \mathcal{F}$. For the estimator $\tilde{f}$ constructed with

$$\begin{aligned} \epsilon &= n^{-\frac{1}{2+p}} \quad && \text{if } p \in (0,2], \\ \epsilon &\ge n^{-\frac{p-1}{p^2}} \quad && \text{if } p \in (2,\infty), \end{aligned}$$

we have

$$V_n(\mathcal{F}) \le \sup_{P_{XY} \in \mathcal{P}} \left\{ \mathbb{E} \left\| \tilde{f} - \eta \right\|^2 - \inf_{f \in \mathcal{F}} \left\| f - \eta \right\|^2 \right\} \le \begin{cases} C_p n^{-\frac{2}{2+p}} & \text{if } p \in (0,2], \\ C_p n^{-\frac{1}{p}} & \text{if } p \in (2,\infty). \end{cases} \tag{14}$$

(ii) If the model is well-specified, then for the estimator $\tilde{f}$ with $\epsilon = n^{-\frac{1}{2+p}}$ we have

$$W_n(\mathcal{F}) \le \sup_{f \in \mathcal{F}} \mathbb{E} \left\| \tilde{f} - f \right\|^2 \le C_p n^{-\frac{2}{2+p}}, \qquad \forall\ p > 0. \tag{15}$$

Proof of Theorem 2 is given in Section 6. An interesting consequence of this theorem is that the minimax risk $W_n(\mathcal{F})$ achieves faster convergence than the minimax regret $V_n(\mathcal{F})$ for $p > 2$, i.e., for classes $\mathcal{F}$ of very high complexity. Theorem 2 provides only an upper bound. However, it turns out to be tight as shown in Section 9.

Observe also that in both cases, $p \in (0,2]$ and $p \in (2,\infty)$, we can use the same value $\epsilon = n^{-1/(2+p)}$ to obtain the rates given in (14). We remark that this ϵ satisfies the bias-variance balance relation

$$n\epsilon^2 \asymp \mathcal{H}_2(\mathcal{F}, \epsilon).$$

We will further comment on this choice in Section 5.

We now turn to the consequences of Theorem 1 for low complexity classes $\mathcal{F}$, such as Vapnik-Chervonenkis (VC) classes and intersections of balls in finite-dimensional spaces. They roughly correspond to the case "$p \approx 0$", and the rates for the minimax risk $W_n(\mathcal{F})$ are the same as for the minimax regret $V_n(\mathcal{F})$.

Assume first that the empirical covering numbers of $\mathcal{F}$ exhibit the growth

$$\sup_{S \in \mathcal{Z}^n} \mathcal{N}_2(\mathcal{F}, \rho, S) \le (A/\rho)^v, \tag{16}$$

$\forall\, \rho > 0$, with some constants $A < \infty$, $v > 0$. Such classes $\mathcal{F}$ are called VC-type classes. Examples include the VC-subgraph classes with VC-dimension v, i.e., classes of functions f whose subgraphs $C_f = \{(x,t) \in \mathcal{X} \times \mathbb{R} : f(x) \geq t\}$ form a Vapnik-Chervonenkis class with VC-dimension v.

Theorem 3. (Bounds for VC-type classes). *Assume that $\mathcal{Y} = [0,1]$ and the empirical covering numbers satisfy (16). Let $0 \leq f \leq 1$ for all $f \in \mathcal{F}$, and let $\tilde{f}$ be a sharp MS-aggregate defined by the above three-stage procedure with $\epsilon = n^{-\frac{1}{2}}$. If $n \geq v$, there exists a constant $C > 0$ depending only on A such that*

$$V_n(\mathcal{F}) \leq \sup_{P_{XY} \in \mathcal{P}} \left\{ \mathbb{E} \left\| \tilde{f} - \eta \right\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\} \leq C \frac{v(1 + \log(n/v))}{n}. \tag{17}$$

The rate of convergence of the excess risk as in (17) for VC-type classes has been obtained previously under the assumption that $L^* = 0$ or for convex classes $\mathcal{F}$ (see discussion in Section 5 below). Theorem 3 does not rely on either of these assumptions.

In Section 9.1 we show that the bound of Theorem 3 is tight: there exists a function class such that, for any estimator, there exists a distribution on which the estimator differs from the regression function by at least $C\frac{v(1+\log(n/v))}{n}$ with positive fixed probability. So, the extra logarithmic term in the rate is necessary, even when the model is well-specified.

The next theorem deals with classes of functions

$$\mathcal{F} = \mathcal{F}_\Theta \triangleq \left\{ \mathsf{f}_\theta = \sum_{i=1}^M \theta_j f_j : \theta = (\theta_1, \ldots, \theta_M) \in \Theta \right\}$$

where $\{f_1, \ldots, f_M\}$ is a given collection of M functions on $\mathcal{X}$ with values in $\mathcal{Y}$, and $\Theta \subseteq \mathbb{R}^M$ is a given set of possible mixing coefficients θ. Such classes arise in the context of aggregation, cf., e.g., [40], [36], where the main problem is to study the behavior of the minimax regret $V_n(\mathcal{F}_\Theta)$ based on the geometry of Θ. For the case of fixed rather than random design, we refer to [36] for a comprehensive treatment. Here, we deal with the random design case and consider several basic examples of sets Θ defined in terms of ℓ_p-balls

$$B_p(r) = \{\theta \in \mathbb{R}^M : |\theta|_p \leq r\}, \quad 0 \leq p < \infty,\ r > 0,$$

where $|\theta|_0$ denotes the number of non-zero components of θ, and $|\theta|_p = \left(\sum_{j=1}^M |\theta_j|^p\right)^{1/p}$ for $0 < p < \infty$. We will also consider the probability simplex

$$\Lambda_M = \left\{ \theta \in \mathbb{R}^M : \sum_{j=1}^M \theta_j = 1,\ \theta_j \geq 0, j = 1 \ldots, M \right\}.$$

Then, model selection type aggregation (or *MS-aggregation*) consists in constructing an estimator $\tilde{f}$ that mimics the best function among $f_1, \ldots, f_M$, i.e., the function that attains the minimum $\min_{j=1,\ldots,M} \|f_j - \eta\|^2$. In this case, $\mathcal{F}_\Theta = \{f_1, \ldots, f_M\}$ or equivalently $\Theta = \Theta^{\mathrm{MS}} \triangleq \{\mathbf{e}_1, \ldots, \mathbf{e}_M\} = \Lambda_M \cap B_0(1)$, where $\mathbf{e}_1, \ldots, \mathbf{e}_M$ are the canonical basis vectors in $\mathbb{R}^M$. *Convex aggregation* (or *C-aggregation*) consists in constructing an estimator $\tilde{f}$ that mimics the best function in the convex hull $\mathcal{F} = \mathrm{conv}(f_1, \ldots, f_M)$, i.e., the function that attains the minimum $\min_{\theta \in \Lambda_M} \|\mathsf{f}_\theta - \eta\|^2$. In this case, $\mathcal{F} = \mathcal{F}_\Theta$ with $\Theta = \Theta^{\mathrm{C}} \triangleq \Lambda_M$. Finally, given an integer $1 \leq s \leq M$, the *s-sparse convex*

aggregation consists in mimicking the best convex combination of at most s among the functions $f_1, \dots, f_M$. This corresponds to the set $\Theta^{\mathrm{C}}(s) = \Lambda_M \cap B_0(s)$. Note that MS-aggregation and convex aggregation are particular cases of s-sparse aggregation: $\Theta^{\mathrm{MS}} = \Theta^{\mathrm{C}}(1)$ and $\Theta^{\mathrm{C}} = \Theta^{\mathrm{C}}(M)$.

For the aggregation setting, we modify the definition of cells $\hat{\mathcal{F}}_i^S$ as discussed in Remark 2. Consider the partition $\Theta^{\mathrm{C}}(s) = \bigcup_{m=1}^{s} \bigcup_{\nu \in I_m} \mathcal{F}_{\nu,m}$ where I_m is the set of all subsets ν of $\{1, \dots, M\}$ of cardinality $|\nu| = m$, and $\mathcal{F}_{\nu,m}$ is the convex hull of f_j's with indices $j \in \nu$. We use the deterministic cells

$$\{\mathcal{F}_1, \dots, \mathcal{F}_N\} = \{\mathcal{F}_{\nu,m},\, m = 1, \dots, s,\, \nu \in I_m\}$$

instead of random ones $\hat{\mathcal{F}}_i^S$. Note that the subsample S is not involved in this construction. We keep all the other ingredients of the estimation procedure as described at the beginning of this section, and we denote the resulting estimator $\tilde{f}$. Then, using the subsample S, we complete the construction by aggregating only two estimators, $\tilde{f}$ and the ERM on Λ_M. The resulting aggregate is denoted by $\tilde{f}^*$.

Theorem 4. (Bounds for aggregation). *Let $\mathcal{Y} = [0,1]$ and $0 \le f_j \le 1$ for $j = 1, \dots, M$. Then there exists an absolute constant $C > 0$ such that*

$$V_n(\mathcal{F}_{\Theta^{\mathrm{C}}(s)}) \le \sup_{P_{XY} \in \mathcal{P}} \left\{ \mathbb{E} \left\| \tilde{f}^* - \eta \right\|^2 - \inf_{\theta \in \Theta^{\mathrm{C}}(s)} \|\mathsf{f}_\theta - \eta\|^2 \right\} \le C \psi_{n,M}(s) \tag{18}$$

where

$$\psi_{n,M}(s) = \frac{s}{n} \log\left(\frac{eM}{s}\right) \wedge \sqrt{\frac{1}{n} \log\left(\frac{eM}{\sqrt{n}}\right)}.$$

for $s \in \{1, \dots, M\}$.

This theorem improves upon the rate of s-sparse aggregation given in Lounici [30] by removing a redundant $(s/n) \log n$ term present there. Note that [30] considers the random design regression model with gaussian errors. Theorem 4 is distribution-free and deals with bounded errors as all the results of this paper and it can be readily extended to the sub-exponential case. By an easy modification of the minimax lower bound given in [30], we get that $\psi_{n,M}(s)$ is the optimal rate for the minimax regret on $\mathcal{F}_{\Theta^{\mathrm{C}}(s)}$ in our setting. Analogous result for gaussian regression with fixed design is proved in [36].

