Bernoulli **23**(2), 2017, 789–824
DOI: 10.3150/14-BEJ679

Empirical entropy, minimax regret and minimax risk

ALEXANDER RAKHLIN[1], KARTHIK SRIDHARAN[2]
and ALEXANDRE B. TSYBAKOV[3]

[1]*Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104, USA.*
E-mail: rakhlin@wharton.upenn.edu
[2]*Department of Computer Science, Cornell University, Ithaca, NY 14853, USA*
[3]*Laboratoire de Statistique, CREST-ENSAE, 92245 Malakoff Cedex, France*

We consider the random design regression model 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 sharp 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 $n^{-2/(2+p)}$ for $p \in (0, 2)$ and $n^{-1/p}$ for $p > 2$ where n is the sample size. 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 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. On the contrary, for $p > 2$ we show that the rates of the minimax regret are, in general, slower than for the minimax risk. Our oracle inequalities also imply the $v \log(n/v)/n$ rates for Vapnik–Chervonenkis type classes of dimension v without the usual 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 [*Math. Methods Statist.* **16** (2007) 246–259] and providing the optimal rate.

Keywords: aggregation; empirical risk minimization; entropy; minimax regret; minimax risk

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 measurable function $f : \mathcal{X} \to \mathcal{Y}$ called the predictor, we define the prediction risk under squared loss:

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

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 on the sample D_n and having a small excess risk

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

1350-7265 © 2017 ISI/BS

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

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

where $\mathbb{E}$ is the generic expectation sign, $\|f\|^2 = \int f^2(x)P_X(\mathrm{d}x)$, 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" [15,25,47], while the object on the right-hand side has been the topic of oracle inequalities in nonparametric statistics [35,42], and in the literature on aggregation [38,41]. Upper bounds on the right-hand side of (2) are called *sharp* 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, that is, they yield upper bounds for the difference

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

with $C > 1$ and not for the excess risk. In this paper, we obtain sharp 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 a minimax point of view, the object of interest 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 probability 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

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

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., for example, [38,41]. 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. 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*, that is, 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*

$$W_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}_{\mathcal{F}}} \mathbb{E}\|\hat{f} - \eta\|^2, \tag{6}$$

where $\mathcal{P}_{\mathcal{F}}$ is the set of all distributions P_{XY} on $\mathcal{X} \times \mathcal{Y}$ such that $\eta \in \mathcal{F}$. 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 the question is whether the two quantities can be of the same order of magnitude for particular $\mathcal{F}$. We show below that the answer is positive for major cases of interest except for very massive classes $\mathcal{F}$, namely, those having the empirical ε-entropy of the order ε^{-p}, $p > 2$, for small ε. We also prove that this entropy condition is tight in the sense that the minimax regret and the minimax risk can have different rates of convergence when it is violated. Furthermore, we show that the optimal rates for the minimax regret and minimax risk are attained by one and the same procedure – the aggregation-of-leaders estimator – that we introduce below.

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 7, 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. In Section 4, we introduce a more general risk measure that realizes a smooth transition between $W_n(\mathcal{F})$ and $V_n(\mathcal{F})$ depending on the magnitude of the approximation error. The minimax risk and the minimax regret appear as the two extremes of this scale.

The paper is organized as follows. In Section 3, we present the aggregation-of-leaders estimator and the upper bounds on its risk. These include the main oracle inequality in Theorem 1 and its consequences for particular classes $\mathcal{F}$ in Theorems 2–4. Section 4 discusses a more general setting allowing for a smooth transition between $W_n(\mathcal{F})$ and $V_n(\mathcal{F})$ in terms of the approximation error. Lower bounds for the minimax risk and minimax regret are proved in Section 5. In Section 6, we compare the aggregation-of-leaders estimator with the two closest competitors – skeleton aggregation and global ERM. Section 7 provides an overview and comparison of our results to those in the literature. Proofs of the theorems are given in Sections 8–10. The Appendix contains some technical results and proofs of the lemmas.

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 of $\mathcal{G}$:

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

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

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

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:

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

Any function $\phi_n : [0, \infty) \mapsto \mathbb{R}$ satisfying

$$\sup_{S\in\mathcal{Z}^n} \hat{\mathfrak{R}}_n\bigl(\mathcal{G}[r, S], S\bigr) \le \phi_n(r) \tag{7}$$

for all $r > 0$ will be called an upper function for the class $\mathcal{G}$. We will sometimes write $\phi_n(r) = \phi_n(r, \mathcal{G})$ to emphasize the dependence on $\mathcal{G}$. It can be shown (cf., e.g., Lemma 8 below) that any class of uniformly bounded functions admits an upper function satisfying the sub-root property: ϕ_n is non-negative, non-decreasing, and $\phi_n(r)/\sqrt{r}$ is non-increasing. We will denote by $r^* = r^*(\mathcal{G})$ the corresponding *localization radius*, that is, an upper bound on the largest solution of the equation $\phi_n(r) = r$. Clearly, r^* is not uniquely defined since we deal here with upper bounds.

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 $\{\ell \circ f\colon f \in \mathcal{F}\}$. Thus,

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

for $S = \{z_1, \dots, z_n\}$ with $z_i = (x_i, y_i)$.

For any bounded measurable function $g : \mathcal{Z} \to \mathbb{R}$, we set $Pg = \mathbb{E}g(Z)$, where $Z = (X, Y)$, and $P_n g = \frac{1}{n}\sum_{i=1}^n g(Z_i)$, where $Z_i = (X_i, Y_i)$. For $S = \{z_1, \dots, z_n\} \in \mathcal{Z}^n$ with $z_i = (x_i, y_i)$ consider the empirical ℓ_2 pseudo-metric

$$d_S(f, g) = \left(\frac{1}{n}\sum_{i=1}^n \bigl|f(x_i) - g(x_i)\bigr|^2\right)^{1/2},$$

and for any $\varepsilon > 0$ denote by $\mathcal{N}_2(\mathcal{F}, \varepsilon, S)$ the ε-covering number of a class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$ with respect to this pseudo-metric. Recall that a covering number at scale ε is the smallest number of balls of radius ε required to cover the set. Denote by $\mathcal{N}_\infty(\mathcal{F}, \varepsilon, S)$ the ε-covering number of the class $\mathcal{F}$ with respect to the supremum norm (over S).

Although not discussed here explicitly, some standard measurability conditions are needed to apply results from the theory of empirical processes as well as to ensure that the ERM estimators we consider below are measurable. This can be done in a very general framework and we assume throughout that these conditions are satisfied. For more details we refer to Chapter 5 of [18], see also [25], page 17.

The minimum risk on the class of functions $\mathcal{F}$ is denoted by

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

Let $\lceil x \rceil$ denote the minimal integer strictly greater than $x \in \mathbb{R}$, and $|\mathcal{F}|$ the cardinality of $\mathcal{F}$. Notation C will be used for positive constants that can vary on different occasions; these are absolute constants unless their dependence on some parameters is explicitly mentioned. We will also assume throughout that $n \geq 5$.

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 pseudo-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 cells 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 the global empirical risk minimization (ERM) over the class $\mathcal{F}$. While the global ERM is, in general, suboptimal (cf. the discussion in Sections 6 and 7 below), the proposed method enjoys the optimal rates. We call our method the aggregation-of-leaders procedure since it aggregates the best solutions obtained in cells of the partition.

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 $\varepsilon > 0$. Let $d_S(f, g)$ be the empirical ℓ_2 pseudo-metric associated with the subsample S of cardinality n, and

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

Clearly, N is finite since $\mathcal{F}$ is included in the set of all functions with values in [0, 1], which is totally bounded with respect to $d_S(\cdot, \cdot)$. 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*, that is, $\hat{c}_i \in \mathcal{F}$ for $i = 1, \dots, N$, and that $N \geq 2$. 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:

$$\hat{\mathcal{F}}_i^S = \hat{\mathcal{F}}_i^S(\varepsilon) = \Bigl\{ f \in \mathcal{F} \colon i \in \operatorname*{argmin}_{j=1,\dots,N} d_S(f, \hat{c}_j) \Bigr\}$$

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':

$$\hat{f}_i^{S,S'} \in \operatorname*{argmin}_{f \in \hat{\mathcal{F}}_i^S} \frac{1}{n} \sum_{(x,y)\in S'} \bigl(f(x) - y\bigr)^2. \tag{8}$$

We will assume that such a minimizer exists; a simple modification of the results is possible if $\hat{f}_i^{S,S'}$ is an approximate solution of (8).

Finally, at the third step we use the subsample S'' to aggregate the estimators $\{\hat{f}_1^{S,S'}, \dots, \hat{f}_N^{S,S'}\}$. 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.

Sharp MS-aggregation property. *There exists a constant $C > 0$ such that, for any $\delta > 0$,*

$$L(\tilde{f}) \le \min_{i=1,\dots,N} L\bigl(\hat{f}_i^{S,S'}\bigr) + C\frac{\log(N/\delta)}{n} \tag{9}$$

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

Note that, in (9), 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], page 5, [30], Theorem B, [31], Theorem A. They are realized as mixtures:

$$\tilde{f} = \sum_{i=1}^N \theta_i g_i = \sum_{i=1}^N \theta_i \hat{f}_i^{S,S'}, \tag{10}$$

where θ_i are some random weights measurable with respect to S''. Either of the aggregates of [2,30,31] satisfy the sharp MS-aggregation property and thus can be used at the third step of our procedure.

Definition 1. *We call an* aggregation-of-leaders *estimator any estimator $\tilde{f}$ defined by the above three-stage procedure with sharp* MS-*aggregation at the third step.*

The next theorem provides the main oracle inequality for aggregation-of-leaders estimators.

Theorem 1. *Let $\mathcal{Y} = [0, 1]$ and $0 \le f \le 1$ for all $f \in \mathcal{F}$. Let $r^* = r^*(\mathcal{G})$ denote a localization radius of $\mathcal{G} = \{(f - g)^2\colon\ f, g \in \mathcal{F}\}$. Consider an aggregation-of-leaders estimator $\tilde{f}$ defined by*

[1] Here, MS-aggregate is an abbreviation for *model selection type aggregate*. The word *sharp* indicates that (9) is an oracle inequality with leading constant 1.

the above three-stage procedure. Then there exists an absolute constant $C > 0$ *such that for any* $\delta > 0$, *with probability at least* $1 - 2\delta$,

$$L(\tilde{f}) \leq \inf_{f \in \mathcal{F}} L(f) + C\left(\frac{\log(\mathcal{N}_2(\mathcal{F}, \varepsilon, S)/\delta)}{n} + \Xi(n, \varepsilon, S')\right), \tag{11}$$

where

$$\Xi(n, \varepsilon, S') = \gamma\sqrt{r^*} + \inf_{\alpha \geq 0}\left\{\alpha + \frac{1}{\sqrt{n}}\int_{\alpha}^{C\gamma} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, S')}\, \mathrm{d}\rho\right\} \tag{12}$$

with $\gamma = \sqrt{\varepsilon^2 + r^* + \beta}$ *and* $\beta = (\log(1/\delta) + \log\log n)/n$.

Remarks.

1. The term $\Xi(n, \varepsilon, S')$ in Theorem 1 is a bound on the rate of convergence of the excess risk of ERM $\hat{f}_i^{S,S'}$ over the cell $\hat{\mathcal{F}}_i^S$. If, in particular instances, there exists a sharper bound for the rate of ERM, one can readily use this bound instead of the expression for $\Xi(n, \varepsilon, S')$ given in Theorem 1.
2. The partition with cells $\hat{\mathcal{F}}_i^S$ 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 $\Xi(n, \varepsilon, S')$. In this case, the partition is non-random. Another example, when $\mathcal{F}$ is isomorphic to a subset of $\mathbb{R}^M$, is a partition of $\mathbb{R}^M$ into a union of linear subspaces of all possible dimensions. In this case, the "cells" are linear subspaces and aggregating the least squares estimators over cells is analogous to sparsity pattern aggregation considered in [38,39].
3. In Theorem 1 we can use the localization radius $r^* = r^*(\hat{\mathcal{G}}_i)$ for $\hat{\mathcal{G}}_i = \{(f - g)^2\colon\ f, g \in \hat{\mathcal{F}}_i^S\}$ instead of the larger quantity $r^*(\mathcal{G})$. Inspection of the proof shows that the oracle inequality (11) generalizes to

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

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

The oracle inequality (11) of Theorem 1 depends on two quantities that should be specified: the entropy $\log \mathcal{N}_2(\mathcal{F}, \cdot, \cdot)$, and the localization radius r^*. The crucial role in determining the rate belongs to the empirical entropies. We further replace in (11) these random entropies by their upper bound

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

and refer to the above quantity as the *empirical entropy*.

The next theorem is a corollary of Theorem 1 in the case of polynomial growth of the empirical entropy characteristic for nonparametric estimation problems. It gives upper bounds on the minimax regret and on the minimax risk.

