Empirical Entropy, Minimax Regret and Minimax Risk

Alexander Rakhlin    Karthik Sridharan    Alexandre B. Tsybakov

April 24, 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 \( \epsilon^{-p} \) growth of the empirical \( \epsilon \)-entropy, the excess risk of the proposed method attains the rate \( n^{-2/p} \) 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 minimax 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 [31] and we show that it yields the optimal rate.

1 Introduction

Let \( D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \) 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 \( f : \mathcal{X} \rightarrow \mathcal{Y} \) 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

\[
L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f)
\]

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

\[ \mathcal{E}_\mathcal{F}(\hat{f}) \triangleq \mathbb{E}L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f) = \mathbb{E}\|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \]  

(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” \([46], [12], [24]\), while the object on the right-hand side has been the topic of oracle inequalities in nonparametric statistics \([33], [41]\), and in the literature on aggregation \([40], [37]\). 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

\[ \mathbb{E}\|\hat{f} - \eta\|^2 - C \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \]  

(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

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

(4)

where \( \mathcal{P} \) is the set of all distributions on \( \mathcal{X} \times \mathcal{Y} \) and \( \inf_f \) 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

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

(5)

The expression in curly brackets in (5) can be viewed as a “distance” between the estimator \( \hat{f} \) and the regression function \( \eta \) (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 misspecified model 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., \([40, 37]\). There, \( \mathcal{F} \) is a subclass of the linear span of \( M \) given functions \( f_1, \ldots, f_M \), for example, their convex hull or sparse linear (convex) hull. Functions \( f_1, \ldots, f_M \) are interpreted as some initial estimators of the regression
function $\eta$ 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 $f_1,\ldots,f_M$ 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 standard nonparametric regression setting, it is assumed that the model is well-specified, i.e., we have $Y_i = f(X_i) + \xi_i$ where the random errors $\xi_i$ satisfy $\mathbb{E}(\xi_i | X_i) = 0$ and $f$ belongs to a given functional class $F$. 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

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

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

$$W_n(F) \leq V_n(F),$$

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 $F$. 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 $F$ 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 $W_n(F)$ and $V_n(F)$. 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 6, 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 4, we compare the results to those in the literature. Section 7 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 $Z = X \times Y$. For $S = \{z_1,\ldots,z_n\} \in Z^n$ and a class $\mathcal{G}$ of real-valued functions on $Z$, consider the Rademacher average on $\mathcal{G}$:

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

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

For any bounded measurable function $g : \mathbb{Z} \to \mathbb{R}$, we set $P g = \mathbb{E}g(Z)$, where $Z = (X, Y)$, and $P_n g = \frac{1}{n} \sum_{i=1}^n g(z_i)$. For any $\epsilon > 0, 1 \leq p < \infty$, $S = \{z_1, \ldots, z_n\} \in \mathbb{Z}^n$ with $z_i = (x_i, y_i)$, we will denote by $\mathcal{N}_p(F, \epsilon, S)$ the empirical $\epsilon$-covering number of the class $F \subseteq \mathcal{Y}^{X}$ with respect to the $L_p$ pseudonorm

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

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

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

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

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

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

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

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

The set $\{1, \ldots, N\}$ is denoted by $[N]$. Let $[x]$ denote the minimal integer strictly greater than $x$, and $|F|$ the cardinality of $F$. 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 $\epsilon$-net on $F$ with respect to the empirical $\ell_2$ metric and to form the induced partition of $F$. 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 $\epsilon$ of the initial net is taken to be large enough, the method reduces to a single empirical risk minimization (ERM) procedure over the class $F$. 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 $\ell_2$ pseudometric associated with the subsample $S$ of cardinality $n$, and

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

Let $\hat{c}_1, \ldots, \hat{c}_N$ be an $\epsilon$-net on $F$ with respect to $d_S(\cdot, \cdot)$. We assume without loss of generality that it is proper, i.e., $\hat{c}_i \in F$ for $i = 1, \ldots, N$, and $N \geq 2$. Let $\mathcal{F}_1^S, \ldots, \mathcal{F}_N^S$ be the following partition of $F$ induced by $\hat{c}_i$'s:

$$\mathcal{F}_i^S = \mathcal{F}_i^S(\epsilon) = \left\{ f \in F : i \in \arg\min_{j=1,\ldots,N} d_S(f, \hat{c}_j) \right\}$$

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

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

Finally, at the third step we use the subsample $S''$ to aggregate the estimators $\{\hat{f}_{1}^{S,S'}, \ldots, \hat{f}_{N}^{S,S'}\}$. We call a function $\hat{f}(x, D_{3n})$ 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 $\delta > 0$,

$$L(\hat{f}) \leq \min_{i=1,\ldots,N} L(\hat{f}_{i}^{S,S'}) + C \frac{\log(N/\delta)}{n}$$

(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 $\hat{f}_{i}^{S,S'} \equiv g_i$ can be considered as fixed (non-random) functions, and $\hat{f}$ as a function of $S''$ only. There exist several examples of sharp $MS$-aggregates of fixed functions $g_1, \ldots, g_N$ [2, 29]. They are realized as mixtures:

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

(8)

where $\theta_i$ are some random weights measurable with respect to $S''$.

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

$$\sup_{S \in \mathcal{Z}^n} \hat{R}_n(\mathcal{G}[r, S], S) \leq \phi_n(r)$$

(9)

\(^1\)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 \( \phi_n \) is non-negative, non-decreasing, and \( \phi_n(r)/\sqrt{r} \) is non-increasing. Examples of such functions will be given below, cf. proof of Lemma 6. Let \( r^* = r^*(\mathcal{G}) \) denote the critical localization radius, i.e., an upper bound on the largest solution of the equation \( \phi_n(r) = r \). Define
\[
\beta = \frac{\log(N_2(\mathcal{F}, \epsilon, S)/\delta)}{n} + \log\log n, \quad \text{and} \quad \gamma = \sqrt{\epsilon^2 + r^* + \beta}.
\]

**Theorem 1.** 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 (see Eq. 8). Fix \( \epsilon > 0 \). Then there exists an absolute constant \( C \) such that for any \( \delta > 0 \), with probability at least \( 1 - 3\delta \),
\[
L(\tilde{f}) \leq \inf_{f \in \mathcal{F}} L(f) + C(\beta + \Xi(n, \epsilon, S')),
\]
where
\[
\Xi(n, \epsilon, S') = \min \left\{ \frac{d}{n}, \frac{1}{\sqrt{n}} \int_0^\gamma \sqrt{\log N_2(\mathcal{F}, \rho, S')} d\rho, \frac{d}{n}, \frac{1}{\sqrt{n}} \int_0^\gamma \sqrt{\log N_2(\mathcal{F}, \rho, S')} d\rho \right\}.
\]

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 (10) holds with the remainder term
\[
\Xi(n, \epsilon, S') = \min \left\{ \frac{d}{n}, \frac{1}{\sqrt{n}} \int_0^\gamma \sqrt{\log N_2(\mathcal{F}, \rho, S')} d\rho, \frac{d}{n}, \frac{1}{\sqrt{n}} \int_0^\gamma \sqrt{\log N_2(\mathcal{F}, \rho, S')} d\rho \right\}.
\]

**Remarks.**

1. The term \( \Xi(n, \epsilon, S') \) in Theorem 1 is a bound on the rate of convergence of the ERM \( \hat{f}_{S,S'}^i \) over the cell \( \mathcal{F}_i^S \). It is the minimum of three possible rates, the first of which (11) we prove in Section 8 (the main crux of this paper), the second (12) is due to [39], while the third term \( d/n \) present only for convex \( d \)-dimensional \( \mathcal{F} \) is due to [24]. We will use only the bound with the first remainder term (11). The second and third terms are included here for completeness, in order to provide a general result that may be useful in other situations. 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 partition with cells \( \mathcal{F}_i^S \) defined above can be viewed as a default option. In some situations, we may better tailor the (possibly overcomplete) partition 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{G}_i}) for {\hat{G}_i} = \{(f - g)^2 : f, g \in \hat{F}_i\} instead of the larger quantity r^*(\mathcal{G})$. Inspection of the proof shows that the oracle inequality (10) generalizes to

$$L(\hat{f}) \leq \min_{i=1,...,N} \inf_{f \in \hat{F}_i} \{L(f) + C(\beta + \Xi_i(n, \epsilon, S'))\},$$

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 $r^*({\hat{G}_i})$.

The oracle inequality (10) of Theorem 1 depends on three quantities that should be specified: the empirical entropy numbers $\log N_2(\mathcal{F}, \cdot, \cdot)$, the localization radius $r^*$ and the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$. The crucial role in determining the rate belongs to the empirical entropies. We further replace in (10)-(13) these random entropies by their upper bound

$$\mathcal{H}_2(\mathcal{F}, \rho) = \sup_{S \in \mathcal{Z}^n} \log N_2(\mathcal{F}, \rho, S).$$

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 (10).

