Localization of VC Classes: Beyond Local Rademacher Complexities

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

Steve Hanneke STEVE.HANNEKE@GMAIL.COM

Abstract

In statistical learning the excess risk of empirical risk minimization (ERM) is controlled by $\left(\frac{\mathrm{COMP}_n(\mathcal{F})}{n}\right)^{\alpha}$, where n is a size of a learning sample, $\mathrm{COMP}_n(\mathcal{F})$ is a complexity term associated with a given class $\mathcal{F}$ and $\alpha \in [\frac{1}{2}, 1]$ interpolates between slow and fast learning rates. In this paper we introduce an alternative localization approach for binary classification that leads to a novel complexity measure: fixed points of the local empirical entropy. We show that this complexity measure gives a tight control over $\mathrm{COMP}_n(\mathcal{F})$ in the upper bounds under bounded noise. Our results are accompanied by a minimax lower bound that involves the same quantity. In particular, we practically answer the question of optimality of ERM under bounded noise for general VC classes.

Keywords: statistical learning, PAC learning, local metric entropy, local Rademacher process, shifted empirical process, offset Rademacher process, ERM, Alexander's capacity, disagreement coefficient, Massart's noise condition

1. Introduction

Since the early days of statistical learning theory understanding of the generalization abilities of empirical risk minimization has been a central question. In 1968, Vapnik and Chervonenkis [38] introduced the combinatorial property of classes of classifiers which we now call the *VC dimension*, which plays a crucial role not only in statistics but in many other areas of mathematics. By now it is strongly believed that the VC-dimension fully characterizes the properties of the empirical risk minimization algorithm. For example, when no restrictions are made on the distributions one can prove that the probability of error of the minimizer of empirical risk is close to the probability of error of the best classifier in the class, up to a term of order $\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}$, with probability at least $1-\delta$, where d is the VC dimension of the class and n is the sample size. One can also prove a minimax lower bound (valid for any learning procedure) matching up to absolute constants. But the fact that VC dimension alone describes the complexity term appears to be true only in the agnostic case, when no assumptions are made on the labelling mechanism. It was noticed several times in the literature, that when considering bounded noise, VC dimension alone is not a right complexity measure of ERM [31; 33; 19]. Until now this phenomenon was discussed only for a small amount of specific classes. In this paper we present this yet unknown combinatorial complexity measure for general VC classes.

In the last twenty years many efforts were made to understand the conditions that imply fast $\frac{1}{n}$ convergence rates, instead of slow $\frac{1}{\sqrt{n}}$ rates. By now these conditions are well understood; we refer for example to van Erven et al. [40] for an extensive survey and related results. At the beginning of the 2000s, so-called *localized* complexities (Bartlett et al. [5], Koltchinskii [23]) were introduced to statistical learning and became popular techniques for proving $\frac{1}{n}$ rates in different scenarios. But in addition to better rates, localization means that *only a small vicinity of the best classifier* really affects the learning complexity. Almost fifty years after the introduction of VC theory this phenomenon is still not fully understood and studied. Specifically, we lack tight error bounds based on localization and expressed in terms of intuitively-simple and calculable combinatorial properties of the class. Existing approaches based on localization (mainly, via *local Rademacher complexities*) are typically difficult to calculate directly, and the simpler relaxations of these bounds in the literature use localization merely to gain improvements due to the *noise conditions*, but fail to maintain the important improvements due to the local *structure of the function class* (i.e., localization of the complexity term in the bound). Moreover, in classification literature there are no known general minimax lower bounds in terms of localized processes.

There does exist one line of results which simultaneously give fast convergence rates and perform direct localization of a class of classifiers, to arrive at simple generalization bounds. Specifically, Massart and Nédélec [31] proved that under Massart's bounded noise condition, generalization of order $\frac{d}{nh}\log(\frac{nh^2}{d}) + \frac{\log(\frac{1}{\delta})}{nh}$ is possible, where h is a margin parameter responsible for the noise level. To derive this bound, Massart and Nédélec use a localized analysis to obtain improved rates under these noise conditions. However, the bound does not reflect this localization in the *complexity term* itself: in this case, the factor $d\log(\frac{nh^2}{d})$. Giné and Koltchinskii [15] refined this bound, establishing generalization of order $\frac{d}{nh}\log(\tau(\frac{d}{nh^2})) + \frac{\log(\frac{1}{\delta})}{nh}$ for empirical risk minimization, where τ is a distribution-dependent quantity they refer to as *Alexander's capacity function* (from the work of Alexander in the 80s [1]). Very recently, Hanneke and Yang [18] introduced a novel combinatorial parameter $\mathfrak{s}$, called the *star number*, which gives perfectly-tight distribution-free control on $\tau(\frac{d}{nh^2})$, and generally cannot be upper bounded in terms of the VC dimension. Thus (as noted by Hanneke [19]), in terms of distribution-free guarantees on the generalization of empirical risk minimization, the implication of Giné and Koltchinskii's result is a bound $\frac{d}{nh}\log(\mathfrak{s}\wedge\frac{nh^2}{d}) + \frac{\log(\frac{1}{\delta})}{nh}$. However, it was noted [19], that this bound is sometimes suboptimal. In this paper we will give a new argument showing potential gaps of this bound.

The aim of this paper is to perform a tight distribution-free localization for VC classes under bounded noise by introducing a new distribution-free complexity measure, thus resolving the existing gap between upper and lower bounds. The complexity measure is a localized empirical entropy measure: essentially, a fixed point of the local empirical entropy. Most of the results will be proved in expectation and in deviation. Although results in expectation can usually be derived by integrating the results in deviation, we will directly prove results in expectation in the main part of the paper. Proofs of standard technical propositions and some results in deviation will be moved to the appendix. This paper is organized as follows:

- In section 2 we introduce the notation, definitions and previous results.
- In section 3 we introduce and further develop the machinery, based on the combination of shifted empirical processes [27] and offset Rademacher complexities [29]. We also obtain a new upper bound on the error rate of empirical risk minimization in the realizable case, involving the star number and the growth function, which refines a recent result of Hanneke [19] in some cases; this bound is a strict improvement over the distribution-free bound implied by the result of Giné and Koltchinskii in the realizable case.
- Section 4 is devoted to an upper bound in terms of fixed point of global metric entropy. Although it gives a fast convergence rate $\frac{1}{n}$, it involves only a global information about the class. Thus, this bound is suboptimal in some interesting cases, as are the other bounds in the literature based solely on global complexities for the class. We include the proof nevertheless, as it cleanly illustrates certain aspects of our approach; for simplicity, we only present this result in the realizable case.
- Section 5 contains our main results. In this section we introduce the local empirical entropy and prove that fixed points of local empirical entropy control the complexity of ERM under bounded noise.
- Section 6 is devoted to a novel lower bound in terms of fixed points of local empirical entropy under mild regularity assumptions.
- Section 7 contains examples of values of fixed points for some standard classes.
- Section 8 is devoted to discussions and some related general results. Specifically, we prove that bounds based on our complexity measure are always not worse than the bounds based on local Rademacher complexities.

2. Notation and Previous Results

We define the *instance space* $\mathcal{X}$ and the *label space* $\mathcal{Y} = \{1, -1\}$. We assume that the set $\mathcal{X} \times \mathcal{Y}$ is equipped with some σ-algebra and a probability measure P on measurable subsets is defined. We also assume that we are given a set of classifiers $\mathcal{F}$; these are measurable functions with respect to the introduced σ-algebra, mapping $\mathcal{X}$ to $\mathcal{Y}$. We may always decompose $P = P_X \times P_{Y|X}$. The risk of a classifier f is its probability of error, denoted $R(f) = P(f(X) \neq Y)$. It is known that among all functions the *Bayes classifier* $f^*(x) = \operatorname{sign}(\eta(x))$, where $\eta(x) = \mathbb{E}[Y|X = x]$, minimizes the risk [11]. Symbol $\wedge$ will denote minimum of two real numbers, $\vee$ will denote maximum of two real numbers and $\mathbb{1}[A]$ will denote 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 $\mathcal{G}$. We will slightly abuse the notation and by $\log(x)$ always 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. Also denote $Z_i = (X_i, Y_i)$. By P_n we will denote expectation with respect to the empirical measure (empirical mean) induced by these samples. *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 $R_n(f) = P_n \mathbb{1}[f(X) \neq Y]$ 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. At times we also refer to a *ghost sample*, which is another n i.i.d. P-distributed samples, independent of the training sample, and we denote by P'_n the empirical mean with respect to the ghost sample. We say a set $\{x_1, \ldots, x_k\} \in \mathcal{X}^k$ is shattered by $\mathcal{F}$ if there are 2^k distinct classifications of $\{x_1, \ldots, x_k\}$ realized by classifiers in $\mathcal{F}$. The *VC dimension* of $\mathcal{F}$ is the largest integer d such that there exists a set $\{x_1, \ldots, x_d\}$ shattered by $\mathcal{F}$ [38]. We define the *growth function* $\mathcal{S}_{\mathcal{F}}(n)$ as the maximum possible number of different classifications of a set of n points realized by classifiers in $\mathcal{F}$ (maximized over the choice of the n points). Throughout the paper n will always denote the size of the training sample, d will denote the VC dimension, and $\hat{f}$ will denote the output of any ERM algorithm. To focus on nontrivial scenarios, we will always suppose $d \geq 1$.

Definition 1 (Massart and Nédélec [31]) $(P, \mathcal{F})$ *is said to satisfy Massart's bounded noise condition if* $f^* \in \mathcal{F}$ *and for some* $h \in [0, 1]$ *it holds* $|\eta(X)| \geq h$ *with probability* 1. *This constant* h *is referred to as the* margin parameter.

For any $\mathcal{F}$, the set of all corresponding distributions satisfying Massart's bounded noise condition will be denoted by $\mathcal{P}(h, \mathcal{F})$. The case $h = 1$ corresponds to the so-called *realizable case*, where $Y = f^*(X)$ almost surely, and $h = 0$ corresponds to a well-specified *agnostic* case. The following result is classic [12; 36; 7]. Let $\mathcal{F}$ be a class with VC-dimension d. For any empirical risk minimizer $\hat{f}$ over n samples, for any $P \in \mathcal{P}(0, \mathcal{F})$, with probability at least $1 - \delta$,

$$R(\hat{f}) - R(f^*) \lesssim \sqrt{\frac{d}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}.$$

Moreover, the following lower bound exists for an output $\tilde{f}$ of *any* algorithm based on n samples: there exists $P \in \mathcal{P}(0, \mathcal{F})$ such that, with probability greater than δ,

$$R(\tilde{f}) - R(f^*) \gtrsim \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right) \wedge 1.$$

Thus we know that the VC-dimension is the right complexity measure for empirical risk minimization, and indeed for optimal learning, when no restrictions are made on the probability distribution. Interestingly, this is not generally the case when $h > 0$. In this paper, we find this yet unknown essentially correct complexity measure, when h is bounded away from 0 and 1. But first, we review a refinement to the above bound for the case $h > 0$, due to Giné and Koltchinskii [15]. Specifically, consider the following definition.

Definition 2 *For $\varepsilon_0 > 0$ fix a set $\mathcal{F}_{\varepsilon_0} = \{f \in \mathcal{F} : P_X(f(X) \neq f^*(X)) \leq \varepsilon_0\}$. For $\varepsilon \in (0,1]$ define*

$$\tau(\varepsilon) = \sup_{\varepsilon_0 \geq \varepsilon} \frac{P_X\{x \in \mathcal{X} \colon \exists f \in \mathcal{F}_{\varepsilon_0} \text{ s.t. } f(x) \neq f^*(x)\}}{\varepsilon_0} \vee 1.$$

This quantity (essentially[1]) was introduced to the empirical processes literature by Alexander [1], and is referred to as *Alexander's capacity* by Giné and Koltchinskii [15]. The same quantity appeared independently in the literature on active learning, where it is referred to as the *disagreement coefficient* [16; 17]. $\tau(\varepsilon)$ is a distribution-dependent measure of the diversity of ways in which classifiers in a relatively small vicinity of f^* can disagree with f^*. Giné and Koltchinskii [15] gave the following upper bound. Let $\mathcal{F}$ be a class of VC dimension d, and $\hat{f}$ the classifier produced by an ERM based on n training samples. For any probability measure $P \in \mathcal{P}(h, \mathcal{F})$, with probability at least $1 - \delta$,

$$R(\hat{f}) - R(f^*) \lesssim \frac{d}{nh} \log\left(\tau\left(\frac{d}{nh^2}\right)\right) + \frac{\log(\frac{1}{\delta})}{nh}. \tag{1}$$

This bound is the best simple, easily calculable upper bound known so far for ERM in the case of binary classification under Massart's bounded noise condition. The proof of this bound is based on the analysis of the localized Rademacher processes. So we may also consider this result as the best relaxation of the local Rademacher analysis.

Recently, Hanneke and Yang [18] introduced a distribution-free complexity measure, called the *star number*, which perfectly captures the worst case value for Alexander's capacity. It is defined as follows.