Theorem 2. *Let $\mathcal{Y} = [0,1]$ and $\mathcal{H}_2(\mathcal{F}, \rho) \le A\rho^{-p}$, $\forall \rho > 0$, for some constants $A < \infty$, $p > 0$. Let $\tilde{f}$ be an aggregation-of-leaders estimator defined by the above three-stage procedure with the covering radius $\varepsilon = n^{-1/(2+p)}$. There exist constants $C_p > 0$ depending only on A and p such that*:

(i) *Let $0 \le f \le 1$ for all $f \in \mathcal{F}$. For the estimator $\tilde{f}$ we have*:

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

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

$$W_n(\mathcal{F}) \le \sup_{P_{XY} \in \mathcal{P}_{\mathcal{F}}} \mathbb{E}\|\tilde{f} - \eta\|^2 \le C_p n^{-2/(2+p)} \qquad \forall p > 0. \tag{15}$$

The proof of Theorem 2 is given in Section 8. The first conclusion of this theorem is that the minimax risk $W_n(\mathcal{F})$ has the same rate of convergence as the minimax regret $V_n(\mathcal{F})$ for $p \in (0,2)$. For example, if $\mathcal{F}$ is a class of functions on $\mathbb{R}^d$ with bounded derivative of order k, the entropy bound required in the theorem holds with exponent $p = d/k$, as follows from [22]. In this case, Theorem 2 yields that, for $k \ge d/2$, both $W_n(\mathcal{F})$ and $V_n(\mathcal{F})$ converge with the usual nonparametric rate $n^{-2k/(2k+d)}$ while for $k < d/2$ (corresponding to very irregular functions) the rate of the minimax regret deteriorates to $n^{-k/d}$. In Section 5, we will show that the bounds of Theorem 2 for $p < 2$ are tight in the sense that there exists a marginal distribution of X and a class $\mathcal{F}$ of regression functions satisfying the above entropy assumptions such that the bounds (14) and (15) cannot be improved for $p < 2$.

The second message of Theorem 2 is that $W_n(\mathcal{F})$ has faster rate than $V_n(\mathcal{F})$ for $p > 2$, that is, for very massive classes $\mathcal{F}$. Note that here we compare only the upper bounds. However, in Section 5 we will provide a lower bound showing that the effect indeed occurs. Namely, we will exhibit a marginal distribution of X and a class $\mathcal{F}$ of regression functions satisfying the above entropy assumptions such that $V_n(\mathcal{F})$ is of the order $n^{-1/(p-1)}$, which is slower than the rate $n^{-2/(2+p)}$ for $W_n(\mathcal{F})$.

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

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

We will further comment on this choice in Section 6.

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* with VC-dimension $\mathrm{VC}(\mathcal{F}) = v$. We will also call them *parametric classes*, as opposed to non-parametric classes considered in Theorem 2. Indeed, entropy bounds as in (16) are associated to compact subsets of v-dimensional Euclidean space. Other example is given by the VC-subgraph classes with VC-dimension v, that is, classes of functions f whose subgraphs $C_f = \{(x,t) \in \mathcal{X} \times \mathbb{R}\colon\ f(x) \ge t\}$ 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 $0 \le f \le 1$ for all $f \in \mathcal{F}$, and let $\tilde{f}$ be an aggregation-of-leaders estimator defined by the above three-stage procedure with $\varepsilon = n^{-1/2}$. If $n \ge C_A v$ for a large enough constant $C_A > 1$ depending only on A, then there exists a constant $C > 0$ depending only on A such that*

$$V_n(\mathcal{F}) \le \sup_{P_{XY} \in \mathcal{P}} \Bigl\{ \mathbb{E}\|\tilde{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \Bigr\} \le C \frac{v}{n} \log\biggl(\frac{en}{v}\biggr). \tag{17}$$

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 7 below). Theorem 3 does not rely on either of these assumptions.

In Section 5, 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 $C(v/n)\log(en/v)$ with positive fixed probability. So, the extra logarithmic factor $\log(en/v)$ in the rate is necessary, even when the model is well-specified.

The next theorem deals with classes of functions

$$\mathcal{F} = \mathcal{F}_\Theta \triangleq \Biggl\{ \mathrm{f}_\theta = \sum_{i=1}^{M} \theta_j f_j\colon\ \theta = (\theta_1, \dots, \theta_M) \in \Theta \Biggr\},$$

where $\{f_1, \dots, f_M\}$ 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., for example, [38,41], 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 [38] for a comprehensive treatment. Here, we deal with the random design case and consider the sets Θ defined as intersections of ℓ_0-balls with the simplex. For an integer $1 \le s \le M$, the ℓ_0-ball with radius s is defined by

$$B_0(s) = \bigl\{ \theta \in \mathbb{R}^M\colon\ |\theta|_0 \le s \bigr\},$$

where $|\theta|_0$ denotes the number of non-zero components of θ. We will also consider the simplex

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

Then, model selection type aggregation (or MS-*aggregation*) consists in constructing an estimator $\tilde{f}$ that mimics the best function among $f_1, \dots, f_M$, that is, the function that attains the minimum $\min_{j=1,\dots,M} \|f_j - \eta\|^2$. In this case, $\mathcal{F}_\Theta = \{f_1, \dots, f_M\}$ or equivalently $\Theta = \Theta^{\mathrm{MS}} \triangleq \{\mathbf{e}_1, \dots, \mathbf{e}_M\} = \Lambda_M \cap B_0(1)$, where $\mathbf{e}_1, \dots, \mathbf{e}_M$ 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 $\mathcal{F} = \mathrm{conv}(f_1, \dots, f_M)$, that is, 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-*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-convex 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 (via a sharp MS-aggregation procedure) only two estimators, $\tilde{f}$ and the least squares estimator on Λ_M. The resulting aggregate is denoted by $\tilde{f}^*$.

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

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

where

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

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

This theorem improves upon the rate of s-convex aggregation given in Lounici [33] by removing a redundant $(s/n)\log n$ term present there. Note that [33] 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; it can be readily extended to the case of sub-exponential errors. By an easy modification of the minimax lower bound given in [33], we get that $\psi_{n,M}(s)$ is the optimal rate for the minimax regret on $\mathcal{F}_{\Theta^{C}(s)}$ in our setting. Analogous result for Gaussian regression with fixed design is proved in [38].

Remark 4. Inspection of the proofs shows that Theorems 2–4 as well as Theorem 5 below provide bounds on the risk not only in expectation but also in deviation. For example, under the assumptions of Theorem 3, along with (17) we obtain that there exists a constant $C > 0$ depending only on A such that, for any $t > 0$,

$$\sup_{P_{XY}\in\mathcal{P}} \mathbb{P}\left\{\|\tilde{f}-\eta\|^2 - \inf_{f\in\mathcal{F}}\|f-\eta\|^2 \ge C\left(\frac{v}{n}\log\left(\frac{en}{v}\right)+\frac{t}{n}\right)\right\} \le e^{-t}. \tag{19}$$

The “in deviation” versions of Theorems 2, 4 and 5 are analogous and we skip them for brevity. We also note that all the results trivially extend to the case $\mathcal{Y}=[a,b]$, $\mathcal{F}\subseteq\{f\colon\ a\le f\le b\}$, where $-\infty < a < b < \infty$.

4. Adapting to approximation error rate of function class

In Theorem 2, we have shown that for $p > 2$ our estimator has the rate of $n^{-2/(2+p)}$ 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. This can be viewed as an intermediate setting between the pure statistical learning and pure estimation. In such situation, one would expect to achieve rates varying between $n^{-1/p}$ and $n^{-2/(2+p)}$ depending on how small the approximation error is. This is indeed the case as described in the next theorem.

Theorem 5. *Let $\mathcal{Y}=[0,1]$, $\mathcal{F}\subseteq\{f\colon\ 0\le f\le 1\}$, and $\mathcal{H}_2(\mathcal{F},\rho)\le A\rho^{-p}$, $\forall\rho>0$, for some constants $A<\infty$, $p\ge 2$. Consider an aggregation-of-leaders estimator $\tilde{f}$ with the covering radius set as $\varepsilon=n^{-1/(2+p)}$. For this estimator and for any joint distribution P_{XY} we have*:

$$\mathbb{E}\|\tilde{f}-\eta\|^2 - \inf_{f\in\mathcal{F}}\|f-\eta\|^2 \le C_p\bar{\psi}_{n,p}(\Delta), \tag{20}$$

where $\Delta^2=\inf_{f\in\mathcal{F}}\|f-\eta\|^2$, $C_p>0$ is a constant depending only on p and A, and

$$\bar{\psi}_{n,p}(\Delta)=\begin{cases} n^{-2/(2+p)}, & \text{if } \Delta^2\le n^{-2/(2+p)},\\ \Delta^2, & \text{if } n^{-2/(2+p)}\le\Delta^2\le n^{-1/p},\\ n^{-1/p}, & \text{if } \Delta^2\ge n^{-1/p}\end{cases} \tag{21}$$

for $p>2$. At $p=2$ the rate $\bar{\psi}_{n,p}(\Delta)$ is $n^{-1/2}\log n$ independently of Δ.

The proof of this theorem is given in Section 8.

For particular cases $\Delta=0$ (well-specified model) or $\Delta=1$ (misspecified model), we recover the result of Theorem 2 for $p>2$. Theorem 5 reveals that there is a smooth transition in terms of approximation error rate in the intermediate regime between these two extremes. Note also that the estimator $\tilde{f}$ in Theorem 5 is the same in all the cases; it is defined by the aggregation-of-leaders procedure with ε fixed as $n^{-1/(2+p)}$. Thus, the estimator is *adaptive* to the approximation error.

Theorem 5 naturally suggests to study a minimax problem which is more general than those considered in Statistical Learning Theory or Nonparametric Estimation. Introduce the class of Δ-misspecified models

$$\mathcal{P}_\Delta(\mathcal{F}) = \Bigl\{P_{XY} \in \mathcal{P}\colon \inf_{f\in\mathcal{F}} \|f-\eta\| \le \Delta\Bigr\}, \qquad \Delta \ge 0,$$

and define the Δ-*misspecified regret* as

$$V_n^\Delta(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_{XY}\in\mathcal{P}_\Delta(\mathcal{F})} \Bigl\{\mathbb{E}\|\hat{f}-\eta\|^2 - \inf_{f\in\mathcal{F}} \|f-\eta\|^2\Bigr\}.$$

Note 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 diameter 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 Δ. Theorem 5 implies that the rate of convergence of Δ-misspecified regret admits the bound $V_n^\Delta(\mathcal{F}) \le C_p\bar{\psi}_{n,p}(\Delta)$.

5. Lower bounds

In this section, we show that the upper bounds obtained in Theorems 2, 3, and 5 cannot be improved. First, we exhibit a VC-subgraph class $\mathcal{F}$ with VC-dimension at most d such that

$$W_n(\mathcal{F}) \ge C\frac{d}{n}\log\biggl(\frac{en}{d}\biggr),$$

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

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

$$\mathcal{F} = \bigl\{f\colon f(x) = a\mathbf{1}\{x \in W\} \text{ for some } W \subset \mathcal{X} \text{ with } |W| \le d\bigr\},$$

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

Theorem 6. *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 μ_X and numerical constants $c, c' > 0$ such that*

$$\inf_{\hat{f}} \sup_{\eta\in\mathcal{F}} \mathbb{P}_\eta\biggl(\|\hat{f}-\eta\|^2 \ge c\frac{d}{n}\log\biggl(\frac{en}{d}\biggr)\biggr) \ge c',$$

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

The proof of Theorem 6 is given in Section 10.

We now exhibit a class $\mathcal{F}$ with polynomial growth of the empirical entropy, for which the rates of minimax risk and minimax regret given in Theorems 2 and 5 cannot be improved on any estimators. To state the result, we need some notation. Let ℓ be the set of all real-valued sequences $(f_k, k = 1, 2, \ldots)$. Denote by $\mathbf{e}_j$ the unit vectors in ℓ: $\mathbf{e}_j = (\mathbf{1}\{k = j\}, k = 1, 2, \ldots)$, $j = 1, 2, \ldots$. For $p > 0$, consider the set $B_p \triangleq \{f \in \ell \colon |f_j| \le j^{-1/p}, j = 1, 2, \ldots\}$.

The next theorem provides lower bounds on $V_n(\mathcal{F})$ and $W_n(\mathcal{F})$ when the ε-entropy of $\mathcal{F}$ behaves as ε^{-p}. It implies that the rates for $V_n(\mathcal{F})$ and $W_n(\mathcal{F})$ in Theorem 2 are tight when $p < 2$.

