

# Localization of VC Classes: Beyond Local Rademacher Complexities

Nikita Zhivotovskiy<sup>1,2(✉)</sup> and Steve Hanneke<sup>3</sup>

<sup>1</sup> Moscow Institute of Physics and Technology, Moscow, Russia  
`nikita.zhivotovskiy@phystech.edu`

<sup>2</sup> Institute for Information Transmission Problems, Moscow, Russia

<sup>3</sup> Princeton, NJ 08542, USA  
`steve.hanneke@gmail.com`

**Abstract.** In statistical learning the excess risk of empirical risk minimization (ERM) is controlled by  $\left(\frac{\text{COMP}_n(\mathcal{F})}{n}\right)^\alpha$ , where  $n$  is a size of a learning sample,  $\text{COMP}_n(\mathcal{F})$  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  $\text{COMP}_n(\mathcal{F})$  in the upper bounds under bounded noise. Our results are accompanied by a novel 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:** PAC learning · Local metric entropy · Local Rademacher process · Shifted empirical process · Offset Rademacher process · Empirical risk minimization · VC dimension · Star number · 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 [23] 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. But this 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 [18, 20]. Until now this phenomenon was discussed only for several specific classes. The main aim of this paper is to present this yet unknown combinatorial complexity measure.

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. At the beginning of the 2000s, so-called *localized* complexities (Bartlett et al. [3], Koltchinskii [12]) 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. We still 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 largely 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). The present work explores this aspect of localization, resulting in a complexity measure, which correctly captures the optimal rates under bounded noise.

## 2 Notation and Previous Results

We define the *instance space*  $\mathcal{X}$  and the *label space*  $\mathcal{Y} = \{1, -1\}$ , and denote  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ . We assume that the set  $\mathcal{Z}$  is equipped with some  $\sigma$ -algebra and a probability measure  $P$  on measurable subsets is defined. We also assume that we are given a set of classifiers  $\mathcal{F}$ . The risk of a classifier  $f$  is its probability of error, denoted  $R(f) = P(f(X) \neq Y)$ . We denote the *Bayes classifier* by  $f^*(x) = \text{sign}(\eta(x))$ , where  $\eta(x) = \mathbb{E}[Y|X = x]$ . 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  $\text{DIS}(B) = \{x \in \mathcal{X} \mid \exists f, g \in B \text{ s. t. } f(x) \neq g(x)\}$ . We will also consider abstract real-valued function 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  $((X_1, Y_1), \dots, (X_n, Y_n))$ , 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 an empirical mean. *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  $f \in \mathcal{F}$ . 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  $\{x_1, \dots, x_k\} \in \mathcal{X}^k$  is shattered by  $\mathcal{F}$  if there are  $2^k$  distinct classifications of  $\{x_1, \dots, x_k\}$  realized by classifiers in  $\mathcal{F}$ . The *VC dimension* of  $\mathcal{F}$  is the largest integer  $d$  such that there exists a set  $\{x_1, \dots, x_d\}$  shattered by  $\mathcal{F}$  [23]. 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  $\mathcal{F}$ .

**Definition 1** (Massart and Nédélec [18]).  $(P, \mathcal{F})$  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 [4]. Let  $\mathcal{F}$  be a class with VC-dimension  $d$ . For any empirical risk minimizer  $\hat{f}$  over  $n$  samples, for any  $P \in \mathcal{P}(0, \mathcal{F})$ , we have  $\mathbb{E}(R(\hat{f}) - R(f^*)) \lesssim \sqrt{\frac{d}{n}}$ . 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  $\mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \sqrt{\frac{d}{n}} \wedge 1$ . 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 [6]. 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  $\tau(\varepsilon) = \sup_{\varepsilon_0 \geq \varepsilon} (\varepsilon_0^{-1} P_X\{x \in \mathcal{X} : \exists f \in \mathcal{F}_{\varepsilon_0} \text{ s.t. } f(x) \neq f^*(x)\})$ .

This quantity was introduced to the empirical processes literature by Alexander [1], and is referred to as *Alexander’s capacity* by Giné and Koltchinskii [6]. The same quantity appeared independently in the literature on active learning, where it is referred to as the *disagreement coefficient* [7].  $\tau(\varepsilon)$  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 [6] 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  $P \in \mathcal{P}(h, \mathcal{F})$ ,

$$\mathbb{E}(R(\hat{f}) - R(f^*)) \lesssim \frac{d}{nh} \log \left( \tau \left( \frac{d}{nh^2} \right) \right). \quad (1)$$

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. Thus we may consider this result as the best known relaxation of the local Rademacher analysis. Very recently, Hanneke and Yang [8] introduced a distribution-free complexity measure, called the *star number*. It is defined as follows.

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

Similar to Alexander's capacity, the star number describes how diverse the small-size disagreements with a fixed classifier  $f_0$  can be. 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:

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

An immediate corollary of this and (1) is that, for any  $P \in \mathcal{P}(h, \mathcal{F})$ ,  $\mathbb{E}(R(\hat{f}) - R(f^*)) \lesssim \frac{d}{nh} \log \left( \frac{nh^2}{d} \wedge \mathbf{s} \right)$ . Since  $\mathbf{s}$  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.