4 ERM in partitions versus ERM and aggregation of centers

The estimation procedure we propose here has three steps. The first is to find an empirical ϵ-net on the first part of the sample and partition the function class based on the centers (the cover functions) using the empirical distance on the first sample. In the next step, using the second sample we find empirical risk minimizers within each partition. Finally, we use the third sample to aggregate over the ERM's within each of the partitions. The estimation procedure for the well-specified case proposed by Yang and Barron [46] consists of steps one and three, but not two. This method directly aggregates centers of the partitions, ie. the covers obtained from the first sample split. While this procedure works for the well-specified case, one cannot expect this to directly work for the misspecified case. The step of finding ERM's in each partition and aggregating over

the ERM's is rather crucial. The reason the procedure of aggregating over the centers works for the well specified case is because for the the partition $\hat{\mathcal{F}}_i^S(\epsilon)$ that contains the regression function η, i.e., $\eta \in \hat{\mathcal{F}}_i^S(\epsilon)$, we have that

$$\begin{aligned}\|\hat{c}_i - \eta\|^2 - \inf_{f\in\mathcal{F}} \|f-\eta\|^2 = \|\hat{c}_i - \eta\|^2 \\ &\leq 2\ d_S^2(\hat{c}_i, \eta) + \tilde{O}(\mathfrak{R}_n^2(\mathcal{F})) \\ &\leq 2\epsilon^2 + \tilde{O}(\mathfrak{R}_n^2(\mathcal{F}))\end{aligned}$$

Hence trading off the ϵ^2 with the $\log \mathcal{N}_2(\mathcal{F}, \epsilon, S)/n$ from aggregation procedure gives the optimal rate for the well-specified case (the Rademacher squared term is a lower order term). The reason why aggregating over the centers of the partitions fail for the misspecified case is because without the assumption that the regression function is in the function class, even for the partition $\hat{\mathcal{F}}_i^S(\epsilon)$ containing the minimizer $\operatorname{argmin}_{f\in\mathcal{F}} \|f-\eta\|^2$, we can at best only have

$$\begin{aligned}\|\hat{c}_i - \eta\|^2 - \inf_{f\in\mathcal{F}} \|f-\eta\|^2 \leq 8\,\|\hat{c}_i - \eta\| \leq 8\sqrt{2d_S^2(\hat{c}_i, f^*) + \tilde{O}(\mathfrak{R}_n^2(\mathcal{F}))} \\ &\leq 8\sqrt{2\epsilon^2 + \tilde{O}(\mathfrak{R}_n^2(\mathcal{F}))} \leq 8\sqrt{2}\epsilon + \tilde{O}(\mathfrak{R}_n(\mathcal{F}))\end{aligned} \tag{19}$$

Hence aggregating over the centers we get the tradeoff of ϵ with the $\log \mathcal{N}_2(\mathcal{F}, \epsilon, S)/n$ which can only give the sub-optimal rate of $n^{-1/(p+1)} + O(\mathfrak{R}_n(\mathcal{F}))$. This indicates that the ERM over partitions step is rather crucial in getting the right rates.

In summary, we have three estimation procedures we can consider. First, simple Empirical Risk Minimization (ERM) procedure, second aggregating over cover centers ([46]) and, finally, the proposed procedure of aggregating over ERM's within each of the partitions. Due to Equation (19), for classes other than finite function class (when centers coincide with all the M function in the function class $\mathcal{F}$) we always pay an additive factor of $\tilde{O}(\mathfrak{R}_n(\mathcal{F}))$ for the procedure of aggregating over centers for the misspecified model. Hence, other than for the finite case, the procedure of aggregating over cover centers always has worse rate than ERM. The following table summarizes the rates for misspecified case (statistical learning).

Regime	Proposed Method	Aggregating centers [46]	ERM
(Finite) $\lvert\mathcal{F}\rvert = M$	$\frac{\log M}{n}$	$\frac{\log M}{n}$	$\sqrt{\frac{\log M}{n}}$
(parametric) $VC(\mathcal{F}) = d$	$\frac{d\log(n)}{n}$	$\sqrt{\frac{d\log(n)}{n}}$	$\sqrt{\frac{d}{n}}$
$\log \mathcal{N}_2(\mathcal{F}, \epsilon) = \epsilon^{-p}$, $p \in (0,2)$	$n^{-\frac{2}{2+p}}$	$n^{-\frac{1}{2}}$	$n^{-\frac{1}{2}}$
$p \in [2, \infty)$	$n^{-\frac{1}{p}}$	$n^{-\frac{1}{p+1}}$	$n^{-\frac{1}{p}}$

Table 1: Summary of Rates for Misspecified case

Notice that in the above we see that for finite class case the proposed method and aggregation over centers are optimal whereas ERM has a suboptimal rate. For the parametric case, while the

proposed method is optimal, both ERM and aggregating over centers is suboptimal. When the log covering number grows polynomially as ϵ^{-p}, for the case when $p \geq 2$ both ERM and proposed method enjoy similar guarantees of rates of order $n^{-1/p}$ while the aggregation over centers only gets a sub-optimal rate of $n^{-1/(p+1)}$. The case of $p \in (0, 2)$, the proposed method is optimal and achieves a rate of $n^{-2/(2+p)}$ while both aggregation over centers and ERM procedures only achieve a rate of $n^{-1/2}$.

Turning to the well-specified case, both the proposed method and the procedure of aggregating over centers both achieve the optimal rates of $n^{-2/(2+p)}$ while ERM is suboptimal in general.

5 Historical Remarks and Comparison to Previous Work

The literature on nonparametric estimation from the statistics community and on excess risk bounds from statistical learning theory is vast, and we will only attempt to briefly describe the results most relevant to this paper.

The role of entropy and capacity [20] in establishing rates of estimation has been recognized for a long time, since the work of Le Cam [25], Ibragimov and Khas'minskii [17] and Birgé [6]. Other early work on this subject involving estimation on ϵ-nets is due to Devroye [10] and Devroye et al. [11]. The common point is that optimal rate is obtained as a solution to the bias-variance balancing equation $n\epsilon^2 = \mathcal{H}(\epsilon)$, with a conveniently chosen non-random entropy $\mathcal{H}(\cdot)$. Van de Geer [41] invokes the empirical entropy rather than the non-random entropy to derive rates of estimation in regression problems. he entropy of the set. In particular, th! Yang and Barron [46] present a general approach to obtain lower bounds from global (rather than local) capacity properties of the parameter set. Once again, the optimal rate is shown to be a solution to the bias-variance balancing equation described above, with a generic notion of a metric on the parameter space. Under the assumption that the regression errors are gaussian, [46] also provides an achievability result, a procedure inspired by information-theoretic considerations. This procedure is quite different from empirical risk minimization: it averages the predictive distributions corresponding to a covering of the parameter space.

In all these works, it is assumed that the unknown density, regression function, or parameter belongs to the given class, i.e., the model is correctly specified. In parallel to these developments, a line of work on pattern recognition that can be traced to Aizerman, Braverman and Rozonoer [1] and Vapnik and Chervonenkis [44] focused on a different objective, which is characteristic for the statistical learning. Without assuming a form of the distribution that encodes the relationship between the predictors and outputs, the goal is formulated as that of performing as well as the best function within a given set of rules, with the excess risk as the measure of performance (rather than distance to the true underlying function). Thus, no assumption is placed on the underlying distribution. In this form, the problem can be cast as a special case of stochastic optimization and can be solved either via recurrent (e.g. gradient descent) meth! ods or via empirical risk minimization. The latter approach leads to the question of uniform convergence of averages to expectations, also called the uniform Glivenko-Cantelli property. This property is, once again, closely related to entropy of the class, and sufficient conditions have been extensively studied (see [14, 33, 13, 15, 12] and references therein).

For "parametric" classes with a polynomial growth of covering numbers, uniform convergence of

averages to expectations has been shown by Vapnik and Chervonenkis [42, 43, 44]. In the context of classification, they also obtained a faster rate showing $O(1/n)$ convergence when the minimal risk $L^* = 0$. For regression problems, similar fast rate has been obtained in [34, 16]. Lee, Bartlett and Williamson [29] showed $O(\log(n)/n)$ rates for the excess risk without the assumption $L^* = 0$. Instead, they assumed that the class $\mathcal{F}$ is convex and has finite pseudo-dimension. Additionally, it was shown that the $n^{-1/2}$ rate cannot be improved if the class is non-convex and the estimator is a selector (that is, forced to take values in $\mathcal{F}$). In particular, the excess risk of ERM and of any selector on a finite class $\mathcal{F}$ the cannot decrease faster than $\sqrt{(\log|\mathcal{F}|)/n}$ [18]. Optimality of ERM for certain problems is still an open question.

Independently of this work on the excess risk in the distribution-free setting of statistical learning, Nemirovskii [32] proposed to study the problem of aggregation, or mimicking the best function in the given class, for regression models. Nemirovskii [32] outlined three problems: model selection, convex aggregation, and linear aggregation. The notion of optimal rates of aggregation is introduced in [40], along with the derivation of the optimal rates for the three problems. In the following decade, much work has been done on understanding these and related aggregation problems [45, 19, 18, 30, 36]. For recent developments and a survey we refer to [27, 37].

In parallel with these developments, the study of the excess risk blossomed with the introduction of Rademacher and local Rademacher averages in [21, 24, 3, 8, 4, 22]. These techniques provided a good understanding of the behavior of the ERM method. In particular, if $\mathcal{F}$ is a *convex* subset of d-dimensional space, Koltchinskii [22, 23] obtained the exact inequality with the correct rate d/n for ERM. However, the convexity assumption appears to be crucial; without this assumption Koltchinskii [23, Theorem 5.2] obtains for ERM only a non-exact inequality with factor $C > 1$ in front of the infimum (see (3)).

Among a few of the estimators considered in the literature for general classes $\mathcal{F}$, empirical risk minimization on $\mathcal{F}$ has been one of the most studied. As mentioned above, ERM and other selector methods are suboptimal when the class $\mathcal{F}$ is finite. Given the optimality of rates for ERM when $\mathcal{F}$ is convex, it was conjectured that the correct rates for a finite $\mathcal{F}$ will be attained by an ERM on the convex hull of $\mathcal{F}$. This was disproved by Lecué and Mendelson [28]. For the regression setting, the approach that was found to achieve the optimal rate for the excess risk in expectation is through exponential weights with averaging of the trajectory. However, Audibert [2] showed that, for the regression with random design, exponential weighting is deviation suboptimal and proposed an alternative method which involved finding an ERM on a star connecting an overall ERM and the other $|\mathcal{F}| - 1$ functions. Thus, the optimal mixt! ure uses two functions. In [28], the authors also exhibited a deviation optimal method which involves sample splitting. The first part of the sample is used to localize a convex subset around ERM and the second – to find an ERM within this subset.

We close this short summary with a connection to a different literature. In the context of prediction of deterministic individual sequences with logarithmic loss, Cesa-Bianchi and Lugosi [9] considered regret with respect to rich classes of "experts". They showed that mixture of densities is suboptimal and proposed a two-level method where the rich set of distributions is divided into small balls, the optimal algorithm is run on these balls, and then the overall output is an aggregate of these outputs. They derived a bound where the upper bound of the Dudley integral is the radius of the balls. This method served as an inspiration for the present work.

6 Proofs of Theorems 2-4

Theorems 2, 3 and 4 follow from Theorem 1, Lemma 8 and the control of the critical radius r^* given in the following lemma.