**Theorem 2.** Let $\mathcal{Y} = [0,1]$ and $\mathcal{H}_2(\mathcal{F}, \rho) \leq A\rho^{-p}$, $\forall \rho > 0$, for some constants $A < \infty$, $p > 0$. Let $\hat{f}$ be a sharp MS-aggregate (see Eq. 8) defined by the above three-stage procedure with the covering radius $\epsilon = n^{-\frac{1}{2p}}$. There exist constants $C_p > 0$ depending only on $A$ and $p$ such that:

(i) Let $0 \leq \rho \leq 1$ for all $f \in \mathcal{F}$. For the estimator $\hat{f}$ we have:

$$V_n(\mathcal{F}) \leq \sup_{f \in \mathcal{F}} \left\{ \mathbb{E} \|\hat{f} - f\|^2 - \inf_{f \in \mathcal{F}} \|f - n\|^2 \right\} \leq \begin{cases} C_p n^{-\frac{2}{2p}} & \text{if } p \in (0,2], \\ C_p n^{-\frac{1}{p}} & \text{if } p \in (2,\infty). \end{cases}$$

(ii) If the model is well-specified, then for the estimator $\tilde{f}$ we have:

$$W_n(\mathcal{F}) \leq \sup_{f \in \mathcal{F}} \mathbb{E} \|\tilde{f} - f\|^2 \leq C_p n^{-\frac{2}{2p}}, \quad \forall \; p > 0.$$
We will further comment on this choice in Section 4.

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

$$\sup_{S \in \mathbb{Z}^n} N_2(\mathcal{F}, \rho, S) \leq \frac{A}{\rho^v},$$

(17)

for any estimator, there exists a distribution on which the estimator differs from the regression function by at least $C v \log(en/v)$ with positive fixed probability. So, the extra logarithmic factor $\log(en/v)$ in the rate is necessary, even when the model is well-specified.

Theorem 3. (Bounds for VC-type classes). Assume that $\mathcal{Y} = [0, 1]$ and the empirical covering numbers satisfy (17). 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 $\epsilon = n^{-2/3}$. If $n \geq v$, there exists a constant $C > 0$ depending only on $A$ such that

$$V_n(\mathcal{F}) \leq \sup_{P_X \in \mathcal{P}} \left\{ E \left\| \tilde{f} - \eta \right\|^2 - \inf_{f \in \mathcal{F}} \left\| f - \eta \right\|^2 \right\} \leq C \frac{v \log(en/v)}{n}.$$  

(18)

The rate of convergence of the excess risk as in (18) for VC-type classes has been obtained previously under the assumption that $L^* = 0$ or for convex classes $\mathcal{F}$ (see discussion in Section 6 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 $C \frac{v \log(en/v)}{n}$ with positive fixed probability. So, the extra logarithmic factor $\log(en/v)$ in the rate is necessary, even when the model is well-specified.

The next theorem deals with classes of functions

$$\mathcal{F} = \mathcal{F}_{\Theta} = \left\{ f_\Theta = \sum_{j=1}^M \theta_j f_j : (\theta_1, \ldots, \theta_M) \in \Theta \right\},$$

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 $\theta$. Such classes arise in the context of aggregation, cf., e.g., [40], [37], where the main problem is to study the behavior of the minimax regret $V_n(\mathcal{F}_{\Theta})$ based on the geometry of $\Theta$. For the case of fixed rather than random design, we refer to [37] for a comprehensive treatment. Here, we deal with the random design case and consider several basic examples of sets $\Theta$ defined in terms of $\ell_p$-balls

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

where $\| \theta \|_0$ denotes the number of non-zero components of $\theta$, 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, \( F_\Theta = \{ f_1, \ldots, f_M \} \) or equivalently \( \Theta = \Theta^{MS} \triangleq \{ e_1, \ldots, e_M \} = \Lambda_M \cap B_0(1) \), where \( e_1, \ldots, e_M \) 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 \( F = \text{conv}(f_1, \ldots, f_M) \), i.e., the function that attains the minimum \( \min_{\theta \in \Lambda_M} \| f_\theta - \eta \|^2 \). In this case, \( F = F_\Theta \) with \( \Theta = \Theta^C \triangleq \Lambda_M \). 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 \( f_1, \ldots, f_M \). This corresponds to the set \( \Theta^C(s) = \Lambda_M \cap B_0(s) \). Note that $MS$-aggregation and convex aggregation are particular cases of $s$-sparse aggregation: \( \Theta^{MS} = \Theta^C(1) \) and \( \Theta^C = \Theta^C(M) \).

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

\[
\{F_1, \ldots, F_N\} = \{F_{\nu,m}, m = 1, \ldots, s, \nu \in I_m\}
\]

instead of random ones \( \hat{F}_i^S \). 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 \( \Lambda_M \). The resulting aggregate is denoted by \( f^* \).

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

\[
V_n(F_{\Theta^C(s)}) \leq \sup_{P_{X,Y} \in P} \left\{ \mathbb{E} \left\| \tilde{f}^* - \eta \right\|^2 - \inf_{\theta \in \Theta^C(s)} \| f_\theta - \eta \| \right\} \leq C \psi_{n,M}(s)
\]

where

\[
\psi_{n,M}(s) = \frac{s}{n} \log \left( \frac{eM}{s} \right) \wedge \sqrt{\frac{1}{n} \log \left( 1 + \frac{M}{\sqrt{n}} \right)} \wedge 1
\]

for \( s \in \{1, \ldots, M\} \).

This theorem improves upon the rate of \( s \)-sparse aggregation given in Lounici [31] by removing a redundant \( (s/n) \log n \) term present there. Note that [31] 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 [31], we get that \( \psi_{n,M}(s) \) is the optimal rate for the minimax regret on \( F_{\Theta^C(s)} \) in our setting. Analogous result for gaussian regression with fixed design is proved in [37].

4 Comparison with global ERM and with skeleton aggregation

The estimation procedure we propose here has three steps. The first is to find an empirical \( \epsilon \)-net (that we will call a skeleton) from the first subsample and partition the function class based on
the skeleton using the empirical distance on this subsample. In the next step, using the second subsample we find empirical risk minimizers within each cell of the partition. Finally, we use the third sample to aggregate these ERM’s. A simpler procedure that we will call skeleton aggregation consists of steps one and three, but not two. This method directly aggregates centers of the cells \( \hat{F}_i^e(\varepsilon) \), i.e., the elements \( \hat{c}_i \) of the \( \varepsilon \)-net obtained from the first subsample \( S \). Such kind of procedure was first proposed by Yang and Barron [48] in the context of well-specified models. The setting in [48] is different from the ours since in that paper the \( \varepsilon \)-net is taken with respect to a non-random metric and the bounds on the minimax risk \( W_n(F) \) are obtained when the regression errors are Gaussian. We argue that, while the skeleton aggregation achieves the desired rates for well-specified models (i.e., for the minimax risk), one cannot expect it to be successful for the misspecified setting. This will explain why aggregating ERM’s in cells of the partition (and not simply the centers of cells) is crucial for the success of our procedure.

Let us first show why the skeleton aggregation yields the correct rates for well-specified models (i.e., when \( \eta \in F \)). Similarly to (8), we define the skeleton aggregate \( \tilde{f}^{sk} = \sum_{i=1}^{N} \theta_i \hat{c}_i \) as a sharp MS-aggregate satisfying a bound analogous to (7): There exists a constant \( C > 0 \) such that, for any \( \delta > 0 \),

\[
L(f^{sk}) \leq \min_{i=1, \ldots, N} L(\hat{c}_i) + C\frac{\log(N/\delta)}{n} \tag{20}
\]

with probability at least \( 1 - \delta \) over the sample \( S'' \), conditionally on \( S \) (the subsample \( S' \) is not used here). If the model is well-specified, \( L^* = L(\eta) \), and \( \|f - \eta\|^2 = L(f) - L^* \), \( \forall f \in F \). Hence, with probability \( 1 - 5\delta \),

\[
\|f^{sk} - \eta\|^2 = L(f^{sk}) - L^* \leq \min_{i=1, \ldots, N} L(\hat{c}_i) - L^* + C\frac{\log(N/\delta)}{n}
\]

\[
= \min_{i=1, \ldots, N} \|\hat{c}_i - \eta\|^2 + C\frac{\log(N/\delta)}{n}
\]

\[
\leq 2\varepsilon^2 + C\left( \frac{H_2(F, \varepsilon)}{n} + \frac{\log(1/\delta)}{n} + r^* + \beta' \right) \tag{21}
\]

for \( \beta' = (\log(1/\delta) + 6\log\log n)/n \), and \( r^* = r^* (G) \), where we have used Lemma 7 with \( f = \hat{c}_i, f' = \eta \) and the fact that \( \min_{i=1, \ldots, N} d_S(\hat{c}_i, \eta) \leq \varepsilon \) for any \( \eta \in F \). The optimal choice of \( \varepsilon \) in (21) is given by the balance equation \( n\varepsilon^2 \geq H_2(F, \varepsilon) \) and it follows from Lemma 6 that \( r^* + \beta' \) is negligible as compared to the leading part \( O(\varepsilon^2 + H_2(F, \varepsilon)/n) \) with this optimal \( \varepsilon \). In particular, we get from (21) that, under the assumptions of Theorem 2, \( \sup_{\eta \in F} \mathbb{E}\|f^{sk} - \eta\|^2 \leq Cn^{-\frac{2}{2p}}, \; \forall \; p > 0. \)

