Proceedings of Machine Learning Research vol 65:1–23, 2017

Optimal learning via local entropies and sample compression

Nikita Zhivotovskiy NIKITA.ZHIVOTOVSKIY@PHYSTECH.EDU
Skolkovo Institute of Science and Technology, Institute for Information Transmission Problems, Moscow.

Abstract

Under margin assumptions, we prove several risk bounds, represented via the distribution dependent local entropies of the classes or the sizes of specific sample compression schemes. In some cases, our guarantees are optimal up to constant factors for families of classes. We discuss limitations of our approach and give several applications. In particular, we provide a new tight PAC bound for the hard-margin SVM, an extended analysis of certain empirical risk minimizers under log-concave distributions, a new variant of an online to batch conversion, and distribution dependent localized bounds in the aggregation framework. As a part of our results, we give a new upper bound for the uniform deviations under Bernstein assumptions, which may be of independent interest. The proofs for the sample compression schemes are based on the moment method combined with the analysis of voting algorithms.

Keywords: Empirical risk minimization, sample compression, local entropy, stability, bracketing conditions, VC classes, hard margin SVM, online to batch conversion.

1. Introduction

One of the most important concepts in statistical learning is the notion of complexity of classes. The complexity defines the statistical properties of learning procedures and depends on not only the structure of a class $\mathcal{F}$, but also on the framework and its intrinsic properties such as the noise of the problem. The complexity also depends on the procedure of interest. It means that given a class $\mathcal{F}$ different statistical procedures appear to have different learning rates. Thus, the questions that are usually asked: what is the complexity of the class, which procedures achieve these rates and is it possible for any algorithm to have better risk bounds? If some statistical performance is known to be (almost) optimal it is interesting whether for a computationally efficient procedure we may achieve this optimal performance.

There are a lot of complexity measures and related risk bounds which occur in statistics/ statistical learning. To name just a few, we can mention: the notion of VC dimension and Growth function [38, 39], that of Fat-shattering dimension [3], empirical and distribution-dependent covering (or packing) numbers [11, 36, 29, 33, 42], the more recent notions of local/global and offset Rademacher complexities [6, 7, 26], Alexander's capacity and the Disagreement coefficient [15, 16], empirical and distribution-dependent local entropies [10, 31, 43], and finally the size of sample compression sets [13].

The above list is far from exhaustive and other complexity measures are occasionally used in the literature. A number of contributions have investigated the relevance of these several complexity measures as well as their connections. In particular, some have been shown to provide a somehow full understanding of some statistical problems.

This paper discusses *two* complexity measures (the interest of which will be argued) together with the learning algorithms and some associated generalization bounds. The first one is a *distribu-*

tion dependent local entropy that is a complexity measure which controls the number of functions that are needed to cover an $L_r(P)$ ball of radius 2ε intersected with $\mathcal{G}$ by $L_r(P)$ balls of radius ε, where $\mathcal{G}$ is a so-called *loss class* associated with $\mathcal{F}$. The second complexity measure is a *size of a sample compression set*, which is the minimal size of a subsample of a given sample, such that the so-called *reconstruction function*, given only this subsample is able to recover the remaining sample. A standard example is that of the optimal hyperplane (hard margin SVM) classifier [39], where the role of the compressed sample is played by the so-called support vectors [39]. However, it will be clear from our discussions why we also require a special property, namely *stability* for compression algorithms.

These two simple complexity measures, proposed above are of very different nature. The first one is connected to the classic condition of learning, namely, the uniform convergence of frequencies to their means [38]. The learning algorithm corresponding to the second complexity measure will be based on two other conditions which are directly connected with sufficient conditions for learning, namely stability [8] and sample compression [13]. Our model will be a standard i.i.d. model under general (β, B)-Bernstein class conditions. However, in the case of sample compression, we will be in the specific case, the so-called realizable classification. For this case, we have in particular a binary loss and $(1, 1)$-Bernstein class condition. We will motivate our choice of the local entropy in a question/answer format. The motivation about the sample compression will be presented in Section 4.

Why using bounds in terms of the local entropy (doubling dimension)?

The local entropy is known to appear in general *lower bounds* in the related statistical models, see [20, 42, 29, 24, 31, 43] and reference therein. By general we mean that these lower bounds hold not for the one specific "worst case" class $\mathcal{F}$, but for a *whole family of classes*, having the same local entropy. This is similar to standard (tight in some regimes) upper and lower bounds in classification in terms of VC dimension as they are valid *for all classes* with the same VC dimension [11, 12, 39] regardless of any other characteristics of these classes.

However, the existing upper bounds, based on certain localized empirical processes are not easily matched by the lower bounds. This means that relaxation of these bounds is known to be matched by the lower bounds only for some specific classes $\mathcal{F}$ [29, 33]. Even more, as noted by Mendelson [31] some bounds represented in terms of localized processes are not tight. The problem of finding general matching upper and lower bounds in regression and classification frameworks has recently been addressed by Lecué and Mendelson [24], in the above-mentioned work [31], by Hanneke [17] and Zhivotovskiy and Hanneke [43]. This work further contributes to this direction.

Why one is interested in complexity measures in a standard i.i.d. statistical learning framework?

Complexity measures are related to certain properties of functional classes, not only the underlying statistical framework. Speaking informally, assume that in the noise free (realizable case) classification case for a class $\mathcal{F}_1$ we have a learning rate of order $\frac{1}{n}$ for any empirical minimizer and for a class $\mathcal{F}_2$ the learning rate of any empirical risk is of order $\frac{\log(n)}{n}$. This seemingly modest difference, however, can not be captured by standard VC bounds and is explained by differences between complexities of $\mathcal{F}_1$ and $\mathcal{F}_2$ [43]. Although, it follows from the recent results of Hanneke and Yang [16] that when applied to the closely related *active learning* framework, the gap between

rates is enormous: $\mathcal{F}_1$ can be learnt with convergence rates exponentially fast in n, while for $\mathcal{F}_2$ no significant improvements compared to i.i.d. model are possible by any active algorithm.

Why using distribution dependent complexity measures?

The initial idea of Vapnik and Chervonenkis [38] was to build a distribution-free theory of statistical learning, that is $\mathcal{X} \times \mathcal{Y}$ is equipped with some unknown distribution, and the only assumption was that data is i.i.d. Later it became clear that some restrictions should be made to make these bounds more realistic. However, popular conditions like Massart or Tsybakov noise conditions are restricting the conditional distribution of $Y|X$ [29, 36], but not the distribution of X. Some recent results in this direction show [5, 17] that sometimes, given a nice distribution of X (together with $Y|X$) one is able to obtain *significant improvements* in different statistical frameworks. Our approach, contrary to the standard statistical learning analysis, will not use *symmetrization techniques*. Symmetrization is usually used in the literature to obtain distribution free upper bounds based only on metric or combinatorial properties of the class. The fact that we do not use symmetrization techniques will be one of the reasons why we are able to simply capture certain improvements by introducing P_X-dependent complexity measures.

2. Notation

We define the *space of predictors* $\mathcal{X}$ and the *space* of *response variables* $\mathcal{Y} \subseteq \mathbb{R}$. We assume that the set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is equipped with some σ-algebra of events and a probability measure P on measurable subsets is defined. We also assume that we are given a set of functions $\mathcal{F}$; these are measurable functions with respect to the introduced σ-algebra, mapping $\mathcal{X}$ to $\mathcal{Y}$. Symbol $\wedge$ denotes minimum of two real numbers, $\vee$ denotes maximum of two real numbers and $\mathbb{1}[A]$ denotes an indicator of the event A. For any subset $B \subseteq \mathcal{F}$ define the *region of disagreement* as $\mathrm{DIS}(B) = \{x \in \mathcal{X} | \ \exists f, g \in B \text{ s. t. } f(x) \neq g(x)\}$. We will also consider abstract real-valued functional classes, which will usually be denoted by $\mathcal{G}$. By $\log(x)$ we mean truncated logarithm: $\ln(\max(x, e))$. The notation $f(n) \lesssim g(n)$ or $g(n) \gtrsim f(n)$ will mean that for some universal constant $c > 0$ it holds that $f(n) \leq cg(n)$ for all $n \in \mathbb{N}$. Similarly, we introduce $f(n) \simeq g(n)$ to be equivalent to $g(n) \lesssim f(n) \lesssim g(n)$.

A *learner* observes $((X_1, Y_1), \ldots, (X_n, Y_n))$, an i.i.d. training sample from an unknown distribution P. By P_n we denote expectation with respect to the empirical measure (empirical mean) induced by these samples. Symbols P and $\mathbb{E}$ denote expectations with respect to the true measure. We introduce the loss function $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ that will measure our losses by predicting $\hat{Y}$ instead of Y. We further assume that for all $y \in \mathbb{R}$ it holds that $\ell(y, y) = 0$. The risk of f is its expected loss, denoted $R(f) = \mathbb{E}\ell(f(X), Y)$. The function $f^* \in \mathcal{F}$ will be the minimizer of $R(f)$. *Empirical risk minimization* (ERM) refers to any learning algorithm with the following property: given a training sample, it outputs a classifier $\hat{f}$ that minimizes $R_n(f) = P_n\ell(f(X), Y)$ among all $f \in \mathcal{F}$. Depending on context we will usually refer to $\hat{f}$ as an empirical risk minimizer and use the same abbreviation. For a class $\mathcal{G} \subseteq L_p(P)$ and $f, g \in \mathcal{G}$ we denote $\|f - g\|_{L_p(P)} = \left(\int\limits_{\mathcal{Z}} |f(z) - g(z)|^p dP(z)\right)^{\frac{1}{p}}$ for $p > 0$. In particular, if $p = 1$, then $\|f - g\|_{L_1(P)} = \mathbb{E}|f - g|$, $\|f - g\|_{L_1(P_n)} = \frac{1}{n}\sum\limits_{i=1}^{n} |f(Z_i) - g(Z_i)|$ and if $p = \infty$ we have the standard $\| \ \|_{L_\infty} = \| \ \|_\infty$ norm.

In the special case when $\mathcal{Y} = \{1, -1\}$ we will consider the binary loss, that is $\ell(Y, \hat{Y}) = \mathbb{1}[Y \neq \hat{Y}]$. In this case we say that a set $\{x_1, \dots, x_k\} \in \mathcal{X}^k$ is shattered by $\mathcal{F}$ if there are 2^k distinct classifications of $\{x_1, \dots, x_k\}$ realized by classifiers in $\mathcal{F}$. The *VC dimension* of $\mathcal{F}$ is the largest integer d such that there exists a set $\{x_1, \dots, x_d\}$ shattered by $\mathcal{F}$ [38]. By the *realizable case classification* we will mean the learning model with the binary loss such that for some $f^* \in \mathcal{F}$ it holds $Y = f^*(X)$.

3. Distribution dependent upper bounds

The aim of this section is to present several simple upper bounds on the performance of learning procedures under $L_1(P)$ entropy conditions. As discussed before, the important part of our analysis is that our technique avoids the symmetrization step. At first we provide several notations. Consider the *excess loss class* $\mathcal{L}_{\mathcal{Y}} = \{(x, y) \to \ell(f(x), y) - \ell(f^*(x), y) \text{ for } f \in \mathcal{F}\}$, and the *loss class* $\mathcal{G}_{\mathcal{Y}} = \{(x, y) \to \ell(f(x), y) \text{ for } f \in \mathcal{F}\}$.