### 3 Preliminaries from Empirical Processes

Given a function class  $\mathcal{G}$  mapping  $\mathcal{Z}$  to  $\mathbb{R}$ , one may consider the following quantity:  $\sup_{g \in \mathcal{G}} (P - P_n)g$ . This random value plays an important role in statistical

learning theory. Since the pioneering paper of Vapnik and Chervonenkis [23], 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 controlled by a supremum of the so-called

*Rademacher process*:  $\frac{1}{n} \mathbb{E}_\varepsilon \max_{g \in \mathcal{G}} \left( \sum_{i=1}^n \varepsilon_i g_i \right)$ , where  $g_i$  denotes  $g(Z_i)$ ,  $\varepsilon_i$  are independent Rademacher variables taking values  $\pm 1$  with equal probabilities, and  $\mathbb{E}_\varepsilon$  denoted the expectation over the  $\varepsilon_i$  random variables (conditioning on the  $Z_i$  variables). We will instead consider different quantities, so-called *shifted empirical processes*, introduced by Lecué and Mitchell [14]. 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 [16]:  $\frac{1}{n} \mathbb{E}_\varepsilon \max_{g \in \mathcal{G}} \left( \sum_{i=1}^n \varepsilon_i g_i - c' g_i^2 \right)$ . This quantity was introduced for the analysis of a specific aggregation procedure under the squared loss and so far has not been related to a shifted process. In this paper, we will investigate some new properties of these processes and will show how they may be used in the classification framework. The following short lemma appears in a more general form in [16] (Lemma 5).

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

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

**Lemma 2 (Shifted Symmetrization in Expectation).** *Let  $\mathcal{G}$  be a function 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.* Denote  $g(Z_i)$  by  $g_i$ . Using the symmetrization trick and Jensen's inequality,

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

□

Let  $\mathbf{s}$  be the star number of a class of binary classifiers  $\mathcal{F}$ . Hanneke [9] recently proved that  $\mathbb{E} P(\text{DIS}(\mathcal{V}_n)) \leq \frac{\mathbf{s}}{n+1}$ , where  $\mathcal{V}_n = \{f \in \mathcal{F} | P_n \mathbb{1}[f(X) \neq f^*(X)] = 0\}$  is a *version space*, and used this fact to bound the risk of ERM. In this same spirit, this inequality will be important in our next theorem, one of the novel contributions of the present work. Its proof is in the appendix.

**Theorem 1.** *Let  $\mathbf{s}$  be the star number of a class of binary classifiers  $\mathcal{F}$ . In the realizable case, for any ERM  $\hat{f}$ ,*

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

*Example 1.* Theorem 1 yields examples showing the gaps in the distribution-free bound (1) in the realizable case. Specifically, suppose  $\mathcal{X} = \{x_1, \dots, x_{\mathbf{s}}\}$ , 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  $\{x_1, \dots, x_{d-1}\}$  and at most one point classified 1 among  $\{x_d, \dots, x_{\mathbf{s}}\}$ . 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 1 gives a bound of order  $\frac{d \log(\frac{\mathbf{s} \wedge n}{d})}{n}$ , 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 [9, 10]. It should be noted, however, that one can also construct examples where Theorem 1 is itself not tight.

## 4 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  $\gamma$ :

$$\mathcal{B}_H(f, \gamma, \{x_1, \dots, x_n\}) = \{g \in \mathcal{F} | \rho_H(f, g) \leq \gamma\},$$

where  $\rho_H(f, g) = |\{i \in \{1, \dots, n\} : f(x_i) \neq g(x_i)\}|$ . When  $x_1, \dots, x_n$  are clear from the context, we sometimes simply write  $\mathcal{B}_H(f, \gamma)$ . We further introduce

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

where  $\mathcal{M}_1(\mathcal{H}, \varepsilon)$  denotes the size of a maximal  $\varepsilon$ -packing of  $\mathcal{H}$  under  $\rho_H$  distance (for the given  $x_1, \dots, x_n$  points). This quantity measures how one can pack a ball in  $\mathcal{F}$  by balls of smaller radius. For any  $h, h' \in (0, 1]$ , define

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

When  $\mathcal{F}$  is clear from the context, we simply write  $\gamma_{h, h'}^{\text{loc}}(n)$  instead of  $\gamma_{h, h'}^{\text{loc}}(n, \mathcal{F})$ . The quantity  $\gamma_{h, h'}^{\text{loc}}(n)$  defines the *fixed point of a local empirical entropy*.

We note that, because  $1 \leq d < \infty$  in this work, when  $h, h' > 0$  the set on the right in this definition is finite and nonempty, so that  $\gamma_{h, h'}^{\text{loc}}(n)$  is a well-defined strictly-positive integer. Indeed, for any  $h, h' \in (0, 1]$ , the value  $\gamma = \lfloor \frac{1}{h} \rfloor$  satisfies  $h\gamma \leq 1$ , so that (because  $\log(\cdot)$  is the truncated logarithm) this  $\gamma$  is contained in the set; in particular, this implies  $h\gamma_{h, h'}^{\text{loc}}(n, \mathcal{F}) \geq h\lfloor \frac{1}{h} \rfloor \geq \frac{1}{2}$  always. The next theorem is the main upper bound of this paper. The rest of this section is devoted to its proof.