Let us now consider the misspecified model setting (i.e., the statistical learning framework). Here, the balance equation for the skeleton aggregation takes the form \( n\varepsilon \geq H_2(F, \varepsilon) \), which yields suboptimal rates unless the class \( F \) is finite. Indeed, without the assumption that the regression function \( \eta \) is in \( F \), we only obtain the bounds

\[
L(\hat{c}_i) - L^* = \|\hat{c}_i - \eta\|^2 - \inf_{f \in F} \|f - \eta\|^2 \leq 2(\|\hat{c}_i - \eta\| - \|\eta_F - \eta\|) + \frac{1}{n} \leq 2\|\hat{c}_i - \eta_F\| + \frac{1}{n} \tag{22}
\]

where \( \eta_F \in F \) is such that \( \|\eta_F - \eta\|^2 \leq \inf_{f \in F} \|f - \eta\|^2 + 1/n. \) The crucial difference from (21) is that here \( L(\hat{c}_i) - L^* \) behaves itself as a norm \( \|\hat{c}_i - \eta_F\| \) and not as a squared norm \( \|\hat{c}_i - \eta\|^2 \). Using (22)
and arguing analogously to (21) we find that for misspecified models, with probability $1 - 5\delta$,

$$L(\hat{f}^sk) - L^* \leq 2 \min_{i=1,\ldots,N} \|\hat{c}_i - \eta_{F}\| + C \frac{\log(N/\delta)}{n}$$

$$\leq 2\sqrt{2\epsilon^2 + C(r^* + \beta')} + C \frac{\log(N/\delta)}{n}$$

$$\leq 2\sqrt{2\epsilon} + C \left(\frac{\mathcal{H}_2(F, \epsilon)}{n} + \frac{\log(1/\delta)}{n} + \sqrt{r^* + \beta'}\right).$$

(23)

Here, the optimal $\epsilon$ is obtained from the tradeoff of $\epsilon$ with $\mathcal{H}_2(F, \epsilon)/n$. As a result, we only get the suboptimal rate $n^{-1/(p+1)} + O(\sqrt{r^* + \beta'})$ for the excess risk of $\hat{f}^sk$ under the assumptions of Theorem 2. This indicates that introducing the ERM’s in the cells of the partition (the second step of our procedure) is crucial in getting the right rates.

We can now compare the following three estimators. First, the ERM over $F$:

$$\hat{f}^{erm} \in \arg\min_{f \in F} \frac{1}{n} \sum_{(x,y) \in S} (f(x) - y)^2,$$

second the skeleton aggregate $\hat{f}^sk$ and, finally, the proposed procedure $\tilde{f}$ which consists in aggregating ERM’s on the cells of the partition. Table 1 summarizes the convergence rates of the expected excess risk $\mathcal{E}_F(\hat{f})$ for $\hat{f} \in \{ \tilde{f}, \hat{f}^sk, \hat{f}^{erm} \}$, in misspecified model setting.

<table>
<thead>
<tr>
<th>Regime</th>
<th>Proposed Method</th>
<th>Skeleton aggregation</th>
<th>ERM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finite: $</td>
<td>F</td>
<td>= M$</td>
<td>$\frac{\log M}{n}$</td>
</tr>
<tr>
<td>Parametric: $VC(F) = v \leq n$</td>
<td>$\frac{v\log(en/v)}{n}$</td>
<td>$\frac{v\log(en/v)}{n}$</td>
<td>$\sqrt{\frac{v}{n}}$</td>
</tr>
<tr>
<td>Nonparametric: $\mathcal{H}_2(F, \epsilon) = \epsilon^{-p}$, $p \in (0, 2)$</td>
<td>$n^{-\frac{2}{2+p}}$</td>
<td>$n^{-\frac{1}{p+1}} \lor n^{-\frac{1}{2}}(\log n)^{\frac{3}{2}}$</td>
<td>$n^{-\frac{1}{2}}$</td>
</tr>
<tr>
<td>$p \in [2, \infty)$</td>
<td>$n^{-\frac{1}{p}}$</td>
<td>$n^{-\frac{1}{p+1}}$</td>
<td>$n^{-\frac{1}{p}}$</td>
</tr>
</tbody>
</table>

Table 1: Summary of rates for misspecified case.

The rates for finite $F$ in Table 1 are obtained in a trivial way by taking the skeleton that coincides with the $M$ functions in the class $F$. In parametric and nonparametric regime, the rates for the proposed method are taken from Theorems 2 and 3, while for the skeleton aggregate they follow from (21) with optimized $\epsilon$ and the bounds on $r^*$ in Lemma 6 and in (28), (31) below. Finally, the rates for the ERM in Table 1 are well-known. For the nonparametric regime, they follow, for example, from Lemma 10 and the bounds on $\mathcal{R}_n(F)$ in (28) and (31) below. Moreover, for finite $F$, it can be shown that the slow rate $\sqrt{\frac{\log M}{n}}$ cannot be improved neither for ERM, nor for any other selector, i.e., any estimator with values in $F$, cf. [19].

In conclusion, for finite class $F$ the proposed method and skeleton aggregation are optimal whereas ERM has a suboptimal rate. For a very massive class $F$, when the empirical entropy grows polynomially as $\epsilon^{-p}$ with $p \geq 2$ both ERM and proposed method enjoy similar guarantees of rates of
order $n^{-1/p}$ while the skeleton aggregation only gets a suboptimal rate of $n^{-1/(p+1)}$. For all other cases, while the proposed method is optimal, both ERM and skeleton aggregation are suboptimal. Note also that, unless $\mathcal{F}$ is finite, the skeleton aggregation does not improve upon the ERM.

Turning to the well-specified case, both the proposed method and the skeleton aggregation achieve the optimal rate for the minimax risk while the ERM is, in general, suboptimal.

## 5 Adapting to Approximation Error Rate of Function Class

Often in statistical learning problems the choice of a function class $\mathcal{F}$ is not fixed and is, in fact, a design choice. The art of picking the right function class $\mathcal{F}$ depends on how best we can trade-off statistical learning error with its approximation error for the unknown regression function $\eta$. Recall that in Theorem 2 we have shown that for $p > 2$ our estimator has the rate of $n^{-2} \sqrt{p}$ when $\eta \in \mathcal{F}$ and achieves the rate of $n^{-1/p}$ if not. A natural question one can ask is what happens if $\eta \notin \mathcal{F}$ but the approximation error $\inf_{f \in \mathcal{F}} \|\eta - f\|^2$ is small. In this case one would like to achieve rates varying between $n^{-1/p}$ and $n^{-2/(2+p)}$ based on how small the approximation error rate is.

**Lemma 5.** Let $\mathcal{Y} = [0,1]$ and $\mathcal{H}_2(\mathcal{F}, \rho) \leq A \rho^{-p}$, $\forall \rho > 0$, for some constants $A < \infty$, $p > 2$. Consider the sharp $MS$-aggregate, $\tilde{f}$, defined by the three-stage procedure with the covering radius set as $\epsilon = n^{-1/2p}$. For this estimator and for any joint distribution $P_{XY}$ we have:

$$
\mathbb{E} \|\tilde{f} - \eta\|^2 - \argmin_{f \in \mathcal{F}} \|f - \eta\|^2 \leq \begin{cases} 
O\left(n^{-\frac{2}{2p}}\right) & \text{if } \Delta^2 \leq n^{-2/(2+p)} \\
O\left(\Delta^2\right) & \text{if } n^{-2/(2+p)} \leq \Delta^2 \leq n^{-1/p} \\
O\left(n^{-1/p}\right) & \text{otherwise}
\end{cases}
$$

(24)

where $\Delta^2 = \inf_{f \in \mathcal{F}} \|\eta - f\|^2$.

**Proof.** Let $f^* = \argmin_{f \in \mathcal{F}} \|f - \eta\|^2$. For any $\epsilon > 0$, for the the partition $\mathcal{F}_i^S(\epsilon)$ that contains $f^*$ we have that

$$
\|\tilde{f}_i^{S,S'} - \eta\|^2 - \|f^* - \eta\|^2 \leq 2 \|\tilde{f}_i^{S,S'} - f^*\|^2 + \|f^* - \eta\|^2
\leq 4 \|d^2_S(\tilde{f}_i^{S,S'}, f^*) + \|f^* - \eta\|^2 + C \log^3(n) \mathcal{R}_n^2(\mathcal{F})
\leq 4\epsilon^2 + \Delta^2 + O\left(\frac{\log^3(n)}{n^{2/p}}\right),
$$

where the last term is not dominating by our choice of $\epsilon$ below. The second inequality is by using Theorem 16 along with Lemma 6 to bound $\|\tilde{f}_i^{S,S'} - f^*\|^2$ in terms of twice the empirical distance $d^2_S(\tilde{f}_i^{S,S'}, f^*)$, on similar lines as the inclusion lemma (Lemma 12). Further, we also have by simple Rademacher bound (see Corollary 11) that

$$
\mathbb{E} \|\tilde{f}_i^{S,S'} - \eta\|^2 - \|f^* - \eta\|^2 \leq C \mathcal{ER}_n(\mathcal{F}_i^S(\epsilon)) \leq O\left(n^{-1/p}\right)
$$

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

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

12
The bound for the excess risk is then given by
\[
E \| \hat{f} - \eta \|^2 - \inf_{f \in \mathcal{F}} \| f - \eta \|^2 \leq \frac{\log(\epsilon^{-p})}{n} + O\left( \min\left( n^{-\frac{2}{2+p}} + \Delta^2, \frac{1}{n^{1/p}} \right) \right)
\]
\[
\leq n^{-\frac{2}{2+p}} + O\left( \min\left( n^{-\frac{2}{2+p}} + \Delta^2, \frac{1}{n^{1/p}} \right) \right)
\]
This gives the interpolation in Eq. (24).