Definition 1 (Bernstein condition [7, 22]) *Class $\mathcal{G}$ satisfies the (β, B)-Bernstein condition if for all $g \in \mathcal{G}$*

$$Pg^2 \leq B(Pg)^{\beta}$$

for some $\beta \in [0, 1]$ and $B > 1$.

This condition naturally generalizes other well known conditions, such as Massart [29] and Tsybakov [36] noise conditions and appears naturally in certain misspecified models under the convexity of $\mathcal{F}$ [7, 22]. Moreover sometimes it holds in the misspecified case even without the convexity assumption (see Mendelson [30]).

Given a class $\mathcal{G} \subset L_1(P)$ we define the covering number $\mathcal{N}(\mathcal{G}, \varepsilon)$ of $\mathcal{G}$ as the minimal number of functions $g_1, \dots, g_N \in \mathcal{G}$ (we will consider only proper coverings) such that for all $g \in \mathcal{G}$ there exists $j \in \{1, \dots, N\}$ such that $\|g - g_j\|_{L_1(P)} \leq \varepsilon$. Let $\mathcal{B}_{L_1}(g, \varepsilon)$ be a ball in $L_1(P)$ of radius ε with the center g. Finally, we introduce

$$\mathcal{D}^{\text{loc}}(\mathcal{G}, \varepsilon) = \sup_{\gamma \geq \varepsilon} \sup_{g \in \mathcal{G}} \log \left(\mathcal{N}(\mathcal{G} \cap \mathcal{B}_{L_1}(g, 2\gamma), \gamma) \right), \tag{1}$$

which will be referred to as a *local entropy*. When the class of interest is clear, we will avoid writing it as an argument. Another quantity of interest will be the local entropy with bracketing. We need several extra definitions. Let $f_1, f_2 \in L_1(P)$ and $f_1 \leq f_2$ with probability one. If $\|f_1 - f_2\|_{L_1(P)} \leq \varepsilon$, then *$\varepsilon$-bracket* consist of all functions, such that $f \in L_1(P)$ and $f_1 \leq f \leq f_2$. Given a class $\mathcal{G} \subset L_1(P)$ we define a bracketing entropy $\mathcal{N}_{[\,]}(\mathcal{G}, \varepsilon)$ as a minimal number of ε-brackets $B_1, \dots, B_N$, such that $\mathcal{G} \subseteq \cup_{i=1}^N B_i$. In the same manner we define the *local entropy with bracketing*:

$$\mathcal{D}^{\text{loc}}_{[\,]}(\mathcal{G}, \varepsilon) = \sup_{\gamma \geq \varepsilon} \sup_{g \in \mathcal{G}} \log \left(\mathcal{N}_{[\,]}(\mathcal{G} \cap \mathcal{B}_{L_1}(g, 2\gamma), \gamma) \right), \tag{2}$$

The local entropy is known in statistical learning theory. It appeared in the analysis of linear half-spaces [10], in the lower bounds for a many statistical problems and more recently in the analysis of convex regression [31]. Interestingly that the upper bounds are provided not for the empirical risk minimizer, but for a specially designed algorithm, namely ERM over the ε-net of the functional class, which is less preferable in terms of computational efficiency. We will discuss this algorithm

later. Before going to our results we introduce two *fixed points* which will be essentially our complexity measures

$$\gamma(\mathcal{G},k,\beta)=\{\inf\varepsilon>0:k\mathcal{D}^{\text{loc}}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}\right)\leq\varepsilon\} \tag{3}$$

and

$$\gamma_{[\,]}(\mathcal{G},k,\beta)=\{\inf\varepsilon>0:k\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}\right)\leq\varepsilon\} \tag{4}$$

The next simple theorem gives an analysis of ERM under the local entropy with bracketing under general margin condition.

In what following we will consider the loss, that is bounded by 1. Our results, however, can be extended to the unbounded losses, since the only concentration tool that will be used is a Bernstein inequality, and versions of it for the unbounded random variables (represented via the Orlicz norms or related moment conditions) will be sufficient for our purposes [1, 9, 25]. However, the extension to heavy-tailed scenarios as in [32] is not so straightforward.

Theorem 2 *Assume that the loss function is bounded by* 1 *and the excess loss class* $\mathcal{L}_{\mathcal{Y}}$ *is a* (β,B)*-Bernstein class. Then with probability at least* $1-\delta$ *over the learning sample for any ERM* $\hat{f}$

$$R(\hat{f})-R(f^*)\lesssim\left(\gamma_{[\,]}\left(\mathcal{G}_{\mathcal{Y}},\frac{B}{n},\beta\right)+\frac{B\log(\frac{1}{\delta})}{n}\right)^{\frac{1}{2-\beta}}.$$

Notice that our bound does not involve convexity or star-shapedness assumptions comparing to the related techniques in the literature [6, 7]. Taking star-hulls may be harmful if one wants to prove the optimal rates for certain small classes (see discussions in [43] and related Lemma 11 in [26]). Moreover, since we do not use the symmetrization step our bound is fully distribution dependent, which will be also important in our examples. The form of the bound, represented via the local entropy, will later be useful as our aim is to match it by the lower bounds, usually represented via a similar quantity. However, there are some important limitations. Namely, the bracketing entropy condition, which is not simple to check in general. Moreover, at least sometimes this bound will give weaker results, compared to *chaining-based* bounds. These issues will be discussed and partially avoided by complementing results below.

Before we start to prove this result we need several lemmas. Versions of the next lemma are known in the literature for the local entropy without bracketing (see Lemma 2 in [10] or Lemma 2.2 in [31]). We simply adapt these arguments to our case. Denote for simplicity $\mathcal{N}(\delta,\varepsilon)=\sup_{g\in\mathcal{G}}\mathcal{N}(\mathcal{G}\cap\mathcal{B}_{L_1}(g,\delta),\varepsilon)$ and $\mathcal{N}_{[\,]}(\delta,\varepsilon)=\sup_{g\in\mathcal{G}}\mathcal{N}_{[\,]}(\mathcal{G}\cap\mathcal{B}_{L_1}(g,\delta),\varepsilon)$.

Lemma 3 *It holds for* $\delta>2\varepsilon$

$$\log(\mathcal{N}_{[\,]}(\delta,\varepsilon))\leq\left(\log_4\left(\frac{16\delta}{\varepsilon}\right)\right)\mathcal{D}^{loc}_{[\,]}(\mathcal{G},\varepsilon)$$

and

$$\log(\mathcal{N}(\delta,\varepsilon))\leq\left(\log_2\left(\frac{4\delta}{\varepsilon}\right)\right)\mathcal{D}^{loc}(\mathcal{G},\varepsilon).$$

The next lemma is a bound on a *shifted-type empirical process* $\sup_{g\in\mathcal{G}}(Pg-(1+c)P_ng)$ in terms of the local entropy with bracketing. These processes are used in the literature in various contexts. For example, there were previously introduced to obtain non-exact oracle inequalities (see Lecué and Mitchel [25] or Wegkamp [41]) or to obtain sharp bounds in the binary classification but using symmetrization techniqes (see Zhivotovskiy and Hanneke [43]). However, contrary to previous results, our bound is represented via a localized complexity measure for an infinite class under general margin conditions, does not involve symmetrization and holds with high probability.

Lemma 4 *Let $\mathcal{G}\subset L_1(P)$ be a class of functions, such that $0\in\mathcal{G}$, $Pg\geq 0$ and $\|g\|_\infty\leq 1$ for all $g\in\mathcal{G}$, the Bernstein condition $Pg^2\leq B(Pg)^\beta$ holds for some constant $B\geq 1$ and $\beta\in[0,1]$. Then for any $c\geq 1$ with probability at least $1-\delta$ it holds*

$$\sup_{g\in\mathcal{G}}(Pg-(1+c)P_ng)\lesssim\left(\gamma_{[\,]}\left(\mathcal{G},\frac{c'B}{n},\beta\right)+\frac{c'B\log(\frac{1}{\delta})}{n}\right)^{\frac{1}{2-\beta}},$$

where $c'=64(1+c)^2$.

Proof Fix $\varepsilon>0$. Given a class $\mathcal{G}$ and a distribution P we construct a $\varepsilon^{\frac{1}{2-\beta}}$-covering with bracketing of the whole set (with respect to $L_1(P)$ metric). Let p denote the projection on the smallest function in the bracket, $p[\mathcal{G}]$ be a set of projections, that is the set of functions $\{p[g]|g\in\mathcal{G}\}$. In what following we assume without the loss of generality that $0\in p[\mathcal{G}]$. Then, since $p[g]\leq g$ with probability 1 we have

$$\sup_{g\in\mathcal{G}}(Pg-(1+c)P_ng)\leq\sup_{g\in\mathcal{G}}(Pg-Pp[g]+Pp[g]-(1+c)P_np[g]) \tag{5}$$

$$\leq\varepsilon^{\frac{1}{2-\beta}}+\sup_{g\in\mathcal{G}}(Pp[g]-(1+c)P_np[g]), \tag{6}$$

We denote $\mathcal{G}_0=p[\mathcal{G}]\cap\mathcal{B}_{L_1}(0,2\varepsilon^{\frac{1}{2-\beta}})$ and $\mathcal{G}_1=\{0\}\cup(p[\mathcal{G}]\setminus\mathcal{B}_{L_1}(0,2\varepsilon^{\frac{1}{2-\beta}}))$, obviously $\mathcal{G}_0\cup\mathcal{G}_1=p[\mathcal{G}]$. We rewrite the last summand as

$$\sup_{g\in\mathcal{G}}(Pp[g]-(1+c)P_np[g])\leq\sup_{g\in\mathcal{G}_0}(Pg-(1+c)P_ng)+\sup_{g\in\mathcal{G}_1}(Pg-(1+c)P_ng).$$

Step 1. At first, we focus on $\sup_{g\in\mathcal{G}_1}(Pg-(1+c)P_ng)$. We estimate the following quantity

$$P(\exists g\in\mathcal{G}_1:P_ng<\frac{1}{1+c}Pg)$$
$$\leq\sum_{j=1}^{\infty}P(\exists g\in p[\mathcal{G}]:Pg\in[2^j\varepsilon^{\frac{1}{2-\beta}},2^{j+1}\varepsilon^{\frac{1}{2-\beta}}]\ \cap\ P_ng<\frac{1}{1+c}Pg)$$

Given a function $g\in p[\mathcal{G}]$ with $Pg\in[2^j\varepsilon^{\frac{1}{2-\beta}},2^{j+1}\varepsilon^{\frac{1}{2-\beta}}]$ we consider

$$P(Pg-(1+c)P_ng>0)=P\left(Pg-P_ng>\frac{cPg}{1+c}\right).$$

Using the Bernstein inequality [9] and simple algebra we have since $Pg > 0$

$$P\left(Pg - P_n g > \frac{cPg}{1+c}\right) \le \exp\left(-\frac{nc^2(Pg)^2}{(1+c)^2(2Pg^2 + \frac{2cPg}{3(1+c)})}\right)$$
$$\le \exp\left(-\frac{nc^2}{4(1+c)}\left(\frac{(Pg)^2}{(1+c)Pg^2} \wedge \frac{3Pg}{c}\right)\right).$$

