Journal of Machine Learning Research 17 (2016) 1-55 Submitted 12/15; Revised 7/16; Published 8/16

Refined Error Bounds for Several Learning Algorithms

Steve Hanneke STEVE.HANNEKE@GMAIL.COM

Editor: John Shawe-Taylor

Abstract

This article studies the achievable guarantees on the error rates of certain learning algorithms, with particular focus on refining logarithmic factors. Many of the results are based on a general technique for obtaining bounds on the error rates of sample-consistent classifiers with monotonic error regions, in the realizable case. We prove bounds of this type expressed in terms of either the VC dimension or the sample compression size. This general technique also enables us to derive several new bounds on the error rates of general sample-consistent learning algorithms, as well as refined bounds on the label complexity of the CAL active learning algorithm. Additionally, we establish a simple necessary and sufficient condition for the existence of a distribution-free bound on the error rates of all sample-consistent learning rules, converging at a rate inversely proportional to the sample size. We also study learning in the presence of classification noise, deriving a new excess error rate guarantee for general VC classes under Tsybakov's noise condition, and establishing a simple and general necessary and sufficient condition for the minimax excess risk under bounded noise to converge at a rate inversely proportional to the sample size.

Keywords: sample complexity, PAC learning, statistical learning theory, active learning, minimax analysis

1. Introduction

Supervised machine learning is a classic topic, in which a learning rule is tasked with producing a classifier that mimics the classifications that would be assigned by an expert for a given task. To achieve this, the learner is given access to a collection of examples (assumed to be i.i.d.) labeled with the correct classifications. One of the major theoretical questions of interest in learning theory is: How many examples are necessary and sufficient for a given learning rule to achieve low classification error rate? This quantity is known as the *sample complexity*, and varies depending on how small the desired classification error rate is, the type of classifier we are attempting to learn, and various other factors. Equivalently, the question is: How small of an error rate can we guarantee a given learning rule will achieve, for a given number of labeled training examples?

A particularly simple setting for supervised learning is the *realizable case*, in which it is assumed that, within a given set $\mathbb{C}$ of classifiers, there resides some classifier that is *always* correct. The optimal sample complexity of learning in the realizable case has recently been completely resolved, up to constant factors, in a sibling paper to the present article (Hanneke, 2016). However, there remains the important task of identifying interesting general families of algorithms achieving this optimal sample complexity. For instance, the best known general upper bounds for the general family of *empirical risk minimization* algorithms differ from the optimal sample complexity by a logarithmic factor, and it is known that there exist spaces $\mathbb{C}$ for which this is unavoidable (Auer and Ortner, 2007). This same logarithmic factor gap appears in the analysis of several other learning methods as well. The present article focuses on this logarithmic factor, arguing that for certain types of learning rules, it can be entirely removed in some cases, and for others it can be somewhat refined. The technique leading to these results is rooted in an idea introduced in the author's doctoral dissertation (Hanneke, 2009). By further exploring this technique, we also obtain new results for the related problem of *active learning*. We also derive interesting new results for learning with classification noise, where again the focus is on a logarithmic factor gap between upper and lower bounds.

1.1 Basic Notation

Before further discussing the results, we first introduce some essential notation. Let $\mathcal{X}$ be any nonempty set, called the *instance space*, equipped with a σ-algebra defining the measurable sets; for simplicity, we will suppose the sets in $\{\{x\} : x \in \mathcal{X}\}$ are all measurable. Let $\mathcal{Y} = \{-1, +1\}$ be the label space. A *classifier* is any measurable function $h : \mathcal{X} \to \mathcal{Y}$. Following Vapnik and Chervonenkis (1971), define the VC dimension of a set $\mathcal{A}$ of subsets of $\mathcal{X}$, denoted $\mathrm{vc}(\mathcal{A})$, as the maximum cardinality $|S|$ over subsets $S \subseteq \mathcal{X}$ such that $\{S \cap A : A \in \mathcal{A}\} = 2^S$ (the power set of S); if no such maximum cardinality exists, define $\mathrm{vc}(\mathcal{A}) = \infty$. For any set $\mathcal{H}$ of classifiers, denote by $\mathrm{vc}(\mathcal{H}) = \mathrm{vc}(\{\{x : h(x) = +1\} : h \in \mathcal{H}\})$ the VC dimension of $\mathcal{H}$. Throughout, we fix a set $\mathbb{C}$ of classifiers, known as the *concept space*, and abbreviate $d = \mathrm{vc}(\mathbb{C})$. To focus on nontrivial cases, throughout we suppose $|\mathbb{C}| \geq 3$, which implies $d \geq 1$. We will also generally suppose $d < \infty$ (though some of the results would still hold without this restriction).