Theorem 7. *Fix any $p > 0$. Let $\mathcal{F} = \{f \in \ell \colon f_j = (1 + g_j)/2, \{g_j\} \in B_p\}$ and let $\mathcal{X} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots\}$ be the set of all unit vectors in ℓ. For any $\varepsilon > 0$ we have*

$$\mathcal{H}_2(\mathcal{F}, \varepsilon) \le \left(\frac{A}{\varepsilon}\right)^p, \tag{22}$$

where A is a constant depending only on p. Furthermore, for this $\mathcal{F}$, there exists an absolute positive constant c such that the minimax risk satisfies, for any $n \ge 1$,

$$W_n(\mathcal{F}) \ge c n^{-2/(2+p)}, \tag{23}$$

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

$$V_n(\mathcal{F}) \ge c n^{-1/(p-1)}. \tag{24}$$

The proof of Theorem 7 is given in Section 10. We remark that the lower bound (24) (for $p > 2$) holds, up to logarithmic factors, for *any* class satisfying the entropy growth $\Omega(\varepsilon^{-p})$, but we omit the longer proof of this fact. We also remark that for $p > 2$, the $n^{-1/p}$ lower bound can be shown for any estimator taking values within the class $\mathcal{F}$. Obtaining such a lower bound for any estimator remains an open problem.

6. Comparison with global ERM and with skeleton aggregation

Among the methods of estimation designed to work under general entropy assumptions on $\mathcal{F}$, the global ERM or the ERM on ε-nets [9,14,34] hold a dominant place in the literature (see an overview in Section 7). Somewhat less studied method is skeleton aggregation [49]. In this section, we discuss the deficiencies of these two previously known methods that motivated us to introduce aggregation-of-leaders.

Recall that the aggregation-of-leaders procedure has three steps. The first one is to find an empirical ε-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 and seemingly intuitive procedure that we will call the *skeleton aggregation* consists of steps one and three, but not two. This method directly aggregates centers of the cells $\hat{\mathcal{F}}_i^S(\varepsilon)$, that is, the elements $\hat{c}_i$ of the ε-net obtained from the first subsample S. Such kind of procedure was studied by Yang and Barron [49] in the context of well-specified models. The setting in [49] is different from ours since in that paper the ε-net is taken with respect to a non-random metric and the bounds on the minimax risk $W_n(\mathcal{F})$ are obtained when the regression errors are Gaussian. Under this model, [49] provides the bounds not for skeleton aggregation but for a more complex procedure that comprises an additional projection in Hellinger metric. 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 aggregating the centers of cells, is crucial for the success of the aggregation-of-leaders procedure.

Let us first show why the skeleton aggregation yields the correct rates for well-specified models (i.e., when $\eta \in \mathcal{F}$). Similarly to (10), we define the skeleton aggregate $\tilde{f}^{\mathrm{sk}} = \sum_{i=1}^N \theta_i \hat{c}_i$ as a sharp MS-aggregate satisfying a bound analogous to (9): there exists a constant $C > 0$ such that, for any $\delta > 0$,

$$L(\tilde{f}^{\mathrm{sk}}) \le \min_{i=1,\dots,N} L(\hat{c}_i) + C\frac{\log(N/\delta)}{n} \tag{25}$$

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 $\|f-\eta\|^2 = L(f) - L^*$, $\forall f \in \mathcal{F}$. Hence, with probability $1-5\delta$,

$$\begin{aligned}\|\tilde{f}^{\mathrm{sk}} - \eta\|^2 &= L(\tilde{f}^{\mathrm{sk}}) - L^* \\ &\le \min_{i=1,\dots,N} L(\hat{c}_i) - L^* + C\frac{\log(N/\delta)}{n} \\ &= \min_{i=1,\dots,N} \|\hat{c}_i - \eta\|^2 + C\frac{\log(N/\delta)}{n} \\ &\le 2\varepsilon^2 + C\left(\frac{\mathcal{H}_2(\mathcal{F},\varepsilon)}{n} + \frac{\log(1/\delta)}{n} + r^* + \beta\right)\end{aligned} \tag{26}$$

for $\beta = (\log(1/\delta) + \log\log n)/n$, and $r^* = r^*(\mathcal{G})$ with $\mathcal{G} = \{(f-g)^2\colon\ f, g \in \mathcal{F}\}$, where we have used Lemma 9 with $f = \hat{c}_i$, $f' = \eta$ and the fact that $\min_{i=1,\dots,N} d_S(\hat{c}_i, \eta) \le \varepsilon$ for any $\eta \in \mathcal{F}$. The optimal choice of ε in (26) is given by the balance relation $n\varepsilon^2 \asymp \mathcal{H}_2(\mathcal{F},\varepsilon)$ and it can be deduced from Lemma 8 that $r^* + \beta$ is negligible as compared to the leading part $\mathrm{O}(\varepsilon^2 + \mathcal{H}_2(\mathcal{F},\varepsilon)/n)$ with this optimal ε. In particular, we get from (26) combined with (30) and (33), (41) that, under the assumptions of Theorem 2, $\sup_{\eta\in\mathcal{F}} \mathbb{E}\|\tilde{f}^{\mathrm{sk}} - \eta\|^2 \le Cn^{-2/(2+p)}$, $\forall p > 0$.

Let us now consider the misspecified model setting (i.e., the statistical learning framework). Here, the balance relation for the skeleton aggregation takes the form $n\varepsilon \asymp \mathcal{H}_2(\mathcal{F},\varepsilon)$, which yields suboptimal rates unless the class $\mathcal{F}$ is finite. Indeed, without the assumption that the regression function η is in $\mathcal{F}$, we only obtain the bounds

$$\begin{aligned}L(\hat{c}_i) - L^* &= \|\hat{c}_i - \eta\|^2 - \inf_{f\in\mathcal{F}} \|f-\eta\|^2 \\ &\le 2\big(\|\hat{c}_i - \eta\| - \|\eta_{\mathcal{F}} - \eta\|\big) + \frac{1}{n} \le 2\|\hat{c}_i - \eta_{\mathcal{F}}\| + \frac{1}{n},\end{aligned} \tag{27}$$

where $\eta_{\mathcal{F}} \in \mathcal{F}$ is such that $\|\eta_{\mathcal{F}} - \eta\|^2 \leq \inf_{f\in\mathcal{F}} \|f - \eta\|^2 + 1/n$. The crucial difference from (26) is that here $L(\hat{c}_i) - L^*$ behaves itself as a norm $\|\hat{c}_i - \eta_{\mathcal{F}}\|$ and not as a squared norm $\|\hat{c}_i - \eta\|^2$. Using (27) and arguing analogously to (26), we find that for misspecified models, with probability $1 - 5\delta$,

$$\begin{aligned} L\big(\tilde{f}^{\mathrm{sk}}\big) - L^* &\leq 2 \min_{i=1,\dots,N} \|\hat{c}_i - \eta_{\mathcal{F}}\| + C\frac{\log(N/\delta)}{n} \\ &\leq 2\sqrt{2\varepsilon^2 + C\big(r^* + \beta\big)} + C\frac{\log(N/\delta)}{n} \qquad (28) \\ &\leq 2\sqrt{2}\varepsilon + C\bigg(\frac{\mathcal{H}_2(\mathcal{F},\varepsilon)}{n} + \frac{\log(1/\delta)}{n} + \sqrt{r^* + \beta}\bigg). \end{aligned}$$

Here, the optimal ε is obtained from the tradeoff of ε with $\mathcal{H}_2(\mathcal{F},\varepsilon)/n$. As a result, we only get the suboptimal rate $n^{-1/(p+1)} + \mathrm{O}(\sqrt{r^* + \beta})$ for the excess risk of $\tilde{f}^{\mathrm{sk}}$ under the assumptions of Theorem 2. While the above argument is based on upper bounds, it is possible to construct a simple scenario where $\eta, \eta_{\mathcal{F}}$ and some $\hat{c}_i$ are on a line, $\|\eta_{\mathcal{F}} - \hat{c}_i\| = \mathrm{O}(\varepsilon)$, and no other element $\hat{c}_j$ is closer to η than $\hat{c}_i$. For such a setup, $L(\hat{c}_i) - L^*$ is of the order of ε and no convex combination of $\hat{c}_j$ can improve upon $\hat{c}_i$. This indicates that introducing least squares estimators over 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, we consider the global ERM over $\mathcal{F}$ defined by

$$\hat{f}^{\mathrm{erm}} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n} \sum_{(x,y)\in S'} \big(f(x) - y\big)^2, \qquad (29)$$

second – the skeleton aggregate $\tilde{f}^{\mathrm{sk}}$ and, finally, the proposed aggregation-of-leaders estimator $\tilde{f}$. Table 1 summarizes the convergence rates of the expected excess risk $\mathcal{E}_{\mathcal{F}}(\hat{f})$ for $\hat{f} \in \{\tilde{f}, \tilde{f}^{\mathrm{sk}}, \hat{f}^{\mathrm{erm}}\}$ in misspecified model setting, that is, upper bounds on the minimax regret.

The rates for finite $\mathcal{F}$ in Table 1 are obtained in a trivial way by taking the skeleton that coincides with the M functions in the class $\mathcal{F}$. In parametric and nonparametric regime, the rates for the proposed method are taken from Theorems 2 and 3, while for the skeleton aggregate

Table 1. Summary of rates for misspecified case

Regime	Aggregation-of-leaders	Skeleton aggregation	ERM
Finite: $\lvert\mathcal{F}\rvert = M$	$\frac{\log M}{n}$	$\frac{\log M}{n}$	$\sqrt{\frac{\log M}{n}}$
Parametric: $\mathrm{VC}(\mathcal{F}) = v \leq n$	$\frac{v\log(en/v)}{n}$	$\sqrt{\frac{v\log(en/v)}{n}}$	$\sqrt{\frac{v}{n}}$
Nonparametric: $\mathcal{H}_2(\mathcal{F},\varepsilon) = \varepsilon^{-p}$,			
$p \in (0,2)$	$n^{-2/(2+p)}$	$n^{-1/(p+1)} \vee n^{-1/2}(\log n)^{3/2}$	$n^{-1/2}$
$p \in (2,\infty)$	$n^{-1/p}$	$n^{-1/(p+1)}$	$n^{-1/p}$

they follow from (28) with optimized ε combined with the bounds on r^* in Lemma 8 and in (33), (41) below. The rate $\sqrt{v/n}$ for the excess risk of ERM in parametric case is well-known, cf., for example, [3,7]. For the nonparametric regime, the rates for ERM in Table 1 follow from Lemma 11 and the bounds on $\mathfrak{R}_n(\mathcal{F})$ in (33) and (41) below. Moreover, for finite $\mathcal{F}$, it can be shown that the slow rate $\sqrt{\frac{\log M}{n}}$ cannot be improved neither for ERM, nor for any other selector, that is, any estimator with values in $\mathcal{F}$, cf. [21].

In conclusion, for finite class $\mathcal{F}$ aggregation-of-leaders and skeleton aggregation achieve the excess risk rate $\frac{\log M}{n}$, which is known to be optimal [41], whereas the global ERM has a suboptimal rate. For a very massive class $\mathcal{F}$, when the empirical entropy grows polynomially as ε^{-p} with $p \geq 2$ both ERM and aggregation-of-leaders 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 aggregation-of-leaders is optimal, both ERM and skeleton aggregation are suboptimal. Thus, in the misspecified case, skeleton aggregation is good only for very meager (finite) classes while ERM enjoys optimality only for the other extreme – massive nonparametric classes. Note also that, unless $\mathcal{F}$ is finite, skeleton aggregation does not improve upon ERM in the misspecified case.

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

7. Historical remarks and comparison with previous work

The role of entropy and capacity [22] in establishing rates of estimation has been recognized for a long time, since the work of Le Cam [27], Ibragimov and Has'minskiĭ [20] and Birgé [6]. This was also emphasized by Devroye [14] and Devroye *et al.* [15] in the study ERM on ε-nets. The common point is that optimal rate is obtained as a solution to the balance equation $n\varepsilon^2 = \mathcal{H}(\varepsilon)$, with an appropriately chosen non-random entropy $\mathcal{H}(\cdot)$. Yang and Barron [49] 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 balance 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, [49] also provides an achievability result via a skeleton aggregation procedure complemented by a Hellinger projection step. Van de Geer [43] invokes the empirical entropy rather than the non-random entropy to derive rates of estimation in regression problems.

In all these studies, it is assumed that the unknown density, regression function, or parameter belongs to the given class, that is, the model is well-specified. In parallel to these developments, a line of work on pattern recognition that can be traced back to Aizerman, Braverman and Rozonoer [1] and Vapnik and Chervonenkis [47] focused on a different objective, which is characteristic for 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 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 [16–19,36] and references therein).

For parametric classes with a polynomial growth of covering numbers, uniform convergence of averages to expectations with the $\sqrt{(\log n)/n}$ rate has been proved by Vapnik and Chervonenkis [45–47]. In the context of classification, they also obtained a faster rate showing $\mathrm{O}((\log n)/n)$ convergence when the minimal risk $L^* = 0$. For regression problems, similar fast rate when $L^* = 0$ can be shown (it can be deduced after some argument from Assertion 2 on page 204 in [44]; an exact formulation is available, e.g., in [40]). Lee, Bartlett and Williamson [32] showed that the excess risk of the least squares estimator on $\mathcal{F}$ can attain the rate $\mathrm{O}((\log n)/n)$ 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 (i.e., forced to take values in $\mathcal{F}$). In particular, the excess risk of ERM and of any selector on a finite class $\mathcal{F}$ cannot decrease faster than $\sqrt{(\log |\mathcal{F}|)/n}$ [21]. 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 [35] proposed to study the problem of aggregation, or mimicking the best function in the given class, for regression models. Nemirovskii [35] outlined three problems: model selection, convex aggregation, and linear aggregation. The notion of optimal rates of aggregation based on the minimax regret is introduced in [41], 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 [21,33,38,48,50]. For recent developments and a survey we refer to [28,39].