Let $g' \in \mathcal{G}$ be any function such that $p[g'] = g$ for $g \in p[\mathcal{G}]$ with $Pg \ge 2\varepsilon^{\frac{1}{2-\beta}}$. Without loss of generality we may assume $\|g\|_\infty \le 1$ and with probability one we have $|g(Z) - g'(Z)| \le 2$. Using the Bernstein assumption we have

$$Pg^2 \le 2Pg'^2 + 2P(g-g')^2 \le 2Pg'^2 + 4\varepsilon^{\frac{1}{2-\beta}} \le 2B(Pg')^\beta + 4\varepsilon^{\frac{1}{2-\beta}}$$
$$\le B(Pg + \varepsilon^{\frac{1}{2-\beta}})^\beta + 2Pg \le 4B(Pg)^\beta.$$

Substituting, we have (provided that $(Pg)^{2-\beta} \le Pg$ and $B \ge 1$)

$$P\left(Pg - P_n g > \frac{cPg}{1+c}\right) \le \exp\left(-\frac{nc^2}{4(1+c)}\left(\frac{(Pg)^{2-\beta}}{4B(1+c)} \wedge \frac{3Pg}{c}\right)\right)$$
$$= \exp\left(-\frac{nc^2(Pg)^{2-\beta}}{16B(1+c)^2}\right)$$
$$\le \exp\left(-\frac{nc^2 2^{j(2-\beta)}\varepsilon}{16B(1+c)^2}\right).$$

We want to estimate the number of functions $g \in p[\mathcal{G}]$ with $Pg \in [2^j\varepsilon^{\frac{1}{2-\beta}}, 2^{j+1}\varepsilon^{\frac{1}{2-\beta}}]$. It is straightforward using Lemma 3. A small technical detail is that we can not guarantee that our global minimal covering is still minimal when restricted on the subset. However, it is almost minimal in a sense that it is enough to consider in what is following $\mathcal{D}^{\mathrm{loc}}_{[\,]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)$ instead of $\mathcal{D}^{\mathrm{loc}}_{[\,]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}\right)$. The argument is standard and is based on relations between minimal coverings and maximal packings and our technique of controlling the entropy, used in the proof of Lemma 3. Now,

$$\sum_{j=1}^{\infty} P(\exists g \in p[\mathcal{G}] : Pg \in [2^j\varepsilon^{\frac{1}{2-\beta}}, 2^{j+1}\varepsilon^{\frac{1}{2-\beta}}] \cap P_n g < \frac{1}{1+c}Pg)$$
$$\le \sum_{j=1}^{\infty} (2^{j+5})^{\mathcal{D}^{\mathrm{loc}}_{[\,]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}\right)/\log(4)} \exp\left(-\frac{nc^2 2^{j(2-\beta)}\varepsilon}{16B(1+c)^2}\right)$$
$$\le \sum_{j=1}^{\infty} \exp\left(\frac{(j+5)\mathcal{D}^{\mathrm{loc}}_{[\,]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)}{\log(4)} - \frac{nc^2(j+5)\varepsilon}{48B(1+c)^2}\right).$$

Provided that $n \ge \frac{48B(1+c)^2}{c^2\log(4)}\left(\frac{\mathcal{D}^{\mathrm{loc}}_{[\,]}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)}{\varepsilon} + \frac{\log(\frac{1}{\delta})}{\varepsilon}\right)$ the last term is upper bounded by $\frac{\delta}{2}$. Therefore, with probability at least $1 - \frac{\delta}{2}$ it holds that for all $g \in \mathcal{G}_1$ we have $Pg - (1+c)P_n g \le 0$. Thus, on this event $\sup_{g \in \mathcal{G}_1}(Pg - (1+c)P_n g) = 0$.

Step 2. Now we work with $\sup_{g\in\mathcal{G}_0}(Pg-(1+c)P_ng)$. We consider only the interesting range $\varepsilon\in[0,1]$. To control this process we use the Bernstein inequality together with the union bound, taking into account that $|\mathcal{G}_0|\leq \exp\left(\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}/2\right)\right)$ and that as before $Pg^2\leq 2Pg'^2+2P(g-g')^2\leq 2Pg'^2+4\varepsilon^{\frac{1}{2-\beta}}\leq 2B(Pg')^{\beta}+4\varepsilon^{\frac{1}{2-\beta}}\leq 8B\varepsilon^{\frac{\beta}{2-\beta}}$

$$
\begin{aligned}
&P(\sup_{g\in\mathcal{G}_0}(Pg-(1+c)P_ng)\geq \varepsilon^{\frac{1}{2-\beta}})\\
&\leq \exp\left(\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}/2\right)-\frac{n}{4(1+c)}\left(\frac{(\varepsilon^{\frac{1}{2-\beta}}+cPg)^2}{(1+c)Pg^2}\wedge 3(\varepsilon^{\frac{1}{2-\beta}}+cPg)\right)\right)\\
&\leq \exp\left(\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}/2\right)-\frac{n}{4(1+c)}\left(\frac{(\varepsilon^{\frac{1}{2-\beta}}+cPg)^2}{16(1+c)B\varepsilon^{\frac{\beta}{2-\beta}}}\wedge 3(\varepsilon^{\frac{1}{2-\beta}}+cPg)\right)\right)\\
&\leq \exp\left(\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}/2\right)-\frac{n}{4(1+c)}\left(\frac{\varepsilon}{16(1+c)B}\wedge 3\varepsilon^{\frac{1}{2-\beta}}\right)\right)\\
&\leq \exp\left(\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}/2\right)-\frac{n}{4(1+c)}\left(\frac{\varepsilon}{16(1+c)B}\right)\right).
\end{aligned}
$$

By taking $n\geq 64B(1+c)^2\left(\frac{\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}/2\right)}{\varepsilon}+\frac{\log(\frac{2}{\delta})}{\varepsilon}\right)$ we obtain that with probability at least $1-\frac{\delta}{2}$ we have $\sup_{g\in\mathcal{G}_0}(Pg-(1+c)P_ng)\leq\varepsilon^{\frac{1}{2-\beta}}$.

Step 3. Using a union bound for events from Steps 1 and 2 with probability at least $1-\delta$, given that $n\geq 64B(1+c)^2\left(\frac{\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}/2\right)}{\varepsilon}+\frac{\log(\frac{2}{\delta})}{\varepsilon}\right)$ it holds

$$\sup_{g\in\mathcal{G}}(Pg-(1+c)P_ng)\leq 2\varepsilon^{\frac{1}{2-\beta}}$$

We denote $c'=64(1+c)^2$. Now taking $\gamma^*(\mathcal{G},k,\beta,\delta)=\{\inf\varepsilon>0: k(\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}/2\right)+\log(\frac{2}{\delta}))\leq\varepsilon\}$ we have that with probability at least $1-\delta$ for a given n it holds $\sup_{g\in\mathcal{G}}(Pg-(1+c)P_ng)\leq 2(\gamma^*(\mathcal{G},c'B/n,\beta,\delta))^{\frac{1}{2-\beta}}$. However, if we take $\gamma_{[\,]}(\mathcal{G},k,\beta)=\{\inf\varepsilon>0: k\mathcal{D}^{\text{loc}}_{[\,]}\left(\mathcal{G},\varepsilon^{\frac{1}{2-\beta}}/2\right)\leq\varepsilon\}$ it is straightforward to see (using monotonicity of $\mathcal{D}^{\text{loc}}_{[\,]}(.)$) that

$$\gamma^*(\mathcal{G},c'B/n,\beta,\delta)\leq\gamma_{[\,]}(\mathcal{G},c'B/n,\beta)+\frac{c'B\log(\frac{2}{\delta})}{n}.$$

The claim follows. ■

Proof [of Theorem 2] With Lemma 4 the proof is rather straightforward. Given an empirical risk minimizer $\hat{f}$ denote the corresponding function in $\mathcal{L}_{\mathcal{Y}}$ as $\hat{g}$. We have $P\hat{g}=R(\hat{f})-R(f^*)$ and

$P_n\hat{g} \le 0$. Then for any $c > 0$ (namely we may take $c = 1$) we have

$$P\hat{g} \le P\hat{g} - (1+c)P_n\hat{g} \le \sup_{g \in \mathcal{L}_\mathcal{Y}} (Pg - (1+c)P_n g)$$

The final step will be to understand that metric properties of $\mathcal{L}_\mathcal{Y}$ are the same as the properties of $\mathcal{G}_\mathcal{Y}$. That means, for example, that $\gamma_{[\,]}\left(\mathcal{L}_\mathcal{Y}, \frac{B}{n}, \beta\right) = \gamma_{[\,]}\left(\mathcal{G}_\mathcal{Y}, \frac{B}{n}, \beta\right)$. Lemma 4 for $\mathcal{L}_\mathcal{Y}$ finishes the proof. ■

The next important question is to understand how to estimate local entropies with bracketing. In some cases, this may be easily done. For example, for numerous nonparametric classes not only the upper bounds on the entropies are known, but it is also true that bracketing entropies and standard entropies are of the same order (see [37, 36, 29] and reference therein). Moreover, following Yang and Barron [42] for these nonparametric classes the local entropies are of the same order as the global entropies. However, controlling the local entropy with the bracketing for smaller classes does not seem trivial. Only recently, Gassiat and van Handel have provided a tight analysis for the local entropies with the bracketing for certain parametric classes of densities [14]. For general VC classes nothing more than the boundedness of entropies with bracketing is known [2].

When we are unable to guarantee that entropies with bracketing are close to the entropies without bracketing we may use the following trick, which in our case will be used *only* to remove the bracketing condition from the local entropy in our bounds. The technique is based on the so-called *skeleton estimates*: these algorithms are ERM over the ε-net of the initial class. Versions of this algorithm appear widely in the literature [11, 42, 36, 10, 31, 33]. This algorithm is more of a theoretical interest since it is unlikely that it will be computationally efficient compared to ERM over the whole class. However, to the best of our knowledge, our next result is the first localized result of this kind under general Bernstein conditions. In the following section we will demonstrate that is some cases this bound may recover the optimal learning rate.

Corollary 5 (Bound for ERM over the ε-net) *Assume that given $\eta \in [0,1]$ one can select functions $f_1, \dots, f_{N_\eta} \in \mathcal{F}$ such that corresponding functions $\ell(f_1(X), Y), \dots, \ell(f_{N_\eta}(X), Y)$ form a minimal $L_1(P)$ η-covering of the loss class $\mathcal{G}_\mathcal{Y}$. Define $\hat{f}_\eta = \operatorname{arg\,min}_{f \in \{f_1, \dots, f_{N_\eta}\}} R_n(f)$. If $\eta \simeq \left(\gamma\left(\mathcal{G}_\mathcal{Y}, \frac{B}{n}, \beta\right) + \frac{B\log(\frac{1}{\delta})}{n}\right)^{\frac{1}{2-\beta}}$, then we have under (β, B)-Bernstein condition on $\mathcal{L}_\mathcal{Y}$ that with probability at least $1-\delta$ it holds*

$$R(\hat{f}_\eta) - R(f^*) \lesssim \left(\gamma\left(\mathcal{G}_\mathcal{Y}, \frac{B}{n}, \beta\right) + \frac{B\log(\frac{1}{\delta})}{n}\right)^{\frac{1}{2-\beta}}.$$

The proof of this result is deferred to appendix. The conditions of this Corollary hold naturally, for example, for the binary loss, since for this case

$$|\mathbb{1}[f(X) \ne Y] - \mathbb{1}[g(X) \ne Y]| = |f(X) - g(X)|/2. \tag{7}$$

Thus, if one wants to cover the loss class it is sufficiently and enough to cover the initial class $\mathcal{F}$.