**Theorem 2.** *Fix any function class  $\mathcal{F}$ ; denote its VC dimension  $d$  and star number  $\mathbf{s}$ . Fix any  $h \in \left(\sqrt{\frac{d}{n}}, 1\right]$  and suppose  $\gamma_{h, h}^{\text{loc}}(n) > 0$ . If  $P \in \mathcal{P}(h, \mathcal{F})$ , then for any ERM  $\hat{f}$ ,*

$$\mathbb{E}(R(\hat{f}) - R(f^*)) \lesssim \frac{\gamma_{h, h}^{\text{loc}}(n)}{n}. \quad (2)$$

Moreover,

$$\frac{d + \log(nh^2 \wedge \mathbf{s})}{h} \lesssim \gamma_{h, h}^{\text{loc}}(n) \lesssim \frac{d \log\left(\frac{nh^2}{d} \wedge \mathbf{s}\right)}{h} + \frac{d \log\left(\frac{1}{h}\right)}{h}. \quad (3)$$

Our complexity term (3) is not worse than the upper bound of Giné and Koltchinskii (1) when  $h$  is bounded from 0 by a constant. Another interesting property is that the bound (2) involves 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. We should mention that the connection between global covering numbers and VC dimension is well known [11].

Consider the *excess loss class*  $\mathcal{G}_Y = \{(x, y) \rightarrow \mathbb{1}[f(x) \neq y] - \mathbb{1}[f^*(x) \neq y] \text{ for } f \in \mathcal{F}\}$  and the *class*  $\mathcal{G}_{f^*} = \{x \rightarrow \mathbb{1}[f(x) \neq f^*(x)] \text{ for } f \in \mathcal{F}\}$ , which may be interpreted as an excess loss class in the realizable case. For any  $g \in \mathcal{G}_Y$  it holds  $g^2(x, y) = \mathbb{1}[f(x) \neq f^*(x)] = \frac{1}{2}|f(x) - f^*(x)| = \frac{1}{4}(f(x) - f^*(x))^2$ . And also for any  $g \in \mathcal{G}_Y$  it holds  $g(x, y) = \frac{y(f^*(x) - f(x))}{2}$  and  $R(f^*) \leq \frac{1}{2}(1 - h)$  [5].

**Lemma 3 (Contraction).** *Let  $\mathcal{G}_Y$  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  $P \in \mathcal{P}(h, \mathcal{F})$ ,*

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

where  $\xi_1, \dots, \xi_n$  are r. v. conditionally independent given  $X_1, \dots, X_n$ , with  $\mathbb{E}[\xi_i | X_1, \dots, X_n] = 0$  and  $\mathbb{E}[\exp(\lambda \xi_i) | X_1, \dots, X_n] \leq \exp(\frac{\lambda^2}{2})$  for all  $\lambda$ .

*Proof.* We will denote  $g(X_i, Y_i)$  by  $g_i$ . First we notice that any  $g \in \mathcal{G}_Y$  may be defined by some  $f \in \mathcal{F}$ . Then note that

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

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  $\xi_1^{(h')}, \dots, \xi_n^{(h')}$  random variables are conditionally independent given  $X_1, \dots, X_n$ , with  $\mathbb{E}[\xi_i^{(h')} | X_1, \dots, X_n] = 0$ . 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  $-\sum_{i=1}^n g_i = -\sum_{i=1}^n h'_i g_i^2 + \sum_{i=1}^n \xi_i^{(h')} g_i^2 \leq -(\min_i h'_i) \sum_{i=1}^n g_i^2 + \sum_{i=1}^n \xi_i^{(h')} g_i^2$ .

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  $h \leq h'_i$  almost surely, we have  $\frac{1}{4} \mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_Y} \left( \sum_{i=1}^n \varepsilon_i g_i^2 - 4cg_i \right) \leq \frac{1}{4} \mathbb{E}\mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_Y} \left( \sum_{i=1}^n \xi'_i g_i^2 - 4hc g_i^2 \right) =$

$\frac{1}{4} \mathbb{E}_X \mathbb{E}_{\xi'} \sup_{g' \in \mathcal{G}_{f^*}} \left( \sum_{i=1}^n \xi'_i g'(X_i) - 4hc g'(X_i) \right)$ . Finally, because  $\varepsilon_i$  and  $\xi_i^{(h')}$  both have zero conditional mean, so does  $\xi'_i$ , and since we also have  $-5 + 4ch'_i \leq \xi'_i \leq 5 + 4ch'_i$ , Hoeffding's lemma ([5] Lemma 8.1) implies  $\mathbb{E}[\exp(\lambda \xi'_i) | X_1, \dots, X_n] \leq \exp(25\lambda^2/2)$ . The lemma easily follows, taking  $\xi_i = \xi'_i/5$ .  $\square$

**Lemma 4 (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*

$\mathbb{E}[\xi_i | X_1, \dots, X_n] = 0$  and  $\mathbb{E}[\exp(\lambda \xi_i) | X_1, \dots, X_n] \leq \exp(\frac{\lambda^2}{2})$  for all  $\lambda$ . Then if  $c\gamma_{c,c}^{\text{loc}}(n, \mathcal{G}) \gtrsim 1$ ,

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

The proof of this lemma is deferred to the appendix.