Definition 3 *The star number $\mathfrak{s}$ is the largest integer such that there exist distinct $x_1, \ldots, x_{\mathfrak{s}} \in \mathcal{X}$ and $f_0, f_1, \ldots, f_{\mathfrak{s}} \in \mathcal{F}$ such that, for all $i \in \{1, \ldots, \mathfrak{s}\}$, DIS$(\{f_0, f_i\}) \cap \{x_1, \ldots, x_{\mathfrak{s}}\} = \{x_i\}$.*

Similar to Alexander's capacity, the star number describes how diverse the small-size disagreements with a fixed classifier f_0 can be. In terms of the one-inclusion graph studied by Haussler, Littlestone, and Warmuth [21], the star number may be described as the maximum possible degree in the data-induced one-inclusion graph. It is easy to see that, for any class of VC dimension d, it always holds that $d \leq \mathfrak{s}$, but the difference may be as large as inifinte. We refer to [18] for examples and further discussions related to the star number. One of the most interesting results about this value is its connection with the worst case of Alexander's capacity. The paper of Hanneke and Yang contains the following equality

$$\sup_{f^* \in \mathcal{F}} \sup_{P_X} \tau(\varepsilon) = \mathfrak{s} \wedge \frac{1}{\varepsilon}. \tag{2}$$

As noted by Hanneke [19], an immediate corollary of this and (1) is that, for any $P \in \mathcal{P}(h, \mathcal{F})$, with probability at least $1 - \delta$,

$$R(\hat{f}) - R(f^*) \lesssim \frac{d}{nh} \log\left(\frac{nh^2}{d} \wedge \mathfrak{s}\right) + \frac{\log(\frac{1}{\delta})}{nh}. \tag{3}$$

In particular, in the realizable case (when $h = 1$), with probability at least $1 - \delta$,

$$R(\hat{f}) \lesssim \frac{d}{n} \log\left(\frac{n}{d} \wedge \mathfrak{s}\right) + \frac{\log(\frac{1}{\delta})}{n}.$$

Since $\mathfrak{s}$ controls Alexander's capacity with equality, there is no room for any kind of improvement using the bound of Giné and Koltchinskii if we consider distribution-free upper bounds. However, the above bound for

1. The original definition did not include the supremum over ε_0, instead taking $\varepsilon_0 = \varepsilon$ directly. However, the results were proven under a very restrictive monotonicity assumption. Taking the supremum allows one to dispense with such assumptions.

the realizable case has recently been refined by Hanneke [19] in the realizable case, establishing that for any $P \in \mathcal{P}(1, \mathcal{F})$, with probability at least $1-\delta$,

$$R(\hat{f}) \lesssim \frac{d}{n} \log\left(\frac{n}{d} \wedge \frac{\mathrm{s}}{d}\right) + \frac{\log(\frac{1}{\delta})}{n}. \tag{4}$$

Even this slight improvement indicates the suboptimality of the bound (3). In this paper we will further refine this bound and discuss in details the following fact: the pair d, s alone is not a right complexity measure for the VC classes when h is bounded away from zero.

3. Preliminaries from Empirical Processes

Given a function class $\mathcal{G}$ mapping $\mathcal{Z}$ to $\mathbb{R}$, one may consider the supremum of the empirical process:

$$\sup_{g \in \mathcal{G}} (P - P_n)\, g.$$

This quantity plays an important role in statistical learning theory. Since the pioneering paper of Vapnik and Chervonenkis [38], the analysis of learning algorithms is usually performed by the tight uniform control over the process $(P - P_n)\, g$ for a special class of functions. The behaviour of the supremum of this empirical process is tightly connected with the supremum of the so-called *Rademacher process*:

$$\frac{1}{n}\mathbb{E}_{\varepsilon} \sup_{g \in \mathcal{G}} \left(\sum_{i=1}^{n} \varepsilon_i g_i\right),$$

where g_i denotes $g(Z_i)$, ε_i are independent Rademacher variables taking values ± 1 with equal probabilities, and $\mathbb{E}_{\varepsilon}$ denoted the expectation over the ε_i random variables (conditioning on the Z_i variables). This approach, however, usually leads to suboptimal upper and lower bounds that are not capturing both improved learning rates due to the noise conditions and the localization of the complexity term.

We will instead consider different quantities, so-called *shifted empirical* processes, introduced by Lecué and Mitchell [27]. Given $c > 0$, we consider

$$\sup_{g \in \mathcal{G}} (P - (1+c)P_n)\, g.$$

The second important quantity is an expected supremum of the *offset Rademacher process*, introduced recently by Liang, Rakhlin, and Sridharan [29]:

$$\frac{1}{n}\mathbb{E}_{\varepsilon} \sup_{g \in \mathcal{G}} \left(\sum_{i=1}^{n} \varepsilon_i g_i - c' g_i^2\right).$$

The last quantity was introduced for the analysis of a specific aggregation procedure under the square loss and so far has not been related to a shifted process [2]. In this paper, we will investigate some new properties of these processes and show how they may be applied in the classification framework. The following short lemma appears in a more general form in [29] (Lemma 5).

Lemma 4 *Let $V \subset \{0,1\}^n$ be a finite set of binary vectors of cardinality N. Then for any $c > 0$,*

$$\frac{1}{n}\mathbb{E}_{\varepsilon} \max_{v \in V} \left(\sum_{i=1}^{n} \varepsilon_i v_i - c v_i\right) \leq \frac{1}{2c}\frac{\log(N)}{n}.$$

2. We should note that shifted processes and related techniques appeared independently earlier in the paper of Wegkamp [42]. He uses the term *desymmetrized* empirical processes for the shifted processes.

Compare this result with an upper bound for Rademacher averages [7] where the best rate is of order $\sqrt{\frac{\log(N)}{n}}$. The next simple lemma is a new symmetrization lemma for the shifted process in expectation.

Lemma 5 (Shifted symmetrization in expectation) *Let $\mathcal{G}$ be a functional class and $c \geq 0$ an absolute constant. Then*

$$\mathbb{E}\sup_{g\in\mathcal{G}}((P-(1+c)P_n)g) \leq \frac{c+2}{n}\mathbb{E}\mathbb{E}_{\varepsilon}\sup_{g\in\mathcal{G}}\left(\sum_{i=1}^{n}\varepsilon_i g(Z_i) - \frac{c}{c+2}g(Z_i)\right).$$

Proof Proof technique is inspired by the proof of Theorem 3 in [29]. Using standard symmetrization trick and Jensen's inequality we have

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathcal{G}}((P-(1+c)P_n)g)\\
&\leq \mathbb{E}\sup_{g\in\mathcal{G}}\left(P'_n g-(1+c)P_n g\right)\\
&= \mathbb{E}\sup_{g\in\mathcal{G}}((1+c/2)(P'_n g-P_n g)-cP'_n g/2-cP_n g/2)\\
&\leq 2\mathbb{E}\mathbb{E}_{\varepsilon}\sup_{g\in\mathcal{G}}\left(\frac{1+c/2}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)-cP_n g/2\right)\\
&= 2(1+c/2)\mathbb{E}\mathbb{E}_{\varepsilon}\sup_{g\in\mathcal{G}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(Z_i)-\frac{c/2}{1+c/2}P_n g\right).
\end{aligned}$$

■

Interestingly, by setting $c = 0$ we immediately obtain the standard symmetrization inequality. The next lemma, which provides a novel symmetrization tool for the shifted processes in deviation requires the following definition. This result is motivated by existing classic symmetrization results [7; 38], but the proof technique is adapted for our shifted case. We say that a functional class $\mathcal{G}$ is a (B,β)-*Bernstein class* if for any $g \in \mathcal{G}$ we have $Pg^2 \leq B\,(Pg)^{\beta}$. The parameter β is called the *Bernstein parametr* and B the *Bernstein constant*.

Lemma 6 (Shifted symmetrization in deviation) *Let $\mathcal{G}$ be a $(B,1)$-Bernstein class, such that for all $g \in \mathcal{G}$ we have $|g| \leq 1$ and $Pg \geq 0$. Fix constants $c_1, c_2 > 0$, such that $c_2 < c_1$ and $\frac{c_2}{3(1+c_2)} \leq B$. Then if $nt \geq 2B\log(2)\frac{(1+c_2)^2}{c_2}$*

$$P\left(\sup_{g\in\mathcal{G}}(P-(1+c_1)P_n)g \geq t\right) \leq 2P\left(\sup_{g\in\mathcal{G}}((1+c_2)P'_n-(1+c_1)P_n)g \geq t/2\right).$$

Proof Given a random sample let $\tilde{g}$ be the function achieving the supremum.

$$\mathbb{1}[(P-(1+c_1)P_n)\tilde{g} > t]\mathbb{1}[(P-(1+c_2)P'_n)\tilde{g} < t/2] \leq \mathbb{1}[((1+c_2)P'_n-(1+c_1)P_n)\tilde{g} > t/2].$$

Taking expectation with respect to the ghost sample we have

$$\mathbb{1}[(P-(1+c_1)P_n)\tilde{g} > t]P'[(P-(1+c_2)P'_n)\tilde{g} < t/2] \leq P'[((1+c_2)P'_n-(1+c_1)P_n)\tilde{g} > t/2].$$

We further have

$$P'\Big[(P-(1+c_2)P'_n)\tilde{g} \geq t/2\Big] = P'\Big[(P-P'_n)\tilde{g} \geq \frac{t/2+c_2P\tilde{g}}{1+c_2}\Big].$$

Using the Bernstein bound [8] we have

$$P'\left[(P - P'_n)\tilde{g} \geq \frac{t/2 + c_2 P\tilde{g}}{1 + c_2}\right] \leq \exp\left(-\frac{n}{2}\left(\frac{t/2 + c_2 P\tilde{g}}{1 + c_2}\right)^2 / \left(P\tilde{g}^2 + \frac{1}{3}\frac{t/2 + c_2 P\tilde{g}}{1 + c_2}\right)\right).$$

Then using Bernstein condition we have

$$\left(\frac{t/2 + c_2 P\tilde{g}}{1 + c_2}\right)^2 / \left(P\tilde{g}^2 + \frac{1}{3}\frac{t/2 + c_2 P\tilde{g}}{1 + c_2}\right) \geq \left(\frac{t/2 + c_2 P\tilde{g}}{1 + c_2}\right)^2 / \left(BP\tilde{g} + \frac{1}{3}\frac{t/2 + c_2 P\tilde{g}}{1 + c_2}\right).$$

A simple analysis shows that if $B \geq \frac{c_2}{3(1+c_2)}$, then

$$\left(\frac{t/2 + c_2 P\tilde{g}}{1 + c_2}\right)^2 / \left(BP\tilde{g} + \frac{1}{3}\frac{t/2 + c_2 P\tilde{g}}{1 + c_2}\right) \geq \frac{18Btc_2}{(3B(1 + c_2) + c_2)^2} \geq \frac{tc_2}{2B(1 + c_2)^2}.$$

Finally, we have that if $\frac{ntc_2}{2B(1+c_2)^2} \geq \log(2)$, then $P'[(P - (1 + c_2)P'_n)\tilde{g} < t/2] \geq \frac{1}{2}$. Taking an expectation with respect to the initial sample finishes the proof. ■

Let $\mathbf{s}$ be the star number of a class of binary classifiers $\mathcal{F}$. Hanneke [19] recently proved that in this case

$$\mathbb{E}P_X(\mathrm{DIS}(\mathcal{V}_n)) \leq \frac{\mathbf{s}}{n+1}, \tag{5}$$

where $\mathcal{V}_n = \{f \in \mathcal{F} | P_n[f(X) \neq f^*(X)] = 0\}$ is the *version space*. That work also established a similar result holding with high probability: with probability at least $1 - \delta$,

$$P_X(\mathrm{DIS}(\mathcal{V}_n)) \leq \frac{21\mathbf{s}}{n} + \frac{16\log(\frac{3}{\delta})}{n}. \tag{6}$$

This result means that if the star number is bounded, then in the realizable case the expected measure of disagreement of the version space has order $\frac{\mathbf{s}}{n}$, where n is the size of the learning sample. A reader familiar with the work of Haussler, Littlestone, and Warmuth [21] may remember that the performance of some learning algorithms can be controlled by the maximum possible out-degree in a corresponding orientation of the data-induced *one-inclusion graph*, and that there exists such an orientation with maximum out-degree at most the VC dimension. The above result has relations to this, except that instead of the out-degree of an oriented one-inclusion graph, the measure of the region of disagreement is controlled by the largest possible value of the (undirected) degree of the data-induced one-inclusion graph.

Since both the ERM and the optimal classifier are contained in $\mathcal{V}_n$ in the realizable case, one consequence of the above results is that, when $\mathbf{s} \approx d$, ERM achieves the optimal order d/n in its error rate. Even more interesting, [19] used the bound (6) in a more subtle way to show that ERM in the realizable case obtains expected error rate of order $\frac{d}{n}\log\frac{n\wedge\mathbf{s}}{d}$, and with probability at least $1 - \delta$ has error rate bounded as in (4): i.e., of order $\frac{d}{n}\log\frac{n\wedge\mathbf{s}}{d} + \frac{1}{n}\log\frac{1}{\delta}$. Via a more sophisticated variant of this argument, we obtain the following theorem, which is one of the novel contributions of this work. It offers interesting general refinements over (4) which we discuss below. Its proof is included in the appendix.

Theorem 7 *Let* $\mathbf{s}$ *be the star number of a class of binary classifiers* $\mathcal{F}$*. In the realizable case, any ERM* $\hat{f}$ *has*

$$\mathbb{E}R(\hat{f}) \lesssim \frac{\log\left(\mathcal{S}_{\mathcal{F}}\left(\mathbf{s} \wedge n\right)\right)}{n}.$$

Moreover, with probability at least $1 - \delta$*,*

$$R(\hat{f}) \lesssim \frac{\log\left(\mathcal{S}_{\mathcal{F}}\left(\mathbf{s} \wedge n\right)\right)}{n} + \frac{\log(\frac{1}{\delta})}{n}.$$

We may prove (due to Vapnik and Chervonenkis's bound on the growth function [38]) that this inequality is an alternative way of recovering the upper bound $\frac{d\log\left(\frac{\mathbf{s}\wedge n}{d}\right)}{n}$ proven by [19] for ERM, which is itself a refinement of the distribution-free bound (3) implied by Giné and Koltchinskii's bound (1) in the realizable case.