Lemma 5. *The following critical radii $r^* = r^*(\mathcal{G})$ satisfy the conditions stated before Theorem 1.*

- *For any class $\mathcal{F}$,*
$$r^* = C \log^3(n) \mathfrak{R}_n^2(\mathcal{F}). \tag{20}$$
- *If the empirical covering numbers exhibit the polynomial growth $\sup_{S \in \mathcal{Z}^n} \mathcal{N}_2(\mathcal{F}, \rho, S) \le \left(\frac{A}{\rho}\right)^v$ for some constants $A < \infty$ and $v > 0$, $n \ge v$, then*
$$r^* = C \frac{v(1 + \log(n/v))}{n}.$$
- *If $\mathcal{F}$ is a finite class,*
$$r^* = C \frac{\log |\mathcal{F}|}{n}.$$

We will also use the following bound on the Rademacher average in terms of the empirical entropy [**?**]:

$$\hat{\mathfrak{R}}_n(\mathcal{F}, S) \le \inf_{\alpha \ge 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, S)} d\rho \right\}. \tag{21}$$

6.1 Proof of Theorem 2

Assume without loss of generality that $A = 1$, i.e., $\sup_{S \in \mathcal{Z}^n} \log \mathcal{N}_2(\mathcal{F}, \rho, S) \le \rho^{-p}$ for some $p > 0$. We consider separately the cases $p \in (0, 2]$ and $p > 2$.

The regime $p \in (0, 2]$. If $p \in (0, 2)$, the bound (21) with $\alpha = 0$ yields

$$\hat{\mathfrak{R}}_n(\mathcal{F}, S) \le \frac{12}{\sqrt{n}(1 - p/2)}$$

and thus $\mathfrak{R}_n^2(\mathcal{F}) \le c/n$ with $c = 144/(1 - p/2)^2$. This and (20) imply that $r^* \le C \log^3(n)/n$ for some absolute constant C. Next,

$$\beta \le \frac{C\left(\epsilon^{-p} + \log\log n + \log(1/\delta)\right)}{n}$$

and

$$\gamma^2 \le C\left(\epsilon^2 + \frac{(\log n)^3}{n} + \frac{\epsilon^{-p} + \log(1/\delta)}{n}\right). \tag{22}$$

So

$$\gamma \sqrt{r^*} \le C (\log n)^{3/2} \left(\frac{\epsilon}{\sqrt{n}} + \frac{(\log n)^{3/2}}{n} + \frac{\epsilon^{-p/2} + \sqrt{\log(1/\delta)}}{n} \right).$$

These inequalities together with (9) and (10) yield that, with probability at least $1-3\delta$,

$$L\left(\tilde{f}\right) - L^* \le C\left(\frac{\epsilon^{-p}}{n} + \frac{\log(1/\delta)}{n} + \gamma\sqrt{r^*} + \frac{\gamma^{1-p/2}}{\sqrt{n}}\right). \tag{23}$$

The value of ϵ minimizing the right hand side in (22) and in (23) is given by solving $\epsilon^2 \asymp 1/(\epsilon^p n)$, so $\epsilon = n^{-1/(2+p)}$ provides the correct rate. Notably, the logarithmic factor arising from r^* only appears together with the lower order terms and the summand $\gamma\sqrt{r^*}$ does not affect the rate. By choosing $\epsilon = n^{-1/(2+p)}$ we guarantee that the right hand side of (23) is $Cn^{-\frac{2}{2+p}}$ ignoring the terms with $\log(1/\delta)$ that disappear when passing from the bound in probability to that in expectation. Thus, the expected excess risk is bounded by $Cn^{-\frac{2}{2+p}}$, which proves (14) for $p \in (0,2)$.

For $p = 2$, the bounds on the Rademacher complexity and on r^* involve an extra logarithmic factor, which does not affect the final rate as it goes with lower order terms.

The regime $p \in (2,\infty)$. For $p \in (2,\infty)$, there is a difference between the rates under well-specified models (15) and misspecified models (14), so we consider the two cases separately.

1. Proof of (14) for $p \in (2,\infty)$. Here, the rate is governed by the Rademacher complexity of the function class. Using (21) we bound $\mathfrak{R}_n(\mathcal{F})$ as follows:

$$\mathfrak{R}_n(\mathcal{F}) \le \inf_{\alpha\ge 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \rho^{-p/2}d\rho\right\} \le \inf_{\alpha\ge 0}\left\{4\alpha + \frac{24}{\sqrt{n}(p-2)}\alpha^{-(p-2)/2}\right\}.$$

Balancing $\alpha = n^{-1/2}\alpha^{-(p-2)/2}$ yields $\alpha = n^{-1/p}$ and

$$\mathfrak{R}_n(\mathcal{F}) \le Cn^{-1/p}. \tag{24}$$

Thus, Lemma 8 implies

$$\mathbb{E}L(\hat{f}_i^{S,S'}) - \inf_{f\in\hat{\mathcal{F}}_i^S} L(f) \le Cn^{-1/p}.$$

The aggregation step (7) adds to this bound the term $\log\mathcal{N}_2(\mathcal{F},\epsilon,S)/n \le 1/(\epsilon^p n)$, so that

$$\mathbb{E}L(\tilde{f}) - \inf_{f\in\mathcal{F}} L(f) \le C(n^{-1/p} + 1/(\epsilon^p n)).$$

We conclude that for the case $p > 2$ we can use any $\epsilon \ge n^{-(p-1)/p^2}$ to obtain the overall rate $n^{-1/p}$.

2. Proof of (15) for $p \in (2,\infty)$. Now consider the case when $\eta \in \mathcal{F}$. We show that the method introduced in this paper enjoys optimal rates for any $p > 0$. Given S, let i^* be the index of the partition containing the regression function: $\eta \in \hat{\mathcal{F}}_{i^*}^S$. By definition, for any $f \in \hat{\mathcal{F}}_{i^*}^S$, $d_S(f,\eta) \le \epsilon$. Consider the set $\mathcal{G} = \{g = (f-\eta)^2 : f \in \mathcal{F}\}$. By Theorem 14 applied to this class, with probability at least $1-4\delta$ for all $g \in \mathcal{G}$

$$Pg \le 2P_n g + C(r^* + \beta')$$

for $\beta' = \frac{\log(1/\delta)+6\log\log n}{n}$ and some constant C. In other words,

$$P(f-\eta)^2 \le 2P_n(f-\eta)^2 + C(r^* + \beta').$$

Under this event, for any $f \in \hat{\mathcal{F}}_{i^*}^S$,

$$\|f - \eta\|^2 \le 2\epsilon^2 + C(r^* + \beta')$$

and hence (intersecting with the event of Eq. (7))

$$\begin{aligned} L\left(\tilde{f}\right) &\le \inf_{i \in [N]} L\left(\hat{f}_i^{S,S'}\right) + \frac{C\log(N/\delta)}{n} \\ &= \inf_{i \in [N]} \|\hat{f}_i^{S,S'} - \eta\|^2 + L(\eta) + \frac{C\log(N/\delta)}{n} \\ &\le L(\eta) + \frac{C(\epsilon^{-p} + \log(1/\delta))}{n} + 2\epsilon^2 + C(r^* + \beta') \end{aligned}$$

with probability at least $1 - 5\delta$. The expression in the last line of this display has the best rate for $\epsilon = n^{-1/(2+p)}$. From (24) and (20) we get that $r^* \le C\log^3(n) n^{-2/p}$. Thus, $r^* + \beta'$ is of smaller order than the other terms when $\epsilon = n^{-1/(2+p)}$. Taking into account that $L\left(\tilde{f}\right) - L(\eta) = \|f - \eta\|^2$ we obtain the overall rate of $n^{-\frac{2}{2+p}}$ for $\mathbb{E}\|f - \eta\|^2$ uniformly over $\eta \in \mathcal{F}$ as claimed.

6.2 Proof of Theorem 3

As shown in Lemma 5,

$$r^* = C\frac{v(1 + \log(n/v))}{n}.$$

Here, we can replace $(1 + \log(n/v))$ by $\log(n/v)$ if $n \ge av$ for $a > 0$ large enough (it is easy to see that it suffices to consider this case). Choosing $\epsilon = n^{-1/2}$, we get $\beta \le Cv\log(n/\delta)/n$, and $\gamma \le C\sqrt{\frac{v\log(n/(v\delta))}{n}}$. The overall rate in expectation is then $O\left(\frac{v\log(n/v)}{n}\right)$.

6.3 Proof of Theorem 4

As shown in [26], the rate of ERM for the simplex in $\mathbb{R}^s$ is

$$O\left(\frac{s}{n} \wedge \sqrt{\frac{\log(es/\sqrt{n})}{n}}\right).$$

The same result yields that the rate is not worse than

$$\sqrt{\frac{\log(eM/\sqrt{n})}{n}}$$

since we add in the aggregation procedure the ERM on the convex hull of all the M functions f_j. Since the number of such subsets is $N = \sum_{j=1}^s \binom{M}{j} \le \left(\frac{eM}{s}\right)^s$, we obtain the overall rate of the order

$$\left[\frac{s\log(eM/s)}{n} \vee \left(\frac{s}{n} \wedge \sqrt{\frac{\log(es/\sqrt{n})}{n}}\right)\right] \wedge \sqrt{\frac{\log(eM/\sqrt{n})}{n}} = \left(\frac{s\log(eM/s)}{n} \wedge \sqrt{\frac{\log(eM/\sqrt{n})}{n}}\right).$$

7 Adapting to Approximation Error Rate of Function Class

Often in statistical learning problems the choice of function class $\mathcal{F}$ is not prefixed and is in fact a design choice. The art of picking the right function class $\mathcal{F}$ to use depends on how best we can trade-off statistical learning rate for the function class with its approximation rate of how well it approximates the bays optimal predictor η. In the Theorem 2 we have shown that our estimator has the rate of $n^{-\frac{2}{2+p}}$ when the regression function η is in the function class $\mathcal{F}$ and achieves the rate $n^{-1/p}$ if not. A natural question one can ask is, what if $\eta \notin \mathcal{F}$ but then the approximation error rate $\inf_{f\in\mathcal{F}} \|\eta - f\|^2$ is small. In this case one would like to get rates varying from $n^{-1/p}$ all the way to $n^{-2/(2+p)}$ based on how small the approximation error rate is.