*Proof (Theorem 2).* Let  $\hat{f}$  be an ERM and  $\hat{g}$  be a corresponding function in the excess loss class  $\mathcal{G}_y$ . We obviously have  $\mathbb{E}(R(\hat{f}) - R(f^*)) = \mathbb{E}P\hat{g}$  and  $P_n\hat{g} \leq 0$ . Then  $\forall c > 0$ ,  $\mathbb{E}(R(\hat{f}) - R(f^*)) \leq \mathbb{E}(P\hat{g} - (1+c)P_n\hat{g}) \leq \mathbb{E} \sup_{g \in \mathcal{G}_y} (Pg - (1+c)P_ng)$ .

Now using the symmetrization lemma (Lemma 2) we have

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

Applying Lemma 3, we have  $\frac{c+2}{n} \mathbb{E} \mathbb{E}_\varepsilon \sup_{g \in \mathcal{G}_y} \left( \sum_{i=1}^n \varepsilon_i g(X_i, Y_i) - \frac{c}{c+2} g(X_i, Y_i) \right) \leq \frac{5(c+2)}{4n} \mathbb{E} \mathbb{E}_\varepsilon \sup_{g' \in \mathcal{G}_{f^*}} \left( \sum_{i=1}^n \xi_i g'(X_i) - \frac{4ch}{5(c+2)} g'(X_i) \right)$ . Now we are ready to apply the localization lemma (Lemma 4). The conditions on the  $\xi_i$  variables required for Lemma 4 are supplied by Lemma 3, and all functions in  $\mathcal{G}_{f^*}$  take only binary values. Thus, for a fixed  $c$ ,

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

The following proposition finishes the proof of Theorem 2. Its proof is in the appendix.

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

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

## 5 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 based on classic lower bound techniques from the literature [18, 20, 25], previously used only for specific classes.

**Definition 4.** *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  $\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma_{h,1}^{\text{loc}}(N), N, 1)$  is achieved at some  $\varepsilon_h(N) \leq c\gamma_{h,1}^{\text{loc}}(N)$ . This class will be referred to as  $c$ -pseudoconvex.*



**Theorem 3.** *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*

$$\mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{d}{nh} + \frac{1}{c_{\mathcal{F}}} \frac{(1-h)\gamma_{h,1}^{\text{loc}}\left(\lceil \frac{nc_{\mathcal{F}}h}{(1-h)} \rceil\right)}{n}. \quad (4)$$

Conditions involving the constant  $c_{\mathcal{F}}$  can be relaxed in different ways. We may remove the pseudoconvexity assumptions by redefining the local empirical entropy (4) by removing the maximum with respect to the radius. Alternatively one can remove the maximum by introducing certain monotonicity assumptions, which were used implicitly in previous papers [6, 20]. In both cases our lower bound holds with  $c_{\mathcal{F}} = 1$ . Finally, we note that these monotonicity problems do not appear for convex classes, as noted by Mendelson in [19]. The next lemma is given in [17] (Corollary 2.18).

**Lemma 5 (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*

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

*Proof (Theorem 3).* 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 [5] that  $R(\tilde{f}) - R(f^*) = \mathbb{E}(|\eta(X)| \mathbb{1}[\tilde{f}(X) \neq f^*(X)] | \tilde{f}) \geq hP((x, y) : \tilde{f}(x) \neq f^*(x))$ , when  $P \in \mathcal{P}(h, \mathcal{F})$ . Furthermore, when  $P_b$  is the data distribution, we have  $P_b((x, y) : \tilde{f}_b(x) \neq f^*(x)) = \frac{\rho_H(\tilde{b}, b)}{N}$ . Thus, we have  $\sup_{P \in \mathcal{P}(h, \mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \geq$

$\max_{b \in \mathcal{B}} \mathbb{E} \left( hP_b((x, y) : \tilde{f}_b(x) \neq f^*(x)) \right) \geq \frac{h}{N} \max_{b \in \mathcal{B}} \mathbb{E}(\rho_H(\tilde{b}, b))$ . 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  $\rho_H(b', b^*) \leq$