In parallel with this research, the study of the excess risk blossomed with the introduction of Rademacher and local Rademacher complexities [4,5,8,23,24,26]. 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 [24,25] obtained a sharp oracle inequality with the correct rate d/n for the excess risk of least squares estimator on $\mathcal{F}$. Also, for convex $\mathcal{F}$ and $p \in (0, 2)$, the least squares estimator on $\mathcal{F}$ attains the correct excess risk rate $n^{-2/(p+2)}$ under the assumptions of Theorem 2. This can be deduced from Theorem 5.1 in [25], remarks after it and in Example 4 on page 87 of [25]. However, the convexity assumption appears to be crucial; without this assumption Koltchinskii [25], Theorem 5.2, obtains for the least squares estimator only a non-sharp inequality with leading constant $C > 1$, cf. (3). As follows from the results in Section 3 our procedure overcomes this problem.

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. For the regression setting with finite $\mathcal{F}$, 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 [10,13,21,49]. However, Audibert [2] showed that, for the regression with random design, exponential weighting is suboptimal when the error is measured by the probability of deviation rather than by the expected risk. He proposed an alternative method, optimal both in probability and in deviation, which involves finding an ERM on a star connecting a global ERM and the other $|\mathcal{F}|-1$ functions. In [30], the authors exhibited another 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. Recently yet another procedure achieving the deviation optimality has been proposed in [31]. It is based on a penalized version of exponential weighting and extends the method of [12] originally proposed for regression with fixed design. The methods of [2,30, 31] provide examples of sharp MS-aggregates that can be used at the third step of our procedure.

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 [11] 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 each of these balls, and then the overall output is an aggregate of outputs on the balls. 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.

8. Proofs of Theorems 2–4 and 5

We first state some auxiliary lemmas.

Lemma 8. *The following values can be taken as localization radii $r^* = r^*(\mathcal{G})$ for $\mathcal{G} = \{(f-g)^2\colon f, g \in \mathcal{F}\}$.*

(i) *For any class $\mathcal{F} \subseteq \{f\colon 0 \le f \le 1\}$, and $n \ge 2$,*

$$r^* = C\log^3(n)\mathfrak{R}_n^2(\mathcal{F}). \tag{30}$$

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

$$r^* = C\frac{v}{n}\log\left(\frac{en}{v}\right)$$

whenever $n \ge C_A v$ with $C_A > 1$ large enough depending only on A.

(iii) *If $\mathcal{F}$ is a finite class with $|\mathcal{F}| \ge 2$,*

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

The proof of this lemma is given in the Appendix. The following lemma is a direct consequence of Theorem 14 proved in the Appendix.

Lemma 9. *For any class $\mathcal{F} \subseteq \{f\colon 0 \le f \le 1\}$ and $\delta > 0$, with probability at least $1-4\delta$,*

$$\|f-f'\|^2 \le 2d_S^2(f,f') + C(r^* + \beta) \qquad \forall f, f' \in \mathcal{F}, \tag{31}$$

where $\beta = (\log(1/\delta) + \log\log n)/n$, *and* $r^* = r^*(\mathcal{G})$ *for* $\mathcal{G} = \{(f-g)^2\colon\ f, g \in \mathcal{F}\}$.

We will also use the following bound on the Rademacher average in terms of the empirical entropy [3,40].

Lemma 10. *For any class* $\mathcal{F} \subseteq \{f\colon\ 0 \le f \le 1\}$,

$$\hat{\mathfrak{R}}_n(\mathcal{F}, S) \le \inf_{\alpha \ge 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_{\alpha}^{1}\sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho, S)}\,\mathrm{d}\rho\right\}. \tag{32}$$

Proof of Theorem 2. Consider the case $p \in (0,2)$. Assume without loss of generality that $A = 1$, that is, $\sup_{S \in \mathcal{Z}^n} \log \mathcal{N}_2(\mathcal{F}, \rho, S) \le \rho^{-p}$. For $p \in (0,2)$, the bound (32) with $\alpha = 0$ combined with (30) yields

$$\mathfrak{R}_n(\mathcal{F}) \le \frac{12}{\sqrt{n}(1-p/2)}, \qquad r^* \le C\frac{(\log n)^3}{n} \tag{33}$$

for some absolute constant C. Thus,

$$\gamma \le C\left(\varepsilon + \frac{(\log n)^{3/2}}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{34}$$

$$\gamma\sqrt{r^*} \le C(\log n)^{3/2}\left(\frac{\varepsilon}{\sqrt{n}} + \frac{(\log n)^{3/2}}{n} + \frac{\sqrt{\log(1/\delta)}}{n}\right). \tag{35}$$

These inequalities together with (11) and (12) yield that for $0 < \delta < 1/2$, with probability at least $1 - 2\delta$,

$$L(\tilde{f}) - L^* \le C\left(\frac{\varepsilon^{-p}}{n} + \frac{\log(1/\delta)}{n} + \gamma\sqrt{r^*} + \frac{\gamma^{1-p/2}}{\sqrt{n}}\right). \tag{36}$$

The value of ε minimizing the right-hand side in (36) is $\varepsilon = n^{-1/(2+p)}$, which justifies the choice made in the theorem. 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. For $\varepsilon = n^{-1/(2+p)}$ the right-hand side of (36) is bounded by $Cn^{-2/(2+p)}$ 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 $Cn^{-2/(2+p)}$, which proves (14) for $p \in (0,2)$.

Next, consider the case $p > 2$. From (32) with $\alpha = n^{-1/p}$, $\hat{\mathfrak{R}}_n(\mathcal{F}, S) \le Cn^{-1/p}$ and $r^* = (\log n)^3 n^{-2/p}$. Choosing $\varepsilon = n^{-1/(2+p)}$,

$$\gamma\sqrt{r^*} \le C\big(\varepsilon\sqrt{r^*} + r^* + \sqrt{\beta r^*}\big) \le Cn^{-1/p}.$$

The first statement of the theorem follows from (12) with the choice $\alpha = n^{-1/p}$ and by noting that $\frac{\varepsilon^{-p}}{n}$ is of the lower order than $n^{-1/p}$. The case of $p = 2$ follows similarly (see proof of Theorem 5). The second part of the theorem follows from Theorem 5. □

Proof of Theorem 3. Throughout this proof, C is a generic notation for positive constants that may depend only on A. Since $\varepsilon = n^{-1/2}$ the expression for r^* in Lemma 8(ii) leads to the bounds $\gamma \le C(\sqrt{\frac{v\log(en/v)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}})$, and $\gamma\sqrt{r^*} \le C(\frac{v\log(en/v)}{n} + \frac{\log(1/\delta)}{n})$. Next, since $\mathcal{N}_2(\mathcal{F}, \rho, S') \le \max\{1, (A/\rho)^v\}$ we get

$$\begin{aligned}\frac{1}{\sqrt{n}}\int_0^{C\gamma}\sqrt{\log\mathcal{N}_2\bigl(\mathcal{F},\rho,S'\bigr)}\,\mathrm{d}\rho &\le \sqrt{\frac{v}{n}}\int_0^{C\gamma/A\wedge 1}\sqrt{\log(1/t)}\,\mathrm{d}t\\ &\le C\sqrt{\frac{v}{n}}\gamma\sqrt{\log(C/\gamma)\vee 1},\end{aligned}$$

where the last inequality is due to (A.9). We assume w.l.o.g. that in the last expression C is large enough to guarantee that the function $\gamma \mapsto \gamma\sqrt{\log(C/\gamma)\vee 1}$ is increasing, so that we can replace γ by the previous upper bound. This yields, after some algebra,

$$\gamma\sqrt{\log(C/\gamma)\vee 1} \le C\biggl(\frac{\sqrt{v}\log(en/v)}{\sqrt{n}} + \frac{\sqrt{\log(1/\delta)\log(en/v)}}{\sqrt{n}}\biggr)$$

if $n \ge Cv$ for C large enough. The above inequalities together with (11) and (12) imply that, with probability at least $1-2\delta$,

$$L(\tilde{f}) - L^* \le C\biggl(\frac{\log\mathcal{N}_2(\mathcal{F},\varepsilon,S)}{n} + \frac{v\log(en/v)}{n} + \frac{\log(1/\delta)}{n}\biggr).$$

Using that $\log\mathcal{N}_2(\mathcal{F}, \varepsilon, S) \le \max\{1, (A/\varepsilon)^v\}$ and integrating over δ we get the desired bound for the expected excess risk $\mathbb{E}L(\tilde{f}) - L^*$. □

Proof of Theorem 4. By definition of the estimator, for any fixed integer $m \le s$ and ν such that $|\nu| = m$ we first construct the least squares estimators over the cells $\mathcal{F}_{\nu,m}$:

$$\hat{f}_{\nu,m}^{S,S'} \in \operatorname*{argmin}_{f\in\mathcal{F}_{\nu,m}} \frac{1}{n}\sum_{(x,y)\in S'}\bigl(f(x)-y\bigr)^2. \tag{37}$$

Since $\mathcal{F}_{\nu,m}$ is a convex hull of m functions we can apply [29] to get that for any $t > 0$, with probability at least $1-\mathrm{e}^{-t}$,

$$L\bigl(\hat{f}_{\nu,m}^{S,S'}\bigr) \le \inf_{f\in\mathcal{F}_{\nu,m}} L(f) + C(\tilde{\psi}_{m,n} + t/n), \tag{38}$$

where

$$\tilde{\psi}_{m,n} \triangleq \frac{m}{n} \wedge \sqrt{\frac{1}{n}\log\biggl(1+\frac{m}{\sqrt{n}}\biggr)}.$$

Thus, the event E where (38) holds simultaneously for all $(m,\nu) \in \mathcal{I} = \{(m,\nu)\colon m = 1,\dots,s, |\nu| = m\}$ is of probability at least $1 - N\mathrm{e}^{-t}$. Here, $N = |\mathcal{I}|$. Choose now $t = \log(N/\delta)$.

Then, on the intersection of E with the event where (9) holds we have that, with probability at least $1-2\delta$,

$$\begin{aligned} L(\tilde{f}) &\le \inf_{f\in\mathcal{F}_{\Theta^{\mathrm{C}}(s)}} L(f) + C\biggl(\tilde{\psi}_{s,n} + \frac{\log(N/\delta)}{n}\biggr) \\ &\le \inf_{f\in\mathcal{F}_{\Theta^{\mathrm{C}}(s)}} L(f) + C\biggl(\frac{s}{n}\log\biggl(\frac{eM}{s}\biggr) + \frac{\log(1/\delta)}{n}\biggr), \end{aligned} \tag{39}$$

where we have used the inequalities $\tilde{\psi}_{m,n} \le \tilde{\psi}_{s,n}$, $\forall m \le s$, and $N = \sum_{m=1}^{s}\binom{M}{m} \le (\frac{eM}{s})^s$. On the other hand, for the least squares estimator $\hat{f}^{\mathrm{C}}$ on the convex hull of all $f_1, \dots, f_M$, using again the result of [29] we have that for any $u > 0$, with probability at least $1-\mathrm{e}^{-u}$,

$$\begin{aligned} L(\hat{f}^{\mathrm{C}}) &\le \inf_{f\in\mathcal{F}_{\Theta^{\mathrm{C}}}} L(f) + C(\tilde{\psi}_{M,n} + u/n) \\ &\le \inf_{f\in\mathcal{F}_{\Theta^{\mathrm{C}}(s)}} L(f) + C\biggl(\sqrt{\frac{1}{n}\log\biggl(1+\frac{M}{\sqrt{n}}\biggr)} + \frac{u}{n}\biggr). \end{aligned} \tag{40}$$

Now, we aggregate only two estimators, $\tilde{f}$ and $\hat{f}^{\mathrm{C}}$ to obtain the final aggregate $\tilde{f}^*$. This yields, in view of (9) with $N=2$, (39), and (40) with $u = \log(1/\delta)$, that with probability at least $1-4\delta$,

$$\begin{aligned} L(\tilde{f}^*) &\le \min\bigl\{L(\tilde{f}), L(\hat{f}^{\mathrm{C}})\bigr\} + C\frac{\log(2/\delta)}{n} \\ &\le \inf_{f\in\mathcal{F}_{\Theta^{\mathrm{C}}(s)}} L(f) + C\biggl(\min\biggl\{\frac{s}{n}\log\biggl(\frac{eM}{s}\biggr), \sqrt{\frac{1}{n}\log\biggl(1+\frac{M}{\sqrt{n}}\biggr)}\biggr\} + \frac{\log(1/\delta)}{n}\biggr), \end{aligned}$$

which immediately implies the desired bound for the expected excess risk $\mathbb{E}L(\tilde{f}^*) - \inf_{f\in\mathcal{F}_{\Theta^{\mathrm{C}}(s)}} L(f)$. □