For any $L_m = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$, and any classifier h, define $\mathrm{er}_{L_m}(h) = \frac{1}{m} \sum_{(x,y) \in L_m} \mathbb{1}[h(x) \neq y]$. For completeness, also define $\mathrm{er}_{\{\}}(h) = 0$. Also, for any set $\mathcal{H}$ of classifiers, denote $\mathcal{H}[L_m] = \{h \in \mathcal{H} : \forall (x, y) \in L_m, h(x) = y\}$, referred to as the set of classifiers in $\mathcal{H}$ *consistent* with L_m; for completeness, also define $\mathcal{H}[\{\}] = \mathcal{H}$. Fix an arbitrary probability measure $\mathcal{P}$ on $\mathcal{X}$ (called the *data distribution*), and a classifier $f^\star \in \mathbb{C}$ (called the *target function*). For any classifier h, denote $\mathrm{er}(h) = \mathcal{P}(x : h(x) \neq f^\star(x))$, the *error rate* of h. Let $X_1, X_2, \ldots$ be independent $\mathcal{P}$-distributed random variables. We generally denote $\mathcal{L}_m = \{(X_1, f^\star(X_1)), \ldots, (X_m, f^\star(X_m))\}$, and $V_m = \mathbb{C}[\mathcal{L}_m]$ (called the *version space*). The general setting in which we are interested in producing a classifier $\hat{h}$ with small $\mathrm{er}(\hat{h})$, given access to the data $\mathcal{L}_m$, is a special case of supervised learning known as the *realizable case* (in contrast to settings where the observed labeling might not be realizable by any classifier in $\mathbb{C}$, due to label noise or model misspecification, as discussed in Section 6).

We adopt a few convenient notational conventions. For any $m \in \mathbb{N}$, denote $[m] = \{1, \ldots, m\}$; also denote $[0] = \{\}$. We adopt a shorthand notation for sequences, so that for a sequence $x_1, \ldots, x_m$, we denote $x_{[m]} = (x_1, \ldots, x_m)$. For any $\mathbb{R}$-valued functions f, g, we write $f(z) \lesssim g(z)$ or $g(z) \gtrsim f(z)$ if there exists a finite numerical constant $c > 0$ such that $f(z) \leq c g(z)$ for all z. For any $x, y \in \mathbb{R}$, denote $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$. For $x \geq 0$, denote $\mathrm{Log}(x) = \ln(x \vee e)$ and $\mathrm{Log}_2(x) = \log_2(x \vee 2)$. We also adopt the conventions that for $x > 0$, $x/0 = \infty$, and $0\mathrm{Log}(x/0) = 0\mathrm{Log}(\infty) = 0 \cdot \infty = 0$. It will also be convenient to use the notation $\mathcal{Z}^0 = \{()\}$ for a set $\mathcal{Z}$, where $()$ is the empty sequence.

Throughout, we also make the usual implicit assumption that all quantities required to be measurable in the proofs and lemmas from the literature are indeed measurable. See, for instance, van der Vaart and Wellner (1996, 2011), for discussions of conditions on $\mathbb{C}$ that typically suffice for this.

1.2 Background and Summary of the Main Results

This work concerns the study of the error rates achieved by various *learning rules*: that is, mappings from the data set $\mathcal{L}_m$ to a classifier $\hat{h}_m$; for simplicity, we sometimes refer to $\hat{h}_m$ itself as a learning rule, leaving dependence on $\mathcal{L}_m$ implicit. There has been a substantial amount of work on bounding the error rates of various learning rules in the realizable case. Perhaps the most basic and natural type of learning rule in this setting is the family of *consistent* learning rules: that is, those that choose $\hat{h}_m \in V_m$. There is a general upper bound for all consistent learning rules $\hat{h}_m$, due to Vapnik and Chervonenkis (1974); Blumer, Ehrenfeucht, Haussler, and Warmuth (1989), stating that with probability at least $1-\delta$,

$$\mathrm{er}\left(\hat{h}_m\right) \lesssim \frac{1}{m}\left(d\mathrm{Log}\left(\frac{m}{d}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right). \tag{1}$$

This is complemented by a general lower bound of Ehrenfeucht, Haussler, Kearns, and Valiant (1989), which states that for any learning rule (consistent or otherwise), there exists a choice of $\mathcal{P}$ and $f^\star \in \mathbb{C}$ such that, with probability greater than δ,

$$\mathrm{er}\left(\hat{h}_m\right) \gtrsim \frac{1}{m}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right). \tag{2}$$