Let us start with defining the approximation error as $\Delta^2 := \inf_{f\in\mathcal{F}} \|f-\eta\|^2$. Further define $f^* = \operatorname{argmin}_{f\in\mathcal{F}} \|f-\eta\|^2$. Note that for any ϵ, for the the partition $\hat{\mathcal{F}}_i^S(\epsilon)$ that contains f^* we have that

$$\begin{aligned}\left\|\hat{f}_i^{S,S'} - \eta\right\|^2 - \|f^* - \eta\|^2 &\le 2\left\|\hat{f}_i^{S,S'} - f^*\right\|^2 + \|f^* - \eta\|^2 \\ &\le 4\ d_S^2(\hat{f}_i^{S,S'}, f^*) + \|f^* - \eta\|^2 + C\log^3(n)\mathfrak{R}_n^2(\mathcal{F}) \\ &\le 4\epsilon^2 + \Delta^2 + \tilde{O}\left(\frac{1}{n^{2/p}}\right)\end{aligned}$$

where the second inequality above is by using Theorem 14 along with Lemma 5 to bound $\left\|\hat{f}_i^{S,S'} - f^*\right\|^2$ in terms of twice the empirical distance $d_S^2(\hat{f}_i^{S,S'}, f^*)$ (on similar lines as the inclusion lemma 10). Of course we also have by simple Rademacher bound (see Corollary 9) that

$$\left\|\hat{f}_i^{S,S'} - \eta\right\|^2 - \|f^* - \eta\|^2 \le C\mathfrak{R}_n(\hat{\mathcal{F}}_i^S(\epsilon)) \le O\left(n^{-1/p}\right)$$

Hence for the choice of $\epsilon = n^{-1/(2+p)}$, we can conclude that for the partition i, containing f^*,

$$\left\|\hat{f}_i^{S,S'} - \eta\right\|^2 - \|f^* - \eta\|^2 \le O\left(\min\left(n^{-\frac{2}{2+p}} + \Delta^2, n^{-1/p}\right)\right)$$

Hence for the choice of $\epsilon = n^{-1/(2+p)}$, if approximation error is Δ, the bound for the excess risk is given by

$$\begin{aligned}\left\|\hat{f} - \eta\right\|^2 - \inf_{f\in\mathcal{F}} \|f-\eta\|^2 &\le \frac{\log \mathcal{N}_2(\mathcal{F},\epsilon,S)}{n} + O\left(\min\left(n^{-\frac{2}{2+p}} + \Delta^2, \frac{1}{n^{1/p}}\right)\right) \\ &\le n^{-\frac{2}{2+p}} + O\left(\min\left(n^{-\frac{2}{2+p}} + \Delta^2, \frac{1}{n^{1/p}}\right)\right)\end{aligned}$$

Hence overall the above bound can be read as,

$$\left\|\hat{f} - \eta\right\|^2 - \operatorname*{argmin}_{f\in\mathcal{F}} \|f-\eta\|^2 \le \begin{cases} O\left(n^{-\frac{2}{2+p}}\right) & \text{if } \Delta^2 \le n^{-2/(2+p)} \\ O\left(\Delta^2\right) & \text{if } n^{-2/(2+p)} \le \Delta^2 \le n^{-1/p} \\ O\left(n^{-1/p}\right) & \text{otherwise}\end{cases}$$

Hence we see a smooth transition in terms of approximation error rate in the regime $\Delta^2 \in (n^{-2/(2+p)}, n^{-1/p})$. Notice that the estimator is still the same proposed algorithm with choice of ϵ fixed at $n^{-1/(2+p)}$, however the estimation procedure automatically adapts to get the rates above.

8 Proof of Theorem 1

We break down the proof of Theorem 1 into several subsections.

8.1 The general scheme

Proposition 6. *Suppose $0 \le f \le 1$ for all $f \in \mathcal{F}$. Then for any $\epsilon > 0$, with probability at least $1 - 2\delta$*

$$L\left(\tilde{f}\right) \le L^* + C\frac{\log(\mathcal{N}_2(\mathcal{F}, \epsilon, S)/\delta)}{n} + \Xi(n, \epsilon, S') \tag{25}$$

where $\Xi(n, \epsilon, S')$ is such that with probability at least $1 - \delta$,

$$L(\hat{f}_{i^*}^{S,S'}) - L(f^*) \;\le\; \Xi(n, \epsilon, S') \tag{26}$$

for $i^* \in [N]$ such that $f^* \in \hat{\mathcal{F}}_{i^*}^S$.

The proof of the Proposition is immediate.

8.2 Excess Risk of ERM

The first component of the analysis is a risk bound for the empirical minimizer over a function class. While similar bounds appeared elsewhere (e.g. [5, 38]), we prove them here for completeness with explicit constants. The proof of this Lemma is deferred to page 27.

Lemma 7 ([38, 7])**.** *Let $\hat{g}$ be an empirical minimizer over a class $\mathcal{G}$,*

$$\hat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} P_n g \ .$$

Suppose $0 \le g \le 1$. For any $x > 0$, with probability at least $1 - 9e^{-x}$,

$$P\hat{g} \le Pg^* + \sqrt{Pg^*}\sqrt{20r^* + 17r_0} + 114r^* + 53r_0$$

where $r_0 = (x + 6\log\log n)/n$ and $r^* = r^*(\mathcal{G})$.

Lemma 8. *Let $\ell \circ \mathcal{F} = \{(x, y) \mapsto (f(x) - y)^2 : f \in \mathcal{F}\}$, $\mathcal{Y} = [0, 1]$ and $\mathcal{F}$ is a class of functions frm $\mathcal{X}$ to $\mathcal{Y}$. Then, for any $x > 0$, with probability at least $1 - 2e^{-x}$, the empirical risk minimizer $\hat{f}^{ERM}$ on $\mathcal{F}$ satisfies*

$$\mathbb{E}(\hat{f}^{ERM}(x) - y)^2 - \mathbb{E}(f^*(x) - y)^2 \le c_2'\mathfrak{R}_n(\mathcal{F}) + \frac{c_3' x}{n}$$

where $c_2' = 1408$ and $c_3' = 830$.

Using these lemmata, we obtain the following Corollary:

Corollary 9. *For any $x > 0$, with probability at least $1 - 11Ne^{-x}$, for all $i \in [N]$,*

$$L\left(\hat{f}_i^{S,S'}\right) \le L_i^* + \min\left\{c_2'\hat{\mathfrak{R}}_n(\ell \circ \hat{\mathcal{F}}_i^S, S') + c_3' r_0, \sqrt{L_i^*}\sqrt{20\tau + 17r_0} + 114\tau + 53r_0\right\}$$

where $r_0 = (x + 6\log\log n)/n$ and $\tau = 12 \cdot 42^2 \log^3(64n)\mathfrak{R}_n^2(\mathcal{F})$.

Proof. Recall that each $\hat{f}_i^{S,S'}$ is an empirical minimizer over the respective set $\hat{\mathcal{F}}_i^S$. Let $\mathcal{E}_1$ be the event (with respect to the draw of S', conditionally on S) under which Lemma 7 and Lemma 8 (applied to $\hat{f}_i^{S,S'}$) hold simultaneously for all $i \in [N]$. We have $P(\mathcal{E}_1) \ge 1 - 11Ne^{-x}$. That is, with probability at least $1 - 11Ne^{-x}$, for all $i \in [N]$,

$$L\left(\hat{f}_i^{S,S'}\right) \le L_i^* + c_2'\hat{\mathfrak{R}}_n(\ell \circ \hat{\mathcal{F}}_i^S, S') + c_3' r_0 \tag{27}$$

where $L_i^* = \operatorname{argmin}_{f \in \hat{\mathcal{F}}_i^S} L(f)$. Under the same event $\mathcal{E}_1$, we also have an alternative *optimistic* bound in terms of the minimal risk, as implied by Lemma 7: for all $i \in [N]$,

$$L\left(\hat{f}_i^{S,S'}\right) \le L_i^* + \sqrt{L_i^*}\sqrt{20\tau + 17r_0} + 114\tau + 53r_0 \ .$$

where $\tau = r^*(\ell \circ \mathcal{F})$ can be taken to be $\tau = 12 \cdot 21^2 \log^3(64n)\mathfrak{R}_n^2(\mathcal{F})$ as shown in [38]. Combining with (27), the result follows. □

8.3 An Inclusion Lemma

We now aim to get a handle on the empirical Rademacher complexity

$$\hat{\mathfrak{R}}_n(\ell \circ \hat{\mathcal{F}}_i^S, S') = \mathbb{E}_\sigma\left[\sup_{f \in \hat{\mathcal{F}}_i^S} \frac{1}{n} \sum_{(x,y) \in S'} \sigma_i (f(x) - y)^2\right].$$

The difficulty lies in the fact that the set $\hat{\mathcal{F}}_i^S$ is defined with respect to d_S while the empirical Rademacher complexity is evaluated on an independent sample S'. To this end, define

$$\hat{\mathcal{F}}_i^{S,S'}(\gamma) = \{f \in \mathcal{F} : d_{S'}(f, \hat{c}_i) \le \gamma\}$$

where the pseudometric $d_{S'}$ is taken with respect to the set S' while the ϵ-net $\{\hat{c}_i\}$ is constructed with respect to S. We will relate $\hat{\mathcal{F}}_i^{S,S'}(\gamma)$ and $\hat{\mathcal{F}}_i^S(\epsilon)$ for an appropriate choice of γ.

Lemma 10. *Fix $x > 0$, $\epsilon > 0$. Let $r^* = r^*(\mathcal{G})$ for $\mathcal{G} = \{(f - g)^2 : f, g \in \mathcal{F}\}$. Define $r_0 = (x + 6\log\log n)/n$ and $\gamma^2 := 4\epsilon^2 + 284r^* + 118r_0$. Then with probability at least $1 - 8Ne^{-x}$ over the draw of $S \cup S'$, for any $i \in [N]$, we have the inclusion*

$$\hat{\mathcal{F}}_i^S(\epsilon) \subseteq \hat{\mathcal{F}}_i^{S,S'}(\gamma)$$

and hence

$$\hat{\mathfrak{R}}_n\left(\ell \circ \hat{\mathcal{F}}_i^S(\epsilon), S'\right) \le \hat{\mathfrak{R}}_n\left(\ell \circ \hat{\mathcal{F}}_i^{S,S'}(\gamma), S'\right) \ ,$$

Proof of Lemma 10. The proof requires relating empirical squared distance $d_S(f,g)^2$ to its expected version $\mathbb{E}(f(x)-g(x))^2$, and then back to the empirical squared distance $d_{S'}(f,g)^2$ on an independent sample. This amounts to working with the class $\{(f-g)^2 : f,g \in \mathcal{F}\}$, which we may treat as a class $\ell \circ \mathcal{H}$ along with the assumption that y's are identically zero. Now, suppose there is a ϕ_n such that $\hat{\mathfrak{R}}_n(\mathcal{G}[r,S],S) \le \phi_n(r)$ and let $r^* = r^*(\mathcal{G})$ be an upper bound on the largest solution $\phi_n(r) = r$. We now appeal to Theorem 14. With probability at least $1-4e^{-x}$, for any $f,g \in \mathcal{F}$

$$P(f-g)^2 \le 2P_n(f-g)^2 + 106r^* + 48r_0$$

and

$$P_n(f-g)^2 \le 2P(f-g)^2 + 72r^* + 22r_0$$

where $r_0 = (x + 6\log\log n)/n$. Let P_n and P_n' denote the empirical average over a sample S and S', respectively. Then with probability at least $1-8e^{-x}$, for all $f,g \in \mathcal{F}$

$$P_n'(f-g)^2 \le 4P_n(f-g)^2 + 284r^* + 118r_0 \ .$$

Taking a union bound over $i \in [N]$ completes the proof. □

8.4 Controlling the Rademacher Complexity

The next result gives an upper bound on the Rademacher Complexity of the set $\ell \circ \hat{\mathcal{F}}_i^{S,S'}(\gamma)$.