We therefore see a smooth transition in terms of approximation error rate in the regime \( \Delta^2 \in \left( n^{-2/(2+p)}, n^{-1/p} \right) \). Notice that the estimator is still the same proposed algorithm with choice of \( \epsilon \) fixed at \( n^{-1/(2+p)} \), however the estimation procedure automatically enjoys the adaptive rate.

6 Historical Remarks and Comparison to Previous Work

The role of entropy and capacity [21] in establishing rates of estimation has been recognized for a long time, since the work of Le Cam [26], Ibragimov and Khas’minskii [18] and Birgé [7]. Other early work on this subject involving estimation on \( \epsilon \)-nets is due to Devroye [11] and Devroye et al. [12]. 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 an appropriately chosen non-random entropy \( \mathcal{H}(\cdot) \). Van de Geer [42] invokes the empirical entropy rather than the non-random entropy to derive rates of estimation in regression problems. Yang and Barron [48] 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 and non-random entropy. Under the assumption that the regression errors are gaussian, [48] 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 [46] 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) methods 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, 34, 16, 15, 13] 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 [44, 45, 46]. 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 [35, 17]. Lee, Bartlett and Williamson [30] 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 $\sqrt{\log |\mathcal{F}|}/n$ [19]. 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 [33] proposed to study the problem of aggregation, or mimicking the best function in the given class, for regression models. Nemirovskii [33] 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 [47, 20, 19, 31, 37]. For recent developments and a survey we refer to [28, 38].

In parallel with these developments, the study of the excess risk blossomed with the introduction of Rademacher and local Rademacher averages in [22, 25, 3, 9, 4, 23]. 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 [23, 24] obtained the exact inequality with the correct rate $d/n$ for ERM. However, the convexity assumption appears to be crucial; without this assumption Koltchinskii [24, 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 [29]. 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 mixture uses two functions. In [29], 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 [10] 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 limit of the Dudley integral is the radius of the balls. This method served as an inspiration for the present work.
7 Proofs of Theorems 2-4

We first state some auxiliary lemmas. Recall that $\mathcal{G} = \{(f - g)^2 : f, g \in \mathcal{F}\}$.

**Lemma 6.** The following critical radii $r^* = r^*(\mathcal{G})$ satisfy the conditions stated before Theorem 1.

(i) For any class $\mathcal{F} \subseteq \{ f : 0 \leq f \leq 1 \}$,

$$r^* = C \log^3(n) \mathfrak{H}^{2}(\mathcal{F}).$$

(ii) If $\mathcal{F} \subseteq \{ f : 0 \leq f \leq 1 \}$ and the empirical covering numbers exhibit the polynomial growth $\sup_{S \in \mathbb{Z}^n} \mathcal{N}_2(\mathcal{F}, \rho, S) \leq \left(\frac{A}{\rho}\right)^v$ for some constants $A < \infty$ and $0 < v \leq n$, then

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

(iii) If $\mathcal{F}$ is a finite class,

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

The following lemma is a direct consequence of Theorem 16.

**Lemma 7.** Let $f, f' \in \mathcal{F}$, and $0 < \delta < 1/4$. Then, with probability at least $1 - 4\delta$,

$$\|f - f'|^2 \leq 2d^2_2(f,f') + C(r^* + \beta')$$

where $\beta' = (\log(1/\delta) + 6 \log \log n)/n$, and $r^* = r^*(\mathcal{G})$.

We will also use the following bound on the Rademacher average in terms of the empirical entropy [5, 39, 43]:

$$\mathfrak{R}_n(\mathcal{F}, S) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, S)} d\rho \right\}.$$  \hspace{1cm} (27)

### 7.1 Proof of Theorem 2

Assume without loss of generality that $A = 1$, i.e., $\sup_{S \in \mathbb{Z}^n} \log \mathcal{N}_2(\mathcal{F}, \rho, S) \leq \rho^{-p}$ 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 (27) with $\alpha = 0$ combined with (25) yields

$$\mathfrak{R}_n(\mathcal{F}) \leq \frac{12}{\sqrt{n}(1 - p/2)}, \quad r^* \leq C \frac{(\log n)^3}{n}$$

for some absolute constant $C$. Next,

$$\beta \leq C \frac{(\epsilon^{-p} + \log \log n + \log(1/\delta))}{n}$$
and
\[ \gamma^2 \leq C \left( \epsilon^2 + \frac{(\log n)^3}{n} + \frac{\epsilon^{-p} + \log(1/\delta)}{n} \right). \] (29)

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

These inequalities together with (10) and (11) yield that, with probability at least 1 - 3\delta,
\[ L(\hat{f}) - L^* \leq C \left( \frac{\epsilon^{-p}}{n} + \frac{\log(1/\delta)}{n} + \gamma \sqrt{r^*} + \frac{\gamma^{1-p/2}}{\sqrt{n}} \right). \] (30)

The value of \( \epsilon \) minimizing the right hand side in (29) and in (30) is given by solving \( \epsilon^2 \approx 1/(e^n) \), 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 (30) is \( C n^{-\frac{2}{2+p}} \) 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 \( C n^{-\frac{2}{2+p}} \), which proves (15) 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 (16) and misspecified models (15), so we consider the two cases separately.

1. **Proof of (15) for \( p \in (2, \infty) \).** Here, the rate is governed by the Rademacher complexity of the function class. Using (27) we bound \( \mathcal{R}_n(\mathcal{F}) \) as follows:
\[ \mathcal{R}_n(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \rho^{-p/2} d\rho \right\} \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{24}{\sqrt{n}(p-2)} \alpha^{-(p-2)/2} \right\}. \]

Balancing \( \alpha = n^{-1/2} \alpha^{-(p-2)/2} \) yields \( \alpha = n^{-1/p} \). This and (25) lead to the bounds
\[ \mathcal{R}_n(\mathcal{F}) \leq C n^{-1/p}, \quad r^* \leq C (\log n)^{3} n^{-2/p}. \] (31)

Thus, Lemma 10 implies
\[ \mathbb{E}L(\hat{f}^{S,S'}_i) - \inf_{f \in \mathcal{F}_i^{S}} L(f) \leq C n^{-1/p}. \]

The aggregation step (7) adds to this bound the term \( \log \mathcal{N}(\mathcal{F}, \epsilon, S)/n \leq 1/(\epsilon^n) \), so that
\[ \mathbb{E}L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f) \leq C (n^{-1/p} + 1/(\epsilon^n)). \]

Now plugging in \( \epsilon = n^{-1/(2+p)} \) we obtain the overall rate \( n^{-1/p} \).

2. **Proof of (16) for \( p \in (2, \infty) \).** Now consider the case when \( \eta \in \mathcal{F} \). Then we have \( L(\eta) = L^* \) and \( L(f) - L^* = \|f - \eta\|^2 \) for all \( f \in \mathcal{F} \). From Lemma 7 we obtain that, for any \( i = 1, \ldots, N \), with probability at least \( 1 - 4\delta \),
\[ \|\hat{f}^{S,S'}_i - \eta\|^2 \leq 2d_2^2(\hat{f}^{S,S'}_i, \eta) + C(r^* + \beta'). \]
Hence, intersecting with the event of inequality (7), we get that, with probability at least $1 - 5\delta$

$$\|\hat{f} - \eta\|^2 = L(\hat{f}) - L^* \leq \min_{i=1,\ldots,N} L(\hat{f}^S_{ij}) - L^* + \frac{C\log(N/\delta)}{n}$$

$$= \min_{i=1,\ldots,N} \|\hat{f}^S_{ij} - \eta\|^2 + \frac{C\log(N/\delta)}{n}$$

$$\leq 8\epsilon^2 + \frac{C(\epsilon^{-p} + \log(1/\delta)) + +C(r^* + \beta')}{n}$$

where we have used that $\min_{i=1,\ldots,N} d_S(f^S_{ij}, \eta) \leq 2\epsilon$. The expression in the last line of this display has the best rate for $\epsilon = n^{-1/(2+p)}$. Using (31) we get that $r^* + \beta'$ is of smaller order than the other terms when $\epsilon = n^{-1/(2+p)}$. This implies that $\mathbb{E}\|\hat{f} - \eta\|^2 = O(n^{-\frac{2}{2+p}})$ uniformly over $\eta \in \mathcal{F}$ as claimed.

### 7.2 Proof of Theorem 3

Choosing $\epsilon = n^{-1/2}$ and using the expression for $r^*$ in Lemma 6 (ii) leads to the bounds $\beta \leq C\sqrt{\log(n/\delta)/n}$, and $\gamma \leq C\sqrt{\log(n/\delta)/n}$. By some algebra, we obtain that the overall rate in expectation is then $O\left(\frac{\sqrt{\log(en/\delta)}}{n}\right)$.