Proof of Theorem 5. Without loss of generality assume in this proof that $A=1$, that is, that $\sup_{S\in\mathcal{Z}^n}\log\mathcal{N}_2(\mathcal{F},\rho,S) \le \rho^{-p}$. Using (32) we bound $\mathfrak{R}_n(\mathcal{F})$ for $p>2$ as follows:

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

For $p>2$, the balance equation $\alpha = n^{-1/2}\alpha^{-(p-2)/2}$ yields $\alpha = n^{-1/p}$. This and (30) lead to the bounds

$$\mathfrak{R}_n(\mathcal{F}) \le Cn^{-1/p}, \qquad r^* \le C(\log n)^3 n^{-2/p}. \tag{41}$$

For $p=2$, choosing $\alpha = n^{-1/2}$,

$$\mathfrak{R}_n(\mathcal{F}) \le Cn^{-1/2}\log n, \qquad r^* \le C(\log n)^5 n^{-2/p}. \tag{42}$$

Consider the case $p > 2$. Let $\eta_{\mathcal{F}} \in \mathcal{F}$ be such that $\|\eta_{\mathcal{F}} - \eta\|^2 \le \inf_{f\in\mathcal{F}} \|f - \eta\|^2 + 1/n$. Lemma 9, (30) and (41) imply that, with probability at least $1 - 4\delta$, for all $i = 1, \dots, N$,

$$\begin{aligned}\|\hat{f}_i^{S,S'} - \eta\|^2 - \inf_{f\in\mathcal{F}} \|f - \eta\|^2 &\le 2\|\hat{f}_i^{S,S'} - \eta_{\mathcal{F}}\|^2 + \|\eta_{\mathcal{F}} - \eta\|^2 + 1/n \\ &\le 4d_S^2(\hat{f}_i^{S,S'}, \eta_{\mathcal{F}}) + \|\eta_{\mathcal{F}} - \eta\|^2 + C(r^* + \beta) + 1/n \\ &\le 4d_S^2(\hat{f}_i^{S,S'}, \eta_{\mathcal{F}}) + \Delta^2 + C\left(\frac{(\log n)^3}{n^{2/p}} + \frac{\log(1/\delta)}{n}\right).\end{aligned}$$

Since $\min_{i=1,\dots,N} d_S(\hat{f}_i^{S,S'}, \eta_{\mathcal{F}}) \le 2\varepsilon$ and $\varepsilon = n^{-1/(2+p)}$ we get that, with probability at least $1 - 4\delta$,

$$\begin{aligned}\min_{i=1,\dots,N} L(\hat{f}_i^{S,S'}) - L^* &= \min_{i=1,\dots,N} \|\hat{f}_i^{S,S'} - \eta\|^2 - \inf_{f\in\mathcal{F}} \|f - \eta\|^2 \\ &\le \Delta^2 + C\left(n^{-2/(2+p)} + \frac{\log(1/\delta)}{n}\right).\end{aligned} \tag{43}$$

Further, Lemma 11 and (41) imply that, with probability at least $1 - 2\delta$,

$$\begin{aligned}\|\hat{f}_i^{S,S'} - \eta\|^2 - \inf_{f\in\hat{\mathcal{F}}_i^S} \|f - \eta\|^2 &\le C\left(\mathfrak{R}_n(\hat{\mathcal{F}}_i^S) + \frac{\log(1/\delta)}{n}\right) \\ &\le C\left(n^{-1/p} + \frac{\log(1/\delta)}{n}\right).\end{aligned}$$

Combining this bound with (43) we can conclude that, with probability at least $1 - 6\delta$,

$$\begin{aligned}\min_{i=1,\dots,N} L(\hat{f}_i^{S,S'}) - L^* &= \min_{i=1,\dots,N} \|\hat{f}_i^{S,S'} - \eta\|^2 - \inf_{f\in\mathcal{F}} \|f - \eta\|^2 \\ &\le C\left(\min(n^{-2/(2+p)} + \Delta^2, n^{-1/p}) + \frac{\log(1/\delta)}{n}\right).\end{aligned}$$

Together with (9), this yields the next bound that holds with probability at least $1 - 7\delta$:

$$\begin{aligned}\|\tilde{f} - \eta\|^2 - \inf_{f\in\mathcal{F}} \|f - \eta\|^2 &= L(\tilde{f}) - L^* \\ &\le C\left(\frac{A\varepsilon^{-p}}{n} + \min(n^{-2/(2+p)} + \Delta^2, n^{-1/p}) + \frac{\log(1/\delta)}{n}\right) \\ &\le C\left(n^{-2/(2+p)} + \min(n^{-2/(2+p)} + \Delta^2, n^{-1/p}) + \frac{\log(1/\delta)}{n}\right),\end{aligned}$$

and (20) follows. For $p = 2$, the above bound gains a factor $\log n$ in front of $n^{-1/p}$ only. □

9. Proof of Theorem 1

We start with the following bound on the risk of least squares estimators in terms of Rademacher complexity.

Lemma 11. *Let $\mathcal{F}$ be a class of measurable functions from $\mathcal{X}$ to* $[0,1]$. *Then, for any* $t>0$, *with probability at least* $1-2\mathrm{e}^{-t}$, *the least squares estimator* $\hat{f}^{\mathrm{erm}}$ *on* $\mathcal{F}$ *based on a sample* S' *of size* n (*cf.* (29)) *satisfies*

$$L(\hat{f}^{\mathrm{erm}}) \le L^* + C\hat{\mathfrak{R}}_n(\ell\circ\mathcal{F},S') + \frac{Ct}{n}.$$

The proof of this lemma is given in the Appendix and is based on combination of results from [4]. Note that here we have both the remainder term of the order $1/n$ and the leading constant 1, which is crucial for our purposes.

Using Lemma 11 with $\mathcal{F}=\hat{\mathcal{F}}_i^S$ and the union bound, we obtain that, with probability at least $1-2N\mathrm{e}^{-t}$, for all $i=1,\dots,N$,

$$L(\hat{f}_i^{S,S'}) \le \inf_{f\in\hat{\mathcal{F}}_i^S} L(f) + C\hat{\mathfrak{R}}_n(\ell\circ\hat{\mathcal{F}}_i^S,S') + Ct/n. \tag{44}$$

Recall that $N=\mathcal{N}_2(\mathcal{F},\varepsilon,S)$. Setting $t=\log(4N/\delta)$ and using (44) and (9) we obtain that, with probability at least $1-(3/2)\delta$,

$$L(\tilde{f}) \le L^* + C\left(\frac{\log(\mathcal{N}_2(\mathcal{F},\varepsilon,S)/\delta)}{n} + \max_{i=1,\dots,N}\hat{\mathfrak{R}}_n(\ell\circ\hat{\mathcal{F}}_i^S,S')\right). \tag{45}$$

To complete the proof of (12) we need to evaluate the Rademacher complexities appearing in (45):

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

The difficulty here is that the set $\hat{\mathcal{F}}_i^S=\hat{\mathcal{F}}_i^S(\varepsilon)$ is defined via the pseudo-metric d_S based on sample S while the empirical Rademacher complexity is evaluated on another sample S'. To match the metrics, we embed $\hat{\mathcal{F}}_i^S(\varepsilon)$ into $d_{S'}$-balls with properly chosen radius $\bar{\gamma}$:

$$\hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma}) \triangleq \{f\in\mathcal{F}: d_{S'}(f,\hat{c}_i)\le\bar{\gamma}\}, \qquad i=1,\dots,N,$$

where the pseudo-metric $d_{S'}$ is taken with respect to the set S' while the ε-net $\hat{c}_1,\dots,\hat{c}_N$ is constructed with respect to d_S. The next lemma shows that, with high probability, $\hat{\mathcal{F}}_i^S(\varepsilon)$ is included into $\hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma})$ for an appropriate choice of $\bar{\gamma}$.

Lemma 12. *Fix* $t>0$, $\varepsilon>0$. *Let* $r^*=r^*(\mathcal{G})$ *for* $\mathcal{G}=\{(f-g)^2: f,g\in\mathcal{F}\}$. *Define* $r_0=(t+6\log\log n)/n$ *and* $\bar{\gamma}=\sqrt{4\varepsilon^2+284r^*+120r_0}$. *Then, with probability at least* $1-8N\mathrm{e}^{-t}$ *with* *respect to the distribution of $S \cup S'$, we have the inclusions*

$$\hat{\mathcal{F}}_i^S(\varepsilon) \subseteq \hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma}), \qquad i = 1, \dots, N,$$

and hence, with the same probability,

$$\hat{\mathfrak{R}}_n\big(\ell \circ \hat{\mathcal{F}}_i^S(\varepsilon), S'\big) \le \hat{\mathfrak{R}}_n\big(\ell \circ \hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma}), S'\big), \qquad i = 1, \dots, N.$$

Proof. Let P_n and P_n' denote the empirical averages over the samples S and S', respectively. By Theorem 14, with probability at least $1 - 4\mathrm{e}^{-t}$,

$$P(f-g)^2 \le 2P_n(f-g)^2 + 106r^* + 48r_0 \qquad \forall f, g \in \mathcal{F},$$

and, with the same probability,

$$P_n'(f-g)^2 \le 2P(f-g)^2 + 72r^* + 24r_0 \qquad \forall f, g \in \mathcal{F}.$$

Therefore, with probability at least $1 - 8\mathrm{e}^{-t}$,

$$P_n'(f-g)^2 \le 4P_n(f-g)^2 + 284r^* + 120r_0 \qquad \forall f, g \in \mathcal{F}.$$

Applying this to $g = \hat{c}_i$ and taking a union bound over $i = 1, \dots, N$, completes the proof. □

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

Lemma 13. *Let $r^* = r^*(\mathcal{G})$ for $\mathcal{G} = \{(f-g)^2\colon\ f, g \in \mathcal{F}\}$. Then, for any $\bar{\gamma} \ge \sqrt{r^*}$ we have*

$$\hat{\mathfrak{R}}_n\big(\ell \circ \hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma}), S'\big) \le \bar{\gamma}\sqrt{r^*} + \inf_{\alpha \ge 0}\left\{4\alpha + \frac{24}{\sqrt{n}}\int_\alpha^{\bar{\gamma}} \sqrt{\log \mathcal{N}_2\big(\mathcal{F}, \rho, S'\big)}\,\mathrm{d}\rho\right\}.$$

Proof. Throughout the proof, we fix the samples S and S'. We have

$$\begin{aligned}\hat{\mathfrak{R}}_n\big(\ell \circ \hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma}), S'\big) &= \mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma})} \frac{1}{n} \sum_{(x_j,y_j)\in S'} \sigma_j\big(f(x_j) - y_j\big)^2 \\ &= \mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma})} \frac{1}{n} \sum_{(x_j,y_j)\in S'} \sigma_j\big(f(x_j) - \hat{c}_i(x_j)\big)^2 \qquad (46) \\ &\quad + 2\mathbb{E}_\sigma \sup_{f \in \hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma})} \frac{1}{n} \sum_{(x_j,y_j)\in S'} \sigma_j\big(f(x_j) - \hat{c}_i(x_j)\big)\big(\hat{c}_i(x_j) - y_j\big),\end{aligned}$$

where we have used the decomposition $(f(x) - y)^2 = (f(x) - \hat{c}_i(x))^2 + (\hat{c}_i(x) - y)^2 + 2(f(x) - \hat{c}_i(x))(\hat{c}_i(x) - y)$, $\forall x, y$, and the fact that $(\hat{c}_i(x) - y)^2$ does not depend on f. Conditionally on the sample S, the functions $\hat{c}_i$ are fixed. Consider the sets of functions

$$\begin{aligned}\mathcal{G}'_i &= \{(f-\hat{c}_i)^2\colon\ f \in \hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma})\} \\ &= \Big\{(f-\hat{c}_i)^2\colon\ f \in \mathcal{F}, \frac{1}{n}\sum_{(x,y)\in S'}\big(f(x)-\hat{c}_i(x)\big)^2 \le \bar{\gamma}^2\Big\}.\end{aligned}$$

Recall that we assume $\hat{c}_i \in \mathcal{F}$ (the ε-net is proper). Thus $\mathcal{G}'_i \subseteq \mathcal{G}[\bar{\gamma}^2, S']$ for $\mathcal{G} = \{(f-g)^2\colon\ f, g \in \mathcal{F}\}$, which implies

$$\begin{aligned}\mathbb{E}_\sigma \sup_{f\in\hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma})} \frac{1}{n}\sum_{(x_j,y_j)\in S'} \sigma_j\big(f(x_j)-\hat{c}_i(x_j)\big)^2 &\le \hat{\mathfrak{R}}_n\big(\mathcal{G}[\bar{\gamma}^2, S'], S'\big) \\ &\le \phi_n\big(\bar{\gamma}^2\big) \le \bar{\gamma}\sqrt{r^*},\end{aligned} \tag{47}$$

where $\phi_n(\bar{\gamma}^2) = \phi_n(\bar{\gamma}^2, \mathcal{G})$ and the last inequality is due to the assumption $\bar{\gamma}^2 > r^*$ and the fact that $\phi_n(r)/\sqrt{r}$ is non-increasing.