Examples

GLOBAL ENTROPY CONDITIONS

Classic rates for global $L_1(P)$ entropies with bracketing, obtained by Tsybakov (see Theorem 1 in [36]) are approached by our general bound. Namely under the condition that $\log(\mathcal{N}_{[\,]}(\mathcal{G}_{\mathcal{Y}}, \varepsilon)) \simeq \varepsilon^{-r}$ for $\varepsilon \in [0, 1)$ and $r > 0$ we provide under (β, B)-Bernstein conditions the bound of order

$$\left(\left(\frac{1}{n} \right)^{\frac{2-\beta}{2-\beta+r}} + \frac{\log(\frac{1}{\delta})}{n} \right)^{\frac{1}{2-\beta}}, \tag{8}$$

where we skipped the dependence on B to maintain the same form as in [36]. The important fact is that we give the same result for ERM (not for ERM over the ε-net as used in [36], which is more favourable in the case when one wants to get an effective learning algorithm). Interestingly, that using Theorem 6 in [29] for the case $\beta = 1$ and the binary loss it is simple to prove the lower bound in terms of the local entropy without bracketing, that is valid *for any class* with the same local entropy (however the dependence on B will be slightly suboptimal). We will demonstrate related techniques for a similar problem in Proposition 6.

HOMOGENEOUS HALFSPACES UNDER ISOTROPIC LOG-CONCAVE DISTRIBUTIONS

This example took a lot of attention in the literature (see [10, 5, 28, 17] and reference therein) and is one of our motivations to consider distribution dependent complexity measures. It follows directly from the result of Hanneke (see section 5.1 in [17]) which is based on recent results of Balcan and Long [5] that for the class $\mathcal{F}$ of homogeneous halfspaces (passing through the origin) in $\mathbb{R}^d$ under isotropic log-concave distributions of X it holds $\gamma(\mathcal{G}_{\mathcal{Y}}, \frac{B}{n}) \lesssim \frac{Bd}{n}$ under the binary loss. Here we also used 7 to relate the loss class to the initial class $\mathcal{F}$. Thus, for ERM over the net we have a rate $R(\hat{f}_\eta) - R(f^*) \lesssim \left(\frac{Bd}{n} + \frac{B\log(\frac{1}{\delta})}{n} \right)^{\frac{1}{2-\beta}}$. Previously, for ERM over the net this result was provided [10] only in the simplest realizable case (in particular, for this case $\beta = B = 1$) and our rate is strictly better than the rate[1] implied by the recent Theorem 19 in [17] under general B and β, which itself is the best known bound so far. However, it remains not clear for us, whether for this problem $\gamma(\mathcal{G}_{\mathcal{Y}}, \frac{B}{n}, \beta) \simeq \gamma_{[\,]}(\mathcal{G}_{\mathcal{Y}}, \frac{B}{n}, \beta)$. If so, we can safely use ERM algorithm to obtain the same rates. Moreover, for the case when $\beta = 1$ we prove a B dependent matching lower bound, showing that the complexity term $\frac{Bd}{n}$ can not be avoided (previously the lower bound was provided only in the realizable case [5, 28]). The important component of the analysis is that the proof of the lower bound is based on the *local entropy*. So, for this specific problem the learning rates are *fully determined* by this complexity measure.

Proposition 6 (Lower bound for log-concave distributions) *Consider the problem of learning the class $\mathcal{F}$ of homogeneous halfspaces in $\mathbb{R}^d$ with the binary loss. Let $\tilde{f}$ be an output of any learning algorithm. Then for any $B \le \sqrt{\frac{n}{d}}$ there exists a distribution $P_{X,Y}$, such that the excess loss class $\mathcal{L}_{\mathcal{Y}}$ is $(1, B)$-Bernstein, P_X is an isotropic log-concave distribution and $\mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{Bd}{n}$.*

The proof of this proposition is deferred to appendix.

1. Interestingly, that this previous bound is based on a different technique and obtained via a different P_X-dependent learning algorithm (taking its roots in the theory of active learning).

EXACT AND NON-EXACT ORACLE INEQUALITIES IN AGGREGATION THEORY

The following two simple examples are instructive corollaries of our results. The non-exact oracle inequalities are the upper bounds on $R(\hat{f}) - (1+a)R(f^*)$ for some $a > 0$. It is known, that this $1+a$ term instead of 1 allows one to obtain under mild conditions the same rates as if the Bernstein condition for $\mathcal{L}_{\mathcal{Y}}$ holds true. As noted by Lecué [22] shifted processes are sufficient for proving this kind of results. Via simple calculations, one can prove (see [22]) that for any ERM $\hat{f}$ it holds $R(\hat{f}) - (1+2c)R(f^*) \le \sup_{g\in\mathcal{G}_{\mathcal{Y}}} (Pg - (1+c)P_n g) + (1+c)\sup_{g\in\mathcal{G}_{\mathcal{Y}}} \left(P_n g - \frac{1+2c}{1+c}Pg\right)$, where as before $\mathcal{G}_{\mathcal{Y}}$ is a loss class. However, our approach to bound these processes is different: namely using our Lemma 4 and an easily obtainable generalization of it for $\sup_{g\in\mathcal{G}_{\mathcal{Y}}} \left(P_n g - \frac{1+2c}{1+c}Pg\right)$. The key point here is that to apply this Lemma we need a (β, B)-Bernstein condition not for the excess loss class $\mathcal{L}_{\mathcal{Y}}$ but for the loss class $\mathcal{G}_{\mathcal{Y}}$ which holds trivially for bounded losses. Thus, for any loss bounded by 1 since $\mathcal{G}_{\mathcal{Y}}$ is a $(1,1)$-Bernstein class it holds (for example for $a=1$) with probability at least $1-\delta$ that for any ERM $\hat{f}$:

$$R(\hat{f}) - 2R(f^*) \lesssim \gamma_{[\,]}\left(\mathcal{G}_{\mathcal{Y}} \cup \{0\}, \frac{1}{n}, 1\right) + \frac{\log(\frac{1}{\delta})}{n}. \tag{9}$$

An instructive case is when $R(f^*)$ is small and is of an order of the right hand side. In this case we have the same guaranties on the excess risk of ERM as if $(1,1)$-Bernstein condition holds true for $\mathcal{L}_{\mathcal{Y}}$. Previously in the literature special aggregation procedures were used to obtain related bounds in this regime (see, for example, Theorem 5 in [33]).

The problem of convex aggregation consists in finding a procedure that has a risk as close as possible to the minimal risk over the convex hull of $\mathcal{F}$, where $\mathcal{F}$ is a finite class of M functions [36, 21]. Here the loss function is square and $0 \le Y \le 1$ and $0 \le f(X) \le 1$ for all $f \in \mathcal{F}$. When, for example, $M \le \sqrt{n}$ for any ERM $\hat{f}$ over $\mathrm{conv}(\mathcal{F})$ it holds with probability at least $1-\delta$ that $R(\hat{f}) - \inf_{f\in\mathrm{conv}(\mathcal{F})} R(f) \lesssim \frac{M}{n} + \frac{\log(\frac{1}{\delta})}{n}$ and this bound is optimal in the sense that it is tight up to constant factors for some $\mathcal{F}$ [21]. However, it seems natural to ask whether for some classes and probability distributions this rate can be improved and to what extent (these questions arose previously in the aggregation literature [23]). Our technique provides a natural step towards this question. For this specific convex problem the excess loss class is known to be $(1,16)$-Bernstein [22] and our Corollary 5 immediately gives us the localized bound $R(\hat{f}_\eta) - \inf_{f\in\mathrm{conv}(\mathcal{F})} R(f) \lesssim \gamma\left(\mathcal{G}_{\mathcal{Y}}, \frac{1}{n}, 1\right) + \frac{\log(\frac{1}{\delta})}{n}$, where $\mathcal{G}_{\mathcal{Y}}$ is the loss class associated with $\mathrm{conv}(\mathcal{F})$. Moreover, under the square loss $|(Y-f(X))^2 - (Y-g(X))^2| \le 2|f(X)-g(X)|$. This immediately leads to the bound

$$R(\hat{f}_\eta) - \inf_{f\in\mathrm{conv}(\mathcal{F})} R(f) \lesssim \gamma\left(\mathrm{conv}(\mathcal{F}), \frac{1}{n}, 1\right) + \frac{\log(\frac{1}{\delta})}{n}, \tag{10}$$

which captures the distribution and a localized complexity of the class $\mathrm{conv}(\mathcal{F})$. It is straightforward to construct examples with $\gamma\left(\mathrm{conv}(\mathcal{F}), \frac{1}{n}, 1\right)$ much smaller than $\frac{M}{n}$.

LOCAL $L_2(P)$ ENTROPIES IN WELL-SPECIFIED REGRESSION MODELS

So far we discussed only $L_1(P)$ entropies. However, for nonpaprametric classes the analysis is usually performed under $L_2(P)$. We consider the bounded regression model with a square loss (as

in the previous example). However, additionally we require that the model is well-specified, so $f^*(X) = \mathbb{E}[Y|X]$, but without assuming the convexity of $\mathcal{F}$. For example, this includes the model $Y = f^*(X) + \xi$, where ξ is independent, bounded and zero mean and $f^* \in \mathcal{F}$. At first we simply adapt our notation to $L_2(P)$ case. We define

$$\mathcal{D}^{\text{loc}}_{L_2}(\mathcal{G}, \varepsilon) = \sup_{\gamma \geq \varepsilon} \sup_{g \in \mathcal{G}} \log \left(\mathcal{N}_{L_2}(\mathcal{G} \cap \mathcal{B}_{L_2}(g, 2\gamma), \gamma) \right),$$

where $\mathcal{N}_{L_2}$ denotes the covering number with respect to $L_2(P)$ norm. Finally,

$$\zeta(\mathcal{G}, k) = \{\inf \varepsilon > 0 : k\mathcal{D}^{\text{loc}}_{L_2}(\mathcal{G}, \varepsilon) \leq \varepsilon^2\}$$

Corollary 7 *Assume that we are in the situation of a well-specified bounded regression model with the square loss as defined above. Given $\eta \in [0,1]$ we choose $f_1, \ldots, f_{N_\eta} \in \mathcal{F}$ that form a minimal η-cover of $\mathcal{F}$ with respect to $L_2(P)$. Define $\hat{f}_\eta = \underset{f \in \{f_1, \ldots, f_{N_\eta}\}}{\arg\min} R_n(f)$. Then if $\eta \simeq \zeta\left(\mathcal{F}, \frac{1}{n}\right) + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}$, then with probability at least $1 - \delta$ it holds*

$$R(\hat{f}_\eta) - R(f^*) \lesssim \left(\zeta\left(\mathcal{F}, \frac{1}{n}\right)\right)^2 + \frac{\log(\frac{1}{\delta})}{n}.$$

Previously realted rates were obtained via global empirical entropies in [33] using so-called *skeleton aggregation* or *aggregation of leaders* procedures. However, in our special case a simpler proper (taking its values in $\mathcal{F}$) procedure is used and our complexity measure is localized. We also note that almost the same $L_2(P)$-based local entropy (see Theorem 4.5 in [31]) appears in general minimax lower bound in the unbounded case for convex $\mathcal{F}$. Finally, our result (still special to this well-specified case) is valid even for the very expressive nonparametric classes with $\log(\mathcal{N}_{L_2}(G, \varepsilon)) \simeq \varepsilon^{-r}$, for $r > 2$ and does not require the convexity or the star-shapedness of $\mathcal{F}$ or $\mathcal{L}_\mathcal{Y}$.