Resolving the logarithmic factor gap between (2) and (1) has been a challenging subject of study for decades now, with many interesting contributions resolving special cases and proposing sometimes-better upper bounds (e.g., Haussler, Littlestone, and Warmuth, 1994; Giné and Koltchinskii, 2006; Auer and Ortner, 2007; Long, 2003). It is known that the lower bound is sometimes *not* achieved by certain consistent learning rules (Auer and Ortner, 2007). The question of whether the lower bound (2) can always be achieved by *some* algorithm remained open for a number of years (Ehrenfeucht, Haussler, Kearns, and Valiant, 1989; Warmuth, 2004), but has recently been resolved in a sibling paper to the present article (Hanneke, 2016). That work proposes a learning rule $\hat{h}_m$ based on a majority vote of classifiers consistent with carefully-constructed subsamples of the data, and proves that with probability at least $1-\delta$,

$$\mathrm{er}\left(\hat{h}_m\right) \lesssim \frac{1}{m}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

However, several avenues for investigation remain open, including identifying interesting general families of learning rules able to achieve this optimal bound under general conditions on $\mathbb{C}$. In particular, it remains an open problem to determine necessary and sufficient conditions on $\mathbb{C}$ for the entire family of consistent learning rules to achieve the above optimal error bound.

The work of Giné and Koltchinskii (2006) includes a bound that refines the logarithmic factor in (1) in certain scenarios. Specifically, it states that, for any consistent learning rule

$\hat{h}_m$, with probability at least $1-\delta$,

$$\operatorname{er}\left(\hat{h}_m\right) \lesssim \frac{1}{m}\left(d\operatorname{Log}\left(\theta\left(\frac{d}{m}\right)\right) + \operatorname{Log}\left(\frac{1}{\delta}\right)\right), \tag{3}$$

where $\theta(\cdot)$ is the *disagreement coefficient* (defined below in Section 4). The doctoral dissertation of Hanneke (2009) contains a simple and direct proof of this bound, based on an argument which splits the data set in two parts, and considers the second part as containing a subsequence sampled from the conditional distribution given the region of disagreement of the version space induced by the first part of the data. Many of the results in the present work are based on variations of this argument, including a variety of interesting new bounds on the error rates achieved by certain families of learning rules.

As one of the cornerstones of this work, we find that a variant of this argument for consistent learning rules with *monotonic* error regions leads to an upper bound that *matches* the lower bound (2) up to constant factors. For such monotonic consistent learning rules to exist, we would need a very special kind of concept space. However, they do exist in some important cases. In particular, in the special case of learning *intersection-closed* concept spaces, the *Closure* algorithm (Natarajan, 1987; Auer and Ortner, 2004, 2007) can be shown to satisfy this monotonicity property. Thus, this result immediately implies that, with probability at least $1-\delta$, the Closure algorithm achieves

$$\operatorname{er}(\hat{h}_m) \lesssim \frac{1}{m}\left(d + \operatorname{Log}\left(\frac{1}{\delta}\right)\right),$$

which was an open problem of Auer and Ortner (2004, 2007); this fact was recently also obtained by Darnstädt (2015), via a related direct argument. We also discuss a variant of this result for monotone learning rules expressible as *compression schemes*, where we remove a logarithmic factor present in a result of Littlestone and Warmuth (1986) and Floyd and Warmuth (1995), so that for $\hat{h}_m$ based on a compression scheme of size n, which has monotonic error regions (and is permutation-invariant), with probability at least $1-\delta$,

$$\operatorname{er}(\hat{h}_m) \lesssim \frac{1}{m}\left(n + \operatorname{Log}\left(\frac{1}{\delta}\right)\right).$$

This argument also has implications for *active learning*. In many active learning algorithms, the *region of disagreement* of the version space induced by m samples, $\operatorname{DIS}(V_m) = \{x \in \mathcal{X} : \exists h, g \in V_m \text{ s.t. } h(x) \neq g(x)\}$, plays an important role. In particular, the label complexity of the CAL active learning algorithm (Cohn, Atlas, and Ladner, 1994) is largely determined by the rate at which $\mathcal{P}(\operatorname{DIS}(V_m))$ decreases, so that any bound on this quantity can be directly converted into a bound on the label complexity of CAL (Hanneke, 2011, 2009, 2014; El-Yaniv and Wiener, 2012). Wiener, Hanneke, and El-Yaniv (2015) have argued that the region $\operatorname{DIS}(V_m)$ can be described as a compression scheme, where the size of the compression scheme, denoted $\hat{n}_m$, is known as the *version space compression set size* (Definition 6 below). By further observing that $\operatorname{DIS}(V_m)$ is monotonic in m, applying our general argument yields the fact that, with probability at least $1-\delta$, letting $\hat{n}_{1:m} = \max_{t \in [m]} \hat{n}_t$,

$$\mathcal{P}(\operatorname{DIS}(V_m)) \lesssim \frac{1}{m}\left(\hat{n}_{1:m} + \operatorname{Log}\left(\frac{1}{\delta}\right)\right), \tag{4}$$

which is typically an improvement over the best previously-known general bound by a logarithmic factor.