### 7.3 Proof of Theorem 4

As shown in [27], the rate of ERM for the simplex in $\mathbb{R}^s$ is

$$O\left(\frac{s \wedge \sqrt{\frac{1}{n} \log\left(1 + \frac{s}{\sqrt{n}}\right)}}{n}\right).$$

The same result yields that the rate is not worse than

$$\sqrt{\frac{1}{n} \log\left(1 + \frac{M}{\sqrt{n}}\right)}$$

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 $N = \sum_{j=1}^s M_j \leq \left(\frac{eM}{s}\right)^s$, we obtain the overall rate of the order

$$\left[\frac{s \log\left(\frac{eM}{s}\right)}{n} \wedge \sqrt{\frac{1}{n} \log\left(1 + \frac{s}{\sqrt{n}}\right)}\right] \wedge \sqrt{\frac{1}{n} \log\left(1 + \frac{M}{\sqrt{n}}\right)} \leq C\left(\frac{s \log\left(\frac{eM}{s}\right)}{n} \wedge \sqrt{\frac{1}{n} \log\left(1 + \frac{M}{\sqrt{n}}\right)}\right).$$

### 8 Proof of Theorem 1

We break down the proof of Theorem 1 into several subsections.
8.1 The general scheme

**Proposition 8.** Let \( 0 \leq f \leq 1 \) for all \( f \in \mathcal{F} \), and \( 0 < \delta < 1/2 \). Then for any \( \epsilon > 0 \), with probability at least \( 1 - 2\delta \)

\[
L(\hat{f}) \leq L^* + C \frac{\log(N_d(\mathcal{F}, \epsilon, S)/\delta)}{n} + \Xi(n, \epsilon, S')
\]

(32)

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

\[
\min_{i=1, \ldots, N} L(\hat{f}^{S, S'}_i) - L^* \leq \Xi(n, \epsilon, S').
\]

(33)

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. \([6, 39]\)), we prove them here for completeness with explicit constants. The proof of this Lemma is deferred to page 28.

**Lemma 9 \(([39, 8])\).** Let \( \hat{g} \) be an empirical minimizer over a class \( \mathcal{G} \),

\[
\hat{g} = \arg\min_{g \in \mathcal{G}} P_n g.
\]

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

\[
P\hat{g} \leq Pg^* + \sqrt{P g^* \sqrt{20r^* + 17r_0}} + 114r^* + 53r_0
\]

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

**Lemma 10.** Let \( \mathcal{Y} = [0, 1] \) and let \( \mathcal{F} \) be a class of functions from \( X \) to \( \mathcal{Y} \). Then, for any \( x > 0 \), with probability at least \( 1 - 2e^{-x} \), the empirical risk minimizer \( \hat{f}_{\text{erm}} \) on \( \mathcal{F} \) satisfies

\[
L(\hat{f}_{\text{erm}}) - L^* \leq c'_2 R_n(\mathcal{F}) + \frac{c'_3 x}{n}
\]

where \( c'_2 = 1408 \) and \( c'_3 = 830 \).

Using these lemmas, we obtain the following corollary:

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

\[
L(\hat{f}^{S, S'}_i) \leq L^*_i + \min \left\{ c'_2 R_n(\ell \circ \hat{F}^S_i, S') + c'_3 r_0, \sqrt{L^*_i} \sqrt{20r + 17r_0} + 114r + 53r_0 \right\}
\]

where \( r_0 = (x + 6 \log \log n)/n \) and \( \tau = 12 \cdot 42^2 \log^3(64n) R^2_n(\mathcal{F}) \).
Proof. Recall that each $\hat{f}_i^{S,S'}$ is an empirical minimizer over the respective set $\hat{F}_i^S$. Let $\mathcal{E}_i$ be the event (with respect to the draw of $S'$, conditionally on $S$) under which Lemma 9 and Lemma 10 (applied to $\hat{f}_i^{S,S'}$) hold simultaneously for all $i \in [N]$. We have $P(\mathcal{E}_i) \geq 1 - 11Ne^{-x}$. That is, with probability at least $1 - 11Ne^{-x}$, for all $i \in [N]$,

$$L\left(\hat{f}_i^{S,S'}\right) \leq L_i^* + c_2\mathcal{R}_n(\ell \circ \hat{F}_i^S, S') + c_3r_0$$

(34)

where $L_i^* = \arg\min_{f \in \hat{F}_i^S} L(f)$. Under the same event $\mathcal{E}_i$, we also have an alternative optimistic bound in terms of the minimal risk, as implied by Lemma 9: for all $i \in [N]$,

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

where $\tau = r^*(\ell \circ \mathcal{F})$ can be taken to be $\tau = 12 \cdot 21^2 \log^3(64n)\mathcal{R}_n^2(\mathcal{F})$ as shown in [39]. Combining with (34), the result follows.

8.3 An Inclusion Lemma

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

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

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

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

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

Lemma 12. 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

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

and hence

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

Proof of Lemma 12. 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 $\phi_n$ such that $\mathcal{R}_n(\mathcal{G}[r,S], S) \leq \phi_n(r)$ and let $r^* = r^*(\mathcal{G})$ be an upper bound on the largest solution $\phi_n(r) = r$. We now appeal to Theorem 16. With probability at least $1 - 4\epsilon^{-x}$, for any $f, g \in \mathcal{F}$

$$P(f - g)^2 \leq 2P_n(f - g)^2 + 106r^* + 48r_0$$
and
\[ P_n(f - g)^2 \leq 2P(f - g)^2 + 72r^* + 22r_0 \]
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 \( f, g \in \mathcal{F} \)
\[ P'_n(f - g)^2 \leq 4P_n(f - g)^2 + 284r^* + 118r_0. \]
Taking a union bound over \( i \in [N] \) completes the proof. \( \square \)

### 8.4 Controlling the Rademacher Complexity

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

**Lemma 13.** Let \( r^* = r^* (\mathcal{G}) \) for \( \mathcal{G} = \{(f - g)^2 : f, g \in \mathcal{F}\} \), and suppose \( \gamma^2 \geq r^* \). Then
\[ \mathcal{R}_n \left( \ell \circ \hat{\mathcal{F}}^{S,S'}_i(\gamma), S' \right) \leq \gamma \sqrt{r^*} + \frac{20}{\sqrt{n}} \int_0^{\gamma} \sqrt{\log N_2(\mathcal{F}, \rho, S')} d\rho \]

**Proof.** We reason conditionally on \( S \cup S' \). We have,
\[
\mathcal{R}_n \left( \ell \circ \hat{\mathcal{F}}^{S,S'}_i(\gamma), S' \right) \\
= \mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}^{S,S'}_i(\gamma)} \frac{1}{n} \sum_{(x, y) \in S'} \sigma_j(f(x) - y_j)^2 \\
= \mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}^{S,S'}_i(\gamma)} \frac{1}{n} \sum_{(x, y) \in S'} \sigma_j(f(x) - \hat{c}_i(x))^2 + \sigma_j(\hat{c}_i(x) - y_j)^2 + 2\sigma_j(f(x) - \hat{c}_i(x))(\hat{c}_i(x) - y_j) \\
\leq \mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}^{S,S'}_i(\gamma)} \frac{1}{n} \sum_{(x, y) \in S'} \sigma_j(f(x) - \hat{c}_i(x))^2 + 2\mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}^{S,S'}_i(\gamma)} \frac{1}{n} \sum_{(x, y) \in S'} \sigma_j(f(x) - \hat{c}_i(x))(\hat{c}_i(x) - y_j) \\
\leq \gamma \sqrt{r^*}
\]

(35)

Consider the first term. Conditioned on the set \( S \), the centers \( \hat{c}_i \) are fixed and we may view the set \( \hat{\mathcal{F}}^{S,S'}_i(\gamma) \) as giving rise to the set of \( \gamma^2 \)-approximate empirical minimizers
\[ \mathcal{G}'_i = \left\{ (f - \hat{c}_i)^2 : f \in \mathcal{F}, \frac{1}{n} \sum_{(x, y) \in S'} (f(x) - \hat{c}_i(x))^2 \right\} \]

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 \( \mathcal{G}'_i \in \mathcal{G}[\gamma^2, S'] \) where \( \mathcal{G} = \{(f - g)^2 : f, g \in \mathcal{F}\} \). Then the first term in (35) is
\[
\mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}^{S,S'}_i(\gamma)} \frac{1}{n} \sum_{(x, y) \in S'} \sigma_j(f(x) - \hat{c}_i(x))^2 \leq \mathcal{R}_n(\mathcal{G}[\gamma^2, S'], S') \leq \phi_n(\gamma^2) \leq \gamma \sqrt{r^*}
\]

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 (35). Define
\[ G_i^{S,S'} = \{ g_f(x,y) = (f(x) - \hat{c}_i(x))(\hat{c}_i(x) - y) : f \in \hat{F}_i^{S,S'}(\gamma) \} \]
First, observe that for any \( g_f \in G_i^{S,S'} \),
\[
\frac{1}{n} \sum_{(x,y) \in S'} g_f(x,y)^2 = \frac{1}{n} \sum_{(x,y) \in S'} (f(x) - \hat{c}_i(x))^2(\hat{c}_i(x) - y)^2 \leq \gamma^2
\]
under the boundedness assumption. Next, let \( M = N_2(\hat{F}_i^{S,S'}, \delta, S') \) be a covering number with respect to \( d_{S'}(f,g) \) and suppose \( C = \{ h_1, \ldots, h_M \} \) is such a \( \delta \)-cover. Pick any \( f \in \hat{F}_i^{S,S'} \) and let \( h \in C \) be a cover center \( \delta \)-close to \( f \) in the above sense. Then
\[
\frac{1}{n} \sum_{(x,y) \in S'} (g_f(x,y) - gh(x,y))^2 = \frac{1}{n} \sum_{(x,y) \in S'} [(f(x) - \hat{c}_i(x))(\hat{c}_i(x) - y) - (h(x) - \hat{c}_i(x))(\hat{c}_i(x) - y)]^2
\]
\[
= \frac{1}{n} \sum_{(x,y) \in S'} (f(x) - h(x))^2(\hat{c}_i(x) - y)^2 \leq \delta^2
\]
implying \( N_2(G_i^{S,S'}, \delta, S') \leq N_2(\hat{F}_i^{S,S'}, \delta, S') \). Hence,
\[
\mathcal{R}_n(G_i^{S,S'}) \leq \frac{10}{\sqrt{n}} \int_0^\gamma \sqrt{\log N_2(\hat{F}_i^{S,S'}, \rho, S')} d\rho \leq \frac{10}{\sqrt{n}} \int_0^\gamma \sqrt{\log N_2(F, \rho, S')} d\rho. \quad (36)
\]
Putting together the results,
\[
\mathcal{R}_n(\ell \circ \hat{F}_i^{S,S'}(\gamma), S') \leq \gamma \sqrt{r^2} + \frac{20}{\sqrt{n}} \int_0^\gamma \sqrt{\log N_2(F, \rho, S')} d\rho.
\]