4. Stability and sample compression schemes

Nevertheless, the approach based on the local entropy is not a panacea. In some cases, related techniques will not provide tight learning rates. One of the simplest examples is the classification when no assumptions are made about the noise (see the related discussions in [43]); even in the nonparametric case when $\beta \neq 1$ our upper bound 8 is slightly worse than the bound based on the *chaining technique* [36, 29]. The other simple model is the realizable case classification, defined above. The optimal sample complexity in this case is known to be $\frac{d}{n} + \frac{\log(\frac{1}{\delta})}{n}$ [18, 19], where d is a VC dimension of the class. However, the principles behind this optimal rate are not well understood. Clearly, it is not a uniform convergence principle since *some ERMs are known to be suboptimal for this problem* [4]. Namely, the achieve exactly $\frac{d\log(\frac{n}{d})}{n} + \frac{\log(\frac{1}{\delta})}{n}$ rate. Moreover, since the local entropies are related to the uniform convergence rates, they do not appear in minimax lower bounds (in general introduced fixed points of local entropies are known to be of order greater than $\frac{d}{n}$ for some distributions [16]).

There are two other principles, that guarantee generalization: *sample compression* and *stability*. At first, we give several formal definitions.

Definition 8 (Sample compression schemes [13]) *Define the sequence of permutation invariant functions* $\kappa_n : (\mathcal{X} \times \mathcal{Y})^n \to \cup_{i=1}^k (\mathcal{X} \times \mathcal{Y})^i$. *These are* compression functions. *The reconstruction function* $\rho : \cup_{i=1}^k (\mathcal{X} \times \mathcal{Y})^i \to \mathcal{Y}^{\mathcal{X}}$. *Functions* κ_n *and* ρ *define a sample compression scheme of size* k *if for any* $n \in \mathbb{N}$ *and* $f \in \mathcal{F}$ *and any sample* $(x_i, f(x_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ *it holds* $\kappa_n((x_i, f(x_i))_{i=1}^n) \subseteq (x_i, f(x_i))_{i=1}^n$ *and denoting* $\hat{f} = \rho(\kappa_n((x_i, f(x_i))_{i=1}^n))$ *we have* $f(x_i) = \hat{f}(x_i)$, *for all* $i = 1, \dots, n$.

A state of the art result for the generalization ability of sample compression schemes is the following.

Lemma 9 (Floyd and Warmuth [13]) *Assume that* ρ *is a reconstruction function of some sample compression scheme of size* k. *Given* f *and any i.i.d. sample* $(X_i, f(X_i))_{i=1}^n$ *it holds with probability at least* $1 - \delta$ *simultaneously for all sets* $A \subset (X_i, f(X_i))_{i=1}^n$ *with* $|A| \leq k$ *and with the property that* $(\rho(A))(X_i) = f(X_i)$ *for* $i = 1, \dots, n$

$$P((\rho(A))(X) \neq f(X)) \leq \frac{k \log(\frac{en}{k})}{n-k} + \frac{\log(\frac{1}{\delta})}{n-k}. \tag{11}$$

An existence of sample compression schemes of size $O(d)$ is a well known open problem [13] . However, even if we are able to construct a sample compression scheme of size $O(d)$ it is known [13] (see the discussion after their Theorem 6) that the rate $O(\frac{d \log(\frac{n}{d})}{n} + \frac{\log(\frac{1}{\delta})}{n})$ as stated by 11 can not be improved for some compression schemes. Simultaneously, ERM bounds based on the local empirical entropies *are always not worse* [43]. Thus, in our framework in terms of general statistical performance sample compression schemes *are not preferable* to ERM over the class $\mathcal{F}$. However, under natural assumptions the sample compression schemes are approaching minimax optimal rates.

Definition 10 (Stable compression scheme) *A sample compression scheme* (κ_n, ρ) *is stable iff for arbitrary* n, $f \in \mathcal{F}$, *sample* $(x_i, f(x_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ *and any* $(x, y) \in (x_i, f(x_i))_{i=1}^n \setminus \kappa_n((x_i, f(x_i))_{i=1}^n)$ *it holds* $\kappa_{n-1}((x_i, f(x_i))_{i=1}^n \setminus (x, y)) = \kappa_n((x_i, f(x_i))_{i=1}^n)$.

This means that removing an element that is not in the compression set never changes the compression set of the subsample. The next general definition is motivated by the similar property, analyzed for the spans of intersection-closed classes [4].

Definition 11 (Homogeneous compression scheme) *A stable sample compression scheme* (κ_n, ρ) *is homogeneous iff for arbitrary* n, $f \in \mathcal{F}$, *sample* $(x_i, f(x_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ *and any* $(x, y) \in \kappa_n((x_i, f(x_i))_{i=1}^n$ *it holds* $\kappa_n((x_i, f(x_i))_{i=1}^n) \setminus (x, y) \subseteq \kappa_{n-1}((x_i, f(x_i))_{i=1}^n \setminus (x, y))$.

Homogeneous compression schemes are stable compression schemes with the property that removing an element (x, y) that is *inside the compression set* $\kappa_n((x_i, f(x_i))_{i=1}^n$ leaves the remaining compression elements inside the new compression set $\kappa_{n-1}((x_i, f(x_i))_{i=1}^n \setminus (x, y))$. As we already mentioned, they are naturally presented by several intersection-closed classes [4].

Theorem 12 *For a stable compression scheme* (κ_n, ρ) *of size* k *it holds with probability at least* $1 - \delta$ *over the learning sample*

$$\mathbb{E}R(\hat{f}) \leq \frac{k}{n+1}, \quad R(\hat{f}) \lesssim \frac{k \log(\frac{1}{\delta})}{n},$$

where $\hat{f} = \rho(\kappa_n((X_i, Y_i)_{i=1}^n))$. *Moreover, if the function* ρ *takes its values in the class of VC dimension* $d \lesssim k$, *then there exists an efficient*[2] *modification of our sample compression scheme (with an output denoted by* $\hat{g}$*) that gives a learning algorithm with*

$$R(\hat{g}) \lesssim \frac{k \log(k)}{n} + \frac{\log(\frac{1}{\delta})}{n}. \tag{12}$$

If (κ_n, ρ) *is also homogeneous, then*

$$R(\hat{f}) \lesssim \frac{k}{n} + \frac{\log(\frac{1}{\delta})}{n}.$$

The proof of this result is deferred to appendix. We describe the modification of the scheme which is used to obtain 12. Given a sample S of size $3n$ we denote $S_{1/3}, S_{2/3}$ the first n and $2n$ elements of the sample respectively. The modification of the compression scheme means that any point X is classified by a major vote of three functions obtained by the compression scheme applied on $S_{1/3}, S_{2/3}, S$ (analogous to a so-called $\mathbf{L}_2$ algorithm, introduced for general empirical risk minimizers in [34]). As a direct corollary of Theorem 12 when the sample compression scheme is of size $k = O(d)$, where d is a VC dimension, homogeneous sample compression schemes have an optimal learning rate up to constant factors by matching the lower bound [12] and all stable sample compression schemes are optimal in expectation [39] up to constant factors. The ideas behind these specific schemes appeared in some form previously in the literature. One of the first applications of sample compression schemes (before the initial general research [13]) appears in [39]. Vapnik and Chervonenkis implicitly used stability and sample compression arguments to prove in expectation bound for the hard margin SVM of order $\frac{d}{n}$, where d is a dimension. However, obtaining tight high probability results based on stability arguments is considered a difficult problem (see [11, 8, 19, 39] or the related open problem [40]). Another trick which was used previously in the literature to obtain $\frac{k \log(\frac{1}{\delta})}{n}$ bound given the $\frac{k}{n}$ bound in expectation is the following [19]: one runs the algorithm $\log(\frac{1}{\delta})$ times on independent samples of a certain size and chooses the best classifier based on a small independent test sample. However, further obtaining of inequalities like 12 does not seem straightforward using this technique.

Examples

IMPROVED HARD MARGIN SVM

Corollary 13 (Tight PAC bound for Halfspaces) *For the class* $\mathcal{F}$ *of linear halfspaces in the realizable case there exists a learning algorithm with polynomial running time and output denoted by* $\hat{f}$, *such that with probability at least* $1 - \delta$

$$R(\hat{f}) \lesssim \frac{d \log(d)}{n} + \frac{\log(\frac{1}{\delta})}{n}. \tag{13}$$

Proof At first we show that in the realizable case the separating hyperplane in $\mathbb{R}^d$ constructed by the SVM defines a stable compression of size at most $d + 1$. This follows from the existence of the

2. Although we do not focus on computational issues, by efficiency we mean that to obtain this rate we would need to run the compression scheme exactly three times. This technique is based on the voting algorithm by Simon [34].

so-called *essential support vectors* (Chapter 14 in [39]). Taking into account that the VC dimension of this class if equal to $d+1$ using Theorem 12 we obtain $\frac{d\log(\frac{1}{\delta})}{n}$ rate for SVM and $\frac{d\log(d)}{n}+\frac{\log(\frac{1}{\delta})}{n}$ for its modification. Finally, we observe that both SVM and its modification (voting over three SVMs) are algorithms with polynomial running time. ■

This bound is a direct corollary of our general result and almost matches the minimax lower bound $\frac{d}{n}+\frac{\log(\frac{1}{\delta})}{n}$ [12] . Previously the polynomial time algorithm with the tight risk bound was known only for the class of homogenious halfspaces and only for log-concave distributions of X. That risk bound was obtained via more involved arguments and both assumptions were crucial [5]. Moreover, 13 compares favourably with the best possible rate that can be obtained via the uniform convergence principle for the problem of learning halfspaces, namely, $\frac{d\log(\frac{n}{d})}{n}+\frac{\log(\frac{1}{\delta})}{n}$ (see [43] for related discussions).

NEW ONLINE TO BATCH CONVERSION

Assume that in the realizable online framework we are given a conservative online learning algorithm making at most k mistakes on any sample (see [13] or [27] for more details on the framework). By conservative we mean that the algorithm does not change its state after a correct classification of the next point.