We now turn to the cross-product term in (46). Define the following sets of functions on $\mathcal{X} \times \mathcal{Y}$:

$$\mathcal{G}_i^{S,S'} = \big\{g_f(x,y) = \big(f(x)-\hat{c}_i(x)\big)\big(\hat{c}_i(x)-y\big)\colon\ f \in \hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma})\big\}.$$

Then,

$$\mathbb{E}_\sigma \sup_{f\in\hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma})} \frac{1}{n}\sum_{(x_j,y_j)\in S'} \sigma_j\big(f(x_j)-\hat{c}_i(x_j)\big)\big(\hat{c}_i(x_j)-y_j\big) = \hat{\mathfrak{R}}_n\big(\mathcal{G}_i^{S,S'}, S'\big). \tag{48}$$

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

$$\frac{1}{n}\sum_{(x,y)\in S'} g_f(x,y)^2 = \frac{1}{n}\sum_{(x,y)\in S'} \big(f(x)-\hat{c}_i(x)\big)^2\big(\hat{c}_i(x)-y\big)^2 \tag{49}$$

$$\le \frac{1}{n}\sum_{(x,y)\in S'} \big(f(x)-\hat{c}_i(x)\big)^2 \le \bar{\gamma}^2 \tag{50}$$

since $\hat{c}_i$ and y take values in $\mathcal{Y} = [0, 1]$. For the same reason,

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

implying $\mathcal{N}_2(\mathcal{G}_i^{S,S'}, \rho, S') \le \mathcal{N}_2(\hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma}), \rho, S')$ for all $\rho > 0$. Hence, by Lemma 10,

$$\begin{aligned} \hat{\mathfrak{R}}_n\bigl(\mathcal{G}_i^{S,S'}, S'\bigr) &\le \inf_{\alpha\ge 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^{\bar{\gamma}} \sqrt{\log\mathcal{N}_2\bigl(\hat{\mathcal{F}}_i^{S,S'}(\bar{\gamma}), \rho, S'\bigr)}\,\mathrm{d}\rho\right\} \\ &\le \inf_{\alpha\ge 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^{\bar{\gamma}} \sqrt{\log\mathcal{N}_2\bigl(\mathcal{F}, \rho, S'\bigr)}\,\mathrm{d}\rho\right\}, \end{aligned} \tag{51}$$

where the integration goes to $\bar{\gamma}$ in view of (49). The lemma now follows from (46)–(48) and (51). □

Combining (45), Lemma 12 with $t = \log(16N/\delta)$, and Lemma 13 we find that, with probability at least $1 - 2\delta$,

$$\begin{aligned} L(\tilde{f}) \le L^* + C\biggl(&\frac{\log(\mathcal{N}_2(\mathcal{F}, \varepsilon, S)/\delta)}{n} + \bar{\gamma}\sqrt{r^*} \\ &+ \inf_{\alpha\ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{\bar{\gamma}} \sqrt{\log\mathcal{N}_2\bigl(\mathcal{F}, \rho, S'\bigr)}\,\mathrm{d}\rho\right\}\biggr), \end{aligned} \tag{52}$$

which yields the bound (11).

10. Proofs of the lower bounds

Proof of Theorem 6. 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 [37]), for $k \ge 2d$ there exists of a subset $\mathcal{C}'$ of $\mathcal{C}$ with the following properties: (a) $\log|\mathcal{C}'| \ge (d/4)\log(k/(6d))$ and (b) $\rho_H(\omega, \omega') \ge d$ for any $\omega, \omega' \in \mathcal{C}'$. Here, $\rho_H(\omega, \omega') = \sum_j \mathbf{1}\{\omega_j \ne \omega'_j\}$ denotes the Hamming distance where ω_j, ω'_j are the components of ω, ω'. To any $\omega \in \mathcal{C}'$ we associate a function g_ω on $\mathcal{X}$ defined by $g_\omega(x^i) = \omega_i$ for $i = 1, \dots, k$ and $g_\omega(x^i) = 0$, $i \ge k + 1$, where ω_i is the ith component of ω.

Let μ_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 $j \ge k + 1$. Denote by $\mathbf{P}_\omega$ the joint distribution of (X, Y) having this marginal μ_X and $Y \in \{0, 1\}$ with the conditional distribution $\mathbb{E}(Y|X = x) = P(Y = 1|X = x) = 1/2 + g_\omega(x)/4 \triangleq \eta_\omega(x)$ for all $x \in \mathcal{X}$.

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

$$\|\eta_\omega - \eta_{\omega'}\|^2 = \rho_H\bigl(\omega, \omega'\bigr)/(16k) \ge \alpha/32 \qquad \forall\omega, \omega' \in \mathcal{C}'. \tag{53}$$

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

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

Using the inequality $-\log(1+u) \le -u + u^2/2$, $\forall u > -1$, and the fact that $1/2 \le \eta_\omega(X) \le 3/4$ 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

$$K(\mathbf{P}_\omega, \mathbf{P}_{\omega'}) \le 2n\mathbb{E}\big(\eta_\omega(X) - \eta_{\omega'}(X)\big)^2 \le \frac{n\|g_\omega - g_{\omega'}\|^2}{8} \le \frac{nd}{8k} \le \frac{n\alpha}{8} \qquad \forall \omega, \omega' \in \mathcal{C}'. \tag{54}$$

From (53), (54) and Theorem 2.7 in [42], the result of Theorem 6 follows if we show that

$$n\alpha/8 \le \log\big(|\mathcal{F}'| - 1\big)/16 \tag{55}$$

with

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

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

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

Using that $x \ge 2\log x$ for $x \ge 0$ 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 $d \le 3$, it is enough to consider $\alpha = (C_1/n)\log(C_2 n)$ and (55) is also satisfied for suitable C_1, C_2. □

Proof of Theorem 7. We first prove the entropy bound (22). It suffices to obtain the same bound for B_p in place of $\mathcal{F}$. Fix $\varepsilon > 0$ and set $J = (2/\varepsilon)^p$. Without loss of generality, assume that J is an integer. Let $\mathcal{M}$ be an ε-net on B_p in ℓ_∞ metric constructed as follows. For all $v \in \mathcal{M}$, the coordinate v_j of v takes discrete values with step ε within the interval $[-j^{-1/p}, j^{-1/p}]$ if $j \le J$, and $v_j = 0$ for all $j > J$. Then,

$$|\mathcal{M}| \le \prod_{j=1}^{J} \left(\frac{2}{\varepsilon j^{1/p}}\right).$$

One can check that

$$\log\left(\prod_{j=1}^{J} j^{-1/p}\right) = -\frac{1}{p}\sum_{j=1}^{J} \log j \le -\frac{1}{p}\int_2^J (\log t)\,\mathrm{d}t \le -\frac{J}{p}(\log J - 1),$$

which implies that $|\mathcal{M}| \le \exp(J/p)$. Thus (22) follows.

Proof of (23). Fix $d = \lceil n^{p/(2+p)} \rceil$. Let $\Omega_d = \{0,1\}^d$ be the set of all binary sequences of length d. Define μ_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 $j \ge d+1$. For any $\omega \in \Omega_d$, denote by $\mathbf{P}_\omega$ the joint distribution of (X, Y) having this marginal μ_X and $Y \in \{0,1\}$ with the conditional distribution defined by the relation

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

for $i = 1, \dots, d$, and $\eta_\omega(\mathbf{e}_i) = 1/2$ for $i \geq d+1$. The regression function corresponding to $\mathbf{P}_\omega$ is then $\eta_\omega = \{\eta_\omega(\mathbf{e}_j)\} \in \ell$. It is easy to see that since $\omega_i \in \{0, 1\}$ for any estimator $\hat{f} = \{\hat{f}_j\} \in \ell$ we have

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

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

$$\|\hat{f} - \eta_\omega\|^2 \geq \frac{1}{d}\sum_{i=1}^{d} \frac{|\hat{\omega}_i - \omega_i|^2}{64d^{2/p}} = \frac{\rho_H(\hat{\omega}, \omega)}{64d^{1+2/p}}, \tag{56}$$

where $\rho_H(\cdot, \cdot)$ is the Hamming distance. From Assouad's lemma (cf. Theorem 2.12(iv) in [42]),

$$\inf_{\hat{\omega}} \max_{\omega \in \Omega_d} \mathbf{E}_\omega^{(n)} \rho_H(\hat{\omega}, \omega) \geq \frac{d}{4}\exp(-\alpha), \tag{57}$$

where $\alpha = \max\{K(\mathbf{P}_\omega, \mathbf{P}_{\omega'})\colon\ \omega, \omega' \in \Omega_d, \rho_H(\omega, \omega') = 1\}$. Here, $\mathbf{E}_\omega^{(n)}$ denotes the distribution of the n-sample D_n when $(X_i, Y_i) \sim \mathbf{P}_\omega$ for all i. Since $1/2 \leq \eta_\omega(X) \leq 3/4$, the Kullback–Leibler divergence can be bounded in the same way as in (54):

$$K(\mathbf{P}_\omega, \mathbf{P}_{\omega'}) \leq 2n\mathbb{E}\big(\eta_\omega(X) - \eta_{\omega'}(X)\big)^2 = \frac{2n}{d}\sum_{i=1}^{d} \frac{(\omega_i - \omega_i')^2}{64d^{2/p}} = \frac{n\rho_H(\hat{\omega}, \omega)}{32d^{1+2/p}} \leq \frac{1}{32}$$

for all $\omega, \omega' \in \Omega_d$ such that $\rho_H(\omega, \omega') = 1$. Combining this result with (56) and (57), we find

$$\inf_{\hat{f}} \max_{\omega \in \Omega_d} \mathbf{E}_\omega^{(n)} \|\hat{f} - \eta_\omega\|^2 \geq \frac{\mathrm{e}^{-1/32}}{128d^{2/p}} \geq c_* n^{-2/(2+p)} \tag{58}$$

for some absolute constant $c_* > 0$. Now, the set $\{\eta_\omega\colon\ \omega \in \Omega_d\}$ is contained in $\mathcal{F}$, so that

$$W_n(\mathcal{F}) \geq \inf_{\hat{f}} \max_{\omega \in \Omega_d} \mathbf{E}_\omega^{(n)} \|\hat{f} - \eta_\omega\|^2 \tag{59}$$

and (23) follows immediately from (58) and (59).

Proof of (24). Set $d = 2\lceil n^{p/(p-1)}\rceil$ and define the joint distribution $\mathbf{P}_\omega$ of (X, Y) as in the proof of (23) with the difference that now we choose the conditional probabilities as follows:

$$\eta_\omega(\mathbf{e}_j) = \frac{1}{2} + \frac{\omega_j}{4}, \qquad j = 1, \dots, d \quad \text{and} \quad \eta_\omega(\mathbf{e}_j) = \frac{1}{2}, \qquad j \geq d+1,$$

where $\omega = (\omega_1, \dots, \omega_d) \in \Omega_d' = \{-1, 1\}^d$. Set $\eta_\omega = \{\eta_\omega(\mathbf{e}_j)\} \in \ell$ with $\omega \in \Omega_d'$. Then

$$\inf_{f \in \mathcal{F}} \|f - \eta_\omega\|^2 \leq \|f_\omega - \eta_\omega\|^2 = \frac{1}{16}\big(1 - d^{-1/p}\big)^2,$$

where $f_\omega = \{f_\omega(\mathbf{e}_j)\} \in \mathcal{F}$ is a sequence with components

$$f_\omega(\mathbf{e}_j) = \frac{1}{2} + \frac{\omega_j}{4d^{1/p}}, \qquad j = 1, \dots, n \quad \text{and} \quad f_\omega(\mathbf{e}_j) = \frac{1}{2}, \qquad j \geq d+1.$$

Hence,

$$\begin{aligned} V_n(\mathcal{F}) &= \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}} \Bigl\{ \mathbb{E}\|\hat{f} - \eta\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta\|^2 \Bigr\} \\ &\geq \inf_{\hat{f}} \max_{\omega \in \Omega'_d} \Bigl\{ \mathbf{E}_\omega^{(n)} \|\hat{f} - \eta_\omega\|^2 - \inf_{f \in \mathcal{F}} \|f - \eta_\omega\|^2 \Bigr\} \\ &\geq \inf_{\hat{f}} \int_{\Omega'_d} \mathbf{E}_\omega^{(n)} \|\hat{f} - \eta_\omega\|^2 \nu(\mathrm{d}\omega) - \frac{1}{16}\bigl(1 - d^{-1/p}\bigr)^2, \end{aligned}$$

where ν is the probability measure on Ω'_d under which $\omega_1, \dots, \omega_d$ are i.i.d. Rademacher random variables. Passing to sequences $\bar{\hat{f}}$, $\bar{\eta}_\omega$ in ℓ with components $\bar{\hat{f}}_j = \hat{f}_j - 1/2$, $\bar{\eta}_\omega(\mathbf{e}_j) = \eta_\omega(\mathbf{e}_j) - 1/2$, respectively, we may write

$$V_n(\mathcal{F}) \geq \inf_{\bar{\hat{f}}} \int_{\Omega'_d} \mathbf{E}_\omega^{(n)} \|\bar{\hat{f}} - \bar{\eta}_\omega\|^2 \nu(\mathrm{d}\omega) - \frac{1}{16}\bigl(1 - d^{-1/p}\bigr)^2.$$

For $j = 1, \dots, d$, denote by $\hat{f}[j]$ and r_j the components of $\bar{\hat{f}}$ and of $\bar{\eta}_\omega$, respectively. We will sometimes write $\hat{f}[j] = \hat{f}[j, D_n]$ to emphasize the dependence on the sample $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. Then, we can rewrite the above integral in the form

$$\int_{\Omega'_d} \mathbf{E}_\omega^{(n)} \|\bar{\hat{f}} - \bar{\eta}_\omega\|^2 \nu(\mathrm{d}\omega) = \mathbb{E}_{r_1, \dots, r_d} \mathbb{E}_{D_n} \left[\frac{1}{d} \sum_{j=1}^{d} \bigl(r_j - \hat{f}[j]\bigr)^2 \right],$$

where $\mathbb{E}_{r_1, \dots, r_d}$ and $\mathbb{E}_{D_n}$ denote the expectation over the joint distribution of $r_1, \dots, r_d$ and over the distribution of D_n given $r_1, \dots, r_d$, respectively.