8.5 Concluding the Proof

Putting together Corollary 11, Lemma 12, and Lemma 13, with probability at least 1 - 19Ne\(^{-x}\) over the draw of \( S \cup S' \), for any \( i \in [N] \),
\[
L\left( \hat{f}_i^{S,S'} \right) \leq L_i^* + \min \left\{ c_2' \left( \gamma \sqrt{r^2} + \frac{20}{\sqrt{n}} \int_0^\gamma \sqrt{\log N_2(F, \rho, S')} d\rho \right) + c_3' \sqrt{\log n} + 114r + 53r_0 \right\}
\]
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 \( i \in [N] \)
\[
L\left( \hat{f}_i^{S,S'} \right) \leq L_i^* + \min \left\{ c_2' \left( \gamma \sqrt{r^2} + \frac{20}{\sqrt{n}} \int_0^\gamma \sqrt{\log N_2(F, \rho, S')} d\rho \right) + c_3' \beta + \sqrt{L_i^* \sqrt{20\tau + 17\beta + 114\tau + 53\beta}} \right\}
\]
where \( \beta = (\log(38N/\delta) + 6 \log \log n)/n \) and \( \tau = 12 \cdot 21^2 \log^3(64n) \mathcal{R}_n^2(F) \).
Next, we appeal to Example 1 in [24], which implies that for any \(i \in [N]\) with probability at least \(1 - ce^{-x}\),
\[
L(f_i^S, S') - L(f_i^*) \leq \frac{K(d + x)}{n}
\]
where \(d\) is the dimensionality of the linear space to which \(F\) belongs. Taking a union bound over \(i \in [N]\) and letting \(\delta/2 = cN e^{-x}\), 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 \(F\) with VC-dimension at most \(d\) such that
\[
W_n(F) \geq C \frac{d \log(en/d)}{n}
\]
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, \(\mathcal{X} = \{x^1, x^2, \ldots\}\) is an infinite countable set of elements and \(F\) is the following set of binary-valued functions on \(\mathcal{X}\):
\[
F = \{f : f(x) = a1 \{x \in W\}, \text{ for some } W \subset \mathcal{X} \text{ with } \text{Card}(W) \leq d\},
\]
where \(a > 0\), \(1 \{\cdot\}\) denotes the indicator function and \(\text{Card}(W)\) is the cardinality of \(W\). It is easy to check that \(F\) is a VC subgraph class with VC-dimension at most \(d\).

**Theorem 14.** Let \(d\) be any integer such that \(n \geq d\), 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
\[
\inf_{f} \sup_{\eta \in F} P_{\eta} \left(\|\hat{f} - \eta\|^2 \geq c \frac{d \log(en/d)}{n}\right) \geq c',
\]
where \(P_\eta\) denotes the distribution of the \(n\)-sample \(D_n\) when \(E(Y|X = x) = \eta(x)\).

**Proof.** Fix some \(0 < \alpha < 1\) and set \(k = \lceil d/\alpha \rceil\). Let \(C\) 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 [36]), for \(k \geq 2d\) there exists a subset \(C' \subset C\) with the following properties: (a) \(\log(\text{Card}(C')) \geq (d/4) \log(k/(6d))\) and (b) \(\rho_H(\omega, \omega') \geq d\) for any \(\omega, \omega' \in C'\). Here, \(\rho_H(\cdot, \cdot)\) denotes the Hamming distance. To any \(\omega \in C'\) we associate a function \(f_\omega\) on \(\mathcal{X}\) defined by \(f_\omega(x^i) = \omega_i\) for \(i = 1, \ldots, k\) and \(f_\omega(x^i) = 0\), \(i \geq k + 1\), where \(\omega_i\) is the \(i\)th component of \(\omega\).

Let \(P_X\) be the distribution on \(\mathcal{X}\) which is uniform on \(\{x^1, \ldots, x^k\}\), putting probability \(1/k\) on each of these \(x^j\) and probability 0 on all \(x^j\) with \(j \geq k + 1\). Denote by \(P_\omega\) the joint distribution of \((X, Y)\) having this marginal \(P_X\) and \(Y \in \{0, 1\}\) with the conditional distribution \(E(Y|X = x) = P(Y = 1|X = x) = 1/2 + f_\omega(x)/4 \equiv \eta_\omega(x)\) 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,
\[
\|\eta_\omega - \eta_\omega'\|^2 = \rho_H(\omega, \omega')/(16k) \geq \alpha/32, \quad \forall \, \omega, \omega' \in \mathcal{C}'.
\] (37)

On the other hand, the Kullback-Leibler divergence between \( P_\omega \) and \( P_{\omega'} \) has the form
\[
K(P_\omega, P_{\omega'}) = nE \left( \eta_\omega(X) \log \frac{\eta_\omega(X)}{\eta_{\omega'}(X)} + (1 - \eta_\omega(X)) \log \frac{1 - \eta_\omega(X)}{1 - \eta_{\omega'}(X)} \right).
\]
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
\[
K(P_\omega, P_{\omega'}) \leq \frac{\|f_\omega - f_{\omega'}\|^2}{8} \leq \frac{nd}{8k} \leq \frac{n\alpha}{8}, \quad \forall \, \omega, \omega' \in \mathcal{C}'.
\] (38)

From (37), (38) and Theorem 2.7 in [41], the result of Theorem 14 follows if we show that
\[
n\alpha/8 \leq \log(\text{Card}(\mathcal{F}') - 1)/16
\] (39)

with
\[
\alpha = C_1 \frac{d}{n} \log \frac{C_2n}{d}
\]
where \( C_1, C_2 > 0 \) are constants. Assume first that \( d \geq 4 \). Then, using the inequalities \( \log(\text{Card}(\mathcal{F}') - 1) \geq \log(\text{Card}(\mathcal{C}')/2) \geq (d/4) \log(k/(6d)) - \log 2 \geq (d/4) \log(1/(12\alpha)) \) 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 \( \alpha = (C_1/n) \log(C_2n) \) and (39) is also satisfied for suitable \( C_1, C_2 \).

\[\square\]

### 9.2 Lower Bound Under Entropy Conditions

Define the \( \Delta \)-misspecified minimax regret as
\[
V_n^\Delta(\mathcal{F}) := \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P} : \inf_{f \in \mathcal{F}} \|f - \eta\|} \left\{ \mathbb{E} \|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \right\}
\]
Notice that by definition, \( V_n^\Delta(\mathcal{F}) = W_n(\mathcal{F}) \) when \( \Delta = 0 \) and \( V_n^\Delta(\mathcal{F}) = V_n(\mathcal{F}) \) when \( \Delta = 1 \) (the radius of \( \mathcal{F} \)). In general \( V_n^\Delta(\mathcal{F}) \) measures the minimax regret when we consider the statistical estimation problem with approximation error at most \( \Delta^2 \). The next theorem not only tells us that when the entropy of \( \mathcal{F} \) behaves as \( \epsilon^{-p} \) for \( p \geq 2 \), both the minimax rates \( V_n(\mathcal{F}) \) and \( W_n(\mathcal{F}) \) are tight, but in fact tells us that the rate with respect to the approximation error obtained in Section 5 is also tight.

Before we state the theorem, let \( \ell_0 \) be the set of all real-valued sequences \( (f_k, k = 1, 2, \ldots) \). Denote by \( e_j \) the unit vectors in \( \ell_0 \): \( e_j = (1 \{k = j\}, k = 1, 2, \ldots), j = 1, 2, \ldots \). For \( p > 0 \), consider the unit \( \ell_p \)-ball \( B_p = \{ f \in \ell : \sum_{j=1}^\infty |f_j|^p \leq 1 \} \).

23
Theorem 15. Fix any \( p > 0 \). Let \( \mathcal{F} = \{ f \in \ell : f_j = 1/2 + g_j, \{ g_j \} \in B_p \} \) and let \( \mathcal{X} = \{ \mathbf{e}_1, \mathbf{e}_2, \ldots \} \) be the set of all unit vectors in \( \ell_0 \). For any \( \epsilon > 0 \) and any \( n \geq (2/\epsilon)^p \), we have

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

Furthermore, for this \( \mathcal{F} \), there exist positive constants \( C, C' \) and \( C'' \) depending only on \( p \) such that the minimax risk satisfies, for any \( n \geq 1 \),

\[
W_n(\mathcal{F}) \geq Cn^{-2/(2+p)},
\]  

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

\[
V_n(\mathcal{F}) \geq C'n^{-1/p}.
\]  

Finally, for any \( p \geq 2 \), and any \( \Delta \) such that \( n^{-2/(2+p)} \leq \Delta^2 \leq n^{-1/p} \) we have that

\[
V_n^\Delta(\mathcal{F}) \geq C'' \Delta^2.
\]

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

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

shatters the set \( \{ \mathbf{e}_1, \ldots, \mathbf{e}_d \} \) at scale \( 2/d^{1/p} = \epsilon \). Thus, the fat-shattering dimension \( \text{fat}_\epsilon(B_p) \geq d = (2/\epsilon)^p \). This yields (40) via an application of Theorem 2.6 in [32].