Our goal will be to convert an online learning algorithm to a learning algorithm in the standard i.i.d. setting. Assume that the set $\mathcal{X}$ is ordered. Consider the following classifier with an output $\hat{f}$ based on the i.i.d sample $S=(X_i,Y_i)_{i=1}^n$. Given a sample S we define S^* to be a set consisting of pairs (X_i,Y_i) sorted according to the order of $\mathcal{X}$ and $S^*_{\preceq x}$ as the subset of S^* with pairs (X_i,Y_i) such that all X_i precede the fixed element x . Now define for any $x\in\mathcal{X}$

- If there exists $j\in\{1,\dots,n\}$ such that $x=X_j$, then define $\hat{f}(x)=Y_j$.
- Otherwise, define $\hat{f}(x)$ as a label of x that we obtain by applying the last classifier that we get after running our conservative algorithm on the set $S^*_{\preceq x}$.

It is straightforward to see that $\hat{f}$ is an output of sample compression scheme of size k [13]. Moreover, due to the fact that the algorithm is conservative, it appears that the corresponding sample compression scheme is stable and thus we have a rate $R(\hat{f})\le\frac{k\log(\frac{1}{\delta})}{n}$. This already improves over the known bounds for the *longest surviving strategy* with the rate $\frac{k\log(\frac{k}{\delta})}{n}$ [27].

However, when we are able to guarantee that the output space of our online algorithm has VC dimension $d\lesssim k$ we may apply the modification 12. As before we construct three samples $S_{\frac{1}{3}},S_{\frac{2}{3}},S$ and corresponding ordered sets $S^*_{\frac{1}{3},\preceq x},S^*_{\frac{2}{3},\preceq x},S^*_{\preceq x}$. Now the modified $\hat{g}(x)$ is defined as the majority vote over three values that we obtain by applying the last classifier that we get after running our conservative algorithm on sets $S^*_{\frac{1}{3},\preceq x},S^*_{\frac{2}{3},\preceq x},S^*_{\preceq x}$. This modification gives the rate $R(\hat{g})\le\frac{k\log(k)}{n}+\frac{\log(\frac{1}{\delta})}{n}$ that almost coincides with the best known guaranties $\frac{k}{n}+\frac{\log(\frac{1}{\delta})}{n}$ from [27]. Interestingly, that the last bound is achieved using a different strategy and martingale-based proof techniques.

Acknowledgments

We would like to thank Steve Hanneke for several helpful discussions and anonymous reviewers for their useful suggestions. The author was supported solely by the Russian Science Foundation grant (project 14-50-00150).

References

[1] *R. Adamczak.* A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. Electron. J. Probab., 1000–1034, 2008.

[2] *T. M. Adams, A. B. Nobel*. Uniform approximation and bracketing properties of VC classes. Bernoulli, 18:1310–1319, 2012.

[3] *M. Anthony, P. L. Bartlett*. Neural Network Learning: Theoretical Foundations. Cambridge University Press, 1999.

[4] *P. Auer, R. Ortner*. A new PAC bound for intersection-closed concept classes. Machine Learning, 66(2-3): 151–163, 2007.

[5] *M.F. Balcan, P. M. Long*. Active and passive learning of linear separators under log-concave distributions. In Proceedings of the 26th Conference on Learning Theory, 2013.

[6] *P. L. Bartlett, O. Bousquet, S. Mendelson*. Local Rademacher Complexities. The Annals of Statistics, 33(4):1497–1537, 08, 2005.

[7] *P. L. Bartlett, S. Mendelson*. Empirical minimization. Probability Theory and Related Fields, 135(3):311–334, 2006.

[8] *O. Bousquet, A. Elisseeff*. Stability and generalization. Journal of Machine Learning Research, 2002.

[9] *S. Boucheron, G. Lugosi, P. Massart*. Concentration inequalities: A nonasymptotic theory of independence. Cambridge, 2013.

[10] *N. H. Bshouty, Y. Li, P. M. Long*. Using the doubling dimension to analyze the generalization of learning algorithms. Journal of Computer and System Sciences, 2009.

[11] *L. Devroye, L. Györfi, G. Lugosi*. A Probabilistic Theory of Pattern Recognition, volume 31 of Applications of Mathematics. Springer–Verlag, New York, 1996.

[12] *A. Ehrenfeucht, D. Haussler, M. Kearns, L. Valiant*. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

[13] *S. Floyd and M. Warmuth*. Sample Compression, learnability, and the Vapnik Chervonenkis Dimension, Machine Learning, 21, 269–304 (1995).

[14] *E. Gassiat, R. van Handel.* The local geometry of finite mixtures, Trans. Amer. Math. Soc. 366, 1047–1072, 2014.

[15] *E. Giné, V. Koltchinskii*. Concentration inequalities and asymptotic results for ratio type empirical processes. The Annals of Probability, 34(3):1143–1216, 2006.

[16] *S. Hanneke, L. Yang*. Minimax analysis of active learning. Journal of Machine Learning Research, 16 (12): 3487–3602, 2015.

[17] *S. Hanneke*. Refined error bounds for several learning algorithms. Journal of Machine Learning Research 17, 1–55, 2016

[18] *S. Hanneke*. The Optimal Sample Complexity of PAC Learning. Journal of Machine Learning Research, 17 (38): 1-15, 2016.

[19] *D. Haussler, N. Littlestone, M. Warmuth*. Predicting $\{0,1\}$-functions on randomly drawn points. Information and Computation, 115:248–292, 1994.

[20] *L. M. Le Cam*. Convergence of estimates under dimensionality restrictions. Ann. Statist. 1, 38–53, 1973.

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

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

[23] *G. Lecué, S. Mendelson* On the optimality of the aggregate with exponential weights for low temperature. Bernoulli, 2013.

[24] *G. Lecué, S. Mendelson*. Learning subgaussian classes: Upper and minimax bounds. http://arxiv.org/abs/1305.4825, 2013.

[25] *G. Lecué, C. Mitchell*. Oracle inequalities for cross-validation type procedures. Electronic Journal of Statistics, 6, 1803–1837, 2012.

[26] *T. Liang, A. Rakhlin, K. Sridharan*. Learning with square loss: Localization through offset Rademacher complexity. Proceedings of The 28th Conference on Learning Theory, 2015.

[27] *N. Littlestone.* From On-line to batch learning. In COLT, 1989.

[28] *P. M. Long.* On the sample complexity of PAC learning halfspaces against the uniform distribution. IEEE Transactions on Neural Networks, 6(6):1556–1559, 1995.

[29] *P. Massart, E. Nédélec*. Risk bounds for statistical learning. Annals of Statistics, 2006.

[30] *S. Mendelson*. Obtaining fast error rates in nonconvex situations. Journal of Complexity. Volume 24, Issue 3, 380–397, 2008.

[31] *S. Mendelson.* 'Local' vs. 'global' parameters – breaking the Gaussian complexity barrier. https://arxiv.org/abs/1504.02191, to appear in Annals of Statisitcs, 2017

[32] *S.Mendelson*. Learning without concentration. Journal of the ACM, Volume 62, Issue 3, 2015.

[33] *A. Rakhlin, K. Sridharan, A. B. Tsybakov*. Empirical entropy, minimax regret and minimax risk. Bernoulli, 2017.

[34] *H. Simon*. An almost optimal PAC-algorithm. Proceedings of The 28th Conference on Learning Theory, pp. 1552–1563, 2015.

[35] *A. B. Tsybakov*. Optimal rates of aggregation. In Computational Learning Theory and Kernel Machines 303–313. Lecture Notes in Artificial Intelligence, 2003.

[36] *A. B. Tsybakov*. Optimal aggregation of classifiers in statistical learning. The Annals of Statistics. Vol. 32, No. 1, 135–166, 2004

[37] *A. W. van der Vaart, J. A. Wellner*. Weak Convergence and Empirical Processes. Springer, 1996.

[38] *V. Vapnik, A. Chervonenkis*. On the uniform convergence of relative frequencies of events to their probabilities. Proc. USSR Acad. Sci. 181(4), 781–783, 1968.

[39] *V. Vapnik, A. Chervonenkis*. Theory of Pattern Recognition. Nauka, Moscow, 1974.

[40] *M. K. Warmuth.* The optimal PAC algorithm. In Proceedings of the 17th Conference on Learning Theory, 2004.

[41] *M. Wegkamp*. Model selection in nonparametric regression. Annals of Statistics, Vol. 31, No. 1, 252–273, 2003.

[42] *Y. Yang, A. Barron.* Information-theoretic determination of minimax rates of convergence. Annals of Statistics, 27, 1564–1599, 1999.

[43] *N. Zhivotovskiy, S. Hanneke.* Localization of VC classes: Beyond Local Rademacher complexities. https://arxiv.org/abs/1606.00922, 2016.

Appendix

Proof [of the Lemma 3] Denote for $\delta > 4\gamma$

$$\mathcal{N}(\delta, \gamma) = \sup_{g \in \mathcal{G}} \mathcal{N}(\mathcal{G} \cap B_P(g, \delta), \gamma)$$

and

$$\mathcal{N}_{[\,]}(\delta, \gamma) = \sup_{g \in \mathcal{G}} \mathcal{N}_{[\,]}(\mathcal{G} \cap B_P(g, \delta), \gamma)$$

Let $t_1, \ldots, t_N$ be centers of the minimal cover of δ-ball intersected with $\mathcal{G}$ by L_1-balls with radius $\delta/4$. The total number of them is bounded by $\mathcal{N}(\delta, \delta/4)$. Now for a given i we want to cover a set $\mathcal{B}_{L_1}(t_i, \delta/4) \cap \mathcal{G}$ by the γ-brackets. Obviously, since $t_i \in \mathcal{G}$ for all i the minimal number of brackets is bounded by $\mathcal{N}_{[\,]}(\delta/4, \gamma)$. Finally,

$$\mathcal{N}_{[\,]}(\delta, \gamma) \leq \mathcal{N}(\delta, \delta/4) \mathcal{N}_{[\,]}(\delta/4, \gamma).$$

Using a standard bound (see [37]) we have $\mathcal{N}(\delta,\delta/4)\leq\mathcal{N}_{[\,]}(\delta,\delta/2)$. Using the definition of the local entropy with bracketing we have

$$\mathcal{N}_{[\,]}(\delta,\gamma)\leq\exp(\mathcal{D}^{\mathrm{loc}}_{[\,]}(\mathcal{G},\gamma))\mathcal{N}_{[\,]}(\delta/4,\gamma). \tag{14}$$

We continue with the term $\mathcal{N}_{[\,]}(\delta/4,\gamma)$ in the same manner. If $\delta/16>\gamma$, then we use the same decomposition 14. Otherwise, if $\delta/16\leq\gamma$

$$\begin{aligned}&\mathcal{N}_{[\,]}(\delta/4,\gamma)\leq\mathcal{N}_{[\,]}(4\gamma,\gamma)\leq\mathcal{N}(4\gamma,2\gamma)\mathcal{N}_{[\,]}(2\gamma,\gamma)\\&\leq\mathcal{N}_{[\,]}(4\gamma,4\gamma)\exp(\mathcal{D}^{\mathrm{loc}}_{[\,]}(\mathcal{G},\gamma))\leq\exp(2\mathcal{D}^{\mathrm{loc}}_{[\,]}(\mathcal{G},\gamma)).\end{aligned}$$