$\varepsilon_h(N)$  and for any two  $b', b'' \in \mathcal{B}^{\text{loc}}$  we have  $\rho_H(b', b'') > \varepsilon_h(N)/2$ . Next, define  $\check{b}$  as the minimizer of  $\rho_H(\check{b}, \tilde{b})$  among  $\mathcal{B}^{\text{loc}}$ . In particular, if  $b \in \mathcal{B}^{\text{loc}}$ , we have  $\rho_H(\check{b}, \tilde{b}) \leq \rho_H(b, \tilde{b})$ , so that  $\rho_H(\check{b}, b) \leq \rho_H(\tilde{b}, \tilde{b}) + \rho_H(\tilde{b}, b) \leq 2\rho_H(\tilde{b}, b)$ . Therefore,  $\frac{h}{N} \max_{b \in \mathcal{B}} \mathbb{E}(\rho_H(\check{b}, b)) \geq \frac{h}{N} \max_{b \in \mathcal{B}^{\text{loc}}} \mathbb{E}(\rho_H(\check{b}, b)) \geq \frac{h}{2N} \max_{b \in \mathcal{B}^{\text{loc}}} \mathbb{E}(\rho_H(\tilde{b}, b))$ . Recalling that  $\check{b}$  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  $\check{b} = b$ . 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  $\mathbb{E}(\rho_H(\check{b}, b)) \geq \frac{\varepsilon_h(N)}{2} P(\check{b} \neq b) = \frac{\varepsilon_h(N)}{2} (1 - P_b^n(A_b))$ . Thus we have that  $\frac{h}{2N} \max_{b \in \mathcal{B}^{\text{loc}}} \mathbb{E}(\rho_H(\check{b}, b)) \geq \frac{h\varepsilon_h(N)}{4N} \left(1 - \min_{b \in \mathcal{B}^{\text{loc}}} P_b^n(A_b)\right)$ . We are interested in using Lemma 5 to upper-bound  $\min_{b \in \mathcal{B}^{\text{loc}}} P_b^n(A_b)$ . Toward this end, note that for any  $b', b'' \in \mathcal{B}^{\text{loc}}$ , simple 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  $\ln(x+1) \leq x$ , it holds that  $h \ln \left( \frac{1+h}{1-h} \right) \leq \frac{2h^2}{1-h}$ . Furthermore, for any  $b', b'' \in \mathcal{B}^{\text{loc}}$  we have  $\rho_H(b', b'') \leq 2\varepsilon_h(N)$ . Therefore,  $\text{KL}(P_{b'}^n \| P_{b''}^n) \leq \frac{4nh^2\varepsilon_h(N)}{N(1-h)}$ . Thus, by Lemma 5,  $\min_{b \in \mathcal{B}^{\text{loc}}} P_b^n(A_b) \leq 0.71 \vee \frac{4nh^2\varepsilon_h(N)}{\log(|\mathcal{B}^{\text{loc}}|) \frac{N(1-h)}{\log(|\mathcal{B}^{\text{loc}}|)}}$ . 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  $\min_{b \in \mathcal{B}^{\text{loc}}} P_b^n(A_b) \leq 0.71$ . Finally, 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 [18].  $\square$

## 6 Discussion and Open Problems

Local entropies are well known in statistics since the early work of Le Cam [13]. Since then local metric entropies appear in minimax lower bounds. Simultaneously, the upper bounds are usually given in terms of global empirical entropies. Interestingly, it is sometimes possible to recover optimal rates by considering only global packings [21, 25]. 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  $\varepsilon^{-p}$  for some  $p > 0$ . It was proved in [25] that for nonparametric classes local and global entropies are of the same order. Thus for such classes localization of the class does not give any significant improvement. We also note that questions similar to ours have been considered recently by Mendelson [19] and by Lecué and Mendelson [15]. Both papers show that in the convex regression setup for subgaussian classes distribution dependent

fixed points of particular local entropies 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.

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 4 (Sudakov Minoration for Bernoulli Process [22]).** *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  $\|v_1 - v_2\|_2 \geq a$  for some  $a > 0$  and for any  $v \in V$  it holds  $\|v\|_\infty \leq b$  for some  $b > 0$ . Then*

$$\mathbb{E}_\varepsilon \sup_{v \in V} \sum_{i=1}^n \varepsilon_i v_i \gtrsim a \sqrt{\log |V|} \wedge \frac{a^2}{b}. \quad (5)$$

For simplicity, we will consider only the realizable case. However we note that similar arguments will also work under bounded noise and general distributions  $P_X$ . Fix a sample  $x_1, \dots, x_n$ . Applying Corollary 5.1 from [3] we have  $\mathbb{E}R(\hat{f}) \lesssim \sup_{x_1, \dots, x_n} r^*$ , where  $r^*$  is a fixed point of the local empirical Rademacher complexity,

that is a solution of the following equality  $\frac{1}{n} \mathbb{E}_\varepsilon \sup_{g \in \text{star}(\mathcal{G}_{f^*}), P_n g \leq 2r} \sum_{i=1}^n \varepsilon_i g(x_i) = r$ ,

where  $\text{star}(\mathcal{G})$  denotes the *star-hull* of a class  $\mathcal{G}$ : that is, the class of functions  $\alpha g$ , where  $g \in \mathcal{G}$  and  $\alpha \in [0, 1]$ . Since  $\text{star}(\mathcal{G}_{f^*})$  is star-shaped, it can be simply proven (see appropriate discussions in [19]) that local empirical entropies are not increasing in its radius. Using this fact together with (5) it can be shown

$\mathbb{E}_\varepsilon \sup_{g \in \text{star}(\mathcal{G}_{f^*}), P_n g \leq \frac{2\gamma}{n}} \sum_{i=1}^n \varepsilon_i g(x_i) \gtrsim \sqrt{\gamma} \sqrt{\log(\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, 1))} \wedge \gamma$ . 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 the local Rademacher complexities.

There are still interesting questions and possible directions that are out of the scope of this paper. At first, we are focusing on a distribution free analysis. At the same time one may obtain a distribution dependent version of Theorem 2. Recently, Balcan and Long [2] have proved that for some special distributions and classes of homogenous linear separators rates of convergence of ERM may be faster than if we consider worst-case distributions. It will be interesting to generalize our results using distribution dependent fixed points of the local empirical entropy and also to miss-specified models, when  $f^* \notin \mathcal{F}$ .