In studying the distribution-free minimax label complexity of active learning, Hanneke and Yang (2015) found that a simple combinatorial quantity $\mathfrak{s}$, which they term the *star number*, is of fundamental importance. Specifically (see also Definition 9), $\mathfrak{s}$ is the largest number s of distinct points $x_1, \ldots, x_s \in \mathcal{X}$ such that $\exists h_0, h_1, \ldots, h_s \in \mathbb{C}$ with $\forall i \in [s]$, $\mathrm{DIS}(\{h_0, h_i\}) \cap \{x_1, \ldots, x_s\} = \{x_i\}$, or else $\mathfrak{s} = \infty$ if no such largest s exists. Interestingly, the work of Hanneke and Yang (2015) also establishes that the largest possible value of $\hat{n}_m$ (over m and the data set) is exactly $\mathfrak{s}$. Thus, (4) also implies a *data-independent* and *distribution-free* bound: with probability at least $1 - \delta$,

$$\mathcal{P}(\mathrm{DIS}(V_m)) \lesssim \frac{1}{m}\left(\mathfrak{s} + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

Now one interesting observation at this point is that the direct proof of (3) from Hanneke (2009) involves a step in which $\mathcal{P}(\mathrm{DIS}(V_m))$ is relaxed to a bound in terms of $\theta(d/m)$. If we instead use (4) in this step, we arrive at a new bound on the error rates of *all* consistent learning rules $\hat{h}_m$: with probability at least $1 - \delta$,

$$\mathrm{er}(\hat{h}_m) \lesssim \frac{1}{m}\left(d\mathrm{Log}\left(\frac{\hat{n}_{1:m}}{d}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right). \tag{5}$$

Since Hanneke and Yang (2015) have shown that the maximum possible value of $\theta(d/m)$ (over m, $\mathcal{P}$, and $f^\star$) is also exactly the star number $\mathfrak{s}$, while $\hat{n}_{1:m}/d$ has as its maximum possible value $\mathfrak{s}/d$, we see that the bound in (5) sometimes reflects an improvement over (3). It further implies a new data-independent and distribution-free bound for any consistent learning rule $\hat{h}_m$: with probability at least $1 - \delta$,

$$\mathrm{er}(\hat{h}_m) \lesssim \frac{1}{m}\left(d\mathrm{Log}\left(\frac{\min\{\mathfrak{s}, m\}}{d}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

Interestingly, we are able to complement this with a *lower bound* in Section 5.1. Though not quite matching the above in terms of its joint dependence on d and $\mathfrak{s}$ (and necessarily so), this lower bound does provide the interesting observation that $\mathfrak{s} < \infty$ is *necessary and sufficient* for there to exist a distribution-free bound on the error rates of all consistent learning rules, converging at a rate $\Theta(1/m)$, and otherwise (when $\mathfrak{s} = \infty$) the best such bound is $\Theta(\mathrm{Log}(m)/m)$.

Continuing with the investigation of general consistent learning rules, we also find a variant of the argument of Hanneke (2009) that refines (3) in a different way: namely, replacing $\theta(\cdot)$ with a quantity based on considering a well-chosen *subregion* of the region of disagreement, as studied by Balcan, Broder, and Zhang (2007); Zhang and Chaudhuri (2014). Specifically, in the context of active learning, Zhang and Chaudhuri (2014) have proposed a general quantity $\varphi_c(\cdot)$ (Definition 15 below), which is never larger than $\theta(\cdot)$, and is sometimes significantly smaller. By adapting our general argument to replace $\mathrm{DIS}(V_m)$ with this well-chosen subregion, we derive a bound for all consistent learning rules $\hat{h}_m$: with probability at least $1 - \delta$,

$$\mathrm{er}(\hat{h}_m) \lesssim \frac{1}{m}\left(d\mathrm{Log}\left(\varphi_c\left(\frac{d}{m}\right)\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

In particular, as a special case of this general result, we recover the theorem of Balcan and Long (2013) that all consistent learning rules have optimal sample complexity (up to constants) for the problem of learning homogeneous linear separators under isotropic log-concave distributions, as $\varphi_c(d/m)$ is bounded by a finite numerical constant in this case. In Section 6, we also extend this result to the problem of learning with *classification noise*, where there is also a logarithmic factor gap between the known general-case upper and lower bounds. In this context, we derive a new general upper bound under the Bernstein class condition (a generalization of Tsybakov's noise condition), expressed in terms of a quantity related to $\varphi_c(\cdot)$, which applies to a particular learning rule. This sometimes reflects an improvement over the best previous general upper bounds (Massart and Nédélec, 2006; Giné and Koltchinskii, 2006; Hanneke and Yang, 2012), and again recovers a result of Balcan and Long (2013) for homogeneous linear separators under isotropic log-concave distributions, as a special case.