Continuing in the same manner we have

$$\mathcal{N}_{[\,]}(\delta,\gamma)\leq\exp(\mathcal{D}^{\mathrm{loc}}_{[\,]}(\mathcal{G},\gamma))^{(\log_4(\frac{\delta}{\gamma})+2)}.$$

■

Proof [of Corollary 5] Define $f^*_\eta=\underset{f\in\{f_1,\ldots,f_{N_\eta}\}}{\arg\min}R(f)$. We have since $R_n(\hat{f}_\eta)-R_n(f^*_\eta)\leq 0$ for any $c\geq 1$

$$\begin{aligned}&R(\hat{f}_\eta)-R(f^*)\\&\leq R(\hat{f}_\eta)-R(f^*)-(1+c)(R_n(\hat{f}_\eta)-R_n(f^*_\eta))\\&=R(\hat{f}_\eta)-R(f^*)-(1+c)(R_n(\hat{f}_\eta)-R_n(f^*))+(1+c)(R_n(f^*_\eta)-R_n(f^*))\\&\leq\sup_{f\in\{f_1,\ldots,f_{N_\eta}\}}(R(f)-R(f^*)-(1+c)(R_n(f)-R_n(f^*)))+(1+c)(R_n(f^*_\eta)-R_n(f^*))\\&=\sup_{g\in\{g_1,\ldots,g_{N_\eta}\}}(Pg-(1+c)P_ng)+(1+c)(R_n(f^*_\eta)-R_n(f^*)),\end{aligned}$$

where $g_1,\ldots,g_{N_\eta}\in\mathcal{L}_{\mathcal{Y}}$ correspond to $f_1,\ldots,f_{N_\eta}$. Now we analyze the second summand separately. Using Bernstein inequality and the fact that $0\leq R(f^*_\eta)-R(f^*)\leq\eta\leq 1$

$$\begin{aligned}&P(R_n(f^*_\eta)-R_n(f^*)\geq R(f^*_\eta)-R(f^*)+\eta)\\&=P\left(R_n(f^*_\eta)-R_n(f^*)-(R(f^*_\eta)-R(f^*))\geq\eta\right)\\&\leq\exp\left(-\frac{n}{4}\left(\frac{\eta^{2-\beta}}{B}\wedge 3\eta\right)\right)\\&\leq\exp\left(-\frac{n\eta^{2-\beta}}{4B}\right).\end{aligned}$$

Given that $n>\frac{B\log(\frac{2}{\delta})}{4\eta^{2-\beta}}$ the last summand is bounded by $\frac{\delta}{2}$. With probability at least $1-\frac{\delta}{2}$ we have $(1+c)(R_n(f^*_\eta)-R_n(f^*))\leq 2(1+c)\eta$. Now denoting $\varepsilon^{\frac{1}{2-\beta}}=\eta$ we have (since the Bernstein condition holds for $g_1,\ldots,g_{N_\eta}$) that with probability at least $1-\frac{\delta}{2}$ it holds (see steps of Lemma 4)

$$\sup_{g\in\{g_1,\ldots,g_{N_\eta}\}}(Pg-(1+c)P_ng)\leq\varepsilon^{\frac{1}{2-\beta}},$$

provided that $n \gtrsim B(1+c)^2 \left(\frac{\mathcal{D}^{\text{loc}}\left(\mathcal{G}, \varepsilon^{\frac{1}{2-\beta}}/2\right)}{\varepsilon} + \frac{\log(\frac{2}{\delta})}{\varepsilon} \right)$. Using the union bound, it follows that with probability at least $1-\delta$ under the same condition on n it holds $R(\hat{f}_\eta) - R(f^*) \leq (3+2c)\varepsilon^{\frac{1}{2-\beta}}$. The claim follows. ■

Proof [of Proposition 6] We mentioned that the lower bounds, based on local entropies are well known. Thus, one may simply use standard techniques from the literature. Our proof is based on the proof of Theorem 6 in Massart and Nédélec [29], which is based itself on the application of Birge's Lemma. Let $h \in [0,1]$. We assume that X has a uniform distribution on the unit ball, which is an isotropic log-concave distribution [5]. Given $f \in \mathcal{F}$ the distribution of $Y|X$ will be defined as follows: $P^f_{Y=1|X} = \frac{1+f(X)h}{2}$. It is known [29] that for this particular distribution of $Y|X$ the class $\mathcal{L}_\mathcal{Y}$ is $(1, \frac{1}{h})$-Bernstein for any choice of P_X, moreover the model is well-specified. This will in particular mean that $R(\tilde{f}) - R(f^*) \geq 0$. Given $\varepsilon = \varepsilon_d \in [0,1]$ we want to construct a set $\mathcal{F}' \subset \mathcal{F}$, such that this set is a ε-packing of $\mathcal{F}$ intersected with the $L_1(P)$ ball or radius 2ε.

It holds $\sup_{\varepsilon \in [0,1]} \sup_{f \in \mathcal{F}} \log\left(\mathcal{N}(\mathcal{F} \cap \mathcal{B}_{L_1}(f, 2\varepsilon), \varepsilon)\right) \gtrsim d$, since otherwise in the realizable case our upper bound will contradict the lower bound [28]. However, from the symmetry of the unit ball and the fact that we have a uniform distribution it easily follows that for any $\varepsilon \in [0,1]$ and any $f \in \mathcal{F}$ it holds $\log\left(\mathcal{N}(\mathcal{F} \cap \mathcal{B}_{L_1}(f, 2\varepsilon), \varepsilon)\right) \gtrsim d$. Following the lines of Theorem 6 we have that if $\varepsilon \in [0,1]$ and $8n\frac{h^2}{1-h}\varepsilon \leq 0.71 \log(|\mathcal{F}'|)$, then $\inf_{\tilde{f}} \sup_{f^* \in \mathcal{F}} \mathbb{E}(R(\tilde{f}) - R(f^*)) \geq \frac{0.29\varepsilon h}{4}$. But since $\log(|\mathcal{F}'|) \gtrsim d$ choosing $\varepsilon \simeq \frac{d(1-h)}{nh^2}$ we have $\inf_{\tilde{f}} \sup_{f^* \in \mathcal{F}} \mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{d(1-h)}{nh}$, provided that $\varepsilon \leq 1$. The last condition holds if $h \gtrsim \sqrt{\frac{d}{n}}$. Combination of this bound with the lower bound $\frac{d}{n}$ for the realizable case [28] gives the bound of order $\frac{d}{nh}$. Finally, we set $h = \frac{1}{B}$. ■

Proof [of Proposition 7] It is known, that in our case for $g \in \mathcal{G}_Y$ it holds $Pg = \|f - f^*\|^2_{L_2(P)}$ (see, e.g., [33]), where f is the function corresponding to g. For every $g_1, g_2 \in \mathcal{G}_Y$ it holds

$$\begin{aligned}\|g_1 - g_2\|_{L_2(P)} &= \sqrt{P((f_1(X) - Y)^2 - (f_2(X) - Y)^2)^2} \\ &\leq 2\sqrt{P((f_1(X) - f_2(X))^2} \\ &= 2\|f_1 - f_2\|_{L_2(P)}.\end{aligned}$$

We repeat the lines and use the same notation as in Corollary 5. We have

$$\begin{aligned}&R(\hat{f}_\eta) - R(f^*) \\ &\leq \sup_{g \in \{g_1, \ldots, g_{N_\eta}\}} (Pg - (1+c)P_n g) + (1+c)(R_n(f^*_\eta) - R_n(f^*)),\end{aligned}$$

Notice that for $g \in \mathcal{L}_\mathcal{Y}$ the Bernstein condition holds since $Pg^2 = P((f(X) - Y)^2 - (f^*(X) - Y)^2)^2 \leq 4P(f(X) - f^*(X))^2 = 4Pg$. Next we understand that $g_1, \ldots, g_{N_\eta}$ form a 2η-cover of $\mathcal{L}_\mathcal{Y}$ (since it holds $\|g_1 - g_2\|_{L_2(P)} \leq 2\|f_1 - f_2\|_{L_2(P)}$). Now we repeat the lines of the proof of Lemma 4, but now with respect to $L_2(P)$ distance. In what following we emphasize only the differences compared to the proof of Lemma 4. We define $\mathcal{G}_\eta = \{0, g_1, \ldots, g_{N_\eta}\}$, $\mathcal{G}_0 = \mathcal{G}_\eta \cap \mathcal{B}_{L_2}(0, 4\eta)$ and

$\mathcal{G}_1 = \{0\} \cup (\mathcal{G}_\eta \setminus \mathcal{B}_{L_2}(0, 4\eta))$. By adding the zero function (which correspond to $f^* \in \mathcal{F}$) our covering numbers are changing by a small constant factor. Now

$$\sup_{g\in\mathcal{G}_\eta}(Pp[g] - (1+c)P_n p[g]) \le \sup_{g\in\mathcal{G}_0}(Pg - (1+c)P_n g) + \sup_{g\in\mathcal{G}_1}(Pg - (1+c)P_n g).$$

As before we analyze these terms one by one. Using Bernstein inequality we have that for $g \in \mathcal{G}_\eta$ it holds $P\left(Pg - P_n g > \frac{cPg}{1+c}\right) \le \exp\left(-\frac{nc^2 Pg}{32(1+c)^2}\right)$. However, since $Pg^2 \le 4Pg$ we have under $\sqrt{Pg^2} \ge 2^j\eta$

$$P\left(Pg - P_n g > \frac{cPg}{1+c}\right) \le \exp\left(-\frac{nc^2 2^{2j}\eta^2}{128(1+c)^2}\right).$$

Next we have to control the size of the subset of $\mathcal{G}_\eta$ consisting of functions with $2^j\eta \le \sqrt{Pg^2} \le 2^{j+1}\eta$. Using our relation $\|g_1 - g_2\|_{L_2(P)} \le 2\|f_1 - f_2\|_{L_2(P)}$ it is straightforward to show that we may control the sizes of these sets by the corresponding subsets of $\mathcal{F} - f^*$. Thus, as is in the proof of Lemma 4 we show that for a fixed c if $n \gtrsim \frac{\mathcal{D}^{\rm loc}_{L_2}(\mathcal{F},\eta)}{\eta^2} + \frac{\log(\frac{1}{\delta})}{\eta^2}$ we have with probability at least $1-\delta/3$ that $\sup_{g\in\mathcal{G}_1}(Pg - (1+c)P_n g) = 0$. Finally, we analyze $\sup_{g\in\mathcal{G}_0}(Pg - (1+c)P_n g)$. As before, we have $P(Pg - P_n g \ge \eta^2) \le \exp\left(-\frac{nc\eta^2}{32(1+c)^2}\right)$. And given $n \gtrsim \frac{\mathcal{D}^{\rm loc}_{L_2}(\mathcal{F},\eta)}{\eta^2} + \frac{\log(\frac{1}{\delta})}{\eta^2}$ we have with probability at least $1-\delta/3$ that $\sup_{g\in\mathcal{G}_1}(Pg - (1+c)P_n g) \le \eta^2$. Finally, since $0 \le R(f^*_\eta) - R(f^*) \le \eta^2 \le 1$

$$P(R_n(f^*_\eta) - R_n(f^*) \ge R(f^*_\eta) - R(f^*) + \eta^2) \le \exp\left(-\frac{n\eta^2}{16}\right)$$

and given $n \gtrsim \frac{\log(\frac{1}{\delta})}{\eta^2}$ the last probability is upper bounded by $\delta/3$. The proof finishes as before. ■