To prove (41) and (42), consider a subset \( \mathcal{F} = \{ f \in \mathcal{F} : f_j = 1/2, \forall j > d \} \) where \( d = [(c_\ast n)^{p'/(2+p')} \} \) and \( c_\ast > 0 \), \( p' \geq p \) 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 \( \{ \mathbf{e}_1, \ldots, \mathbf{e}_d \} \), 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

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

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

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

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

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

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,

\[
\| \hat{f} - \eta_\omega \|^2 \geq \frac{1}{d} \sum_{i=1}^d \left| \frac{\hat{\omega}_i - \omega_i}{8d^{1/p'}} \right|^2 = \frac{\rho_H(\hat{\omega}, \omega)}{64d^{1+2/p'}}
\]  

(44)
where \(\rho_H(\cdot, \cdot)\) is the Hamming distance. From Assouad’s lemma (Theorem 2.12 (iv) in [41]) we find that
\[
\max_{\omega \in \Omega} \mathbb{E}_\omega \rho_H(\hat{\omega}, \omega) \geq \frac{d}{4} \exp(-\alpha) \tag{45}
\]
where \(\alpha = \max\{\chi^2(P_\omega, P_{\omega'}) : \omega, \omega' \in \Omega, \rho_H(\omega, \omega') = 1\}\) and \(\chi^2(P_\omega, P_{\omega'})\) is the chi-squared divergence between \(P_\omega\) and \(P_{\omega'}\). Here, \(\mathbb{E}_\omega\) denotes the distribution of the \(n\)-sample \(D_n\) when \((X_i, Y_i) \sim P_\omega\) for all \(i\). Since \(1/2 \leq \eta_\omega(X) \leq 3/4\), the chi-squared divergence is bounded as follows:
\[
\chi^2(P_\omega, P_{\omega'}) = n \mathbb{E} \left[ (\eta_\omega(X) - \eta_{\omega'}(X))^2 \left( \frac{1}{\eta_\omega(X)} + \frac{1}{1 - \eta_\omega(X)} \right) \right] = 6n \mathbb{E}(\eta_\omega(X) - \eta_{\omega'}(X))^2 \leq \frac{3}{8c_*}
\]
for all \(\omega, \omega' \in \Omega\) such that \(\rho_H(\omega, \omega') = 1\). Combining this result with (44) and (45) we find
\[
\inf_{\hat{f}} \max_{\omega \in \Omega} \mathbb{E}_\omega \| \hat{f} - \eta_\omega \|^2 \geq \frac{\exp(-3/(8c_*))}{256 d^{2/p'}}. \tag{46}
\]
Now, to prove (41) 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 \(W_n(\mathcal{F}) \geq \inf_{f} \max_{\omega \in \Omega} \mathbb{E}_\omega \| \hat{f} - \eta_\omega \|^2\) and (41) follows immediately from (46).

We now prove (42) and (44). Consider the vector
\[
f^* = \left( \frac{1}{2} + \frac{\omega_1}{4d^{1/p'}}, \frac{1}{2} + \frac{\omega_d}{4d^{1/p'}}, \frac{1}{2}, \ldots \right)
\]
Note that \(f^* \in \mathcal{F}\) and
\[
\| f^* - \eta_\omega \|^2 = \frac{1}{d} \sum_{i=1}^{d} \left( \frac{r_i}{d^{1/p'}} - \frac{r_i}{d^{1/p'}} \right)^2 = \left( 1 - \frac{1}{d^{1/p-1/p'}} \right)^2 \frac{1}{d^{2/p'}} \leq \frac{1}{4d^{2/p'}} \tag{47}
\]
assuming \(n\) is large enough \((n \geq n_0(p)\) where \(n_0(p)\) depends only on \(p\) and \(c_*\)). Hence combining the above inequality with (46), we conclude that
\[
\inf_{\hat{f}} \max_{\omega \in \Omega} \left\{ \mathbb{E}_\omega \| \hat{f} - \eta \|^2 - \inf_{f \in \mathcal{F}} \| f - \eta \|^2 \right\} \geq \left( \frac{\exp(-3/(8c_*))}{64} - 1 \right) \frac{1}{4d^{2/p'}} \tag{48}
\]
Now to prove (42), we simply set \(p' = 2(p-1)\), so that \(2/(2 + p') = 1/p\). Hence we have
\[
V_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_{X,Y} \in \mathcal{P}} \left\{ \mathbb{E}_\omega \| \hat{f} - \eta \|^2 - \inf_{f \in \mathcal{F}} \| f - \eta \|^2 \right\} \geq \inf_{\hat{f}} \max_{\omega \in \Omega} \left\{ \mathbb{E}_\omega \| \hat{f} - \eta_\omega \|^2 - \inf_{f \in \mathcal{F}} \| f - \eta_\omega \|^2 \right\} \geq \left( \frac{\exp(-3/(8c_*))}{64} - 1 \right) \frac{1}{4d^{2/p'}}
\]

Hence choosing $c_*>0$ small enough we obtain for appropriate constant $C'$, (42) for $n \geq n_0$. For $n < n_0(p)$ the result trivially follows from the positivity of $V_n(F)$.

To prove (43), we set $p'$ differently in (48). Specifically, given $\Delta$ such that $n^{-2/(2+\nu)} \leq \Delta^2 \leq n^{-1/\nu}$, we set $p' = \frac{2 \log d}{\log(\frac{1}{\Delta^2})}$. With this $p'$ it holds that $\frac{1}{4d^2 p'} = \Delta^2$. Hence, for this $p'$

$$\inf_{\phi} \max_{\omega \in \Omega} \left\{ \mathbb{E}_\omega \| f - \phi \|^2 - \inf_{f \in F} \| f - \phi \|^2 \right\} \geq \left( \frac{\exp(-3/(8c_*))}{64} - 1 \right) \Delta^2$$

In view of (47), for the choice of $p'$, the approximation error is bounded as $\inf_{f \in F} \| f - \phi \|^2 \leq \| f^* - \phi \|^2 \leq \Delta^2$. Hence we conclude that

$$V_n^\Delta(F) \geq \left( \frac{\exp(-3/(8c_*))}{64} - 1 \right) \Delta^2$$

Choosing $c_*>0$ small enough we obtain for appropriate constant $C''$, (43) for $n \geq n_0$.

\[\square\]

10 Technical Results: Localization

The following is a modification of Theorem 6.1 in [8]. We include part of that Theorem verbatim and make additional changes.

**Theorem 16** (Based on [8]). Let $\mathcal{G}$ be a class of non-negative functions almost surely bounded by $b$. Let $\phi_n$ be a function that is non-negative, non-decreasing, and $\phi_n(r)/\sqrt{r}$ non-increasing, satisfying

$$\mathcal{H}_n(\mathcal{G}[r,S],S) \leq \phi_n(r)$$

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 $g \in \mathcal{G}$

$$P_g \leq 2P_ng + 106r^* + 48r_0$$

and

$$P_ng \leq 2P_g + 72r^* + 22r_0$$

and

$$P_g \leq P_ng + \sqrt{P_ng(x + x(\delta_k))} + 108r^* + 42r_0$$

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

**Proof of Theorem 16.** Define $\delta_k = b2^{-k}$ for $k \geq 0$ and let $\mathcal{G}_k = \{ g \in \mathcal{G} : \delta_{k+1} \leq P_g \leq \delta_k \}$. Denote the empirical Rademacher averages of $\mathcal{G}_k$ by $R_k$. Observe that for $g \in \mathcal{G}_k$, $P_g^2 \leq b\delta_k$. Then Lemma 6.2 in [8] implies that with probability at least $1 - e^{-x}$ for all $k \geq 0$ and $g \in \mathcal{G}_k$,

$$|P_ng - P_g| \leq 6R_k + \frac{2b\delta_k(x + x(\delta_k))}{n} + \frac{6b(x + x(\delta_k))}{n} \quad (49)$$
where \( x(\delta) = 2\log\left(\frac{\pi}{\sqrt{2}} \log_2 \frac{2\delta}{\delta}\right) \). 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 \( P_n g \leq U_k \). This implies that \( R_k \leq \phi_n(U_k) \). Putting together the terms,

\[
U_k \leq \delta_k + 6\phi_n(U_k) + \sqrt{\frac{2b\delta_k(x + x(\delta_k))}{n} + \frac{6b(x + x(\delta_k))}{n}}
\]

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

\[
U_k \leq \delta_k + 6\phi_n(U_k) + 7r_0.
\]

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 \( \phi_n(r) \leq \sqrt{r \cdot r^*} \) for any \( r \geq r^* \). Hence,

\[
U_k \leq 6\sqrt{U_k r^*} + 2\delta_k + 7r_0.
\]