Example 1 *Theorem 7 yields simple examples showing the gaps in the distribution-free bound (3) in the realizable case. Specifically, suppose $\mathcal{X} = \{x_1, \ldots, x_{\mathbf{s}}\}$, define class $\mathcal{F}_1$ as the classifiers on this $\mathcal{X}$ with at most d points classified 1, and class $\mathcal{F}_2$ as the classifiers having at most $d-1$ points classified 1 among $\{x_1, \ldots, x_{d-1}\}$ and at most one point classified 1 among $\{x_d, \ldots, x_{\mathbf{s}}\}$. For both $\mathcal{F}_1$ and $\mathcal{F}_2$, the VC dimension is d and the star number is $\mathbf{s}$. However, for $\mathcal{F}_1$ Theorem 7 gives a bound of order $\frac{d\log\left(\frac{\mathbf{s}\wedge n}{d}\right)}{n}$, but for $\mathcal{F}_2$ it gives a smaller bound of order $\frac{d+\log(\mathbf{s}\wedge n)}{n}$. In both cases, these are known to be tight characterizations of ERM in the realizable case [21; 19]. It should be noted, however, that one can also construct examples where Theorem 7 is itself not tight.*

4. Bounds in Terms of a Global Packing

The main aim of this section is to give a simple bound in terms of a fixed point of global packings. We will further significantly improve this result in the next section, and therefore for simplicity here we will consider only the realizable case. We note that a similar result may be derived from classic results on ratio type empirical processes (see Section 19.6 of [2]). We include the details of our proof here anyway, as it also serves to illustrate certain aspects of our approach in simplified form.

Given a set of n points we define for any two $f, g \in \mathcal{F}$ where $\rho_H(f,g) = |\{i \in \{1,\ldots,n\} : f(x_i) \neq g(x_i)\}|$. We further introduce

$$\mathcal{M}_1^*(\mathcal{F},\gamma,n) = \max_{x_1,\ldots,x_n\in\mathcal{X}} \mathcal{M}_1(\mathcal{F}(\{x_1,\ldots,x_n\}),\gamma),$$

where $\mathcal{M}_1(\mathcal{H},\varepsilon)$ denotes the size of a maximal ε-packing of $\mathcal{H}$ under ρ_H distance (for the given $x_1,\ldots,x_n$ points) and $\mathcal{F}(\{x_1,\ldots,x_n\})$ is a set of projections of $\mathcal{F}$ on $\{x_1,\ldots,x_n\}$.

In many statistical frameworks optimal rates are usually obtained when one carefully balances the radius and the logarithm of a packing number with respect to the same radius (for example, Yang and Barron [43]). It will be shown that in our bounds it is natural to choose γ such that $c\gamma \approx \log(\mathcal{M}_1^*(\mathcal{F},\gamma,n))$ for some $c \in [0,1]$. So we define

$$\gamma_c^*(n,\mathcal{F}) = \max\{\gamma \in \mathbb{N} : c\gamma \leq \log\left(\mathcal{M}_1^*(\mathcal{F},\gamma,n)\right)\}.$$

The value $\gamma_c^*(n,\mathcal{F})$ will be referred to as a *fixed point of empirical entropy*. When $\mathcal{F}$ is clear from the context, we simply write $\gamma_c^*(n)$ instead of $\gamma_c^*(n,\mathcal{F})$. Note that $\gamma_c^*(n,\mathcal{F})$ is a well-defined strictly positive-valued quantity, since we are using the truncated logarithm.

Proposition 8 *Fix any function class $\mathcal{F}$; denote its VC dimension d. If $P \in \mathcal{P}(1,\mathcal{F})$ (realizable case), then for any ERM $\hat{f}$,*

$$\mathbb{E}R(\hat{f}) \lesssim \frac{\gamma_{\frac{1}{2}}^*(n)}{n}.$$

Moreover with probability at least $1-\delta$,

$$R(\hat{f}) \lesssim \frac{\gamma_{\frac{1}{2}}^*(n)}{n} + \frac{\log\frac{1}{\delta}}{n},$$

and

$$\gamma_{\frac{1}{2}}^*(n) \lesssim d\log(n/d). \tag{7}$$

To prove this proposition we need a technical lemma, which may be considered as a modification of Lemma 6 in [29].

Lemma 9 *Let $\mathcal{G}$ be a set of functions taking binary values, and let $c \in [0,1]$ be a constant. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables. Then*

$$\frac{1}{n}\mathbb{E}_{\varepsilon} \max_{g \in \mathcal{G}} \left(\sum_{i=1}^{n} \varepsilon_i g(X_i) - c g(X_i) \right) \leq \frac{7\gamma_c^*(n)}{n}.$$

Proof Given $X_1, \dots, X_n$, let $V = \{(g(X_1), \dots, g(X_n)) : g \in \mathcal{G}\}$ denote the set of binary vectors corresponding to the values of functions in $\mathcal{G}$. As above, for a fixed γ and fixed minimal γ-covering subset $\mathcal{N}_\gamma \subseteq V$, for each $v \in V$, $p(v)$ will denote the closest vector to v in $\mathcal{N}_\gamma$. First we follow the decomposition proposed by Liang, Rakhlin, and Sridharan [29]:

$$\frac{1}{n}\mathbb{E}_{\varepsilon} \max_{v \in V} \left(\sum_{i=1}^{n} \varepsilon_i v_i - c v_i \right)$$
$$\leq \frac{1}{n}\left(\mathbb{E}_{\varepsilon} \max_{v \in V} \left(\sum_{i=1}^{n} \varepsilon_i \left(v_i - p(v)_i\right) \right) + \max_{v \in V} \left(\sum_{i=1}^{n} \frac{c}{4} p(v)_i - c v_i \right) + \mathbb{E}_{\varepsilon} \max_{v \in V} \left(\sum_{i=1}^{n} \varepsilon_i p(v)_i - \frac{c}{4} p(v)_i \right) \right).$$

Since $p(v)$ is within Hamming distance γ of v, we know $\sum_{i=1}^{n} p(v)_i \leq \gamma + \sum_{i=1}^{n} v_i$, and therefore the second summand in the above expression is at most $\max\limits_{v \in V} \left(\frac{c}{4}\gamma - \frac{3c}{4} \sum_{i=1}^{n} v_i \right) \leq \frac{c}{4}\gamma$. The third summand is upper bounded by $\frac{2}{c}\log(\mathcal{N}_1(\mathcal{F}, \gamma))$ by Lemma 4 and the first term is upper bounded by γ by the γ-cover property of the $p(v)$ vectors. Then we use the standard relation that the size a of minimal covering is less than or equal to the size of a maximal packing [10] to conclude that

$$\frac{1}{n}\mathbb{E}_{\varepsilon} \max_{v \in V} \left(\sum_{i=1}^{n} \varepsilon_i v_i - c v_i \right) \leq \frac{(1 + c/4)\gamma}{n} + \frac{2}{c}\frac{\log(\mathcal{M}_1(V, \gamma))}{n}.$$

By choosing $\gamma = \gamma_c^*(n) + 1$ we have

$$\frac{(1 + c/4)\gamma}{n} + \frac{2}{c}\frac{\log(\mathcal{M}_1(V, \gamma))}{n} \leq \frac{(1 + c/4)(\gamma_c^*(n) + 1)}{n} + \frac{2(\gamma_c^*(n) + 1)}{n} \leq \frac{7\gamma_c^*(n)}{n}.$$

■

Proof [Proposition 8] First we introduce a *loss class* $\mathcal{G}_{f^*} = \{x \to \mathbb{1}[f(x) \neq f^*(x)] \text{ for } f \in \mathcal{F}\}$. Let $\hat{f}$ be any ERM and $\hat{g}$ be a corresponding function in the loss class $\mathcal{G}_{f^*}$. We obviously have $\mathbb{E}R(\hat{f}) = P\hat{g}$ and $P_n\hat{g} = 0$. Then for any $c > 0$

$$\mathbb{E}R(\hat{f}) = \mathbb{E}(R(\hat{f}) - (1 + c)R_n(\hat{f})) \leq \mathbb{E} \sup_{g \in \mathcal{G}_{f^*}} (Pg - (1 + c)P_n g).$$

By Lemma 5 we have

$$\mathbb{E} \sup_{g \in \mathcal{G}_{f^*}} (Pg - (1 + c)P_n g) \leq \frac{c + 2}{n}\mathbb{E}\mathbb{E}_{\varepsilon} \sup_{g \in \mathcal{G}_{\mathcal{Y}}} \left(\sum_{i=1}^{n} \varepsilon_i g(X_i) - \frac{c}{c + 2} g(X_i) \right)$$

Applying the Lemma 9 and fixing $c = 2$ we finish the proof of the bound on the expectation. ■

Example 2 *Consider the class of threshold classifiers, that is $\mathcal{F} = \{x \to 2\mathbb{1}[x \le t] - 1 : t \in \mathbb{R}\}$. Using the definition of the star number it easy to see that it is equal to 2 in this case and Theorem 7 gives an optimal $\frac{1}{n}$ upper bound for ERM. At the same time the worst case packing numbers $\mathcal{M}_1^*(\mathcal{F}, \gamma, n)$ are of order $\frac{n}{\gamma}$. A simple analysis of the fixed point gives us $\gamma^*_{\frac{1}{2}}(n) \simeq \log(n)$ and thus Proposition 8 will give us suboptimal $\frac{\log n}{n}$ distribution free upper bound. Although we captured that the rate is faster than $\frac{1}{\sqrt{n}}$, our analysis of the complexity term is suboptimal. The next section discusses a correction for this, which also yields optimal rates under moderate bounded noise in general.*

5. Local Metric Entropy

This section presents our main result. Toward this end, we introduce a new complexity measure: the *worst-case local empirical packing numbers*. Given a set of n points we fix some $f \in \mathcal{F}$ and construct a Hamming ball of the radius γ. So, $\mathcal{B}_H(f, \gamma, \{x_1, \ldots, x_n\}) = \{g \in \mathcal{F} | \rho_H(f, g) \le \gamma\}$ and define

$$\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, h) = \max_{x_1,\ldots,x_n} \max_{f \in \mathcal{F}} \max_{\varepsilon \ge \gamma} \mathcal{M}_1(\mathcal{B}_H(f, \varepsilon/h, \{x_1, \ldots, x_n\}), \varepsilon/2), \tag{8}$$

where once again $\mathcal{M}_1(\mathcal{H}, \varepsilon)$ denotes the size of a maximal ε-packing of $\mathcal{H}$ under ρ_H distance (for the given $x_1, \ldots, x_n$ points). Fix any $h, h' \in (0, 1]$ and define

$$\gamma^{\text{loc}}_{h,h'}(n, \mathcal{F}) = \max\{\gamma \in \mathbb{N} : h\gamma \le \log(\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, h'))\}.$$

When $\mathcal{F}$ is clear from the context, we simply write $\gamma^{\text{loc}}_{h,h'}(n)$ instead of $\gamma^{\text{loc}}_{h,h'}(n, \mathcal{F})$. The quantity $\gamma^{\text{loc}}_{h,h'}(n)$ defines the *fixed point of a local empirical entropy*. We note that, because $1 \le d < \infty$ in this work, when $h, h' > 0$ the set on the right in this definition is finite and nonempty, so that $\gamma^{\text{loc}}_{h,h'}(n)$ is a well-defined strictly-positive integer. Indeed, for any $h, h' \in (0, 1]$, the value $\gamma = \lfloor \frac{1}{h} \rfloor$ satisfies $h\gamma \le 1$, so that (because $\log(\cdot)$ is the truncated logarithm) this γ is contained in the set; in particular, this implies $h\gamma^{\text{loc}}_{h,h'}(n, \mathcal{F}) \ge h\lfloor \frac{1}{h} \rfloor \ge \frac{1}{2}$ always.

The next theorem is the main upper bound of this paper.

Theorem 10 *Fix any function class $\mathcal{F}$; denote its VC dimension d and star number $\mathfrak{s}$. Fix any $h \in \left(\sqrt{\frac{d}{n}}, 1\right]$. If $P \in \mathcal{P}(h, \mathcal{F})$, then for any ERM $\hat{f}$,*

$$\mathbb{E}(R(\hat{f}) - R(f^*)) \lesssim \frac{\gamma^{\text{loc}}_{h,h}(n)}{n}. \tag{9}$$

Also, with probability at least $1 - \delta$,

$$R(\hat{f}) - R(f^*) \lesssim \frac{\gamma^{\text{loc}}_{h,h}(n)}{n} + \frac{\log(\frac{1}{\delta})}{nh}. \tag{10}$$

Moreover

$$\frac{d + \log\left(nh^2 \wedge \mathfrak{s}\right)}{h} \lesssim \gamma^{\text{loc}}_{h,h}(n) \lesssim \frac{d\log\left(\frac{nh^2}{d} \wedge \mathfrak{s}\right)}{h} + \frac{d\log\left(\frac{1}{h}\right)}{h}. \tag{11}$$

where d is the VC dimension of $\mathcal{F}$ and $\mathfrak{s}$ is its star number.

Our complexity term (11) is not worse than the distribution-free upper bound (3) implied by the bound (1) of Giné and Koltchinskii when h is bounded from 0 by a constant. In the last section we will discuss potential suboptimality when h is small, due to the term $\frac{d\log\left(\frac{1}{h}\right)}{h}$ in (11). Another interesting property is that the bounds (9) and (10) involve neither the VC dimension nor the star number explicitly. At the same time one can control the complexity term with both of them from below and above.

For any given $f \in \mathcal{F}$, denote $g_f(x, y) = \mathbb{1}[f(x) \ne y] - \mathbb{1}[f^*(x) \ne y]$. Consider the *excess loss class* $\mathcal{G}_{\mathcal{Y}} = \{g_f | f \in \mathcal{F}\}$, and the previously introduced class $\mathcal{G}_{f^*} = \{x \to \mathbb{1}[f(x) \ne f^*(x)]$ for $f \in \mathcal{F}\}$, which may be interpreted as an excess loss class in the realizable case. The following properties are well known.