Consider the random vector composed of indicators $\zeta = (I(X_1 = e_j), \dots, I(X_n = e_j))$. For any j and any fixed $r_1, \dots, r_d$,

$$\begin{aligned} \mathbb{E}_{D_n}\bigl[\bigl(r_j - \hat{f}[j, D_n]\bigr)^2\bigr] &= \mathbb{E}_\zeta \mathbb{E}_{D_n}\bigl[\bigl(r_j - \hat{f}[j, D_n]\bigr)^2 | \zeta\bigr] \\ &\geq \mathbf{P}(\zeta = 0) \mathbb{E}_{D_n}\bigl[\bigl(r_j - \hat{f}[j, D_n]\bigr)^2 | \zeta = 0\bigr] \\ &\geq \mathbf{P}(\zeta = 0)\bigl(r_j - \mathbb{E}_{D_n}\bigl[\hat{f}[j, D_n] | \zeta = 0\bigr]\bigr)^2, \end{aligned}$$

where we have used Jensen's inequality. We may write $\mathbb{E}_{D_n}[\hat{f}[j, D_n] | \zeta = 0]$ in the form

$$\mathbb{E}_{D_n}\bigl[\hat{f}[j, D_n] | \zeta = 0\bigr] = G\bigl(\{r_k \colon k \neq j\}\bigr),$$

where G is some measurable function. Indeed, under the condition $\zeta = 0$ the distribution of D_n coincides with that of $\{(X_i, Y_i)\colon X_i \neq e_j\}$, which is entirely defined by $\{r_k\colon k \neq j\}$. Thus,

$$\begin{aligned}
&\mathbb{E}_{r_j}\mathbb{E}_{D_n}\bigl[\bigl(r_j - \hat{f}[j, D_n]\bigr)^2\bigr] \\
&\quad \geq \mathbf{P}(\zeta = 0)\mathbb{E}_{r_j}\bigl[\bigl(r_j - G\bigl(\{r_k\colon k \neq j\}\bigr)\bigr)^2\bigr] \\
&\quad = \mathbf{P}(\zeta = 0)\biggl[\frac{1}{2}\biggl(\frac{1}{4} - G\bigl(\{r_k\colon k \neq j\}\bigr)\biggr)^2 + \frac{1}{2}\biggl(-\frac{1}{4} - G\bigl(\{r_k\colon k \neq j\}\bigr)\biggr)^2\biggr] \\
&\quad \geq \frac{1}{16}\mathbf{P}(\zeta = 0) = \frac{1}{16}\biggl(1 - \frac{1}{d}\biggr)^n,
\end{aligned}$$

where $\mathbb{E}_{r_j}$ denotes the expectation over the distribution of r_j and we have used that r_j takes values $1/4$ and $-1/4$ with probabilities $1/2$. This implies

$$\inf_{\hat{f}} \mathbb{E}_{r_1,\dots,r_d}\mathbb{E}_{D_n}\Biggl[\frac{1}{d}\sum_{j=1}^{d}\bigl(r_j - \hat{f}[j]\bigr)^2\Biggr] \geq \frac{1}{16}\biggl(1 - \frac{1}{d}\biggr)^n,$$

so that

$$V_n(\mathcal{F}) \geq \frac{1}{16}\biggl[\biggl(1 - \frac{1}{d}\biggr)^n - \bigl(1 - d^{-1/p}\bigr)^2\biggr].$$

Using that $1 - x \geq \exp(-3x/2)$ for $0 < x \leq 1/2$ we have $(1 - \frac{1}{d})^n \geq \exp(-3n/(2d)) \geq 1 - 3n/(2d)$ for $d \geq 2n$. Since $d = 2\lceil n^{p/(p-1)}\rceil$ we find

$$\begin{aligned}
V_n(\mathcal{F}) &\geq 1 - 3n/(2d) - \bigl(1 - d^{-1/p}\bigr)^2 = -3n/(2d) + 2d^{-1/p} - d^{-2/p} \\
&\geq -3n/(2d) - d^{-1/p} \geq d^{-1/p}/4 \geq cn^{-1/(p-1)}
\end{aligned}$$

for some absolute constant $c > 0$. □

Appendix

The following result is a modification of Theorem 6.1 in [7].

Theorem 14. *Let $\mathcal{G}$ be a class of non-negative functions bounded by b and admitting a localization radius $r^* = r^*(\mathcal{G})$. Then for all $n \geq 5$ and $t > 0$, with probability at least $1 - 4\mathrm{e}^{-t}$, for all $g \in \mathcal{G}$ we have*

$$Pg \leq 2P_n g + 106r^* + 48r_0, \tag{A.1}$$

$$P_n g \leq 2Pg + 72r^* + 24r_0, \tag{A.2}$$

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

Proof of Theorem 14. The fact that for $n \geq 5$ inequality (A.1) holds with probability at least $1 - \mathrm{e}^{-t}$ for all $g \in \mathcal{G}$ is proved in Theorem 6.1 in [7]. Moreover, it is shown in the proof of that theorem that, on the same event of probability at least $1 - \mathrm{e}^{-t}$ (denote this event by $\mathcal{B}$),

$$Pg \leq P_n g + \sqrt{Pg}\bigl(\sqrt{8r^*} + \sqrt{4r_0}\bigr) + 45r^* + 20r_0 \qquad \forall g \in \bigcup_{k=0}^{k_0} \mathcal{G}_k, \tag{A.3}$$

where $\mathcal{G}_k = \{g \in \mathcal{G}\colon\ \delta_{k+1} \leq Pg \leq \delta_k\}$, $\delta_k = b2^{-k}$ for $k \geq 0$, and $k_0 > 0$ be the largest integer such that $\delta_{k_0+1} \geq b/n$. A straightforward modification of the argument in [7] leading to (A.3) yields that, on the event $\mathcal{B}$,

$$|Pg - P_n g| \leq \sqrt{Pg}\bigl(\sqrt{8r^*} + \sqrt{4r_0}\bigr) + 45r^* + 20r_0 \qquad \forall g \in \bigcup_{k=0}^{k_0} \mathcal{G}_k, \tag{A.4}$$

so that

$$\begin{aligned} P_n g &\leq Pg + \sqrt{Pg}\bigl(\sqrt{8r^*} + \sqrt{4r_0}\bigr) + 45r^* + 20r_0 \\ &\leq 2Pg + 53r^* + 24r_0 \qquad \forall g \in \bigcup_{k=0}^{k_0} \mathcal{G}_k, \end{aligned} \tag{A.5}$$

proving (A.2) for $g \in \bigcup_{k=0}^{k_0} \mathcal{G}_k$ with probability at least $1 - \mathrm{e}^{-t}$.

Now, consider $g \in \mathcal{G}_* = \mathcal{G} \setminus \bigcup_{k=0}^{k_0} \mathcal{G}_k$. First, for any $g \in \mathcal{G}_*$, $Pg \leq \delta_k \leq \delta_{k_0} \leq 4b/n$. Hence $\mathcal{G}_* \subseteq \mathcal{G}' = \{g \in \mathcal{G}\colon\ Pg < 4b/n\}$. By Lemma 6.1 in [7], with probability at least $1 - 3\mathrm{e}^{-t}$,

$$\begin{aligned} |P_n g - Pg| &\leq 6\hat{\mathfrak{R}}_n\bigl(\mathcal{G}', S\bigr) + \frac{b}{n}(\sqrt{2t} + 6t) \\ &\leq 6\hat{\mathfrak{R}}_n\bigl(\mathcal{G}', S\bigr) + \frac{b(7t+1)}{n} \qquad \forall g \in \mathcal{G}'. \end{aligned} \tag{A.6}$$

Denote the event where (A.6) holds by $\mathcal{B}'$, and define

$$U' = 6\hat{\mathfrak{R}}_n\bigl(\mathcal{G}', S\bigr) + Pg + \frac{b(7t+1)}{n}.$$

On the event $\mathcal{B}'$ we have $P_n g \leq U'$ for any $g \in \mathcal{G}'$, so that

$$\hat{\mathfrak{R}}_n\bigl(\mathcal{G}', S\bigr) \leq \hat{\mathfrak{R}}_n\bigl(\bigl\{g \in \mathcal{G}\colon\ P_n g \leq U'\bigr\}, S\bigr) \leq \phi_n\bigl(U'\bigr),$$

where $\phi_n(\cdot) = \phi_n(\cdot, \mathcal{G})$ is an upper function for $\mathcal{G}$ satisfying the sub-root property. In view of this property,

$$U' \leq 6\phi_n\bigl(U'\bigr) + Pg + \frac{b(7t+1)}{n} \leq 6\sqrt{U'}\sqrt{r^*} + Pg + \frac{b(7t+1)}{n}.$$

Solving for $\sqrt{U'}$ we get

$$\sqrt{U'} \le 6\sqrt{r^*} + \sqrt{Pg + \frac{b(7t+1)}{n}}$$

and thus, on the event $\mathcal{B}'$,

$$P_n g \le U' \le 2Pg + 72r^* + \frac{2b(7t+1)}{n} \le 2Pg + 72r^* + 14r_0 \qquad \forall g \in \mathcal{G}', \tag{A.7}$$

where the last inequality is due to the fact that $7t + 1 \le 7(t + 6\log\log n)$ for all $n \ge 3$. Combining (A.5) and (A.7), we then obtain (A.2) holds for all $g \in \mathcal{G}$ on the event $\mathcal{B} \cap \mathcal{B}'$ of probability at least $1 - 4\mathrm{e}^{-t}$. □

Proof of Lemma 8. *Proof of* (i). We apply Lemma 2.2 in [40] for the loss function defined by $\varphi(t, y) = t^2, \forall t, y \in \mathbb{R}$. The second derivative of this function with respect to the first argument is $H = 2$, that is, the function is 2-smooth in the terminology of [40]. Consider the class of differences $\mathcal{H} = \{f - g\colon f, g \in \mathcal{F}\}$. Then Lemma 2.2 in [40] provides the following bound for the Rademacher complexity of the set $\mathcal{L} = \{(x, y) \mapsto \varphi(h(x), y)\colon h \in \mathcal{H}, n^{-1}\sum_{(x,y)\in S} h^2(x) \le r\}$:

$$\hat{\mathfrak{R}}_n(\mathcal{L}, S) \le 21\sqrt{12r}\log^{3/2}(64n)\mathfrak{R}_n(\mathcal{H}).$$

On the other hand, $\mathcal{L} = \mathcal{G}[r, S]$, and $\mathfrak{R}_n(\mathcal{H}) \le 2\mathfrak{R}_n(\mathcal{F})$, so that

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

Now define the function $\phi_n(r)$ as the right-hand side of this inequality. This immediately yields a localization radius

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

and (30) follows.