For many of these results, we also state bounds on the *expected* error rate: $\mathbb{E}\left[\mathrm{er}(\hat{h}_m)\right]$. In this case, the optimal distribution-free bound is known to be within a constant factor of d/m (Haussler, Littlestone, and Warmuth, 1994; Li, Long, and Srinivasan, 2001), and this rate is achieved by the one-inclusion graph prediction algorithm of Haussler, Littlestone, and Warmuth (1994), as well as the majority voting method of Hanneke (2016). However, there remain interesting questions about whether other algorithms achieve this optimal performance, or require an extra logarithmic factor. Again we find that *monotone* consistent learning rules indeed achieve this optimal d/m rate (up to constant factors), while a distribution-free bound on $\mathbb{E}\left[\mathrm{er}(\hat{h}_m)\right]$ with $\Theta(1/m)$ dependence on m is achieved by all consistent learning rules if and only if $\mathfrak{s} < \infty$, and otherwise the best such bound has $\Theta(\mathrm{Log}(m)/m)$ dependence on m.

As a final interesting result, in the context of learning with classification noise, under the *bounded* noise assumption (Massart and Nédélec, 2006), we find that the condition $\mathfrak{s} < \infty$ is actually *necessary and sufficient* for the *minimax optimal* excess error rate to decrease at a rate $\Theta(1/m)$, and otherwise (if $\mathfrak{s} = \infty$) it decreases at a rate $\Theta(\mathrm{Log}(m)/m)$. This result generalizes several special-case analyses from the literature (Massart and Nédélec, 2006; Raginsky and Rakhlin, 2011). Note that the "necessity" part of this statement is significantly stronger than the above result for consistent learning rules in the realizable case, since this result applies to the best error guarantee achievable by *any* learning rule.

2. Bounds for Consistent Monotone Learning

In order to state our results for monotonic learning rules in an abstract form, we introduce the following notation. Let $\mathcal{Z}$ denote any space, equipped with a σ-algebra defining the measurable subsets. For any collection $\mathcal{A}$ of measurable subsets of $\mathcal{Z}$, a *consistent monotone rule* is any sequence of functions $\psi_t : \mathcal{Z}^t \to \mathcal{A}$, $t \in \mathbb{N}$, such that $\forall z_1, z_2, \ldots \in \mathcal{Z}$, $\forall t \in \mathbb{N}$, $\psi_t(z_1, \ldots, z_t) \cap \{z_1, \ldots, z_t\} = \emptyset$, and $\forall t \in \mathbb{N}$, $\psi_{t+1}(z_1, \ldots, z_{t+1}) \subseteq \psi_t(z_1, \ldots, z_t)$. We begin with the following very simple result, the proof of which will also serve to introduce, in its simplest form, the core technique underlying many of the results presented in later sections below.

Theorem 1 *Let $\mathcal{A}$ be a collection of measurable subsets of $\mathcal{Z}$, and let $\psi_t : \mathcal{Z}^t \to \mathcal{A}$ (for $t \in \mathbb{N}$) be any consistent monotone rule. Fix any $m \in \mathbb{N}$, any $\delta \in (0,1)$, and any probability measure P on $\mathcal{Z}$. Letting $Z_1, \ldots, Z_m$ be independent P-distributed random variables, and denoting $A_m = \psi_m(Z_1, \ldots, Z_m)$, with probability at least $1-\delta$,*

$$P(A_m) \leq \frac{4}{m}\left(17\mathrm{vc}(\mathcal{A}) + 4\ln\left(\frac{4}{\delta}\right)\right). \tag{6}$$

Furthermore,

$$\mathbb{E}[P(A_m)] \leq \frac{68(\mathrm{vc}(\mathcal{A})+1)}{m}. \tag{7}$$

The overall structure of this proof is based on an argument of Hanneke (2009). The most-significant novel element here is the use of monotonicity to further refine a logarithmic factor. The proof relies on the following classic result. Results of this type are originally due to Vapnik and Chervonenkis (1974); the version stated here features slightly better constant factors, due to Blumer, Ehrenfeucht, Haussler, and Warmuth (1989).