1. For any $g_f \in \mathcal{G}_\mathcal{Y}$ it holds $g_f^2(x,y) = \mathbb{1}[f(x) \neq f^*(x)] = \frac{1}{2}|f(x) - f^*(x)| = \frac{1}{4}(f(x) - f^*(x))^2$.
2. For any $g_f \in \mathcal{G}_\mathcal{Y}$ it holds $g_f(x,y) = \frac{y(f^*(x) - f(x))}{2}$.
3. For any $P \in \mathcal{P}(h, \mathcal{F})$ the class $\mathcal{G}_\mathcal{Y}$ is a $(\frac{1}{h}, 1)$-Bernstein class [7] and $R(f^*) \leq \frac{1}{2}(1-h)$ [11].

Lemma 11 (Contraction) *Let $\mathcal{G}_\mathcal{Y}$ be an excess loss class associated with a given class $\mathcal{F}$, and fix any $h \in [0,1]$. For any $c \in [0,1]$ and any distribution $P \in \mathcal{P}(h, \mathcal{F})$ we have*

$$\mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g(X_i, Y_i) - c g(X_i, Y_i) \right) \leq \frac{5}{4} \mathbb{E}\mathbb{E}_\xi \sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^n \xi_i g'(X_i) - \frac{4}{5} h c g'(X_i) \right),$$

where $\xi_1, \ldots, \xi_n$ are random variables conditionally independent given $X_1, \ldots, X_n$, with $|\xi_i| \lesssim 1$, and with $\mathbb{E}[\xi_i | X_1, \ldots, X_n] = 0$ and $\mathbb{E}[\exp(\lambda \xi_i) | X_1, \ldots, X_n] \leq \exp(\frac{\lambda^2}{2})$ for all λ.

Proof First we notice that any $g \in \mathcal{G}_\mathcal{Y}$ may be defined by some $f \in \mathcal{F}$.

$$\begin{aligned}
&\mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g(X_i, Y_i) - c g(X_i, Y_i) \right) \\
&= \mathbb{E}\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^n \frac{1}{2} \varepsilon_i Y_i (f(X_i) - f^*(X_i)) - c g_f(X_i, Y_i) \right) \\
&= \mathbb{E}\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^n \frac{1}{2} \varepsilon_i (f(X_i) - f^*(X_i)) - c g_f(X_i, Y_i) \right) \\
&= \frac{1}{4} \mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g^2(X_i, Y_i) - 4 c g(X_i, Y_i) \right).
\end{aligned}$$

Now consider the term $-\sum_{i=1}^n g(X_i, Y_i)$. Denoting $h_i' = 1 - 2P(f^*(X_i) \neq Y_i | X_i)$ (an X_i-dependent random variable), we know that $1 \geq h_i' \geq h$ almost surely. Furthermore, the event that $f^*(X_i) \neq Y_i$ has conditional probability (given X_i) equal $\frac{1}{2}(1 - h_i')$, and on this event we have $g^2(X_i, Y_i) = -g(X_i, Y_i)$. Similarly, the event that $f^*(X_i) = Y_i$ occurs with conditional probability (given X_i) equal $\frac{1}{2}(1 + h_i')$, and on this event we have $g^2(X_i, Y_i) = g(X_i, Y_i)$. Thus, defining $\xi_i^{(h')} = h_i' + \mathbb{1}[f^*(X_i) \neq Y_i] - \mathbb{1}[f^*(X_i) = Y_i]$, these $\xi_1^{(h')}, \ldots, \xi_n^{(h')}$ random variables are conditionally independent given $X_1, \ldots, X_n$, with $\mathbb{E}[\xi_i^{(h')} | X_1, \ldots, X_n] = 0$. In particular, if $h_i' = 0$ for all i, these are Rademacher random variables, while if $h_i' = 1$ these random variables are equal to 0 with probability 1. Now note that, by the above reasoning about these events,

$$\begin{aligned}
-\sum_{i=1}^n g(X_i, Y_i) &= -\sum_{i=1}^n h_i' g^2(X_i, Y_i) + \sum_{i=1}^n \xi_i^{(h')} g^2(X_i, Y_i) \\
&\leq -(\min_i h_i') \sum_{i=1}^n g^2(X_i, Y_i) + \sum_{i=1}^n \xi_i^{(h')} g^2(X_i, Y_i).
\end{aligned}$$

Therefore, denoting $\xi_i' = \varepsilon_i + 4c\xi_i^{(h')}$ (which are also conditionally independent over i given $X_1, \dots, X_n$) and using the fact that $h \le h_i'$ almost surely, we have

$$\begin{aligned}
&\frac{1}{4}\mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g^2(X_i, Y_i) - 4cg(X_i, Y_i) \right) \\
\le &\frac{1}{4}\mathbb{E}\mathbb{E}_{\xi'} \sup_{g \in \mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \xi_i' g^2(X_i, Y_i) - 4hcg^2(X_i, Y_i) \right) \\
= &\frac{1}{4}\mathbb{E}_X\mathbb{E}_{\xi'} \sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^n \xi_i' g'(X_i) - 4hcg'(X_i) \right).
\end{aligned}$$

Finally, because ε_i and $\xi_i^{(h')}$ both have zero conditional mean, so does ξ_i', and since we also have $-5+4ch_i' \le \xi_i' \le 5 + 4ch_i'$, Hoeffding's lemma ([11] Lemma 8.1) implies $\mathbb{E}[\exp(\lambda\xi_i')|X_1, \dots, X_n] \le \exp(25\lambda^2/2)$. The lemma easily follows, taking $\xi_i = \xi_i'/5$. ■

Lemma 12 (Localization) *Let $\mathcal{G}$ be a set of functions taking binary values, containing the zero function, and let $c \in [0, \frac{1}{4}]$ be a constant. Let $\xi_1, \dots, \xi_n$ be any random variables conditionally independent given $X_1, \dots, X_n$, with $|\xi_i| \lesssim 1$, and with $\mathbb{E}[\xi_i|X_1, \dots, X_n] = 0$ and $\mathbb{E}[\exp(\lambda\xi_i)|X_1, \dots, X_n] \le \exp(\frac{\lambda^2}{2})$ for all λ. Then*

$$\frac{1}{n}\mathbb{E}\sup_{g \in \mathcal{G}} \left(\sum_{i=1}^n \xi_i g(X_i) - 4cg(X_i) \right) \lesssim \frac{\gamma_{c,c}^{\text{loc}}(n, \mathcal{G})}{n}.$$

Proof This proof is deferred to the appendix. ■

Proof [Theorem 10] Let $\hat{f}$ be an ERM and $\hat{g}$ be a corresponding function in the excess loss class $\mathcal{G}_\mathcal{Y}$. We obviously have $\mathbb{E}(R(\hat{f}) - R(f^*)) = \mathbb{E}P\hat{g}$ and $P_n\hat{g} \le 0$. Then for any $c > 0$,

$$\mathbb{E}(R(\hat{f}) - R(f^*)) \le \mathbb{E}(P\hat{g} - (1+c)P_n\hat{g}) \le \mathbb{E} \sup_{g \in \mathcal{G}_\mathcal{Y}} (Pg - (1+c)P_n g).$$

Now using the symmetrization lemma (Lemma 5) we have

$$\mathbb{E} \sup_{g \in \mathcal{G}_\mathcal{Y}} (Pg - (1+c)P_n g) \le \frac{c+2}{n}\mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g(X_i, Y_i) - \frac{c}{c+2} g(X_i, Y_i) \right).$$

Applying the contraction lemma (Lemma 11)

$$\begin{aligned}
&\frac{c+2}{n}\mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_\mathcal{Y}} \left(\sum_{i=1}^n \varepsilon_i g(X_i, Y_i) - \frac{c}{c+2} g(X_i, Y_i) \right) \\
\le &\frac{5(c+2)}{4n}\mathbb{E}\mathbb{E}_\xi \sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^n \xi_i g'(X_i) - \frac{4ch}{5(c+2)} g'(X_i) \right).
\end{aligned}$$

We are ready to apply the localization lemma (Lemma 12). The conditions on the ξ_i variables required for Lemma 12 are supplied by Lemma 11, and all functions in $\mathcal{G}_{f^*}$ take only binary values. Thus, for a fixed c,

$$\frac{5(c+2)}{4n}\mathbb{E}\mathbb{E}_\xi \sup_{g \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^n \xi_i g'(X_i) - \frac{4ch}{5(c+2)} g'(X_i) \right) \lesssim \frac{\gamma_{h,h}^{\text{loc}}(n)}{n}.$$

The proof of the deviation bound is analogous, and is presented in the appendix. The claimed bounds on $\gamma_{h,h}^{\text{loc}}(n)$ are established in Proposition 13 below. ■

The following proposition finishes the proof of Theorem 10.

Proposition 13 *Let d be the VC-dimension and $\mathbf{s}$ be the star number of $\mathcal{F}$. For any $h \in (0, 1]$, it holds*

$$\frac{d + \log\left(nh^2 \wedge \mathbf{s}\right)}{h} \wedge \sqrt{dn} \lesssim \gamma^{\mathrm{loc}}_{h,h}(n) \lesssim \frac{d \log\left(\frac{nh^2}{d} \wedge \mathbf{s}\right)}{h} + \frac{d\log(\frac{1}{h})}{h}.$$

Proof The first part of the proof closely follows the proof of Theorem 17 in [18], with slight modifications, to arrive at an upper bound on $\mathcal{M}^{\mathrm{loc}}_1(\mathcal{F}, \gamma, n, h)$. The suprema in the definition of local empirical entropy are achieved at some set $\{x_1, \ldots, x_n\}$, some function $f \in \mathcal{F}$, and some $\varepsilon \in [\gamma, n]$. Letting $r = \varepsilon/n$, denote by $\mathcal{M}_r$ the maximal $(rn/2)$-packing (under ρ_H) of $\mathcal{B}_H(f, rn/h, \{x_1, \ldots, x_n\})$, so that $|\mathcal{M}_r| = \mathcal{M}^{\mathrm{loc}}_1(\mathcal{F}, \gamma, n, h)$. Also introduce a uniform probability measure P_X on $\{x_1, \ldots, x_n\}$ and fix $m = \left\lceil \frac{4}{r}\log(|\mathcal{M}_r|)\right\rceil$. Let $X_1, \ldots, X_m$ be m independent P_X-distributed random variables, and let A denote the event that, for all $g, g' \in \mathcal{M}_r$ with $g \neq g'$, there exists an $i \in \{1, \ldots, n\}$ such that $g(X_i) \neq g'(X_i)$. For a given pair of distinct functions $g, g' \in \mathcal{M}_r$, they disagree on some X_i with probability

$$1 - (1 - P_X(g(X) \neq g'(X)))^m > 1 - \exp(-rm/2) \geq 1 - \frac{1}{|\mathcal{M}_r|^2}.$$

Using a union bound and summing over all possible unordered pairs $g, g' \in \mathcal{M}_r$ will give us that $\mathbb{P}(A) > \frac{1}{2}$. On the event A, functions in $\mathcal{M}_r$ realize distinct classifications of $X_1, \ldots, X_m$. For any

$$X_i \notin \mathrm{DIS}(\mathcal{B}_H(f, rn/h, \{x_1, \ldots, x_n\}),$$

all classifiers in $\mathcal{M}_r$ agree. Thus, $|\mathcal{M}_r|$ is bounded by the number of different classifications $\{X_1, \ldots, X_m\} \cap \mathrm{DIS}(\mathcal{B}_H(f, rn/h))$ realized by classifiers in $\mathcal{F}$. By the multiplicative Chernoff bound, on an event B with $\mathbb{P}(B) \geq \frac{1}{2}$ we have $|\{X_1, \ldots, X_m\} \cap \mathrm{DIS}(\mathcal{B}_H(f, rn/h))| \leq 1 + 2eP_X(\mathrm{DIS}(\mathcal{B}_H(f, rn/h))m$. Using the definition of $\tau(\cdot)$ (Definition 2) we have

$$1 + 2eP_X(\mathrm{DIS}(\mathcal{B}_H(f, rn/h)))m \leq 1 + 2e\tau\left(\frac{r}{h}\right)\frac{r}{h}m \leq 11e\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}.$$

With probability at least $\frac{1}{2}$,

$$|\{X_1, \ldots, X_m\} \cap \mathrm{DIS}(\mathcal{B}_H(f, rn/h))| \leq 11e\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}.$$

Using the union bound, we have that with probability greater than zero there exists a sequence of at most $11e\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}$ elements, such that all functions in $\mathcal{M}_r$ classify this sequence distinctly. By the Vapnik and Chervonenkis lemma, we therefore have that

$$|\mathcal{M}_r| \leq \left(\frac{11e^2\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}}{d}\right)^d.$$

Using Corollary 4.1 from [41] we have

$$\log(|\mathcal{M}_r|) \leq 2d\log\left(11e^2\tau\left(\frac{r}{h}\right)\frac{1}{h}\right).$$

Using $\tau\left(\frac{r}{h}\right) \leq \mathbf{s} \wedge \frac{h}{r} \leq \mathbf{s} \wedge \frac{nh}{\gamma}$ (Theorem 10 in [18]) we finally have

$$\log(\mathcal{M}^{\mathrm{loc}}_1(\mathcal{F}, \gamma, n, h)) \leq 2d\log\left(11e^2\left(\frac{n}{\gamma} \wedge \frac{\mathbf{s}}{h}\right)\right).$$

Now we upper bound $\gamma^{\mathrm{loc}}_{h,h}(n)$, knowing that

$$h\gamma^{\mathrm{loc}}_{h,h}(n) \leq 2d\log\left(11e^2\left(\frac{n}{\gamma^{\mathrm{loc}}_{h,h}(n)} \wedge \frac{\mathbf{s}}{h}\right)\right).$$