Lemma 11. *Let $r^* = r^*(\mathcal{G})$ for $\mathcal{G} = \{(f-g)^2 : f,g \in \mathcal{F}\}$, and suppose $\gamma^2 \ge r^*$. Then*

$$\hat{\mathfrak{R}}\left(\ell \circ \hat{\mathcal{F}}_i^{S,S'}(\gamma), S'\right) \le \gamma\sqrt{r^*} + \frac{20}{\sqrt{n}} \int_0^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{F},\rho,d_{S'})}\, d\rho$$

Proof. We reason conditionally on $S \cup S'$. We have,

$$\begin{aligned}
&\hat{\mathfrak{R}}\left(\ell \circ \hat{\mathcal{F}}_i^{S,S'}(\gamma), S'\right) \\
&= \mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}_i^{S,S'}(\gamma)} \frac{1}{n} \sum_{(x_j,y_j)\in S'} \sigma_j (f(x_j)-y_j)^2 \\
&= \mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}_i^{S,S'}(\gamma)} \frac{1}{n} \sum_{(x_j,y_j)\in S'} \sigma_j (f(x_j)-\hat{c}_i(x_j))^2 + \sigma_j(\hat{c}_i(x_j)-y_j)^2 + 2\sigma_j(f(x_j)-\hat{c}_i(x_j))(\hat{c}_i(x_j)-y_j) \\
&\le \mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}_i^{S,S'}(\gamma)} \frac{1}{n} \sum_{(x_j,y_j)\in S'} \sigma_j (f(x_j)-\hat{c}_i(x_j))^2 + 2\mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}_i^{S,S'}(\gamma)} \frac{1}{n} \sum_{(x_j,y_j)\in S'} \sigma_j (f(x_j)-\hat{c}_i(x_j))(\hat{c}_i(x_j)-y_j)
\end{aligned} \tag{28}$$

Consider the first term. Conditioned on the set S, the centers $\hat{c}_i$ are fixed and we may view the set $\hat{\mathcal{F}}_i^{S,S'}(\gamma)$ as giving rise to the set of γ^2-approximate empirical minimizers

$$\mathcal{G}_i' = \left\{ (f-\hat{c}_i)^2 : f \in \mathcal{F}, \quad \frac{1}{n} \sum_{(x_j,y_j)\in S'} (f(x_j)-\hat{c}_i(x_j))^2 \right\}$$

For simplicity, we assume that $\hat{c}_i \in \mathcal{F}$ (all the results hold in the case of an improper cover as well), and thus $\mathcal{G}_i' \subseteq \mathcal{G}[\gamma^2, S']$ where $\mathcal{G} = \{(f-g)^2 : f, g \in \mathcal{F}\}$. Then the first term in (28) is

$$\mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}_i^{S,S'}(\gamma)} \frac{1}{n} \sum_{(x_j,y_j)\in S'} \sigma_j (f(x_j) - \hat{c}_i(x_j))^2 \le \hat{\mathfrak{R}}_n(\mathcal{G}[\gamma^2, S'], S') \le \phi_n(\gamma^2) \le \gamma\sqrt{r^*}$$

where the last inequality follows from the fact that $\gamma^2 > r^*$ by our assumption and $\phi_n(r)/\sqrt{r}$ is non-increasing.

We now turn to the Rademacher complexity of the cross-product term in (28). Define

$$\mathcal{G}_i^{S,S'} = \left\{ g_f(x,y) = (f(x) - \hat{c}_i(x))(\hat{c}_i(x) - y) : f \in \hat{\mathcal{F}}_i^{S,S'}(\gamma) \right\}$$

First, observe that for any $g_f \in \mathcal{G}_i^{S,S'}$,

$$\frac{1}{n} \sum_{(x,y)\in S'} g_f(x,y)^2 = \frac{1}{n} \sum_{(x,y)\in S'} (f(x) - \hat{c}_i(x))^2 (\hat{c}_i(x) - y)^2 \le \gamma^2$$

under the boundedness assumption. Next, let $M = \mathcal{N}_2(\hat{\mathcal{F}}_i^{S,S'}, \delta, d_{S'})$ be a covering number with respect to $d_{S'}(f,g)$ and suppose $\mathcal{C} = \{h_i^1, \ldots, h_i^M\}$ is such a δ-cover. Pick any $f \in \hat{\mathcal{F}}_i^{S,S'}$ and let $h \in \mathcal{C}$ be a cover center δ-close to f in the above sense. Then

$$\begin{aligned}
\frac{1}{n} \sum_{(x,y)\in S'} (g_f(x,y) - g_h(x,y))^2 &= \frac{1}{n} \sum_{(x,y)\in S'} \left[(f(x) - \hat{c}_i(x))(\hat{c}_i(x) - y) - (h(x) - \hat{c}_i(x))(\hat{c}_i(x) - y) \right]^2 \\
&= \frac{1}{n} \sum_{(x,y)\in S'} (f(x) - h(x))^2 (\hat{c}_i(x) - y)^2 \\
&\le \delta^2
\end{aligned}$$

implying $\mathcal{N}_2(\mathcal{G}_i^{S,S'}, \delta, d_{S'}) \le \mathcal{N}_2(\hat{\mathcal{F}}_i^{S,S'}, \delta, d_{S'})$. Hence,

$$\hat{\mathfrak{R}}_n(\mathcal{G}_i^{S,S'}) \le \frac{10}{\sqrt{n}} \int_0^\gamma \sqrt{\log \mathcal{N}_2(\hat{\mathcal{F}}_i^{S,S'}, \rho, d_{S'})} d\rho \le \frac{10}{\sqrt{n}} \int_0^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, d_{S'})} d\rho \tag{29}$$

Putting together the results,

$$\hat{\mathfrak{R}}\left(\ell \circ \hat{\mathcal{F}}_i^{S,S'}(\gamma), S' \right) \le \gamma\sqrt{r^*} + \frac{20}{\sqrt{n}} \int_0^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, d_{S'})} d\rho$$

□

8.5 Concluding the Proof

Putting together Corollary 9, Lemma 10, and Lemma 11, with probability at least $1 - 19Ne^{-x}$ over the draw of $S \cup S'$, for any $i \in [N]$,

$$L\left(\hat{f}_i^{S,S'}\right) \le L_i^* + \min\left\{ c_2' \left(\gamma\sqrt{r^*} + \frac{20}{\sqrt{n}} \int_0^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, d_{S'})} d\rho \right) + c_3' r_0, \sqrt{L_i^*}\sqrt{20\tau + 17 r_0} + 114\tau + 53 r_0 \right\}$$

where $r_0 = (x + 6\log\log n)/n$. We now re-write this inequality by setting $19Ne^{-x} = \delta/2$. With probability at least $1 - \delta/2$, for all $i \in [N]$

$$L\left(\hat{f}_i^{S,S'}\right) \le L_i^* + \min\left\{c_2'\left(\gamma\sqrt{r^*} + \frac{20}{\sqrt{n}}\int_0^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{F},\rho,d_{S'})}d\rho\right) + c_3'\beta, \sqrt{L_i^*}\sqrt{20\tau + 17\beta} + 114\tau + 53\beta\right\}$$

where $\beta = (\log(38N/\delta) + 6\log\log n)/n$ and $\tau = 12\cdot 21^2\log^3(64n)\mathfrak{R}_n^2(\mathcal{F})$.

Next, we appeal to Example 1 in [23], which implies that for any $i \in [N]$ with probability at least $1 - ce^{-x}$,

$$L(\hat{f}_i^{S,S'}) - L(f_i^*) \le \frac{K(d+x)}{n}$$

where d is the dimensionality of the linear space to which $\mathcal{F}$ belongs. Taking a union bound over $i \in [N]$ and letting $\delta/2 = cNe^{-x}$, we obtain the desired statement.

This concludes the proof of Theorem 1.

9 Lower Bounds

9.1 Lower bound for VC subgraph classes

In this section we exhibit a VC subgraph class $\mathcal{F}$ with VC-dimension at most d such that

$$W_n(\mathcal{F}) \ge C\frac{d(1+\log(n/d))}{n}$$

where $C > 0$ is a numerical constant. We will, in fact, prove a more general lower bound, for the risk in probability rather than in expectation.

In this section, $\mathcal{X} = \{x^1, x^2, \dots,\}$ is an infinite countable set of elements and $\mathcal{F}$ is the following set of binary-valued functions on $\mathcal{X}$:

$$\mathcal{F} = \{f : f(x) = a\mathbf{1}\{x \in W\}, \text{ for some } W \subset \mathcal{X} \text{ with } \mathrm{Card}(W) \le d\},$$

where $a > 0$, $\mathbf{1}\{\cdot\}$ denotes the indicator function and $\mathrm{Card}(W)$ is the cardinality of W. It is easy to check that $\mathcal{F}$ is a VC subgraph class with VC-dimension at most d.

Theorem 12. *Let d be any integer such that $n \ge d$, and $a = 3/4$. Let the random pair (X,Y) take values in $\mathcal{X}\times\{0,1\}$. Then there exist a marginal distribution P_X and numerical constants $c, c' > 0$ such that*

$$\inf_{\hat{f}}\sup_{\eta\in\mathcal{F}} P_\eta\left(\left\|\hat{f} - \eta\right\|^2 \ge c\frac{d(1+\log(n/d))}{n}\right) \ge c',$$

where P_η denotes the distribution of the n-sample D_n when $\mathbb{E}(Y|X=x) = \eta(x)$.

Proof. Fix some $0 < \alpha < 1$ and set $k = \lceil d/\alpha\rceil$. Let $\mathcal{C}$ be the set of all binary sequences $\omega \in \{0,1\}^k$ with at most d non-zero components. By the d-selection lemma (see, e.g., Lemma 4 in [35]), for $k \ge 2d$ there exists of a subset $\mathcal{C}'$ of $\mathcal{C}$ with the following properties: (a) $\log(\mathrm{Card}(\mathcal{C}')) \ge (d/4)\log(k/(6d))$ and (b) $\rho_H(\omega,\omega') \ge d$ for any $\omega,\omega' \in \mathcal{C}'$. Here, $\rho_H(\cdot,\cdot)$ denotes the Hamming distance. To any $\omega \in \mathcal{C}'$

we associate a function f_ω on $\mathcal{X}$ defined by $f_\omega(x^i) = \omega_i$ for $i = 1, \dots, k$ and $f_\omega(x^i) = 0,\ i \geq k+1$, where ω_i is the ith component of ω.