Proof [of Theorem 12] The in-expectation form of the bound is well-known. Under slightly different notation if follows from Lemma 2.2 in [19] or similar classic derivations in [39]. Now we continue with the proof in deviation. Given an i.i.d. sample $(X_i, Y_i)_{i=1}^n$ we define $\hat{f} = \rho(\kappa_n((X_i, Y_i)_{i=1}^n)$. We proceed by using the method of moments. For any $\varepsilon > 0$ and $p \in \mathbb{N}$ using Markov inequality we have $P(R(\hat{f}) \ge \varepsilon) \le \frac{\mathbb{E}(R(\hat{f}))^p}{\varepsilon^p}$. Following the same trick as Auer and Ortner [4] (Theorem 4) we have that $\mathbb{E}(R(\hat{f}))^p$ is equal to the expected probability that $\hat{f}$ will be wrong on p independent samples. Using the symmetrization argument, since all permutations of $n+p$ independent points have the same distributions and our sample compression scheme is permutation invariant we upper bound $\frac{\mathbb{E}(R(\hat{f}))^p}{\varepsilon^p}$ by $\frac{\psi(n,p)}{\varepsilon^p\binom{n+p}{p}}$, where $\psi(n,p)$ is maximum possible number of ways (given $(X_i, Y_i)_{i=1}^{n+p}$) to choose p out of $n+p$ points, such that $\hat{f}$ calculated on the remaining n points misclassifies these p points.

Now we bound $\psi(n,p)$ for stable sample compression schemes. We prove that $\psi(n,p) \le k^p$. For $\psi(n,1)$ the only point that is not in the learning sample must be one of the points inside that compression set of the $n+1$ points, because otherwise, $\hat{f}$ will correctly classify the remaining point. Thus, $\psi(n,1) \le k$. We do the similar argument for $\psi(n,p)$ for general p. We will enumerate the points that are inside the wrongly classified subset one by one. We choose the first element that is in the compression set of the sample of $n+p$ elements. This element is one of at most k

possible. Otherwise, if none of the elements of this set is chosen into the wrongly classified set $\hat{f}$ will correctly classify the whole sample. After the first point is chosen we have a $n+p-1$ subsample having its own compression set of size at most k. At least one of its elements must be chosen in its compression set because otherwise, $\hat{f}$ will make at most one misclassification (namely, on the point that was removed on the first stage). And so on, we simply have $\psi(n,p) \leq k^p$. Thus, we have

$$P(R(\hat{f}) \geq \varepsilon) \leq \frac{k^p}{\varepsilon^p \binom{n+p}{p}} \leq \frac{(kp)^p}{(\varepsilon n)^p}.$$

We are interested in two values of p. The first one is $p=k$. Denoting $\delta = \frac{(k^2)^k}{(\varepsilon n)^k}$ we have that with probability at least $1-\delta$ it holds

$$R(\hat{f}) \leq \frac{k^2}{n\delta^{\frac{1}{k}}}. \tag{15}$$

The second value is $p = \lceil \log(\frac{1}{\delta}) \rceil$. For $\varepsilon = \frac{ed\log(\frac{1}{\delta})}{n}$ we have that with probability at least $1-\delta$ it holds $R(\hat{f}) \leq \frac{ed\log(\frac{1}{\delta})}{n}$, so the first claim of the theorem is established.

Now we prove the bound for the modified algorithm 12. Without loss of generality we assume that we are given the sample S of size $3n$ and define $S_{1/3}, S_{2/3}$ to be the first n and $2n$ elements of S respectively. We define $\hat{f}_1 = \rho(\kappa_n(S_{1/3}))$, $\hat{f}_2 = \rho(\kappa_{2n}(S_{2/3}))$ and $\hat{f}_3 = \rho(\kappa_{3n}(S))$ and $\hat{g}$ to be the major voting over $\hat{f}_1, \hat{f}_2, \hat{f}_3$, namely $\hat{g} = \text{sign}(\hat{f}_1 + \hat{f}_2 + \hat{f}_3)$. We also denote $E_i = \{x \in \mathcal{X} : \hat{f}_i(x) \neq f^*(x)\}$. Using the same technique as in the proof of Theorem 5 in [34] we have $P(\hat{g}(X) \neq f^*(X)) \leq 3 \max\limits_{1\leq i<j\leq 3} P(E_i \cap E_j)$. So it is sufficient to control $P(E_i \cap E_j)$, as the proof will be the same. We choose without loss of generality E_1 and E_2. We have $P(E_1 \cap E_2) = P(E_2|E_1)P(E_1)$. We define $N = \sum\limits_{i=n+1}^{2n} \mathbb{1}[X_i \in E_1]$. Conditionally on $S_{1/3}$ the random variable N is binomial with mean $nP(E_1)$. Moreover, (X_i, Y_i) for $i \in n+1, \ldots, 2n$ with $X_i \in E_1$ (that are the elements in $S_{\frac{2}{3}} \cap E_1$) are conditionally independent given $S_{1/3}$.

Now we want to prove that with probability at least $1-\delta$ it holds

$$P(E_2|E_1) \lesssim \frac{k\log(N/k)}{N} + \frac{\log(\frac{1}{\delta})}{N}. \tag{16}$$

To show this we notice that due to our assumption ρ outputs classifiers from the VC class $\mathcal{F}'$ of dimension $d \lesssim k$. Thus, we may consider E_2 as an error set of an empirical risk minimizer over $\mathcal{F}'$. Using Theorem 2 in [34] we have simultaneously for all empirical minimizers $\hat{h}$ over $\mathcal{F}'$ with respect to the sample $S_{\frac{2}{3}}$ that with probability at least $1-\delta$ it holds $P(E_2) \lesssim \frac{k\log(n/k)}{n} + \frac{\log(\frac{1}{\delta})}{n}$. Since the set of all empirical minimzers with respect to the sample $S_{\frac{2}{3}}$ is the subset of the set of all empirical minimzers with respect to the sample $S_{\frac{2}{3}} \cap E_1$, applying the same Theorem 2 for the learning sample $S_{\frac{2}{3}} \cap E_1$ (given $S_{1/3}$) we obtain 16.

If $P(E_1) \geq C(\frac{k\log k}{n} + \frac{\log(\frac{1}{\delta})}{n})$ for large enough C, then using Chernoff bound for N with probability at least $1-\delta$ we have $N \geq \frac{1}{2}P(E_1)n$ and with probability at least $1-\delta$ we have $N \leq 2P(E_1)n$. If, otherwise, $P(E_1) \leq C(\frac{k\log k}{n} + \frac{\log(\frac{1}{\delta})}{n})$ then we have a desired bound since

$P(E_1 \cap E_2) \leq P(E_1)$. Finally, using monotonicity of $\log(x)/x$ we have with probability at least $1-3\delta$

$$P(E_2|E_1)P(E_1) \lesssim \frac{k\log(P(E_1)n/k)}{n} + \frac{\log(\frac{1}{\delta})}{n}.$$

Using 15 we have with probability at least $1-\delta$ over $S_{1/3}$ that $P(E_1) \lesssim \frac{k^2}{n\delta^{\frac{1}{k}}}$. Thus, with probability at least $1-4\delta$

$$P(E_1 \cap E_2) \lesssim \frac{k\log(k/\delta^{\frac{1}{k}})}{n} + \frac{\log(\frac{1}{\delta})}{n} \lesssim \frac{k\log(k)}{n} + \frac{\log(\frac{1}{\delta})}{n}.$$

The inequality 12 follows. We should note that using the same technique we may continue refining the term $k\log(k)$. For example, in the proof we may use the above bound $R(\hat{f}) \leq \frac{ed\log(\frac{1}{\delta})}{n}$ to simply obtain $R(\hat{g}) \lesssim \frac{k(1\vee\log(\log(\frac{1}{\delta})))}{n} + \frac{\log(\frac{1}{\delta})}{n}$. The term $\log(\log(\frac{1}{\delta}))$ is small for any reasonable value of δ.

Finally, we prove the statement for homogeneous sample compression schemes. We are generalizing the counting argument of Auer and Ortner [4]. As before we have to upper bound $\psi(n,p)$. Assume that all $n+p$ elements are ordered and denote this sample by S_{n+p}. Consider a function $\hat{f}$ that misclassifies exactly p elements and is constructed based on the remaining n elements. We denote this current learning sample of n elements by S_n. Consider the compression set of $n+p$ points (that is $\kappa_{n+p}(S)$) and choose the first element x_1 (the first component of the pair (x_1,y_1)) in it according to the order. For this element there are only two possibilities:

1. This element x_1 is misclassified by $\hat{f}$. In this case we will encode this element by 1.
2. This element x_1 is correctly classified by $\hat{f}$. In this case x_1 is in the compression set of the learning sample of n elements. We will be encode this x_1 by 0.

There are only two possible situations because a learning sample of n elements is the subset of the set of $n+p$ elements and due to homogeneous property its compression set contains all elements from the intersection of S_n with $\kappa(S_{n+p})$, formally $\kappa(S_{n+p}) \cap S_n \subseteq \kappa(S_n)$. Then, since x_1 is correctly classified we have $x_1 \in S_n$ and thus $x_1 \in \kappa(S_n)$

Now, after the first element x_1 is chosen we proceed to the second element x_2. There are two options: if on the first step the element x_1 was correctly classified by $\hat{f}$ we choose the next element (according to the order) in $\kappa_{n+p}(S_{n+p})$. Otherwise, if x_1 was misclassified we consider the set $S_{n+p} \setminus x_1$ and its compression set $\kappa_{n+p-1}(S_{n+p} \setminus x_1)$ and choose x_2 from this compression set (once again according to the order).

As we had for the first element x_1 we now have two options for x_2 depending on whether $\hat{f}$ classifies x_2 correctly or not. We encode x_2 by 0 or 1 depending on this and proceed analogously for x_3 as we did for x_2. Given $\hat{f}$, after at most $s \leq k+p$ steps we will encode the set of elements $x_1,\ldots,x_s$ that consists of the set $\kappa_n(S_n)$ and p misclassified elements. Finally, we easily observe that this encoding scheme relates a unique ordered sequence of at most $k+p$ zeroes and ones for every $\hat{f}$ that misclassifies exactly p elements and is constructed based on the remaining n elements. Since these classifiers make p errors there are at most $\binom{k+p}{p}$ of these ordered sequences. Thus, $\psi(n,p) \leq \binom{k+p}{p}$ and we have $P(R(\hat{f}) \geq \varepsilon) \leq \frac{\binom{k+p}{k}}{\varepsilon^p\binom{n+p}{p}}$. By choosing $p = \lceil\log(\frac{1}{\delta})\rceil$ we easily obtain that with probability at least $1-\delta$ we have $R(\hat{f}) \leq \frac{ek}{n} + \frac{e\log(\frac{1}{\delta})}{n}$. ■