**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 for their helpful comments. NZ was supported solely by the Russian Science Foundation grant (project 14-50-00150).

## Appendix

*Proof (Theorem 1).* Let  $\text{DIS}_0$  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  $E_1 = \{x \in \mathcal{X} \mid \hat{f}(x) \neq f^*(x)\}$ . Using symmetrization Lemmas 2 and 1 we have  $\mathbb{E}P(E_1) = \mathbb{E}R(\hat{f}) \leq \mathbb{E} \sup_{g \in \mathcal{G}_{f^*}} (Pg - (1+c)P_n g) \leq \frac{2(1+\frac{\epsilon}{2})^2 \log(\mathcal{S}_{\mathcal{F}}(n))}{c} \text{ for } c > 0$ .

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  $R(\hat{f}) = P(E_1 \mid \text{DIS}_0)P(\text{DIS}_0)$ . Let  $\xi = |\text{DIS}_0 \cap \{X_{\lfloor n/2 \rfloor + 1}, \dots, X_n\}|$ . Conditionally on the first  $\lfloor n/2 \rfloor$  instances  $\xi$  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  $\{x_1, \dots, x_{\lfloor n/2 \rfloor}\}$  we introduce two events:  $A_1 : \xi < \frac{nP(\text{DIS}_0)}{4}$  and  $A_2 : \xi > \frac{3nP(\text{DIS}_0)}{4}$ . Using Chernoff bounds we have  $P(A_1) \leq \exp\left(-\frac{nP(\text{DIS}_0)}{16}\right)$  and  $P(A_2) \leq \exp\left(-\frac{nP(\text{DIS}_0)}{16}\right)$ . Denote  $A = A_1 \cup A_2$ . Then  $\mathbb{E}'P(E_1 \mid \text{DIS}_0) = \mathbb{E}'\left[P(E_1 \mid \text{DIS}_0) \mid \overline{A}\right]P(\overline{A}) + \mathbb{E}'\left[P(E_1 \mid \text{DIS}_0) \mid A\right]P(A)$ . For the first term we have  $\mathbb{E}'\left[P(E_1 \mid \text{DIS}_0) \mid \overline{A}\right]P(\overline{A}) \leq \frac{16 \log\left(\mathcal{S}_{\mathcal{F}}\left(\frac{3nP(\text{DIS}_0)}{4}\right)\right)}{nP(\text{DIS}_0)}$ . We can directly prove for the second term that  $\mathbb{E}'\left[P(E_1 \mid \text{DIS}_0) \mid A\right]P(\text{DIS}_0)P(A) \leq \frac{12}{n}$ . It easy to see, that for all natural  $k, r$  we have  $(\mathcal{S}_{\mathcal{F}}(kr))^{\frac{1}{r}} \leq \mathcal{S}_{\mathcal{F}}(k)$ . Finally,  $\mathbb{E}R(\hat{f}) \leq \mathbb{E} \frac{16 \log\left(\mathcal{S}_{\mathcal{F}}\left(\frac{3nP(\text{DIS}_0)}{4}\right)\right)}{n} + \frac{12}{n} \leq \frac{40 \log(\mathcal{S}_{\mathcal{F}}(s))}{n} + \frac{12}{n}$ .  $\square$

*Proof (Lemma 4).* Once again, 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  $\gamma$  and fixed minimal  $\gamma$ -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  $\xi_i$  variables, given  $X_1, \dots, X_n$ . We follow the decomposition proposed by Liang, Rakhlin, and Sridharan [16]:

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

The first term is  $\lesssim \frac{\gamma}{n}$  by the  $\gamma$ -cover property and the fact that  $|\xi_i| \lesssim 1$ . Furthermore it is easy to show that 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  $W[a, b] = \{w \in W \mid a \leq \rho_H(w, 0) < b\}$ .

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

The first term is upper bounded by  $\frac{2 \log(\mathcal{M}_1^{\text{loc}}(V, \gamma, n, c))}{cn}$  by Lemma 1 and by noting that  $|\mathcal{N}_\gamma[0, 2\gamma/c]| \leq \mathcal{M}_1(\mathcal{B}_H(0, (2\gamma)/c, \{X_1, \dots, X_n\}), (2\gamma)/2) \leq \mathcal{M}_1^{\text{loc}}(V, \gamma, n, c)$ . Now we upper-bound the second term. We start with an arbitrary summand. For any  $\lambda > 0$ , we have

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

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

$$\begin{aligned}
&\ln (\exp \{2^{k+1} \log (\mathcal{M}_1^{\text{loc}}(V, 2K\gamma_{c,c}^{\text{loc}}(n), n, c)) + 2^{k-2} K\gamma_{c,c}^{\text{loc}}(n)(4\lambda^2 - \lambda c)/c\} + 1) \\
&\leq \ln (\exp \{2^{k+1} c(\gamma_{c,c}^{\text{loc}}(n) + 1) + 2^{k-2} K\gamma_{c,c}^{\text{loc}}(n)(4\lambda^2 - \lambda c)/c\} + 1).
\end{aligned}$$