Let P_X be the distribution on $\mathcal{X}$ which is uniform on $\{x^1, \dots, x^k\}$, putting probability $1/k$ on each of these x^j and probability 0 on all x^j with $j \geq k+1$. Denote by $\mathbf{P}_\omega$ the joint distribution of (X, Y) having this marginal P_X and $Y \in \{0, 1\}$ with the conditional distribution $\mathbb{E}(Y|X = x) = P(Y = 1|X = x) = 1/2 + f_\omega(x)/4 \triangleq \eta_\omega(x)$ for all $x \in \mathcal{X}$.

Consider now a set of functions $\mathcal{F}' = \{\eta_\omega : \omega \in \mathcal{C}'\} \subset \mathcal{F}$. Observe that, by construction,

$$\|\eta_\omega - \eta_{\omega'}\|^2 = \rho_H(\omega, \omega')/(16k) \geq \alpha/32, \quad \forall\ \omega, \omega' \in \mathcal{C}'. \tag{30}$$

On the other hand, the Kullback-Leibler divergence between $\mathbf{P}_\omega$ and $\mathbf{P}_{\omega'}$ has the form

$$K(\mathbf{P}_\omega, \mathbf{P}_{\omega'}) = n\mathbb{E}\left(\eta_\omega(X) \log \frac{\eta_\omega(X)}{\eta_{\omega'}(X)} + (1 - \eta_\omega(X)) \log \frac{(1 - \eta_\omega(X))}{(1 - \eta_{\omega'}(X))}\right).$$

Using the inequality $-\log(1 + u) \leq -u + u^2/2, \forall u > -1$, and the fact that $1/2 \leq \eta_\omega(X) \leq 3/4$ for all $\omega \in \mathcal{C}'$ we obtain that the expression under the expectation in the previous display is bounded by $2(\eta_\omega(X) - \eta_{\omega'}(X))^2$, which implies

$$K(\mathbf{P}_\omega, \mathbf{P}_{\omega'}) \leq \frac{\|f_\omega - f_{\omega'}\|^2}{8} \leq \frac{nd}{8k} \leq \frac{n\alpha}{8}, \quad \forall\ \omega, \omega' \in \mathcal{C}'. \tag{31}$$

From (30), (31) and Theorem 2.7 in [39], the result of Theorem 12 follows if we show that

$$n\alpha/8 \leq \log(\mathrm{Card}(\mathcal{F}') - 1)/16 \tag{32}$$

with

$$\alpha = C_1 \frac{d}{n} \log \frac{C_2 n}{d}$$

where $C_1, C_2 > 0$ are constants. Assume first that $d \geq 4$. Then, using the inequalities $\log(\mathrm{Card}(\mathcal{F}') - 1) \geq \log(\mathrm{Card}(\mathcal{C}')/2) \geq (d/4)\log(k/(6d)) - \log 2 \geq (d/4)\log(1/(12\alpha))$ it is enough to show that

$$n\alpha \leq \frac{d}{8} \log \frac{1}{12\alpha}.$$

Using that $x \geq 2\log x$ for $x \geq 0$ it is easy to check that the inequality in the last display holds if we choose, for example, $C_1 = 1/16, C_2 = 1/(12C_1)$. In the case $d \leq 3$ it is enough to consider $\alpha = (C_1/n)\log(C_2 n)$ and (32) is also satisfied for suitable C_1, C_2. □

9.2 Lower Bound Under Entropy Conditions

Let ℓ_0 be the set of all real-valued sequences $(f_k, k = 1, 2, \dots)$. Denote by $\mathbf{e}_j$ the unit vectors in ℓ_0: $\mathbf{e}_j = (\mathbf{1}\{k = j\}, k = 1, 2, \dots)$, $j = 1, 2, \dots$. For $p > 0$, consider the unit ℓ_p-ball $B_p = \{f \in \ell : \sum_{j=1}^\infty |f_j|^p \leq 1\}$.

Theorem 13. *Fix any $p > 0$. Let $\mathcal{F} = \{f \in \ell : f_j = 1/2 + g_j, \{g_j\} \in B_p\}$ and let $\mathcal{X} = \{\mathbf{e}_1, \mathbf{e}_2, \dots\}$ be the set of all unit vectors in ℓ_0. For any $\epsilon > 0$ and any $n \geq (2/\epsilon)^p$, we have*

$$\sup_{S \in \mathcal{Z}^n} \log \mathcal{N}_1(\mathcal{F}, \epsilon/16, S) \geq \frac{1}{8}\left(\frac{2}{\epsilon}\right)^p. \tag{33}$$

Furthermore, there exist positive constants C, C' depending only on p such that the minimax risk satisfies, for any $n \geq 1$,

$$W_n(\mathcal{F}) \geq C n^{-2/(2+p)}, \tag{34}$$

and the minimax regret satisfies, for any $p \geq 2$ and any $n \geq 1$,

$$V_n(\mathcal{F}) \geq C' n^{-1/p}. \tag{35}$$

Proof. First, fix $\epsilon > 0$. Let $d = (2/\epsilon)^p$ and observe that the set of vectors

$$\left\{ \left(\frac{r_1}{d^{1/p}}, \dots, \frac{r_d}{d^{1/p}}, 0, \dots \right) : (r_1, \dots, r_d) \in \{\pm 1\}^d \right\} \subset B_p$$

shatters the set $\{\mathbf{e}_1, \dots, \mathbf{e}_d\}$ at scale $2/d^{1/p} = \epsilon$. Thus, the fat-shattering dimension $\mathrm{fat}_\epsilon(B_p) \geq d = (2/\epsilon)^p$. This yields (33) via an application of Theorem 2.6 in [31].

To prove (34) and (35), consider a subset $\bar{\mathcal{F}} = \{f \in \mathcal{F} : f_j = 1/2, \ \forall j > d\}$ where $d = \lceil (c_* n)^{p'/(2+p')} \rceil$ and $c_* > 0$, $p' \geq p$ are constants that will be chosen later. Let $\Omega = \{0,1\}^d$ be the set of all binary sequences of length d. Define P_X as the distribution on $\mathcal{X}$ which is uniform on $\{\mathbf{e}_1, \dots, \mathbf{e}_d\}$, putting probability $1/d$ on each of these $\mathbf{e}_j$ and probability 0 on all $\mathbf{e}_j$ with $j \geq d+1$. For any $\omega \in \Omega$, denote by $\mathbf{P}_\omega$ the joint distribution of (X, Y) having this marginal P_X and $Y \in \{0,1\}$ with the conditional distribution defined by

$$\mathbb{E}(Y|X = \mathbf{e}_i) = P(Y = 1|X = \mathbf{e}_i) = \frac{1}{2} + \frac{\omega_i}{4d^{1/p'}} \triangleq \eta_\omega(\mathbf{e}_i)$$

for $i = 1, \dots, d$, and arbitrary for $i \geq d+1$. The regression function corresponding to $\mathbf{P}_\omega$ is then

$$\eta_\omega = (\eta_\omega(\mathbf{e}_1), \dots, \eta_\omega(\mathbf{e}_d), \frac{1}{2}, \dots) = \left(\frac{1}{2} + \frac{\omega_1}{4d^{1/p'}}, \dots, \frac{1}{2} + \frac{\omega_d}{4d^{1/p'}}, \frac{1}{2}, \dots \right).$$

It is easy to see that since $\omega_i \in \{0,1\}$, for any estimator $\hat{f} = (\hat{f}(\mathbf{e}_1), \hat{f}(\mathbf{e}_2), \dots)$ we have

$$|\hat{f}(\mathbf{e}_i) - \eta_\omega(\mathbf{e}_i)| \geq \frac{1}{2} \left| \frac{1}{2} + \frac{\hat{\omega}_i}{4d^{1/p'}} - \eta_\omega(\mathbf{e}_i) \right| = \frac{|\hat{\omega}_i - \omega_i|}{8d^{1/p'}}, \quad i = 1, 2, \dots,$$

where $\hat{\omega}_i$ is the closest to $4d^{1/p'}(\hat{f}(\mathbf{e}_i) - 1/2)$ element of the set $\{0,1\}$. Therefore,

$$\left\| \hat{f} - \eta_\omega \right\|^2 \geq \frac{1}{d} \sum_{i=1}^{d} \frac{|\hat{\omega}_i - \omega_i|^2}{64\, d^{2/p'}} = \frac{\rho_H(\hat{\omega}, \omega)}{64\, d^{1+2/p'}} \tag{36}$$

where $\rho_H(\cdot,\cdot)$ is the Hamming distance. From Assouad's lemma (Theorem 2.12 (iv) in [39]) we find that

$$\max_{\omega \in \Omega} \mathbb{E}_\omega \rho_H(\hat{\omega}, \omega) \geq \frac{d}{4} \exp(-\alpha) \tag{37}$$

where $\alpha = \max\{\chi^2(\mathbf{P}_\omega, \mathbf{P}_{\omega'}) : \omega, \omega' \in \Omega, \rho_H(\omega, \omega') = 1\}$ and $\chi^2(\mathbf{P}_\omega, \mathbf{P}_{\omega'})$ is the chi-squared divergence between $\mathbf{P}_\omega$ and $\mathbf{P}_{\omega'}$. Here, $\mathbb{E}_\omega$ denotes the distribution of the n-sample D_n when

$(X_i, Y_i) \sim \mathbf{P}_\omega$ for all i. Since $1/2 \le \eta_\omega(X) \le 3/4$, the chi-squared divergence is bounded as follows:

$$\chi^2(\mathbf{P}_\omega, \mathbf{P}_{\omega'}) = n\mathbb{E}\left((\eta_\omega(X) - \eta_{\omega'}(X))^2 \left(\frac{1}{\eta_{\omega'}(X)} + \frac{1}{1-\eta_{\omega'}(X)}\right)\right)$$
$$\le 6n\mathbb{E}(\eta_\omega(X) - \eta_{\omega'}(X))^2 = \frac{6n}{d}\sum_{i=1}^{d}\frac{(\omega_i - \omega_i')^2}{16\, d^{2/p'}} \le \frac{3}{8c_*}$$

for all $\omega, \omega' \in \Omega$ such that $\rho_H(\omega, \omega') = 1$. Combining this result with (36) and (37) we find

$$\inf_{\hat{f}} \max_{\omega\in\Omega} \mathbb{E}_\omega \left\|\hat{f} - \eta_\omega\right\|^2 \ge \frac{\exp(-3/(8c_*))}{256\, d^{2/p'}}\,. \tag{38}$$