We obviously have $\gamma^{\text{loc}}_{h,h}(n) \leq \frac{2d\log\left(11e^2\frac{\mathbf{s}}{h}\right)}{h}$. For $\gamma = \frac{2d\log\left(11e^2\frac{nh}{d}\right)}{h}$ we have $h\gamma = 2d\log\left(11e^2\frac{nh}{d}\right)$, but $2d\log\left(11e^2\frac{n}{\gamma}\right) \leq 2d\log\left(11e^2\frac{nh}{d}\right)$ if $h > \frac{d}{11en}$. Finally, we have

$$\gamma^{\text{loc}}_{h,h}(n) \leq \frac{2d\log\left(11e^2\left(\frac{nh}{d} \wedge \frac{\mathbf{s}}{h}\right)\right)}{h}.$$

Now we prove the lower bound. From (9) established above, we know that $\frac{\gamma^{\text{loc}}_{h,h}(n)}{n}$ is, up to an absolute constant, a distribution-free upper bound for $\mathbb{E}(R(\hat{f}) - R(f^*))$, holding for all ERM learners $\hat{f}$. Then any lower bound on $\sup\limits_{P \in \mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\hat{f}) - R(f^*))$ holding for any ERM learner is also a lower bound for $\frac{\gamma^{\text{loc}}_{h,h}(n)}{n}$. In particular, it is known [31; 19] that for any learning procedure $\tilde{f}$, if $h \geq \sqrt{\frac{d}{n}}$, then $\sup\limits_{P \in \mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{d + (1-h)\log(nh^2 \wedge \mathbf{s})}{nh}$, while if $h < \sqrt{\frac{d}{n}}$ then $\sup\limits_{P \in \mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \sqrt{\frac{d}{n}}$. Furthermore, in the particular case of ERM, [19] proves that any upper bound on $\sup\limits_{P \in \mathcal{P}(1,\mathcal{F})} \mathbb{E}(R(\hat{f}) - R(f^*))$ holding for all ERM learners $\hat{f}$ must have size, up to an absolute constant, at least $\frac{\log(n \wedge \mathbf{s})}{n}$. Together, these lower bounds imply $\gamma^{\text{loc}}_{h,h}(n) \gtrsim \frac{d + \log(nh^2 \wedge \mathbf{s})}{h} \wedge \sqrt{dn}$. ■

6. Minimax Lower Bound

In this section we prove that under Massart's bounded noise condition, fixed points of the local empirical entropy appear in minimax lower bounds. Results are in expectation and generally use classic lower bound techniques from the literature [31; 33; 43], previously used only for specific classes. We will need the following definition, which will be motivated below.

Definition 14 *Fix a class of classifiers $\mathcal{F}$. Assume that there exists a positive constant $c \geq 1$ such that for any N the supremum with respect to the radius in $\mathcal{M}^{\text{loc}}_1(\mathcal{F}, \gamma^{\text{loc}}_{h,1}(N), N, 1)$ is achieved at some $\varepsilon_h(N) \leq c\gamma^{\text{loc}}_{h,1}(N)$. This class will be referred to as c-*pseudoconvex.

Theorem 15 *Let $\tilde{f}$ be the output of any learning algorithm. Fix any $c_{\mathcal{F}}$-pseudoconvex class $\mathcal{F}$ and any h satisfying $\sqrt{\frac{d}{n}} \leq h \leq 1$. Then there exists a $P \in \mathcal{P}(h, \mathcal{F})$ such that*

$$\mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{d}{nh} + \frac{1}{c_{\mathcal{F}}} \frac{(1-h)\gamma^{\text{loc}}_{h,1}\left(\lceil \frac{nc_{\mathcal{F}}h}{(1-h)} \rceil\right)}{n}. \tag{12}$$

Conditions involving the constant $c_{\mathcal{F}}$ can be relaxed in different ways. It will be clear from our proof that we may remove the pseudoconvexity assumptions by redefining the local empirical entropy (8) by removing the supremum with respect to the radius. Alternatively one can remove the supremum by introducing certain monotonicity assumptions. We note that such assumptions were used implicitly in previous papers [15; 33]. In both relaxations our lower bound holds with $c_{\mathcal{F}} = 1$. Moreover, the bound (12) is valid for an arbitrary class $\mathcal{F}$ as we may always consider $c_{\mathcal{F}}(N)$ instead of $c_{\mathcal{F}}$, which is a minimal natural number satisfying $\varepsilon_h(N) \leq c_{\mathcal{F}}(N)\gamma^{\text{loc}}_{h,1}(N)$. Finally, we note that these monotonicity problems do not appear for convex classes, as noted by Mendelson in [32]. This is our motivation for the name of the condition in Definition 14: local entropy of the class has almost the same monotonicity properties as in the convex case. In the next section we will present examples of natural pseudoconvex classes.

The next lemma is given in [30] (Corollary 2.18).

Lemma 16 (Birgé) *Let $\{P_i\}_{i=0}^N$ be a finite family of distributions defined on the same measurable space and $\{A_i\}_{i=0}^N$ be a family of disjoint events. Then*

$$\min_{0\le i\le N} P_i(A_i) \le 0.71 \vee \frac{\sum_{i=1}^{N} KL(P_i\|P_0)}{N\log(N+1)}.$$

Proof [Theorem 15] First we consider the value $\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma_{h,1}^{\text{loc}}(N), N, 1)$. Recall that the definition of this value considers suprema over $f \in \mathcal{F}$ and over N-element subsets of $\mathcal{X}^n$. Without loss of generality we assume that these suprema are achieved at some classifier $g \in \mathcal{F}$, some $\varepsilon_h(N) \in [\gamma_{h,1}^{\text{loc}}(N), N]$ and at some particular set $\mathcal{X}_N = \{x_1, \dots, x_N\}$. Let k_i define the number of copies of x_i in $\mathcal{X}_N$. We define $P_{\mathcal{X}_N}(\{x_i\}) = \frac{k_i}{N}$. If all elements are distinct this measure is just a uniform measure on $\mathcal{X}_N$. We introduce a natural parametrization: any classifier is represented by an N-dimensional binary vector and two vectors (for classifiers g, f) disagree only on a set corresponding to $\text{DIS}(\{g, f\}) \cap \mathcal{X}_N$. The set of binary vectors corresponding to classifiers in $\mathcal{F}$ will be denoted by $\mathcal{B}$. For a given binary vector b define $P_b = P_{\mathcal{X}_N} \times P_{Y|X}^b$, where $P_{Y=1|X_i}^b = \frac{1+(2b_i-1)h}{2}$. Let $\tilde{f}_b$ denote the classifier $\tilde{f}$ produced by the learning algorithm when P_b is the data distribution, and let $\tilde{b}$ denote the binary vector corresponding to $\tilde{f}_b$; thus, $\tilde{b}$ is a random vector, which depends on the parameter b only through the n data points having distribution P_b. It is known [11] that $R(\tilde{f}) - R(f^*) = \mathbb{E}(|\eta(X)|\mathbb{1}[\tilde{f}(X) \ne f^*(X)]|\tilde{f}) \ge hP((x,y) : \tilde{f}(x) \ne f^*(x))$, when $P \in \mathcal{P}(h, \mathcal{F})$. Furthermore, when P_b is the data distribution, we have $P_b((x,y) : \tilde{f}_b(x) \ne f^*(x)) = \frac{\rho_H(\tilde{b}, b)}{N}$. Thus, we have

$$\sup_{P\in\mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \ge \max_{b\in\mathcal{B}} \mathbb{E}\left(hP_b((x,y) : \tilde{f}_b(x) \ne f^*(x))\right) \ge \frac{h}{N}\max_{b\in\mathcal{B}} \mathbb{E}(\rho_H(\tilde{b}, b)).$$

Let b^* be the binary vector in $\mathcal{B}$ corresponding to the classifier g defined above, and fix a maximal subset $\mathcal{B}^{\text{loc}} \subset \mathcal{B}$ satisfying the properties that for any $b' \in \mathcal{B}^{\text{loc}}$ we have $\rho_H(b', b^*) \le \varepsilon_h(N)$ and for any two $b', b'' \in \mathcal{B}^{\text{loc}}$ we have $\rho_H(b', b'') > \varepsilon_h(N)/2$. Next, define $\breve{b}$ as the minimizer of $\rho_H(\breve{b}, \tilde{b})$ among $\mathcal{B}^{\text{loc}}$. In particular, if $b \in \mathcal{B}^{\text{loc}}$, we have $\rho_H(\breve{b}, \tilde{b}) \le \rho_H(b, \tilde{b})$, so that $\rho_H(\breve{b}, b) \le \rho_H(\breve{b}, \tilde{b}) + \rho_H(\tilde{b}, b) \le 2\rho_H(\tilde{b}, b)$. Therefore,

$$\frac{h}{N}\max_{b\in\mathcal{B}} \mathbb{E}(\rho_H(\tilde{b}, b)) \ge \frac{h}{N}\max_{b\in\mathcal{B}^{\text{loc}}} \mathbb{E}(\rho_H(\tilde{b}, b)) \ge \frac{h}{2N}\max_{b\in\mathcal{B}^{\text{loc}}} \mathbb{E}(\rho_H(\breve{b}, b)).$$

Recalling that $\breve{b}$ is a deterministic function of $\tilde{f}$, which itself is a function of the n data points, we may define disjoint subsets A_b of $(\mathcal{X} \times \mathcal{Y})^n$, for $b \in \mathcal{B}^{\text{loc}}$, where A_b corresponds to the collection of data sets that would yield $\breve{b} = b$.[3] Now, from Markov's inequality and the fact that the vectors in $\mathcal{B}^{\text{loc}}$ are $\frac{\varepsilon_h(N)}{2}$-separated, we have $\mathbb{E}(\rho_H(\breve{b}, b)) \ge \frac{\varepsilon_h(N)}{2}\mathbb{P}(\breve{b} \ne b) = \frac{\varepsilon_h(N)}{2}(1 - P_b^n(A_b))$. Thus we have that

$$\frac{h}{2N}\max_{b\in\mathcal{B}^{\text{loc}}} \mathbb{E}(\rho_H(\breve{b}, b)) \ge \frac{h\varepsilon_h(N)}{4N}\left(1 - \min_{b\in\mathcal{B}^{\text{loc}}} P_b^n(A_b)\right).$$

We are interested in using Lemma 16 to upper-bound $\min\limits_{b\in\mathcal{B}^{\text{loc}}} P_b^n(A_b)$. Toward this end, note that for any $b', b'' \in \mathcal{B}^{\text{loc}}$, standard calculations show that

$$\text{KL}(P_{b'}^n \| P_{b''}^n) = \frac{n}{N} h \ln\left(\frac{1+h}{1-h}\right) \rho_h(b', b'').$$

Because for $x > 0$ we have $\ln(x+1) \le x$, it holds that $h\ln\left(\frac{1+h}{1-h}\right) \le \frac{2h^2}{1-h}$. Furthermore, for any $b', b'' \in \mathcal{B}^{\text{loc}}$ we have $\rho_H(b', b'') \le 2\varepsilon_h(N)$. Therefore,

$$\text{KL}(P_{b'}^n \| P_{b''}^n) \le \frac{4nh^2\varepsilon_h(N)}{N(1-h)}.$$

3. For simplicity, we are supposing the learning algorithm is not randomized; the argument easily extends to randomized algorithms by conditioning on the internal randomness in this step.

Thus, by Lemma 16,

$$\min_{b \in \mathcal{B}^{\text{loc}}} P_b^n(A_b) \leq 0.71 \vee \frac{\frac{4nh^2\varepsilon_h(N)}{N(1-h)}}{\log(|\mathcal{B}^{\text{loc}}|)}. \tag{13}$$

Noting that $\log(|\mathcal{B}^{\text{loc}}|) = \log(\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \varepsilon_h(N), N, 1)) \geq h\gamma_{h,1}^{\text{loc}}(N) \geq h\varepsilon_h(N)/c_{\mathcal{F}}$, choosing $N = \left\lceil \frac{6nc_{\mathcal{F}}h}{(1-h)} \right\rceil$ yields

$$\frac{4nh^2\varepsilon_h(N)}{N(1-h)} \leq \frac{2h\varepsilon_h(N)}{3c_{\mathcal{F}}} \leq \frac{2}{3}\log(|\mathcal{B}^{\text{loc}}|),$$

so that the right hand side of (13) is 0.71. Altogether, we have that for $h < 1$,

$$\sup_{P \in \mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \geq 0.29\frac{h\varepsilon_h(N)}{4N} \geq \frac{0.29}{48c_{\mathcal{F}}}\frac{(1-h)\varepsilon_h(N)}{n} \geq \frac{0.29}{48c_{\mathcal{F}}}\frac{(1-h)\gamma_{h,1}^{\text{loc}}(N)}{n}.$$

The term $\frac{d}{nh}$ for $h > \sqrt{\frac{d}{n}}$ is a part of the classic lower bound of [31] and. ■

The following observation is an important consequence of our analysis.

Corollary 17 *Consider a $c_{\mathcal{F}}$-pseudoconvex class $\mathcal{F}$. Let $0 < C_0 \leq C_1 < 1$. Then if the margin parameter h is such that $C_0 \vee \sqrt{\frac{d}{n}} \leq h \leq C_1$, then for any VC class $\mathcal{F}$ the ERM upper bound (9) and the lower bound (12) match up to the constant factors (also appearing possibly in the argument of the fixed point), which may depend only on C_0, C_1 and $c_{\mathcal{F}}$.*

It is known [31] that ERM is minimax optimal up to constant factors if $0 \leq h < \sqrt{\frac{d}{n}}$. Interestingly, our corollary is certainly not valid for $C_1 = 1$. The optimal bound in the realizable case is of order $\frac{d}{n} + \frac{\log(\frac{1}{\delta})}{n}$ [20], but ERM cannot generally have this convergence rate in the realizable case [21; 3; 35]. This fact is perfectly reflected in our lower bound. When h is close to 1 the term $\frac{(1-h)\gamma_{h,1}^{\text{loc}}}{nc_{\mathcal{F}}}$ disappears and we have a classic $\frac{d}{n}$ lower bound from [13].