Proof of (ii). Let $(f - g)^2$ and $(\bar{f} - \bar{g})^2$ be two elements of $\mathcal{G}$, where $f, g, \bar{f}, \bar{g} \in \mathcal{F}$. Since all these functions take values in [0, 1] we get that, for any $x \in \mathcal{X}$,

$$\bigl(\bigl(f(x) - g(x)\bigr)^2 - \bigl(\bar{f}(x) - \bar{g}(x)\bigr)^2\bigr)^2 \le 8\bigl(\bigl(f(x) - \bar{f}(x)\bigr)^2 + \bigl(g(x) - \bar{g}(x)\bigr)^2\bigr).$$

Thus, if $d_S(f, \bar{f}) \le \varepsilon$ and $d_S(g, \bar{g}) \le \varepsilon$ for some $\varepsilon > 0$, then $d_S((f - g)^2, (\bar{f} - \bar{g})^2) \le 4\varepsilon$. This implies the relation between the empirical entropies: $\mathcal{N}_2(\mathcal{G}, \rho, S) \le \mathcal{N}_2(\mathcal{F}, \rho/4, S)$ for all $\rho > 0$. Using it together with the bound $\mathcal{N}_2(\mathcal{F}, \rho/4, S) \le \max\{1, (4A/\rho)^v\}$ and applying Lemma 10 we obtain

$$\begin{aligned}\hat{\mathfrak{R}}_n\bigl(\mathcal{G}[r, S], S\bigr) &\le \frac{12}{\sqrt{n}}\int_0^{\sqrt{r}} \sqrt{\log \mathcal{N}_2(\mathcal{G}, \rho, S)}\,\mathrm{d}\rho \\ &\le \frac{12}{\sqrt{n}}\int_0^{\sqrt{r}\wedge 1/(4A)} \sqrt{v\log(4A/\rho)}\,\mathrm{d}\rho \end{aligned}\tag{A.8}$$

$$\begin{aligned}&\le \frac{48A\sqrt{v}}{\sqrt{n}} \int_0^{\sqrt{r}/(4A)\wedge 1} \sqrt{\log(1/t)}\, \mathrm{d}t \\ &\le 24\sqrt{\frac{vr}{n}} \bigl(\log(4eA/\sqrt{r}) \vee 1\bigr)^{1/2},\end{aligned}$$

where we have used that, integrating by parts,

$$\begin{aligned}\int_0^b \sqrt{\log(e/t)}\, \mathrm{d}t &= b\sqrt{\log(e/b)} + (b/2)\bigl(\log(e/b)\bigr)^{-1/2} \\ &\le 2b\sqrt{\log(e/b)} \qquad \forall 0 < b \le 1.\end{aligned} \tag{A.9}$$

In view of (A.8), we can take $\phi_n(r) = 24\sqrt{\frac{vr}{n}}(\log(4eA/\sqrt{r}) \vee 1)^{1/2}$ as an upper function in (7). Now, we are looking for r^*, which is an upper bound on the solution of the equation $\phi_n(r) = r$. Since the function $u \mapsto (a/u)(\log(b/u) \vee 1)^{1/2}$, for $a, b > 0$, is decreasing when $u > 0$ one can check that $u^* = a(\log(b/a) \vee 1)^{1/2}$ as an upper bound on the solution of $(a/u)(\log(b/u) \vee 1)^{1/2} = 1$ whenever $b \ge ea > 0$. That is, for $n \ge Cv$ with $C > 0$ large enough depending only on A, we can take

$$r^* = \left[24\sqrt{\frac{v}{n}}\left(\log\left(\frac{eA}{6}\sqrt{\frac{n}{v}}\right) \vee 1\right)^{1/2}\right]^2 \le C\frac{v}{n}\log\left(\frac{en}{v}\right) \tag{A.10}$$

for some constant $C > 0$ depending only on A.

Proof of (iii). For a finite class $\mathcal{F}$, the covering numbers satisfy $\mathcal{N}_2(\mathcal{F}, \varepsilon, S) \le |\mathcal{F}|$ for all $\varepsilon > 0$ and, along the lines of (A.8),

$$\hat{\mathfrak{R}}_n\bigl(\mathcal{G}[r, S], S\bigr) \le \frac{12}{\sqrt{n}} \int_0^{\sqrt{r}} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \rho/4, S)}\, \mathrm{d}\rho \le \frac{12\sqrt{r \log|\mathcal{F}|}}{\sqrt{n}} \triangleq \phi_n(r),$$

so that we can take $r^* = 144(\log|\mathcal{F}|)/n$. □

Proof of Lemma 11. Assume that there exists $f^* \in \mathcal{F}$ such that $L(f^*) = \min_{f\in\mathcal{F}} L(f)$ (if this is not the case, an easy modification of the proof is possible by considering an approximate minimizer). We apply Theorem 3.3 in [4] to the class of functions $\mathcal{G} = \ell \circ \mathcal{F} - \ell \circ f^*$. Observe that, for any $f \in \mathcal{F}$, the variance of the random variable $\ell \circ f(X, Y) - \ell \circ f^*(X, Y)$ satisfies

$$\mathrm{Var}\bigl(\ell \circ f - \ell \circ f^*\bigr) \le \mathbb{E}\bigl[\bigl(\bigl(f(X) - Y\bigr)^2 - \bigl(f^*(X) - Y\bigr)^2\bigr)^2\bigr] \le 2\bigl(L(f) - L\bigl(f^*\bigr)\bigr)$$

and thus the assumption of Theorem 3.3 in [4] holds with $B = 2$. Applying that theorem with $K = 2$ we get that, for any $t > 0$, with probability at least $1 - \mathrm{e}^{-t}$, for any $g \in \mathcal{G}$,

$$Pg \le 2P_n g + c_1'' \bar{r}^* + \frac{t(22 + c_2'')}{n},$$

where $c_1'' = 704$, $c_2'' = 104$, and $\bar{r}^*$ is the solution of fixed point equation $\psi(r) = r$, for a function ψ satisfying the sub-root property and the inequality $\psi(r) \ge 2\mathbb{E}\hat{\mathfrak{R}}_n(\mathcal{G} \cap \{2Pg \le r\}, S')$. Choose now a constant function $\psi(r) \equiv 2\mathbb{E}\hat{\mathfrak{R}}_n(\mathcal{G}, S')$, which trivially satisfies the sub-root property and has the fixed point $\bar{r}^* = 2\mathbb{E}\hat{\mathfrak{R}}_n(\mathcal{G}, S')$. Since $\mathbb{E}\hat{\mathfrak{R}}_n(\mathcal{G}, S') = \mathbb{E}\hat{\mathfrak{R}}_n(\ell \circ \mathcal{F}, S')$, and $P_n g \le 0$ for $g = \ell \circ \hat{f}^{\mathrm{emp}} - \ell \circ f^*$, we obtain that, with probability at least $1 - \mathrm{e}^{-t}$,

$$L(\hat{f}^{\mathrm{emp}}) - L(f^*) \le 2c_1''\mathbb{E}\hat{\mathfrak{R}}_n(\ell \circ \mathcal{F}, S') + \frac{t(22 + c_2'')}{n},$$

where we have used that $L(f) = P(\ell \circ f)$. Next, by Lemma A.4 in [4], with probability at least $1 - \mathrm{e}^{-t}$,

$$\mathbb{E}\hat{\mathfrak{R}}_n(\ell \circ \mathcal{F}, S') \le 2\hat{\mathfrak{R}}_n(\ell \circ \mathcal{F}, S') + \frac{t}{n}.$$

Combining the results of the last two displays we find that, with probability at least $1 - 2\mathrm{e}^{-t}$,

$$L(\hat{f}^{\mathrm{emp}}) - L(f^*) \le 4c_1''\hat{\mathfrak{R}}_n(\ell \circ \mathcal{F}, S') + \frac{t(22 + c_2'' + 2c_1'')}{n}. \qquad \square$$

Acknowledgements

We gratefully acknowledge the support of NSF under Grants CAREER DMS-0954737 and CCF-1116928. The work of the third author was supported by GENES, and by the French National Research Agency (ANR) under the Grants Idex ANR-11-IDEX-0003-02, Labex ECODEC (ANR-11-LABEX-0047), and IPANEMA (ANR-13-BSH1-0004-02).

References

[1] Aizerman, M.A., Braverman, E.M. and Rozonoer, L.I. (1970). *The Method of Potential Functions in the Theory of Machine Learning*. Moscow: Nauka. (in Russian).

[2] Audibert, J.Y. (2007). Progressive mixture rules are deviation suboptimal. *Adv. Neural Inf. Process. Syst.* **20**. 41–48.

[3] Bartlett, P.L. (2006). *CS* 281*B Statistical Learning Theory Course Notes*, *U.C.* Berkeley: Berkeley, CA.

[4] Bartlett, P.L., Bousquet, O. and Mendelson, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554

[5] Bartlett, P.L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. MR1984026

[6] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129

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

[8] Bousquet, O., Koltchinskii, V. and Panchenko, D. (2002). Some local measures of complexity of convex hulls and generalization bounds. In *Computational Learning Theory* (*Sydney*, 2002). *Lecture Notes in Computer Science* **2375** 59–73. Berlin: Springer. MR2040405

[9] Buescher, K.L. and Kumar, P.R. (1996). Learning by canonical smooth estimation. I. Simultaneous estimation. *IEEE Trans. Automat. Control* **41** 545–556. MR1385325

[10] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. *Lecture Notes in Math*. **1851**. Berlin: Springer. MR2163920

[11] Cesa-Bianchi, N. and Lugosi, G. (2001). Worst-case bounds for the logarithmic loss of predictors. *Mach*. *Learn*. **43** 247–264.

[12] Dai, D., Rigollet, P. and Zhang, T. (2012). Deviation optimal learning using greedy Q-aggregation. *Ann*. *Statist*. **40** 1878–1905. MR3015047

[13] Dalalyan, A.S. and Tsybakov, A.B. (2012). Mirror averaging with sparsity priors. *Bernoulli* **18** 914–944. MR2948907

[14] Devroye, L. (1987). *A Course in Density Estimation*. *Progress in Probability and Statistics* **14**. Boston, MA: Birkhäuser. MR0891874

[15] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. New York: Springer. MR1383093

[16] Dudley, R.M. (1978). Central limit theorems for empirical measures. *Ann*. *Probab*. **6** 899–929 (1979). MR0512411

[17] Dudley, R.M. (1987). Universal Donsker classes and metric entropy. *Ann*. *Probab*. **15** 1306–1326. MR0905333

[18] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics* **63**. Cambridge: Cambridge Univ. Press. MR1720712

[19] Dudley, R.M., Giné, E. and Zinn, J. (1991). Uniform and universal Glivenko–Cantelli classes. *J*. *Theoret*. *Probab*. **4** 485–510. MR1115159

[20] Ibragimov, I.A. and Has'minskiĭ, R.Z. (1980). An estimate of the density of a distribution. *Zap*. *Nauchn*. *Sem*. *Leningrad*. *Otdel*. *Mat*. *Inst*. *Steklov*. (*LOMI*) **98** 61–85. MR0591862

[21] Juditsky, A., Rigollet, P. and Tsybakov, A.B. (2008). Learning by mirror averaging. *Ann*. *Statist*. **36** 2183–2206. MR2458184

[22] Kolmogorov, A.N. and Tihomirov, V.M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspehi Mat*. *Nauk* **14** 3–86. MR0112032

[23] Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans*. *Inform*. *Theory* **47** 1902–1914. MR1842526

[24] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann*. *Statist*. **34** 2593–2656. MR2329442

[25] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math*. **2033**. Heidelberg: Springer. MR2829871

[26] Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability*, *II* (*Seattle*, *WA*, 1999). *Progress in Probability* **47** 443–457. Boston, MA: Birkhäuser. MR1857339

[27] LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann*. *Statist*. **1** 38–53. MR0334381

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

[29] Lecué, G. (2013). Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli* **19** 2153–2166. MR3160549

[30] Lecué, G. and Mendelson, S. (2009). Aggregation via empirical risk minimization. *Probab*. *Theory Related Fields* **145** 591–613. MR2529440

[31] Lecué, G. and Rigollet, P. (2014). Optimal learning with Q-aggregation. *Ann*. *Statist*. **42** 211–224. MR3178462

[32] Lee, W.S., Bartlett, P.L. and Williamson, R.C. (1998). The importance of convexity in learning with squared loss. *IEEE Trans*. *Inform*. *Theory* **44** 1974–1980. MR1664079

[33] Lounici, K. (2007). Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.* **16** 246–259. MR2356820

[34] Lugosi, G. and Nobel, A.B. (1999). Adaptive model selection using empirical complexities. *Ann. Statist.* **27** 1830–1864. MR1765619

[35] Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math.* **1738** 85–277. Berlin: Springer. MR1775640

[36] Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer. MR0762984

[37] Raginsky, M. and Rakhlin, A. (2011). Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems* 24 1026–1034.

[38] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. MR2816337

[39] Rigollet, P. and Tsybakov, A.B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575. MR3025134

[40] Srebro, N., Sridharan, K. and Tewari, A. (2010). Smoothness, low-noise and fast rates. In *NIPS*. Available at arXiv:1009.3896.

[41] Tsybakov, A.B. (2003). Optimal rates of aggregation. In *Proceedings of COLT-*2003 303–313. Springer: Berlin.

[42] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer. MR2724359

[43] van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343

[44] Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. New York: Springer. MR0672244

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

[46] Vapnik, V.N. and Chervonenkis, A.Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

[47] Vapnik, V.N. and Chervonenkis, A.Ya. (1974). *Theory of Pattern Recognition*. Moscow: Nauka.

[48] Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** 25–47. MR2044592

[49] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500

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

Received March 2014 and revised July 2014