Now, to prove (34) it suffices to take here $p' = p$. With this choice of p', the set $\{\eta_\omega : \omega \in \Omega\}$ is contained in $\mathcal{F}$, so that $W_n(\mathcal{F}) \ge \inf_{\hat{f}} \max_{\omega\in\Omega} \mathbb{E}_\omega \left\|\hat{f} - \eta_\omega\right\|^2$ and (34) follows immediately from (38). We now prove (35). Set $p' = 2(p-1)$, so that $2/(2+p') = 1/p$. Introduce the vector

$$f^* = \left(\frac{1}{2} + \frac{\omega_1}{4d^{1/p}}, \ldots, \frac{1}{2} + \frac{\omega_d}{4d^{1/p}}, \frac{1}{2}, \ldots\right)$$

Note that $f^* \in \bar{\mathcal{F}}$ and

$$\|f^* - \eta_\omega\|^2 = \frac{1}{d}\sum_{i=1}^{d}\left(\frac{r_i}{d^{1/p}} - \frac{r_i}{d^{1/p'}}\right)^2 = \left(1 - \frac{1}{d^{1/p-1/p'}}\right)^2 \frac{1}{d^{2/p'}} \le \frac{1}{4d^{2/p'}}$$

assuming n is large enough ($n \ge n_0(p)$ where $n_0(p)$ depends only on p and c_*). We then have

$$\begin{aligned}
V_n(\mathcal{F}) &= \inf_{\hat{f}} \sup_{P_{XY}\in\mathcal{P}} \left\{\mathbb{E}_\omega \left\|\hat{f} - \eta\right\|^2 - \inf_{f\in\mathcal{F}} \|f - \eta\|^2\right\} \\
&\ge \inf_{\hat{f}} \max_{\omega\in\Omega} \left\{\mathbb{E}_\omega \left\|\hat{f} - \eta_\omega\right\|^2 - \inf_{f\in\bar{\mathcal{F}}} \|f - \eta_\omega\|^2\right\} \\
&\ge \inf_{\hat{f}} \max_{\omega\in\Omega} \left\{\mathbb{E}_\omega \left\|\hat{f} - \eta_\omega\right\|^2 - \|f^* - \eta_\omega\|^2\right\} \\
&\ge \inf_{\hat{f}} \max_{\omega\in\Omega} \mathbb{E}_\omega \left\|\hat{f} - \eta_\omega\right\|^2 - \frac{1}{4d^{2/p'}}\,.
\end{aligned}$$

Combining this with (38) and choosing $c_* > 0$ small enough we obtain (35) for $n \ge n_0$. For $n < n_0(p)$ the result trivially follows from the positivity of $V_n(\mathcal{F})$. □

10 Technical Results: Localization

The following is a modification of Theorem 6.1 in [7]. We include part of that Theorem verbatim and make additional changes.

Theorem 14 (Based on [7]). *Let $\mathcal{G}$ be a class of non-negative functions almost surely bounded by b. Let ϕ_n be a function that is non-negative, non-decreasing, and $\phi_n(r)/\sqrt{r}$ non-increasing, satisfying*

$$\hat{\mathfrak{R}}_n\Big(\mathcal{G}[r,S],S\Big) \le \phi_n(r)$$

for all $r > 0$. Let $r^* = r^*(\mathcal{G})$ be an upper bound on the largest solution $\phi_n(r) = r$. Then for all $x > 0$, with probability at least $1 - 4e^{-x}$ for all $g \in \mathcal{G}$

$$Pg \le 2P_n g + 106r^* + 48r_0$$

and

$$P_n g \le 2Pg + 72r^* + 22r_0$$

and

$$Pg \le P_n g + \sqrt{P_n g}(\sqrt{8r^*} + \sqrt{4r_0}) + 108r^* + 42r_0$$

where $r_0 = b(x + 6\log\log n)/n$.

Proof of Theorem 14. Define $\delta_k = b2^{-k}$ for $k \ge 0$ and let $\mathcal{G}_k = \{g \in \mathcal{G} : \delta_{k+1} \le Pg \le \delta_k\}$. Denote the empirical Rademacher averages of $\mathcal{G}_k$ by R_k. Observe that for $g \in \mathcal{G}_k$, $Pg^2 \le b\delta_k$. Then Lemma 6.2 in [7] implies that with probability at least $1 - e^{-x}$ for all $k \ge 0$ and $g \in \mathcal{G}_k$,

$$|P_n g - Pg| \le 6R_k + \sqrt{\frac{2b\delta_k(x + x(\delta_k))}{n}} + \frac{6b(x + x(\delta_k))}{n} \tag{39}$$

where $x(\delta) = 2\log\Big(\frac{\pi}{\sqrt{2}}\log_2\frac{2b}{\delta}\Big)$. We now condition on this event. Let

$$U_k = \delta_k + 6R_k + \sqrt{\frac{2b\delta_k(x + x(\delta_k))}{n}} + \frac{6b(x + x(\delta_k))}{n}$$

and observe that $P_n g \le U_k$. This implies that $R_k \le \phi_n(U_k)$. Putting together the terms,

$$U_k \le \delta_k + 6\phi_n(U_k) + \sqrt{\frac{2b\delta_k(x + x(\delta_k))}{n}} + \frac{6b(x + x(\delta_k))}{n}$$

Let $k_0 > 0$ be the smallest integer such that $\delta_{k_0+1} \ge b/n$. For any $k \le k_0$ and $n \ge 5$, $x(\delta_k) \le 6\log\log n$ and

$$U_k \le \delta_k + 6\phi_n(U_k) + 7r_0 \ .$$

We assume $U_k > r^*$, for otherwise we immediately obtain the theorem statement. The fact that $\phi_n(r)/\sqrt{r}$ is non-increasing implies $\phi_n(r) \le \sqrt{r \cdot r^*}$ for any $r \ge r^*$. Hence,

$$U_k \le 6\sqrt{U_k r^*} + 2\delta_k + 7r_0.$$

Solving the quadratic equation,

$$U_k \le 36r^* + 4\delta_k + 14r_0 \le 36r^* + 8Pg + 14r_0$$

because $\delta_k \le 2Pg$. We thus have for all $k \le k_0$ and $g \in \mathcal{G}_k$

$$|Pg - P_n g| \le 6\phi_n(36r^* + 8Pg + 14r_0) + \sqrt{4r_0 Pg} + 6r_0 \tag{40}$$
$$\le 6\sqrt{r^*}\sqrt{36r^* + 8Pg + 14r_0} + \sqrt{4r_0 Pg} + 6r_0 \tag{41}$$
$$\le 45r^* + \sqrt{8r^* Pg} + \sqrt{4r_0 Pg} + 20r_0 \tag{42}$$

Solving the equation

$$Pg \le P_n g + \sqrt{Pg}(\sqrt{8r^*} + \sqrt{4r_0}) + 45r^* + 20r_0 \tag{43}$$

yields

$$Pg \le 2P_n g + 106r^* + 48r_0 \ .$$

Alternatively, using (43) and the implication $A \le B + C\sqrt{A} \Rightarrow A \le B + C^2 + \sqrt{B}C$ for non-negative A, B and C, we obtain

$$Pg \le P_n g + \sqrt{P_n g}(\sqrt{8r^*} + \sqrt{4r_0}) + 108r^* + 42r_0, \tag{44}$$

proving the last statement of the theorem. For the second statement, we solve the equation (40) in terms of the variable $\sqrt{Pg}$:

$$P_n g \le Pg + \sqrt{Pg}(\sqrt{8r^*} + \sqrt{4r_0}) + 45r^* + 20r_0.$$

The roots are found to be

$$-\frac{\sqrt{8r^*} + \sqrt{4r_0}}{2} \pm \sqrt{\left(\frac{\sqrt{8r^*} + \sqrt{4r_0}}{2}\right)^2 + (P_n g - 45r^* - 20r_0)}$$

If $P_n g < 45r^* + 20r_0$, the statement of the theorem holds. Otherwise, we take the positive root and conclude

$$\sqrt{Pg} \ge -\frac{\sqrt{8r^*} + \sqrt{4r_0}}{2} + \sqrt{\left(\frac{\sqrt{8r^*} + \sqrt{4r_0}}{2}\right)^2 + (P_n g - 45r^* - 20r_0)}$$

leading to

$$P_n g - 45r^* - 20r_0 \le 2Pg + \left(\frac{\sqrt{8r^*} + \sqrt{4r_0}}{2}\right)^2$$

and thus

$$P_n g \le 2Pg + 49r^* + 22r_0 \ .$$

Now consider the case $k \ge k_0$. First, for any $g \in \mathcal{G}_k$, $Pg \le \delta_k \le \delta_{k_0} \le 4b/n$. Hence $\mathcal{G}' = \{g \in \mathcal{G} : Pg < 4b/n\} \supseteq \mathcal{G}_k$ for any $k \ge k_0$. By Lemma 6.1 in [7], with probability at least $1 - 3e^{-x}$,

$$|P_n g - Pg| \le 6\mathfrak{R}_n(\mathcal{G}') + \frac{b}{n}(\sqrt{8x} + 4x) \le 6\mathfrak{R}_n(\mathcal{G}') + \frac{8bx}{n}$$

whenever $x > 1/2$. Now reason on this event. Defining

$$U' = 6\mathfrak{R}_n(\mathcal{G}') + Pg + \frac{8bx}{n}$$

we have $P_n g \le U'$ for any $g \in \mathcal{G}'$, and so

$$\mathfrak{R}_n(\mathcal{G}') \le \mathfrak{R}_n(\{g \in \mathcal{G} : P_n g \le U') \le \phi_n(U')$$

Since ϕ_n is sub-root,

$$U' \le 6\phi_n(U') + Pg + \frac{8bx}{n} \le 6\sqrt{U'}\sqrt{r^*} + Pg + \frac{8bx}{n}$$

Solving for $\sqrt{U'}$,

$$\sqrt{U'} \le 6\sqrt{r^*} + \sqrt{Pg + \frac{8bx}{n}}$$

and thus

$$P_n g \le U' \le 2Pg + 72r^* + 16r_0$$

□

Proof of Lemma 7. By Theorem 14, for any ϕ_n satisfying $\mathfrak{R}_n(\{g : P_n g \le r\}) \le \phi_n(r)$ and appropriate growth conditions, for all $x > 0$, with probability at least $1 - 4e^{-x}$ for all $g \in \mathcal{G}$

$$Pg \le P_n g + \sqrt{P_n g}(\sqrt{8r^*} + \sqrt{4r_0}) + 108r^* + 42r_0, \tag{45}$$

where r^* is the largest solution of $\phi_n(r) = r$. Under the above event, for $g^* = \operatorname{argmin}_{g\in\mathcal{G}} Pg$,

$$P\hat{g} \le P_n g^* + \sqrt{P_n g^*}(\sqrt{8r^*} + \sqrt{4r_0}) + 108r^* + 42r_0$$