Then we have for  $\lambda = \frac{c}{8}$ ,

$$\begin{aligned}
&\sum_{k=1}^{\infty} \frac{8}{c} \ln (\exp (2^{k+1} \log (\mathcal{M}_1^{\text{loc}}(\mathcal{G}, 2K\gamma_{c,c}^{\text{loc}}(n), n))) \exp (-2^{k-6} Kc\gamma_{c,c}^{\text{loc}}(n)) + 1) \\
&\leq \sum_{k=1}^{\infty} \frac{8}{c} \ln (\exp (2^{k+2} c\gamma_{c,c}^{\text{loc}}(n) - 2^{k-6} Kc\gamma_{c,c}^{\text{loc}}(n)) + 1).
\end{aligned}$$

We set  $K = 2^9$  and have  $\sum_{k=1}^{\infty} \ln (\exp (2^{k+2} c\gamma_{c,c}^{\text{loc}}(n) - 2^{k-6} Kc\gamma_{c,c}^{\text{loc}}(n)) + 1) \leq C$ , where  $C > 0$  is an absolute constant. Here we used that  $\ln(x+1) \leq x$  for  $x > 0$  and  $c\gamma_{c,c}^{\text{loc}} \gtrsim 1$ . Combining with the first two terms we finish the proof.  $\square$

*Proof (Proposition 1).* The first part of the proof closely follows the proof of Theorem 17 in [8], with slight modifications, to arrive at an upper bound on  $\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, h)$ . The suprema in the definition of local empirical entropy are

achieved at some set  $\{x_1, \dots, x_n\}$ , 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  $\rho_H$ ) of  $\mathcal{B}_H(f, rn/h, \{x_1, \dots, x_n\})$ , so that  $|\mathcal{M}_r| = \mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, h)$ . Also introduce a uniform probability measure  $P_X$  on  $\{x_1, \dots, x_n\}$  and fix  $m = \lceil \frac{d}{r} \log(|\mathcal{M}_r|) \rceil$ . Let  $X_1, \dots, X_m$  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  $i \in \{1, \dots, n\}$  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  $X_1, \dots, X_m$ . For any  $X_i \notin \text{DIS}(\mathcal{B}_H(f, rn/h, \{x_1, \dots, x_n\}))$ , all classifiers in  $\mathcal{M}_r$  agree. Thus,  $|\mathcal{M}_r|$  is bounded by the number of classifications  $\{X_1, \dots, X_m\} \cap \text{DIS}(\mathcal{B}_H(f, rn/h))$  realized by classifiers in  $\mathcal{F}$ . By the Chernoff bound, on an event  $B$  with  $\mathbb{P}(B) \geq \frac{1}{2}$  we have  $|\{X_1, \dots, X_m\} \cap \text{DIS}(\mathcal{B}_H(f, rn/h))| \leq 1 + 2eP_X(\text{DIS}(\mathcal{B}_H(f, rn/h)))m$ . Using the definition of  $\tau(\cdot)$  (Definition 2) we have  $1 + 2eP_X(\text{DIS}(\mathcal{B}_H(f, rn/h)))m \leq 1 + 2e\tau\left(\frac{r}{h}\right)\frac{r}{h}m \leq 11e\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}$ . With probability at least  $\frac{1}{2}$ ,  $|\{X_1, \dots, X_m\} \cap \text{DIS}(\mathcal{B}_H(f, rn/h))| \leq 11e\tau\left(\frac{r}{h}\right)\frac{\log(|\mathcal{M}_r|)}{h}$ . Using the union bound, we have that with positive probability 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 VC lemma [23], 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 [24] we have  $\log(|\mathcal{M}_r|) \leq 2d \log(11e^2\tau\left(\frac{r}{h}\right)\frac{1}{h})$ . Using  $\tau\left(\frac{r}{h}\right) \leq \mathbf{s} \wedge \frac{h}{r} \leq \mathbf{s} \wedge \frac{nh}{\gamma}$  (Theorem 10 in [8]) we finally have  $\log(\mathcal{M}_1^{\text{loc}}(\mathcal{F}, \gamma, n, h)) \leq 2d \log\left(11e^2\left(\frac{n}{\gamma} \wedge \frac{\mathbf{s}}{h}\right)\right)$ . Observe that  $h\gamma_{h,h}^{\text{loc}}(n) \leq 2d \log\left(11e^2\left(\frac{n}{\gamma_{h,h}^{\text{loc}}(n)} \wedge \frac{\mathbf{s}}{h}\right)\right)$ . We have  $\gamma_{h,h}^{\text{loc}}(n) \leq \frac{2d \log(11e^2\frac{\mathbf{s}}{h})}{h}$ . If  $\gamma = \frac{2d \log(11e^2\frac{nh}{h})}{h}$ , then  $h\gamma = 2d \log(11e^2\frac{nh}{d})$ , but  $2d \log\left(11e^2\frac{n}{\gamma}\right) \leq 2d \log(11e^2\frac{nh}{d})$  if  $h > \frac{d}{11en}$ . Finally, we have  $\gamma_{h,h}^{\text{loc}}(n) \leq \frac{2d \log(11e^2(\frac{nh}{d} \wedge \frac{\mathbf{s}}{h}))}{h}$ . Now we prove the lower bound. From (2) established above, we know that  $\frac{\gamma_{h,h}^{\text{loc}}(n)}{n}$  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 a lower bound on  $\sup_{P \in \mathcal{P}(h, \mathcal{F})} \mathbb{E}(R(\hat{f}) - R(f^*))$  holding for any ERM learner is also a lower bound for  $\frac{\gamma_{h,h}^{\text{loc}}(n)}{n}$ . In particular, it is known [9, 18] that for any learning procedure  $\tilde{f}$ , if  $h \geq \sqrt{\frac{d}{n}}$ , then  $\sup_{P \in \mathcal{P}(h, \mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \frac{d+(1-h)\log(nh^2 \wedge \mathbf{s})}{nh}$ , while if  $h < \sqrt{\frac{d}{n}}$  then  $\sup_{P \in \mathcal{P}(h, \mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \gtrsim \sqrt{\frac{d}{n}}$ . Furthermore, in the particular case of ERM, [9] proves that any upper bound on  $\sup_{P \in \mathcal{P}(1, \mathcal{F})} \mathbb{E}(R(\hat{f}) - R(f^*))$  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  $\gamma_{h,h}^{\text{loc}}(n) \gtrsim \frac{d+\log(nh^2 \wedge \mathbf{s})}{h} \wedge \sqrt{dn}$ .  $\square$