Lemma 2 *For any collection $\mathcal{A}$ of measurable subsets of $\mathcal{Z}$, any $\delta \in (0,1)$, any $m \in \mathbb{N}$, and any probability measure P on $\mathcal{Z}$, letting $Z_1, \ldots, Z_m$ be independent P-distributed random variables, with probability at least $1-\delta$, every $A \in \mathcal{A}$ with $A \cap \{Z_1, \ldots, Z_m\} = \emptyset$ satisfies*

$$P(A) \leq \frac{2}{m}\left(\mathrm{vc}(\mathcal{A})\mathrm{Log}_2\left(\frac{2em}{\mathrm{vc}(\mathcal{A})}\right) + \mathrm{Log}_2\left(\frac{2}{\delta}\right)\right).$$

We are now ready for the proof of Theorem 1.

Proof of Theorem 1 Fix any probability measure P, let $Z_1, Z_2, \ldots$ be independent P-distributed random variables, and for each $m \in \mathbb{N}$ denote $A_m = \psi_m(Z_1, \ldots, Z_m)$. We begin with the inequality in (6). The proof proceeds by induction on m. If $m \leq 200$, then since $\log_2(400e) < 34$ and $\log_2\left(\frac{2}{\delta}\right) < 8\ln\left(\frac{4}{\delta}\right)$, and since the definition of a consistent monotone rule implies $A_m \cap \{Z_1, \ldots, Z_m\} = \emptyset$, the stated bound follows immediately from Lemma 2 for any $\delta \in (0,1)$. Now, as an inductive hypothesis, fix any integer $m \geq 201$ such that, $\forall m' \in [m-1]$, $\forall \delta \in (0,1)$, with probability at least $1-\delta$,

$$P(A_{m'}) \leq \frac{4}{m'}\left(17\mathrm{vc}(\mathcal{A}) + 4\ln\left(\frac{4}{\delta}\right)\right).$$

Now fix any $\delta \in (0,1)$ and define

$$N = \left|\left\{Z_{\lfloor m/2 \rfloor+1}, \ldots, Z_m\right\} \cap A_{\lfloor m/2 \rfloor}\right|,$$

and enumerate the elements of $\{Z_{\lfloor m/2 \rfloor+1}, \ldots, Z_m\} \cap A_{\lfloor m/2 \rfloor}$ as $\hat{Z}_1, \ldots, \hat{Z}_N$ (retaining their original order).

Note that $N = \sum_{t=\lfloor m/2 \rfloor+1}^{m} \mathbb{1}_{A_{\lfloor m/2 \rfloor}}(Z_t)$ is conditionally Binomial$(\lceil m/2 \rceil, P(A_{\lfloor m/2 \rfloor}))$-distributed given $Z_1, \ldots, Z_{\lfloor m/2 \rfloor}$. In particular, with probability one, if $P(A_{\lfloor m/2 \rfloor}) = 0$, then $N = 0$. Otherwise, if $P(A_{\lfloor m/2 \rfloor}) > 0$, then note that $\hat{Z}_1, \ldots, \hat{Z}_N$ are conditionally independent and $P(\cdot|A_{\lfloor m/2 \rfloor})$-distributed given $Z_1, \ldots, Z_{\lfloor m/2 \rfloor}$ and N. Thus, since $A_m \cap \{\hat{Z}_1, \ldots, \hat{Z}_N\} \subseteq A_m \cap \{Z_1, \ldots, Z_m\} = \emptyset$, applying Lemma 2 (under the conditional distribution given N and $Z_1, \ldots, Z_{\lfloor m/2 \rfloor}$), combined with the law of total probability, we have that on an event E_1 of probability at least $1 - \delta/2$, if $N > 0$, then

$$P(A_m | A_{\lfloor m/2 \rfloor}) \leq \frac{2}{N}\left(\mathrm{vc}(\mathcal{A})\mathrm{Log}_2\left(\frac{2eN}{\mathrm{vc}(\mathcal{A})}\right) + \log_2\left(\frac{4}{\delta}\right)\right).$$

Additionally, again since N is conditionally Binomial$(\lceil m/2 \rceil, P(A_{\lfloor m/2 \rfloor}))$-distributed given $Z_1, \ldots, Z_{\lfloor m/2 \rfloor}$, applying a Chernoff bound (under the conditional distribution given $Z_1, \ldots, Z_{\lfloor m/2 \rfloor}$), combined with the law of total probability, we obtain that on an event E_2 of probability at least $1 - \delta/4$, if $P(A_{\lfloor m/2 \rfloor}) \geq \frac{16}{m} \ln\left(\frac{4}{\delta}\right)$, then

$$N \geq P(A_{\lfloor m/2 \rfloor}) \lceil m/2 \rceil / 2 \geq P(A_{\lfloor m/2 \rfloor}) m/4.$$