7. Estimation of a Fixed Point of Local Empirical Entropy for Specific Classes

In this section we provide two examples of exact estimation of fixed points of local empirical entropies. First we consider threshold classifiers, introduced in Example 2. For this particular class, $d = 1$ and $\mathbf{s} = 2$. From Theorem 10 we have $\frac{1}{h} \lesssim \gamma_{h,h}^{\text{loc}}(n) \lesssim \frac{\log(\frac{1}{h})}{h}$, and explicit calculation for this special class reveals $\gamma_{h,h}^{\text{loc}}(n) \simeq \frac{\log(\frac{1}{h})}{h}$. In particular, in the realizable case $\gamma_{1,1}^{\text{loc}}(n) \simeq 1$.

Another example will be a class of linear separators in $\mathbb{R}^k$ for $k \geq 2$. This class is known to have VC dimension $d = k+1$. It is easy to verify that for this particular class $\mathbf{s} = \infty$ [18].

Proposition 18 *For the set $\mathcal{F}$ of linear separators in $\mathbb{R}^d$, if $d \geq 2$, then for any $h > \sqrt{\frac{d}{n}}$*

$$\frac{d\log\left(\frac{nh^2}{d}\right)}{h} \lesssim \gamma_{h,h}^{\text{loc}}(n) \lesssim \frac{d\log\left(\frac{nh}{d}\right)}{h}.$$

In particular, $\gamma_{1,1}^{\text{loc}}(n) \simeq d\log(\frac{n}{d})$.

Proof The upper bound follows directly from the Theorem 10. At first we select a special set of points $x_1, \ldots, x_n \in \mathbb{R}^d$. It is known (Theorem 6.5 in [14]) that in $\mathbb{R}^d$ there exists a so called *cyclic polytope* with n vertices, such that it has exactly $\binom{n}{k}$ $(k-1)$-dimensional faces for any $k \leq \lfloor\frac{d}{2}\rfloor$. We choose $x_1, \ldots, x_n$, such that x_i is a vertex of the cyclic polytope. We fix any linear separator f_1 such that all $x_i, \ldots, x_n$ are in the same half-space with respect to this linear separator. Without loss of generality we may assume that $f_1(x_1) = \ldots = f_1(x_n) = -1$. In this notation using the property of cyclic polytopes we see that $\mathcal{F}$ contains all classifiers with at most $\lfloor \frac{d}{2} \rfloor$ ones. We denote this set by $\mathcal{F}_{d/2}$. Analysis of this particular set by Massart and Nédélec (Theorem 5 in [31]) gives a $\frac{(1-h)d\log(\frac{nh^2}{d})}{nh}$ lower bound for $R(\hat{f}) - R(f^*)$ provided that $h > \sqrt{\frac{d}{n}}$. From Theorem 10 we know that this lower bound is also a lower bound for $\gamma^{\mathrm{loc}}_{h,h}(n)$. Thus $\frac{(1-h)d\log(\frac{nh^2}{d})}{h} \lesssim \gamma^{\mathrm{loc}}_{h,h}(n)$. Simultaneously, we have $\gamma^{\mathrm{loc}}_{1,1}(n) \leq \gamma^{\mathrm{loc}}_{h,h}(n)$. So, it is enough to lower bound $\gamma^{\mathrm{loc}}_{1,1}(n)$, which may be derived as a lower bound for ERM in the realizable case. It is known (theorem 6 in [35], or theorem 5 in [3]) that for this particular class $\mathcal{F}_{d/2}$ in the realizable case there exists ERM such that with probability at least $\frac{1}{2}$ we have $\frac{d\log(\frac{n}{d})}{n} \lesssim R(\hat{f})$. This implies that $\frac{d\log(\frac{n}{d})}{n} \lesssim \mathbb{E}R(\hat{f})$ and thus $d\log(\frac{n}{d}) \lesssim \gamma^{\mathrm{loc}}_{1,1}(n)$. Summerizing, we have $d\log(\frac{n}{d}) \vee \frac{(1-h)d\log(\frac{nh^2}{d})}{h} \lesssim \gamma^{\mathrm{loc}}_{h,h}(n)$. We finish the proof by noticing that $\frac{d\log\left(\frac{nh^2}{d}\right)}{h} \lesssim d\log(\frac{n}{d}) \vee \frac{(1-h)d\log(\frac{nh^2}{d})}{h}$. ■

We note that the lower bound 12 may be applied for both classes.

8. Discussion and Open Problems

Local entropies are well known in statistics since the early work of Le Cam [25]. Since then local metric entropies have appeared in minimax lower bounds [43; 32; 28] and in the necessary and sufficient conditions for consistency of ERM estimator in nonparametric regression [39]. Simultaneously, the upper bounds are usually given in terms of global entropies. Interestingly, it is sometimes possible to recover optimal rates by considering only global packings [43; 34]. Generally, empirical covering numbers of classes in statistics have two types of behaviour. There are *parametric* and *VC-type* classes where the logarithm of covering numbers scales as $\log(\frac{1}{\varepsilon})$ and expressive *nonparametric classes* where it scales as ε^{-p} for some $p > 0$. It was proven in [43] that for these expressive nonparametric classes local and global entropies are of the same order. Thus for such classes localization of class does not give any significant improvement and minimax rates are usually obtained using only global entropies [34]. The case of parametric and especially VC-type classes is more delicate and this paper is a first attempt to analyze the last tightly under bounded noise [4]. Our results and examples show that localization of the class is usually needed for VC classes, but definitely not always. Some parametric classes have the features of nonparametric classes: their local entropies are of the same order as their global entropies, and for them bounds in terms of global entropies are essentially optimal. It is not difficult to show that, in the proof of Proposition 18, we gave an example of such a VC class $\mathcal{F}_{d/2}$. Not surprisingly, Massart and Nédélec [31] named this class *rich*. This class appears in almost all class-specific lower bounds [33; 31; 34], which are matched by global upper bounds. In contrast, there are still many interesting classes, for example, threshold classifiers, which are out of the scope of upper bounds based on global entropy.

We should note that a distribution-dependent local entropy has already appeared in the upper bounds in the classification literature under the name of the *doubling dimension*. Given a class of classifiers $\mathcal{F}$ and a probability distribution P_X, define the doubling dimension by

$$D(\mathcal{F}, \gamma) = \max_{f \in \mathcal{F}} \max_{\varepsilon \geq \gamma} \log(\mathcal{N}(\mathcal{B}_{P_X}(f, \varepsilon), \varepsilon/2)), \tag{14}$$

where $\mathcal{B}_{P_X}(f, \varepsilon) = \{g \in \mathcal{F} | P_X(f(X) \neq g(X)) \leq \varepsilon\}$ and $\mathcal{N}(\mathcal{G}, \varepsilon)$ is the ε-covering number of $\mathcal{G}$ with respect to the pseudo-metric $P_X(g(X) \neq g'(X))$. It was proved by Bshouty, Li, and Long [9] that in the realizable case, for any $\varepsilon > 0$, if

$$n \gtrsim \frac{d + D(\mathcal{F}, \varepsilon_0)}{\varepsilon} \sqrt{\log\left(\frac{1}{\varepsilon}\right)} + \frac{\log(\frac{1}{\delta})}{\varepsilon},$$

4. We note that for some parametric classes, specifically for a bounded subset of finite dimensional linear space in L_2, optimal rates were obtained in [24]

then with probability at least $1-\delta$, for any ERM $\hat{f}$ we have $R(\hat{f}) \le \varepsilon$. Here $\varepsilon_0 = \varepsilon \exp\left(-\sqrt{\log(\frac{1}{\varepsilon})}\right)$. It is easy to show that when considering the distribution-free setting, this bound is weaker than ours at least because it contains a square root of an extra logarithmic factor. The following simple inequality compares distribution-free doubling dimension and the local empirical entropy. For any $\gamma \in \mathbb{N}$,

$$\log(\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, 1)) \le 2 \sup_{P_X} D(\mathcal{F}, \gamma/n). \tag{15}$$

To prove this inequality one may consider the uniform probability measure P_X on the n points maximizing the local packing number on the left hand side, in which case the pseudo-metric $P_X(g(X) \neq g'(X))$ is merely $1/n$ times the Hamming distance of the projections to these n points. The constant 2 appears simply due to the fact that empirical local entropies involve packing numbers while the doubling dimension involves covering numbers. Bshouty, Li, and Long [9] also study a non-ERM distribution-dependent learning algorithm in the realizable case, and obtain an error rate guarantee essentially bounded by a fixed point $\varepsilon \approx \frac{D(\mathcal{F},\varepsilon/4)}{n} + \frac{\log(\frac{1}{\delta})}{n}$, with probability at least $1-\delta$. In light of (15), we see that in the worst case over distributions this is essentially no better than our Theorem 10 (with $h=1$), which holds for the much-simpler learning algorithm ERM.

We also note that questions similar to ours have been considered recently by Mendelson [32] and by Lecué and Mendelson [28]. Both papers introduce distribution dependent fixed points of local entropies and show that in the convex regression setup for subgaussian classes they give optimal upper and lower bounds. However, the direct comparison with their results is problematic due to the fact that in the VC case we do not have convexity assumptions: they are replaced by noise assumptions and specifically used by our approach. Moreover, since in the realizable case ERM is not minimax optimal, it can be easily seen from our results that there may not exist a lower bound in terms of fixed points of the local empirical entropy in this case.

We have compared our bound with some of the best known relaxations of the bounds based on local Rademacher processes (1). However, the title of our paper demands also a direct comparison with the bounds based *solely* on local Rademacher complexities. For this, we need the following result.

Theorem 19 (Sudakov minoration for Bernoulli process [37]) *Let $V \subset \mathbb{R}^n$ be a finite set such that for any $v_1, v_2 \in V$ if $v_1 \neq v_2$ then $\|v_1 - v_2\|_2 \ge a$ for some $a > 0$ and for any $v \in V$ it holds $\|v\|_\infty \le b$ for some $b > 0$. Then*

$$\mathbb{E}_\varepsilon \sup_{v \in V} \sum_{i=1}^n \varepsilon_i v_i \gtrsim a\sqrt{\log|V|} \wedge \frac{a^2}{b}. \tag{16}$$

For simplicity we will consider only the realizable case, and distribution-free setting. However we note that similar arguments will also work under bounded noise and general distributions P_X. Fix a sample $x_1, \ldots, x_n$. Applying Corollary 5.1 from [5] we have

$$\mathbb{E}R(\hat{f}) \lesssim \sup_{x_1,\ldots,x_n} r^*,$$

where r^* is a fixed point of the local empirical Rademacher complexity, that is a solution of the following equality

$$\frac{1}{n}\mathbb{E}_\varepsilon \sup_{g \in \text{star}(\mathcal{G}_{f^*}), P_n g \le 2r} \sum_{i=1}^n \varepsilon_i g(x_i) = r,$$

where star$(\mathcal{G})$ denotes the *star-hull* of a class $\mathcal{G}$: that is, the class of functions αg, where $g \in \mathcal{G}$ and $\alpha \in [0,1]$. Since star$(\mathcal{G}_{f^*})$ is star-shaped, it can be simply proven (see appropriate discussions in [32]) that local empirical entropies are not increasing in its radius. Using this fact together with (16) it can be shown $\mathbb{E}_\varepsilon \sup\limits_{g \in \text{star}(\mathcal{G}_{f^*}), P_n g \le \frac{2\gamma}{n}} \sum\limits_{i=1}^n \varepsilon_i g(x_i) \gtrsim \sqrt{\gamma}\sqrt{\log(\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, 1))} \wedge \gamma$. From this it easily follows that $\frac{\gamma_{1,1}^{\text{loc}}(n)}{n} \lesssim r^*$. Thus our bounds are not generally worse than the bounds based *solely* on local Rademacher complexities. Conceptually we are looking for fixed points of the right hand side of (16), while Rademacher analysis works directly with the fixed points of the suprema of localized processes.

There are still interesting questions and possible directions that are out of the scope of this paper:

1. We are focusing on a distribution-free analysis. At the same time by just leaving the expectations with respect to the learning sample we may simply obtain a distribution-dependent version of Theorem 10. Recently, Balcan and Long [4] have proven that for some special distributions P_X, the class of homogenous linear separators admits faster rates of convergence of ERM, compared to worst-case distributions. It may be interesting to generalize our results using distribution-dependent fixed points of the local empirical entropy (based on random data, rather than worst-case data), and specifically to determine whether this yields rates as fast as [4] under similar conditions on P_X.

2. Our approach here makes use of shifted processes and offset Rademacher processes, in place of explicit diameter-localization arguments such as used by [23]. It seems a natural direction to develop a more general theory of this use to understand the limitations of the approach. For example, so far our analysis is specific to the well-specified case when $f^* \in \mathcal{F}$. It would be interesting to generalize our results to more general noise conditions and a miss-specified case.