By Bernstein's inequality, with probability at least $1 - e^{-x}$,

$$P_n g^* \le Pg^* + \sqrt{\frac{4xPg^*}{n}} + \frac{4x}{n}$$

which implies, in particular, $P_n g^* \le 2Pg^* + \frac{5x}{n}$. Together with the previous inequality, we obtain

$$P\hat{g} \le Pg^* + \sqrt{\frac{4xPg^*}{n}} + \frac{4x}{n} + \sqrt{2Pg^* + \frac{5x}{n}}(\sqrt{8r^*} + \sqrt{4r_0}) + 108r^* + 42r_0$$

Simplifying and over-bounding,

$$P\hat{g} \le Pg^* + \sqrt{Pg^*}\sqrt{20r^* + 17r_0} + 114r^* + 53r_0$$

□

Proof of Lemma 8. We apply Theorem 3.3 in [5] to $\mathcal{G} = \ell \circ \mathcal{F} - \ell \circ f^*$. Observe that

$$\operatorname{Var}(\ell \circ f - \ell \circ f^*) \le \mathbb{E}\left((f(x) - y)^2 - (f^*(x) - y)^2\right)^2 \le 2\mathbb{E}\left((f(x) - y)^2 - (f^*(x) - y)^2\right)$$

and thus the requirement of the theorem is satisfied with $B = 2$. Let us take $\phi(r) = \mathbb{E}\mathfrak{R}_n(\mathcal{G})$, a constant which trivially satisfies the subroot property and has fixed point $\mathbb{E}\mathfrak{R}_n(\mathcal{G})$. Then, for any $x > 0$, with probability at least $1 - e^{-x}$, for any $g \in \mathcal{G}$,

$$Pg \le P_n g + c_1'' \mathbb{E}\mathfrak{R}_n(\mathcal{G}) + \frac{x(22 + c_2'')}{n}$$

where $c_1'' = 704$ and $c_2'' = 104$. Choosing $\hat{f}$ to be the minimizer of empirical risk, this implies

$$\mathbb{E}(\hat{f}(x) - y)^2 - \mathbb{E}(f^*(x) - y)^2 \le c_1'' \mathbb{E}\mathfrak{R}_n(\ell \circ \mathcal{F}) + \frac{x(22 + c_2'')}{n}$$

where we passed to the Rademacher averages of $\ell \circ \mathcal{F}$. Now, by Lemma A.4 in [5], with probability at least $1 - e^{-x}$,

$$\mathbb{E}\mathfrak{R}_n(\ell \circ \mathcal{F}) \le 2\mathfrak{R}_n(\ell \circ \mathcal{F}) + \frac{x}{n}$$

Combining, with probability at least $1 - 2e^{-x}$,

$$\mathbb{E}(\hat{f}(x) - y)^2 - \mathbb{E}(f^*(x) - y)^2 \le 2c_1'' \mathfrak{R}_n(\ell \circ \mathcal{F}) + \frac{x(22 + c_2'' + c_1'')}{n}$$

□

11 Proof of Lemma 5

To prove the first estimate for r^* in lemma, we need a result for smooth losses, proved in [38] in the context of supervised learning:

Lemma 15 ([38])**.** *Let ℓ be an H-smooth non-negative loss. Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$. Then for any set $S \in (\mathcal{X} \times \mathcal{Y})^n$,*

$$\hat{\mathfrak{R}}_n((\ell \circ \mathcal{H})[r, S], S) \le 21\sqrt{6Hr}\log^{3/2}(64n)\mathfrak{R}_n(\mathcal{H}) \tag{46}$$

Proof of Lemma 5. We first claim that we may always take $r^* = 21168\log^3(64n)\mathfrak{R}_n^2(\mathcal{F})$. Since the result of Lemma 15 holds for any distribution on $\mathcal{X} \times \mathcal{Y}$, we may apply it for the class $\mathcal{H} = \{f - g : f, g \in \mathcal{F}\}$ of differences, with Y being identically zero. Since the square loss $\ell(y, y') = y^2$ is 2-smooth, we obtain

$$\hat{\mathfrak{R}}_n(\mathcal{G}[r, S], S) = \hat{\mathfrak{R}}_n((\ell \circ \mathcal{H})[r, S], S) \le 21\sqrt{12r}\log^{3/2}(64n)\mathfrak{R}_n(\mathcal{H}) \le 42\sqrt{12r}\log^{3/2}(64n)\mathfrak{R}_n(\mathcal{F}).$$

Now define the right-hand side as the function $\phi_n(r)$. This immediately leads to a fixed-point

$$r^* = 12 \cdot 42^2 \log^3(64n)\mathfrak{R}_n^2(\mathcal{F}),$$

as claimed.

For the second part, fix some $\delta > 0$ and let $c_1, \dots, c_M$ be any minimal δ-cover of $\mathcal{F}$ with respect to d_S with $M = \mathcal{N}_2(\mathcal{F}, \delta, S)$. Without loss of generality, assume $0 \le c_i(x) \le 1$ for all $x \in \mathcal{X}$, $i \in [M]$. Take any $g \in \mathcal{G}$ and express it as $(f' - f'')^2$ with $f', f'' \in \mathcal{F}$. Let c', c'' be the elements of the cover δ-close to f' and f'' respectively. Since

$$\frac{1}{n}\sum_{i=1}^n \left[(f'(x_i) - f''(x_i))^2 - (c'(x_i) - c''(x_i))^2\right]^2 \le 4\frac{1}{n}\sum_{i=1}^n (f'(x_i) - c'(x_i) + f''(x_i) - c''(x_i))^2 \le 16\delta^2$$

we have that $\mathcal{N}_2(\mathcal{G}, \delta, S) \le \mathcal{N}_2(\mathcal{F}, \delta/4, S)$. Hence, the ϕ_n in (8) can be taken to be

$$\begin{aligned}\hat{\mathfrak{R}}_n(\mathcal{G}[r,S],S) &\le \frac{12}{\sqrt{n}}\int_0^{\sqrt{r}}\sqrt{\log \mathcal{N}_2(\mathcal{G},\delta,S)}d\delta \le \frac{12}{\sqrt{n}}\int_0^{\sqrt{r}}\sqrt{v\log(4c/\delta)}d\delta \\ &\le \frac{12\cdot 4c\sqrt{v}}{\sqrt{n}}\int_0^{\sqrt{r}/4c}\sqrt{\log 1/\rho}\; d\rho \le 24\sqrt{\frac{vr}{n}}\log^{1/2}(4c/\sqrt{r}) := \phi_n(r)\end{aligned} \tag{47}$$

We would like to find an upper bound r^* on the fixed point of $\phi_n(r) = r$. Observe that for

$$\phi(x) = a\log^q(b/x)$$

with $q \in (0,1]$, we may take $x^* = a\log^q(b/a)$ as an upper bound on the fixed point of ϕ whenever $b \ge a > 0$. That is, for n large enough,

$$r^* = \left(24\sqrt{\frac{v}{n}}\log^{1/2}\left(\frac{c\sqrt{n}}{6\sqrt{v}}\right)\right)^2 = C\frac{v}{n}\log(n/v) \tag{48}$$

for some constant C.

Finally, for a finite class $\mathcal{F}$, the covering numbers are $\mathcal{N}_2(\mathcal{F},\epsilon,S) \le |\mathcal{F}|$ and, trivially,

$$r^* = C\frac{\log|\mathcal{F}|}{n}$$

along the lines of (47). □

References

[1] MA Aizerman, EM Braverman, and LI Rozonoer. *The Method of Potential Functions in the Theory of Machine Learning*. Nauka, Moscow, 1970.

[2] J.Y. Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 20(2), 2007.

[3] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

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

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

[6] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete.*, 65(2):181–237, 1983.

[7] O. Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, PhD thesis, Ecole Polytechnique, 2002.

[8] O. Bousquet, V. Koltchinskii, and D. Panchenko. Some local measures of complexity of convex hulls and generalization bounds. pages 59–73, Sydney, Australia, July 8-10, 2002.

[9] N. Cesa-Bianchi and G. Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, 43(3):247–264, 2001.

[10] L. Devroye. *A Course in Density Estimation.* Birkhäuser, Boston, 1987.

[11] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

[12] R. M. Dudley, E. Gine, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. 4:485–510, 1991.

[13] Richard M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, 1999.

[14] R.M. Dudley. Central limit theorems for empirical measures. *Ann. Prob.*, 6:899–929, 1978.

[15] R.M. Dudley. Universal Donsker classes and metric entropy. *Ann. Prob.*, 14(4):1306–1326, 1987.

[16] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[17] I.A. Ibragimov and R.Z. Khas' minskii. On estimate of the density function. *Zapiski Nauchnykh Seminarov POMI*, 98:61–85, 1980.

[18] A. Juditsky, P. Rigollet, and A. Tsybakov. Learning by mirror averaging. *Annals of Statistics*, 2008.

[19] A. B. Juditsky, A. V. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by the mirror descent algorithm with averaging. *Problems of Information Transmission*, 41(4):368–384, 2005.

[20] A.N. Kolmogorov and V.M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

[21] V. Koltchinskii. Rademacher penalties and structural risk minimization. 47(5):1902–1914, 2001.

[22] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[23] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011.

[24] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability*, II:443–459, 2000.

[25] L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[26] G. Lecué. Empirical risk minimization is optimal for the convex aggregation problem.

[27] G. Lecué. *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis.* Habilitation thesis., Université Paris-Est., 2011.

[28] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3):591–613, 2009.

[29] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

[30] K. Lounici. *Generalized mirror averaging and D-convex aggregation.*, volume 16. Mathematical methods of statistics edition, 2007.

[31] S. Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. 48(1):251–263, 2002.

[32] A. Nemirovski. Topics in non-parametric statistics. lectures on probability theory and statistics (saint-flour, 1998). *Lecture Notes in Math*, 1738:85–277, 2000.

[33] D. Pollard. *Convergence of stochastic processes.* Springer-Verlag, Berlin, 1984.

[34] D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22(3):271–278, 1995.

[35] M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *NIPS*, 2011.

[36] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[37] P. Rigollet and A. Tsybakov. *Sparse estimation by exponential weighting*, volume 27. Statistical Science, 2012.

[38] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. *arXiv preprint arXiv:1009.3896*, 2010.

[39] A.B. Tsybakov. *Introduction to Nonparametric Estimation.* 2009.

[40] Alexandre B. Tsybakov. Optimal rates of aggregation. In *COLT*, pages 303–313, 2003.

[41] S. van de Geer. Estimating a regression function. 18(2):907–924, 1990.

[42] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[43] V. N. Vapnik and A. Ya. Chervonenkis. Uniform convergence of frequencies of occurrence of events to their probabilities. *Dokl. Akad. Nauk USSR*, 181:915–918, 1968.

[44] V.N. Vapnik and A.J. Chervonenkis. Theory of pattern recognition. 1974.

[45] Y. Yang. *Aggregating regression procedures to improve performance.*, volume 10. 2004.

[46] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.