Solving the quadratic equation,

\[
U_k \leq 36r^* + 4\delta_k + 14r_0 \leq 36r^* + 8Pg + 14r_0
\]

because \( \delta_k \leq 2Pg \). We thus have for all \( k \leq k_0 \) and \( g \in G_k \)

\[
|Pg - P_n g| \leq 6\phi_n(36r^* + 8Pg + 14r_0) + \sqrt{4r_0 Pg} + 6r_0 \tag{50}
\]

\[
\leq 6\sqrt{r^*} \sqrt{36r^* + 8Pg + 14r_0} + \sqrt{4r_0 Pg} + 6r_0 \tag{51}
\]

\[
\leq 45r^* + \sqrt{8r^* Pg} + \sqrt{4r_0 Pg} + 20r_0 \tag{52}
\]

Solving the equation

\[
Pg \leq P_n g + \sqrt{Pg(\sqrt{8r^* + \sqrt{4r_0}})} + 45r^* + 20r_0 \tag{53}
\]

yields

\[
Pg \leq 2P_n g + 106r^* + 48r_0 \; .
\]

Alternatively, using (53) and the implication \( A \leq B + C \sqrt{A} \Rightarrow A \leq B + C^2 + \sqrt{BC} \) for non-negative \( A, B \) and \( C \), we obtain

\[
Pg \leq P_n g + \sqrt{P_n g(\sqrt{8r^* + \sqrt{4r_0}})} + 108r^* + 42r_0, \tag{54}
\]

proving the last statement of the theorem. For the second statement, we solve the equation (50) in terms of the variable \( \sqrt{Pg} \):

\[
P_n g \leq Pg + \sqrt{Pg(\sqrt{8r^* + \sqrt{4r_0}})} + 45r^* + 20r_0.
\]

The roots are found to be

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

27
If \( P_n g < 45 r^* + 20 r_0 \), the statement of the theorem holds. Otherwise, we take the positive root and conclude

\[
\sqrt{P g} \geq \frac{\sqrt{8 r^* + 4 r_0}}{2} + \sqrt{\left(\frac{\sqrt{8 r^* + 4 r_0}}{2}\right)^2 + (P_n g - 45 r^* - 20 r_0)}
\]

leading to

\[
P_n g - 45 r^* - 20 r_0 \leq 2 P g + \left(\frac{\sqrt{8 r^* + 4 r_0}}{2}\right)^2
\]

and thus

\[
P_n g \leq 2 P g + 49 r^* + 22 r_0.
\]

Now consider the case \( k \geq k_0 \). First, for any \( g \in G_k \), \( P g \leq \delta_k \leq \delta_{k_0} \leq 4 b / n \). Hence \( G' = \{ g \in G : P g < 4 b / n \} \supseteq G_k \) for any \( k \geq k_0 \). By Lemma 6.1 in [8], with probability at least \( 1 - 3 e^{-x} \),

\[
|P_n g - P g| \leq 6 \mathfrak{R}_n(G') + \frac{b}{n}(\sqrt{8 x + 4 x}) \leq 6 \mathfrak{R}_n(G') + \frac{8 b x}{n}
\]

whenever \( x > 1/2 \). Now reason on this event. Defining

\[
U' = 6 \mathfrak{R}_n(G') + P g + \frac{8 b x}{n}
\]

we have \( P_n g \leq U' \) for any \( g \in G' \), and so

\[
\mathfrak{R}_n(G') \leq \mathfrak{R}_n(\{ g \in G : P_n g \leq U' \}) \leq \phi_n(U')
\]

Since \( \phi_n \) is sub-root,

\[
U' \leq 6 \phi_n(U') + P g + \frac{8 b x}{n} \leq 6 \sqrt{U'} \sqrt{r^*} + P g + \frac{8 b x}{n}
\]

Solving for \( \sqrt{U'} \),

\[
\sqrt{U'} \leq 6 \sqrt{r^*} + \sqrt{P g + \frac{8 b x}{n}}
\]

and thus

\[
P_n g \leq U' \leq 2 P g + 72 r^* + 16 r_0
\]

Proof of Lemma 9. By Theorem 16, for any \( \phi_n \) satisfying \( \mathfrak{R}_n(\{ g : P_n g \leq r \}) \leq \phi_n(r) \) and appropriate growth conditions, for all \( x > 0 \), with probability at least \( 1 - 4 e^{-x} \) for all \( g \in G \)

\[
P g \leq P_n g + \sqrt{P_n g(\sqrt{8 r^* + 4 r_0}) + 108 r^* + 42 r_0}, \tag{55}
\]

where \( r^* \) is the largest solution of \( \phi_n(r) = r \). Under the above event, for \( g^* = \arg \min_{g \in G} P g \),

\[
P g^* \leq P_n g^* + \sqrt{P_n g^*(\sqrt{8 r^* + 4 r_0}) + 108 r^* + 42 r_0}
\]
By Bernstein’s inequality, with probability at least $1 - e^{-x}$,

$$P_n g^* \leq P g^* + \sqrt{\frac{4x P g^*}{n}} + \frac{4x}{n}$$

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

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

Simplifying and over-bounding,

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

\[\Box\]

**Proof of Lemma 10.** We apply Theorem 3.3 in [6] to $G = \ell \circ \mathcal{F} - \ell \circ f^*$. Observe that

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

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

$$P g \leq P_n g + c_1'\mathbb{E} r_n(G) + \frac{x(22 + c_2'')}{n}$$

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

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

where we passed to the Rademacher averages of $\ell \circ \mathcal{F}$. Now, by Lemma A.4 in [6], with probability at least $1 - e^{-x}$,

$$\mathbb{E} r_n(\ell \circ \mathcal{F}) \leq 2\mathbb{E} r_n(\ell \circ \mathcal{F}) + \frac{x}{n}$$

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

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

\[\Box\]

**11 Proof of Lemma 6**

To prove the first estimate for $r^*$ in lemma, we need a result for smooth losses, proved in [39] in the context of supervised learning:
Lemma 17 ([39]). Let $\ell$ 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 $S \in (\mathcal{X} \times \mathcal{Y})^n$,
\[
\hat{R}_n((\ell \circ \mathcal{H})[r,S],S) \leq 21\sqrt{6Hr}\log^{3/2}(64n)\mathcal{R}_n(\mathcal{H})
\] (56)

Proof of Lemma 6. We first claim that we may always take $r^* = 21168 \log^3(64n)\mathcal{R}_n^2(\mathcal{F})$. Since the result of Lemma 17 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
\[
\hat{R}_n(\mathcal{G}[r,S],S) = \hat{R}_n((\ell \circ \mathcal{H})[r,S],S) \leq 21\sqrt{12r}\log^{3/2}(64n)\mathcal{R}_n(\mathcal{H}) \leq 42\sqrt{12r}\log^{3/2}(64n)\mathcal{R}_n(\mathcal{F}).
\]
Now define the right-hand side as the function $\phi_n(r)$. This immediately leads to a fixed-point
\[
r^* = 12 \cdot 42^2 \log^3(64n)\mathcal{R}_n^2(\mathcal{F}),
\]
as claimed.

For the second part, fix some $\delta > 0$ and let $c_1, \ldots, c_M$ be any minimal $\delta$-cover of $\mathcal{F}$ with respect to $d_S$ with $M = \mathcal{N}_2(\mathcal{F}, \delta, S)$. Without loss of generality, assume $0 \leq c_i(x) \leq 1$ 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 $\delta$-close to $f'$ and $f''$ respectively. Since
\[
\frac{1}{n} \sum_{i=1}^{n} \left( (f'(x_i) - f''(x_i))^2 - (c'(x_i) - c''(x_i))^2 \right)^2 \leq \frac{1}{n} \sum_{i=1}^{n} (f'(x_i) - c'(x_i) + f''(x_i) - c''(x_i))^2 \leq 16\delta^2
\]
we have that $\mathcal{N}_2(\mathcal{G}, \delta, S) \leq \mathcal{N}_2(\mathcal{F}, \delta/4, S)$. Hence, the $\phi_n$ in (9) can be taken to be
\[
\hat{R}_n(\mathcal{G}[r,S],S) \leq \frac{12}{\sqrt{n}} \int_{0}^{\sqrt{r}} \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta, S)} d\delta \leq \frac{12}{\sqrt{n}} \int_{0}^{\sqrt{r}} \sqrt{v} \log(4c/\delta) d\delta
\] (57)
\[
\leq \frac{12 \cdot 4c\sqrt{v}}{\sqrt{n}} \int_{0}^{\sqrt{r}/4c} \sqrt{\log 1/\rho} d\rho \leq 24\sqrt{\frac{v}{n}} \log^{1/2}(4c/\sqrt{r}) =: \phi_n(r)
\]
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 $\phi$ whenever $b \geq a > 0$. That is, for $n$ large enough,
\[
r^* = \left( 24\sqrt{\frac{v}{n}} \log^{1/2} \left( \frac{c\sqrt{n}}{6\sqrt{v}} \right) \right)^2 = C \frac{v}{n} \log(n/v)
\] (58)
for some constant $C$.

Finally, for a finite class $\mathcal{F}$, the covering numbers are $\mathcal{N}_2(\mathcal{F}, \epsilon, S) \leq |\mathcal{F}|$ and, trivially,
\[
r^* = C \frac{\log |\mathcal{F}|}{n}
\]
along the lines of (57). \qed
References