3. It is also interesting to refine our bounds in situations when h is close to zero: i.e., when the noise levels are high. It is known [31] that when $h < \sqrt{\frac{d}{n}}$ the control of Rademacher processes based on the *Dudley integral* [12] give minimax optimal $\sqrt{\frac{d}{n}}$ convergence rate. Moreover, it is known that bounds based on just one covering are suboptimal in this case. If we fix $h = \sqrt{\frac{d}{n}}$, then the bound of Giné and Koltchinskii (1) (also based on the Dudley integral) will give us an optimal $\sqrt{\frac{d}{n}}$ rate in expectation. Simultaneously, we know that their bound is suboptimal when h is close to 1. Due to an extra term $\frac{d\log\left(\frac{1}{h}\right)}{h}$ in (11) our bound (9) can guarantee only a suboptimal $\sqrt{\frac{d}{n}}\log\left(\frac{n}{d}\right)$ rate when $h = \sqrt{\frac{d}{n}}$, but we know that for many other values of h our bound is significantly better. Nonetheless, we believe that there is a transition, continuous in h, from the Dudley integral regime when $h < \sqrt{\frac{d}{n}}$ to the regime when the local empirical entropy provides the optimal characterization of the rates obtained by ERM.

4. We have already discussed that ERM may be suboptimal in the realizable case. Thus, when considering minimax optimality there is a third regime, when we have almost no noise. However, since ERM is such a natural and frequently-used method, it remains an interesting question to precisely characterize its risk. Recall that the case when h is bounded away from 0 and 1 is partially covered by our Corollary 17. We hypothesize that in the realizable case (and even in a more general regime when h is close to 1) our bound (10) also characterizes the *best possible* bound on the risk of the worst-case choice of empirical risk minimizer $\hat{f}$, up to an absolute constant factor. It follows directly from our discussions that our hypothesis it true for the classes presented in Section 7. Partial analysis of the complexity of ERM has recently been performed by Hanneke [19]. Specifically, he finds that the correct characterization of the risk of ERM is somewhere between the upper bounds (3), (4) and a lower bound

$$R(\hat{f}) - R(f^*) \gtrsim \frac{d}{nh} + \frac{\log(nh^2 \wedge \mathbf{s})}{nh} + \frac{\log(\frac{1}{\delta})}{nh}, \tag{17}$$

holding with probability greater than δ for a worst-case choice of $P \in \mathcal{P}(h, \mathcal{F})$ (and worst-case choice of ERM). We know that in the realizable case, for the class presented in Example 1, the bound (17) is matched. At the same time, for the class of linear separators presented in Section 7, this lower bound is not tight. This, in particular, leads to the obvious conclusion that d and $\mathbf{s}$ are also not sufficient to fully characterize the risk of ERM, even in the realizable case.

Acknowledgments

The authors would like to thank Sasha Rakhlin for his suggestion to use offset Rademacher processes to analyze binary classification under Tsybakov noise conditions and anonymous reviewers of the short version of this paper for their helpful comments. NZ was supported solely by the Russian Science Foundation grant (project 14-50-00150).

References

[1] *K. S. Alexander*. Rates of growth and sample moduli for weighted empirical processes indexed by sets. Probability Theory and Related Fields, 75:379–423, 1987.

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

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

[4] *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.

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

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

[7] *S. Boucheron, O. Bousquet, G. Lugosi*. Theory of classification: a survey of recent advances. ESAIM: Probability and Statistics, 9:323–375, 2005.

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

[9] *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.

[10] *L. Devroye, G. Lugosi*. Combinatorial Methods in Density Estimation. Springer, New York, 2001.

[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] *R.M. Dudley*. Empirical processes. In Ecole de Probabilité de St. Flour 1982. Lecture Notes in Mathematics 1097, Springer Verlag, New York, 1984.

[13] *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.

[14] *H. Edelsbrunner*. Algorithms in Combinatorial Geometry. Springer, Berlin. 1987.

[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*. A bound on the label complexity of agnostic active learning. In Proceedings of the 24th Annual International Conference on Machine Learning, 2007.

[17] *S. Hanneke*. Theory of Disagreement-Based Active Learning. Foundations and Trends in Machine Learning, 7 (2-3): 131-309, 2014.

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

[19] *S. Hanneke*. Refined error bounds for several learning algorithms. http://arxiv.org/abs/1512.07146, 2015.

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

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

[22] *D. Haussler*. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. J. Comb. Theory Ser. A, 69(2):217–232, 1995.

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

[24] *V. Koltchinskii*. Oracle inequalities in empirical risk minimization and sparse recovery problems. St. Flour Lecture Notes, 2011.

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

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

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

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

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

[30] *P. Massart*. Concentration Inequalties and Model Selection. Ecole dEté de Probabilités, Saint Flour. Springer, New York, 2003.

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

[32] *S. Mendelson*. 'Local' vs. 'global' parameters – breaking the Gaussian complexity barrier. http://arxiv.org/abs/1504.02191, 2015.

[33] *M. Raginsky, A. Rakhlin*. Lower Bounds for Passive and Active Learning. Advances in Neural Information Processing Systems 24, NIPS, 2011.

[34] *A. Rakhlin, K. Sridharan, A. B. Tsybakov*. Empirical entropy, minimax regret and minimax risk. Bernoulli, 2015 (Forthcoming).

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

[36] *M. Talagrand*. Sharper bounds for Gaussian and empirical processes. The Annals of Probability, 22(1): 28–76, 1994.

[37] *M. Talagrand*. Upper and lower bounds for stochastic processes. Springer, Berlin, vol. 60, 2014.

[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. English tranlation: Soviet Math. Dokl. 9, 915–918, 1968.

[39] *S. van de Geer, M. Wegkamp*. Consistency for the least squares estimator in nonparametric regression-Annals of Statistics, Vol. 24, No. 6, 2513–2523, 1996.

[40] *T. van Erven , P. Grünwald, N. Mehta, M. Reid, R. Williamson*. Fast rates in statistical and online learning. Journal of Machine Learning Research, 16: 1793–1861, 2015.

[41] *M. Vidyasagar*. Learning and Generalization with Applications to Neural Networks. Springer-Verlag, 2nd edition, 2003.

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

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

Appendix A. Proofs

Proposition 20 (Multiplicative Chernoff bounds) *Let Z have binomial distribution with parameters p, n. Then for any $\eta \in (0,1)$*

$$P\left[Z > (1+\eta)\mathbb{E}Z\right] \leq \exp\left(-\eta^2 pn/3\right), P\left[Z < (1-\eta)\mathbb{E}Z\right] \leq \exp\left(-\eta^2 pn/2\right)$$

and for particular $\eta = \frac{1}{2}$

$$P\left[Z < \mathbb{E}Z/2\right] \leq \exp\left(-pn/8\right), P\left[Z > 3\mathbb{E}Z/2\right] \leq \exp\left(-pn/8\right).$$

Next we have the proof of Theorem 7.

Proof [Theorem 7] Let DIS_0 be a disagreement set of the version space of first $\lfloor n/2 \rfloor$ instances of the learning sample. The random error set will be denoted by $E_1 = \{x \in \mathcal{X} | \hat{f}(x) \neq f^*(x)\}$. Using symmetrization Lemma 5 and Lemma 4 we have for any $c > 0$

$$\mathbb{E}P(E_1) = \mathbb{E}R(\hat{f}) \leq \mathbb{E} \sup_{g \in \mathcal{G}_{f^*}} (Pg - (1+c)P_n g) \leq \frac{2\left(1+\frac{c}{2}\right)^2}{c} \frac{\log\left(\mathcal{S}_{\mathcal{F}}(n)\right)}{n}.$$

We fix $c = 2$ and prove that for any distribution $\mathbb{E}P(E_1) \leq \frac{4\log(\mathcal{S}_{\mathcal{F}}(n))}{n}$. Now we use

$$R(\hat{f}) = P(E_1 | \mathrm{DIS}_0) P(\mathrm{DIS}_0).$$

Let $\xi = |\mathrm{DIS}_0 \cap \{X_{\lfloor n/2 \rfloor + 1}, \ldots, X_n\}|$. Conditionally on the first $\lfloor n/2 \rfloor$ instances ξ has binomial distribution. Expectations with respect to the first and the last parts of the sample will be denoted respectively by $\mathbb{E}$ and $\mathbb{E}'$. Conditionally on $\{X_1, \ldots, X_{\lfloor n/2 \rfloor}\}$ we introduce two events

$$A_1 : \xi < \frac{nP(\mathrm{DIS}_0)}{4},$$
$$A_2 : \xi > \frac{3nP(\mathrm{DIS}_0)}{4}.$$

Using multiplicative Chernoff bounds we have $P(A_1) \leq \exp\left(-\frac{nP(\mathrm{DIS}_0)}{16}\right)$ and $P(A_2) \leq \exp\left(-\frac{nP(\mathrm{DIS}_0)}{16}\right)$. Denote $A = A_1 \cup A_2$. Then

$$\mathbb{E}' P(E_1 | \mathrm{DIS}_0) = \mathbb{E}'\left[P(E_1|\mathrm{DIS}_0) \middle| \overline{A}\right] P(\overline{A}) + \mathbb{E}'\left[P(E_1|\mathrm{DIS}_0) \middle| A\right] P(A).$$

For the first term we have

$$\mathbb{E}'\left[P(E_1|\text{DIS}_0)\Big|\overline{A}\right]P(\overline{A}) \le \mathbb{E}'\left[P(E_1|\text{DIS}_0)\Big|\overline{A}\right] \le \frac{16\log\left(\mathcal{S}_{\mathcal{F}}\left(\frac{3nP(\text{DIS}_0)}{4}\right)\right)}{nP(\text{DIS}_0)}.$$

For the second term multiplied by $P(\text{DIS}_0)$ we have

$$\begin{aligned}&\mathbb{E}'\left[P(E_1|\text{DIS}_0)\Big|A\right]P(\text{DIS}_0)P(A) \le 2\mathbb{E}'P(\text{DIS}_0)\exp\left(-\frac{nP(\text{DIS}_0)}{16}\right)\\ &= 2P(\text{DIS}_0)\exp\left(-\frac{nP(\text{DIS}_0)}{16}\right) \le 2\exp\left(-\frac{n}{16}\right) \le \frac{12}{n}.\end{aligned}$$

Combining previous results we have

$$\mathbb{E}'P(E_1|\text{DIS}_0)P(\text{DIS}_0) \le \frac{16\log\left(\mathcal{S}_{\mathcal{F}}\left(\frac{3nP(\text{DIS}_0)}{4}\right)\right)}{n} + \frac{12}{n}.$$

It easy to see that for all natural k, r

$$(\mathcal{S}_{\mathcal{F}}(kr))^{\frac{1}{r}} \le \mathcal{S}_{\mathcal{F}}(k).$$

We have

$$\begin{aligned}&\mathbb{E}R(\hat{f}) \le \mathbb{E}\left(\frac{16\log\left(\mathcal{S}_{\mathcal{F}}\left(\frac{3nP(\text{DIS}_0)}{4}\right)\right)}{n} + \frac{12}{n}\right) \le \mathbb{E}\frac{16\log\left(\mathcal{S}_{\mathcal{F}}\left(\mathbf{s}\max\left\{1, \frac{3nP(\text{DIS}_0)}{4\mathbf{s}}\right\}\right)\right)}{n} + \frac{12}{n}\\ &\le \frac{16\mathbb{E}\max\left\{1, \frac{3nP(\text{DIS}_0)}{4\mathbf{s}}\right\}\log\left(\mathcal{S}_{\mathcal{F}}(\mathbf{s})\right)}{n} + \frac{12}{n} \le \frac{16\left(1+\frac{3}{2}\right)\log\left(\mathcal{S}_{\mathcal{F}}(\mathbf{s})\right)}{n} + \frac{12}{n} = \frac{40\log\left(\mathcal{S}_{\mathcal{F}}(\mathbf{s})\right)}{n} + \frac{12}{n}.\end{aligned}$$

The proof of the deviation bound is completely analogous, but slightly more technical. We refer to the proof of Theorem 11 in [19], which can be easily generalized to our case. ■

We now present the proof of the deviation bound in Proposition 8.