## References

1. Alexander, K.S.: Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probab. Theory Relat. Fields* **75**, 379–423 (1987)
2. Balcan, M.F., Long, P.M.: Active and passive learning of linear separators under log-concave distributions. In: 26th Conference on Learning Theory (2013)
3. Bartlett, P.L., Bousquet, O., Mendelson, S.: Local Rademacher complexities. *Ann. Stat.* **33**(4), 1497–1537 (2005)
4. Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: a survey of recent advances. *ESAIM: Probab. Stat.* **9**, 323–375 (2005)
5. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Applications of Mathematics, vol. 31. Springer, New York (1996)
6. Giné, E., Koltchinskii, V.: Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34**(3), 1143–1216 (2006)
7. Hanneke, S.: Theory of disagreement-based active learning. *Found. Trends Mach. Learn.* **7**(2–3), 131–309 (2014)
8. Hanneke, S., Yang, L.: Minimax analysis of active learning. *J. Mach. Learn. Res.* **16**(12), 3487–3602 (2015)
9. Hanneke, S.: Refined error bounds for several learning algorithms (2015). <http://arXiv.org/abs/1512.07146>
10. Haussler, D., Littlestone, N., Warmuth, M.: Predicting  $\{0, 1\}$ -functions on randomly drawn points. *Inf. Comput.* **115**, 248–292 (1994)
11. Haussler, D.: Sphere packing numbers for subsets of the Boolean  $n$ -cube with bounded Vapnik–Chervonenkis dimension. *J. Combin. Theory Ser. A* **69**, 217–232 (1995)
12. Koltchinskii, V.: Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Stat.* **34**(6), 2593–2656 (2006)
13. Le Cam, L.M.: Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1**, 38–53 (1973)
14. Lecué, G., Mitchell, C.: Oracle inequalities for cross-validation type procedures. *Electron. J. Stat.* **6**, 1803–1837 (2012)
15. Lecué, G., Mendelson, S.: Learning subgaussian classes: upper and minimax bounds (2013). <http://arXiv.org/abs/1305.4825>
16. Liang, T., Rakhlin, A., Sridharan, K.: Learning with square loss: localization through offset Rademacher complexity. In: Proceedings of The 28th Conference on Learning Theory (2015)
17. Massart, P.: Concentration Inequalities and Model Selection. Ecole d’Eté de Probabilités, Saint Flour. Springer, New York (2003)
18. Massart, P., Nédélec, E.: Risk bounds for statistical learning. *Ann. Stat.* **34**(5), 2326–2366 (2006)
19. Mendelson, S.: ‘Local’ vs. ‘global’ parameters – breaking the Gaussian complexity barrier (2015). <http://arXiv.org/abs/1504.02191>
20. Raginsky, M., Rakhlin, A.: Lower bounds for passive and active learning. In: Advances in Neural Information Processing Systems 24, NIPS (2011)
21. Rakhlin, A., Sridharan, K., Tsybakov, A.B.: Empirical entropy, minimax regret and minimax risk. *Bernoulli* (2015, forthcoming)
22. Talagrand, M.: Upper and Lower Bounds for Stochastic Processes. Springer, Heidelberg (2014)
23. Vapnik, V., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. *Proc. USSR Acad. Sci.* **181**(4), 781–783 (1968). English translation: *Soviet Math. Dokl.* **9**, 915–918

24. Vidyasagar, M.: Learning and Generalization with Applications to Neural Networks, 2nd edn. Springer, Heidelberg (2003)
25. Yang, Y., Barron, A.: Information-theoretic determination of minimax rates of convergence. *Ann. Stat.* **27**, 1564–1599 (1999)



Algorithmic Learning Theory

27th International Conference, ALT 2016, Bari, Italy,

October 19-21, 2016, Proceedings

Ortner, R.; Simon, H.U.; Zilles, S. (Eds.)

2016, XIX, 371 p. 21 illus., Softcover

ISBN: 978-3-319-46378-0