In particular, if $P(A_{\lfloor m/2 \rfloor}) \geq \frac{16}{m} \ln\left(\frac{4}{\delta}\right)$, then $P(A_{\lfloor m/2 \rfloor}) m/4 > 0$, so that if this occurs with E_2, then we have $N > 0$. Noting that $\mathrm{Log}_2(x) \leq \mathrm{Log}(x)/\ln(2)$, then by monotonicity of $x \mapsto \mathrm{Log}(x)/x$ for $x > 0$, we have that on $E_1 \cap E_2$, if $P(A_{\lfloor m/2 \rfloor}) \geq \frac{16}{m} \ln\left(\frac{4}{\delta}\right)$, then

$$P(A_m | A_{\lfloor m/2 \rfloor}) \leq \frac{8}{P(A_{\lfloor m/2 \rfloor}) m \ln(2)}\left(\mathrm{vc}(\mathcal{A})\mathrm{Log}\left(\frac{eP(A_{\lfloor m/2 \rfloor})m}{2\mathrm{vc}(\mathcal{A})}\right) + \ln\left(\frac{4}{\delta}\right)\right).$$

The monotonicity property of ψ_t implies $A_m \subseteq A_{\lfloor m/2 \rfloor}$. Together with monotonicity of probability measures, this implies $P(A_m) \leq P(A_{\lfloor m/2 \rfloor})$. It also implies that, if $P(A_{\lfloor m/2 \rfloor}) > 0$, then $P(A_m) = P(A_m | A_{\lfloor m/2 \rfloor}) P(A_{\lfloor m/2 \rfloor})$. Thus, on $E_1 \cap E_2$, if $P(A_m) \geq \frac{16}{m} \ln\left(\frac{4}{\delta}\right)$, then

$$P(A_m) \leq \frac{8}{m \ln(2)}\left(\mathrm{vc}(\mathcal{A})\mathrm{Log}\left(\frac{eP(A_{\lfloor m/2 \rfloor})m}{2\mathrm{vc}(\mathcal{A})}\right) + \ln\left(\frac{4}{\delta}\right)\right).$$

The inductive hypothesis implies that, on an event E_3 of probability at least $1 - \delta/4$,

$$P(A_{\lfloor m/2 \rfloor}) \leq \frac{4}{\lfloor m/2 \rfloor}\left(17\mathrm{vc}(\mathcal{A}) + 4\ln\left(\frac{16}{\delta}\right)\right).$$

Since $m \geq 201$, we have $\lfloor m/2 \rfloor \geq (m-2)/2 \geq (199/402)m$, so that the above implies

$$P(A_{\lfloor m/2 \rfloor}) \leq \frac{4 \cdot 402}{199m}\left(17\mathrm{vc}(\mathcal{A}) + 4\ln\left(\frac{16}{\delta}\right)\right).$$

Thus, on $E_1 \cap E_2 \cap E_3$, if $P(A_m) \geq \frac{16}{m} \ln\left(\frac{4}{\delta}\right)$, then

$$P(A_m) \leq \frac{8}{m \ln(2)}\left(\mathrm{vc}(\mathcal{A})\mathrm{Log}\left(\frac{2 \cdot 402e}{199}\left(17 + \frac{4}{\mathrm{vc}(\mathcal{A})}\ln\left(\frac{16}{\delta}\right)\right)\right) + \ln\left(\frac{4}{\delta}\right)\right).$$

Lemma 20 in Appendix A allows us to simplify the logarithmic term here, revealing that the right hand side is at most

$$\begin{aligned}
&\frac{8}{m \ln(2)}\left(\mathrm{vc}(\mathcal{A})\mathrm{Log}\left(\frac{2 \cdot 402e}{199}\left(17 + 4\ln(4) + \frac{4}{\ln(4/e)}\right)\right) + \left(1 + \ln\left(\frac{4}{e}\right)\right)\ln\left(\frac{4}{\delta}\right)\right) \\
&\leq \frac{4}{m}\left(17\mathrm{vc}(\mathcal{A}) + 4\ln\left(\frac{4}{\delta}\right)\right).
\end{aligned}$$

Since $\frac{16}{m}\ln\left(\frac{4}{\delta}\right) \leq \frac{4}{m}\left(17\mathrm{vc}(\mathcal{A}) + 4\ln\left(\frac{4}{\delta}\right)\right)$, we have that, on $E_1 \cap E_2 \cap E_3$, regardless of whether or not $P(A_m) \geq \frac{16}{m}\ln\left(\frac{4}{\delta}\right)$, we have

$$P(A_m) \leq \frac{4}{m}\left(17\mathrm{vc}(\mathcal{A}) + 4\ln\left(\frac{4}{\delta}\right)\right).$$

Noting that, by the union bound, the event $E_1 \cap E_2 \cap E_3$ has probability at least $1-\delta$, this extends the inductive hypothesis to $m' = m$. By the principle of induction, this completes the proof of the first claim in Theorem 1.