Proof [Proposition 8 Deviation Bound] The proof in deviation is based on the symmetrization Lemma 6. At first we notice that $\mathcal{G}_{f^*}$ is a $(1,1)$-Bernstein class. Thus fixing any c_1, c_2, such that $0 < c_2 < c_1$ and $\frac{c_2}{3(1+c_2)} \le 1$ we have for any $t \ge 2\log(2)\frac{(1+c_2)^2}{nc_2}$

$$P\left(\sup_{g\in\mathcal{G}_{f^*}}(P-(1+c_1)P_n)g \ge t\right) \le 2P\left(\sup_{g\in\mathcal{G}_{f^*}}(P'_n-(1+c')P_n)g \ge t'\right),$$

where $(1+c') = \frac{1+c_1}{1+c_2}$ and $t' = \frac{t}{2(1+c_2)}$. Introducing the Rademacher averages we have that $\sup\limits_{g\in\mathcal{G}_{f^*}}(P'_n - (1+c')P_n)g$ has the same distribution as $\sup\limits_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum\limits_{i=1}^{n}\varepsilon_i(g_i-g'_i) - c'P_ng/2 - c'P'_ng/2\right)$. We may represent the last quantity as a sum of two random variables

$$\begin{aligned}&\sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i(g_i-g'_i) - c'P_ng/2 - c'P'_ng/2\right)\\ &\le \sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i g_i - c'P_ng/2\right) + \sup_{g\in\mathcal{G}_{f^*}}\left(-\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i g'_i - c'P'_ng/2\right).\end{aligned}$$

Both summands have the same distribution. We consider $P\left(\sup\limits_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum\limits_{i=1}^{n}\varepsilon_i g_i - c'P_n g/2\right)\geq x\right)$ for a fixed $x>0$. Using the same decomposition as in Lemma 9 and Chernoff bound [8] we have for a fixed $\lambda>0$

$$\begin{aligned}
&P\left(\sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i g_i - c'P_n g/2\right)\geq x+\frac{\gamma(1+c'/2)}{n}\right)\\
&\leq P\left(\frac{\gamma(1+c'/2)}{n}+\sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i p(g)_i - c'P_n p(g)/2\right)\geq x+\frac{\gamma(1+c'/2)}{n}\right)\\
&\leq \exp(-\lambda x)\mathbb{E}\mathbb{E}_\varepsilon \exp\left(\lambda(1+c'/2)\sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i p(g)_i - \frac{c'}{c'+2}P_n p(g)\right)\right),
\end{aligned}$$

where, as in Lemma 9, the operator p denotes the nearest element in the γ-covering. By denoting $c''=\frac{c'}{c'+2}$ and $\lambda'=\lambda(1+c'/2)$ we have

$$\mathbb{E}_\varepsilon \exp\left(\lambda'\sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i p(g)_i - c''P_n p(g)\right)\right)\leq \mathcal{M}_1^*(\mathcal{F},\gamma,n)\exp\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda'^2}{2}p(g)_i-\lambda' c'' p(g)_i\right)$$

Setting $\lambda'=2c''$ we have

$$P\left(\sup_{g\in\mathcal{G}_{f^*}}\left(\frac{1+c'/2}{n}\sum_{i=1}^{n}\varepsilon_i g_i - c'P_n g/2\right)\geq x+\frac{\gamma(1+c'/2)}{n}\right)\leq \exp\left(-\frac{4c'x}{(2+c')^2}\right)\mathcal{M}_1^*(\mathcal{F},\gamma,n).$$

We set $x=\frac{(2+c')^2}{4c'}\left(\log(\mathcal{M}_1^*(\mathcal{F},\gamma,n))+\frac{\log(\frac{1}{\delta})}{n}\right)$ and choose $c_1=3$ and $c_2=1$. Then with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}_{f^*}}(P-(1+c_1)P_n)g\lesssim\frac{\gamma}{n}+\frac{\log(\mathcal{M}_1^*(\mathcal{F},\gamma,n))}{n}+\frac{\log(\frac{1}{\delta})}{n}.$$

We finish the proof by setting $\gamma=\gamma^*_{\frac{1}{2}}(n)+1$. The upper bound (7) easily follows from the general bound on packing numbers for VC classes [22]. ∎

Next, we have the proof of Lemma 12.

Proof [Lemma 12] Once again, given $X_1,\ldots,X_n$, let $V=\{(g(X_1),\ldots,g(X_n)):g\in\mathcal{G}\}$ denote the set of binary vectors corresponding to the values of functions in $\mathcal{G}$. As above, for a fixed γ and fixed minimal γ-covering subset $\mathcal{N}_\gamma\subseteq V$, for each $v\in V$, $p(v)$ will denote the closest vector to v in $\mathcal{N}_\gamma$. We will denote by $\mathbb{E}_\xi$ the conditional expectation over the ξ_i variables, given $X_1,\ldots,X_n$. Note that

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\xi\max_{v\in V}\left(\sum_{i=1}^{n}\xi_i v_i - c v_i\right)\\
&\leq\frac{1}{n}\mathbb{E}_\xi\max_{v\in V}\left(\sum_{i=1}^{n}\xi_i(v_i-p(v)_i)\right)+\frac{1}{n}\mathbb{E}_\xi\max_{v\in V}\left(\sum_{i=1}^{n}\frac{c}{4}p(v)_i-cv_i\right)+\frac{1}{n}\mathbb{E}_\xi\max_{v\in V}\left(\sum_{i=1}^{n}\xi_i p(v)_i-\frac{c}{4}p(v)_i\right).
\end{aligned}$$

The first term is $\lesssim\frac{\gamma}{n}$ by the γ-cover property and the fact that $|\xi_i|\lesssim 1$. Furthermore, as in the proof of Lemma 9, the second term is at most $\frac{c}{4}\frac{\gamma}{n}$. Now we analyze the last term carefully. First we use the standard peeling argument. Given a set W of binary vectors we define $W[a,b] = \{w \in W | a \le \rho_H(w,0) < b\}$.

$$\frac{1}{n}\mathbb{E}_\xi \max_{v\in V}\left(\sum_{i=1}^n \xi_i p(v)_i - \frac{c}{4}p(v)_i\right) = \frac{1}{n}\mathbb{E}_\xi \max_{v\in\mathcal{N}_\gamma}\left(\sum_{i=1}^n \xi_i v_i - \frac{c}{4}v_i\right)$$
$$\le \frac{1}{n}\mathbb{E}_\xi \max_{v\in\mathcal{N}_\gamma[0,2\gamma/c]}\left(\sum_{i=1}^n \xi_i v_i - \frac{c}{4}v_i\right) + \frac{1}{n}\sum_{k=1}^\infty \mathbb{E}_\xi \max_{\mathcal{N}_\gamma[2^k\gamma/c,2^{k+1}\gamma/c]}\left(\sum_{i=1}^n \xi_i v_i - \frac{c}{4}v_i\right)_+$$

The first term is upper bounded by $\frac{2\log(\mathcal{M}_1^{\mathrm{loc}}(V,\gamma,n,c))}{cn}$ by Lemma 4 and by noting that $|\mathcal{N}_\gamma[0,2\gamma/c]| \le \mathcal{M}_1(\mathcal{B}_H(0,(2\gamma)/c,\{X_1,\dots,X_n\}),(2\gamma)/2) \le \mathcal{M}_1^{\mathrm{loc}}(V,\gamma,n,c)$. Now we upper-bound the second term. We start with an arbitrary summand. For $\lambda = \frac{c}{8}$, we have

$$\mathbb{E}_\xi \max_{v\in\{0\}\cup\mathcal{N}_\gamma[2^k\gamma/c,2^{k+1}\gamma/c]}\left(\sum_{i=1}^n \xi_i v_i - \frac{c}{4}v_i\right)$$
$$\le \frac{1}{\lambda}\ln \mathbb{E}_\xi \max_{v\in\{0\}\cup\mathcal{N}_\gamma[2^k\gamma/c,2^{k+1}\gamma/c]} \exp\left\{\sum_{i=1}^n \lambda\xi_i v_i - \frac{\lambda c}{4}v_i\right\}$$
$$\le \frac{1}{\lambda}\ln\left(\sum_{v\in\mathcal{N}_\gamma[2^k\gamma/c,2^{k+1}\gamma/c]} \mathbb{E}_\xi \exp\left\{\sum_{i=1}^n \lambda\xi_i v_i - \frac{\lambda c}{4}v_i\right\} + 1\right)$$
$$\le \frac{1}{\lambda}\ln\left(\left|\mathcal{N}_\gamma\left[2^k\gamma/c, 2^{k+1}\gamma/c\right]\right| \exp\left\{2^{k-2}\gamma(4\lambda^2 - \lambda c)/c\right\} + 1\right)$$
$$\le \frac{1}{\lambda}\ln\left(\left|\mathcal{N}_\gamma\left[0, 2^{k+1}\gamma/c\right]\right| \exp\left\{2^{k-2}\gamma(4\lambda^2 - \lambda c)/c\right\} + 1\right)$$
$$\le \frac{1}{\lambda}\ln\left(\left(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{G},2\gamma,n,c)\right)^{2^{k+1}} \exp\left\{2^{k-2}\gamma(4\lambda^2 - \lambda c)/c\right\} + 1\right).$$

Here we used that any minimal covering is also a packing, and $\left|\mathcal{M}_\gamma\left[0,2^{k+1}\gamma/c\right]\right| \le \left|\mathcal{M}_1^{\mathrm{loc}}(\mathcal{G},2\gamma,n,c)\right|^{2^{k+1}}$, where $\mathcal{M}_\gamma$ is a γ-packing. We fix $\gamma = K\gamma_{c,c}^{\mathrm{loc}}(n)$ for some $K > 2$. Observe that local entropy is nonincreasing and $K\gamma_{c,c}^{\mathrm{loc}}(n) > 2\gamma_{c,c}^{\mathrm{loc}}(n) \ge \gamma_{c,c}^{\mathrm{loc}}(n) + 1$. Thus,

$$\frac{1}{\lambda}\ln\left(\exp\left\{2^{k+1}\log\left(\mathcal{M}_1^{\mathrm{loc}}(V,2K\gamma_{c,c}^{\mathrm{loc}}(n),n,c)\right)\right\}\exp\left\{2^{k-2}K\gamma_{c,c}^{\mathrm{loc}}(n)(4\lambda^2-\lambda c)/c\right\} + 1\right)$$
$$\le \frac{1}{\lambda}\ln\left(\exp\left\{2^{k+1}c(\gamma_{c,c}^{\mathrm{loc}}(n)+1) + 2^{k-2}K\gamma_{c,c}^{\mathrm{loc}}(n)(4\lambda^2-\lambda c)/c\right\} + 1\right).$$

Then we have

$$\sum_{k=1}^\infty \frac{8}{c}\ln\left(\exp\left(2^{k+1}\log\left(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{G},2K\gamma_{c,c}^{\mathrm{loc}}(n),n)\right)\right)\exp\left(-2^{k-6}Kc\gamma_{c,c}^{\mathrm{loc}}(n)\right)+1\right)$$
$$\le \sum_{k=1}^\infty \frac{8}{c}\ln\left(\exp\left(2^{k+2}c\gamma_{c,c}^{\mathrm{loc}}(n) - 2^{k-6}Kc\gamma_{c,c}^{\mathrm{loc}}(n)\right)+1\right).$$

We set $K = 2^9$ and have $\sum_{k=1}^\infty \ln\left(\exp\left(2^{k+2}c\gamma_{c,c}^{\mathrm{loc}}(n) - 2^{k-6}Kc\gamma_{c,c}^{\mathrm{loc}}(n)\right)+1\right) \le C$, where $C > 0$ is an absolute constant. Here we used that $\ln(x+1) \le x$ for $x > 0$ and $c\gamma_{c,c}^{\mathrm{loc}} \gtrsim 1$. Finally, we have

$$\frac{1}{n}\mathbb{E}_\xi \max_{v\in V}\left(\sum_{i=1}^n \xi_i v_i - cv_i\right) \lesssim \frac{\gamma_{c,c}^{\mathrm{loc}}(n)}{n} + \frac{\log(\mathcal{M}_1^{\mathrm{loc}}(\mathcal{G},\gamma_{c,c}^{\mathrm{loc}}(n),n,c))}{cn} + \frac{1}{cn} \lesssim \frac{\gamma_{c,c}^{\mathrm{loc}}(n)}{n}.$$

■

Finally, we present the proof of the deviation bound in Theorem 10.

Proof [Theorem 10 Deviation Bound] We will provide a detailed outline of the proof. This proof technically repeats the arguments from our previous results. The constants will be denoted by c_i for $i \in \mathbb{N}$. The idea is to combine the technique we previously used for Theorem 10 in expectation with the symmetrization lemma (Lemma 6). Once again, let $\hat{f}$ be any ERM and $\hat{g}$ be a corresponding function in the excess loss class $\mathcal{G}_{\mathcal{Y}}$. We have $R(\hat{f}) - R(f^*) = P\hat{g}$ and $P_n\hat{g} \leq 0$. Then for any $c > 0$

$$R(\hat{f}) - R(f^*) \leq P\hat{g} - (1+c)P_n\hat{g} \leq \sup_{g \in \mathcal{G}_{\mathcal{Y}}} (Pg - (1+c)P_n g).$$

Now due to the fact that $\mathcal{G}_{\mathcal{Y}}$ is a $(\frac{1}{h}, 1)$-Bernstein class we have, using Lemma 6,

$$P\left(\sup_{g \in \mathcal{G}_{\mathcal{Y}}} (P - (1+c_1)P_n)g \geq t\right) \leq 2P\left(\sup_{g \in \mathcal{G}_{\mathcal{Y}}} ((1+c_2)P'_n - (1+c_1)P_n)g \geq t/2\right),$$

provided that $0 < c_2 < c_1$, $\frac{c_2}{3(1+c_2)} \leq \frac{1}{h}$ and $t \geq \frac{2\log(2)}{nh}\frac{(1+c_2)^2}{c_2}$. Now we use the same argument as in the proof of Proposition 8. Specifically, to control the deviation of the value $\sup\limits_{g \in \mathcal{G}_{\mathcal{Y}}} (P'_n - (1+c_3)P_n)g$ it is enough to control the deviation of

$$\sup_{g \in \mathcal{G}_{\mathcal{Y}}} \left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(X_i, Y_i) - c_4 P_n g\right). \tag{18}$$

Now we use the argument from Lemma 11. To control (18) it is enough to control the deviation of the term $\sup\limits_{g' \in \mathcal{G}_{f^*}} \left(\sum\limits_{i=1}^{n} \xi_i g'(X_i) - hc_5 g'(X_i)\right)$. We fix $\gamma \in \mathbb{N}$ and use the decomposition as in the beginning of the proof of Lemma 12. Now the problem is reduced to the analysis a γ-covering

$$\begin{aligned}
&\sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^{n} \xi_i g'(X_i) - hc_5 g'(X_i)\right) \\
&\leq \sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^{n} \xi_i (g'(X_i) - p(g'(X_i)))\right) + \sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^{n} \xi_i p(g'(X_i)) - \frac{hc_5}{4} p(g'(X_i))\right) \\
&\leq \gamma + \sup_{g' \in \mathcal{G}_{f^*}} \left(\sum_{i=1}^{n} \xi_i p(g'(X_i)) - \frac{hc_5}{4} p(g'(X_i))\right).
\end{aligned}$$

The concentration of the last term is given by a combination of Chernoff bound (as in Proposition 8) and an upper bound for the exponential moment of $\sup\limits_{g' \in \mathcal{G}_{f^*}} \left(\sum\limits_{i=1}^{n} \xi_i p(g'(X_i)) - \frac{hc_5}{4} p(g'(X_i))\right)$ from the proof of Lemma 12. ■