For the bound on the expectation in (7), we note that, letting $\varepsilon_m = \frac{4}{m}(17\mathrm{vc}(\mathcal{A}) + 4\ln(4))$, by setting the bound in (6) equal to a value ε and solving for δ, the value of which is in $(0,1)$ for any $\varepsilon > \varepsilon_m$, the result just established can be restated as: $\forall \varepsilon > \varepsilon_m$,

$$\mathbb{P}\left(P(A_m) > \varepsilon\right) \leq 4\exp\left\{(17/4)\mathrm{vc}(\mathcal{A}) - \varepsilon m/16\right\}.$$

Furthermore, for any $\varepsilon \leq \varepsilon_m$, we of course still have $\mathbb{P}\left(P(A_m) > \varepsilon\right) \leq 1$. Therefore, we have that

$$\begin{aligned}
\mathbb{E}\left[P(A_m)\right] &= \int_0^\infty \mathbb{P}\left(P(A_m) > \varepsilon\right) \mathrm{d}\varepsilon \leq \varepsilon_m + \int_{\varepsilon_m}^\infty 4\exp\left\{(17/4)\mathrm{vc}(\mathcal{A}) - \varepsilon m/16\right\} \mathrm{d}\varepsilon \\
&= \varepsilon_m + \frac{4 \cdot 16}{m}\exp\left\{(17/4)\mathrm{vc}(\mathcal{A}) - \varepsilon_m m/16\right\} = \frac{4}{m}\left(17\mathrm{vc}(\mathcal{A}) + 4\ln(4)\right) + \frac{16}{m} \\
&= \frac{4}{m}\left(17\mathrm{vc}(\mathcal{A}) + 4\ln(4e)\right) \leq \frac{68\mathrm{vc}(\mathcal{A}) + 39}{m} \leq \frac{68(\mathrm{vc}(\mathcal{A}) + 1)}{m}.
\end{aligned}$$

■

We can also state a variant of Theorem 1 applicable to *sample compression schemes*, which will in fact be more useful for our purposes below. To state this result, we first introduce the following additional terminology. For any $t \in \mathbb{N}$, we say that a function $\psi : \mathcal{Z}^t \to \mathcal{A}$ is *permutation-invariant* if every $z_1, \ldots, z_t \in \mathcal{Z}$ and every bijection $\kappa : [t] \to [t]$ satisfy $\psi(z_{\kappa(1)}, \ldots, z_{\kappa(t)}) = \psi(z_1, \ldots, z_t)$. For any $n \in \mathbb{N} \cup \{0\}$, a *consistent monotone sample compression rule of size* n is a consistent monotone rule ψ_t with the additional properties that, $\forall t \in \mathbb{N}$, ψ_t is permutation-invariant, and $\forall z_1, \ldots, z_t \in \mathcal{Z}$, $\exists n_t(z_{[t]}) \in [\min\{n, t\}] \cup \{0\}$ such that

$$\psi_t(z_1, \ldots, z_t) = \phi_{t, n_t(z_{[t]})}(z_{i_{t,1}(z_{[t]})}, \ldots, z_{i_{t, n_t(z_{[t]})}(z_{[t]})}),$$

where $\phi_{t,k} : \mathcal{Z}^k \to \mathcal{A}$ is a permutation-invariant function for each $k \in [\min\{n,t\}] \cup \{0\}$, and $i_{t,1}, \ldots, i_{t,n}$ are functions $\mathcal{Z}^t \to [t]$ such that $\forall z_1, \ldots, z_t \in \mathcal{Z}$, $i_{t,1}(z_{[t]}), \ldots, i_{t, n_t(z_{[t]})}(z_{[t]})$ are all distinct. In words, the element of $\mathcal{A}$ mapped to by $\psi_t(z_1, \ldots, z_t)$ depends only on the unordered (multi)set $\{z_1, \ldots, z_t\}$, and can be specified by an unordered subset of $\{z_1, \ldots, z_t\}$ of size at most n. Following the terminology from the literature on sample compression schemes, we refer to the collection of functions $\{(n_t, i_{t,1}, \ldots, i_{t,n_t}) : t \in \mathbb{N}\}$ as the *compression function* of ψ_t, and to the collection of permutation-invariant functions $\{\phi_{t,k} : t \in \mathbb{N}, k \in [\min\{n,t\}] \cup \{0\}\}$ as the *reconstruction function* of ψ_t.

This kind of ψ_t is a type of sample compression scheme (see Littlestone and Warmuth, 1986; Floyd and Warmuth, 1995), though certainly not all permutation-invariant compression schemes yield consistent monotone rules. Below, we find that consistent monotone sample compression rules of a quantifiable size arise naturally in the analysis of certain learning algorithms (namely, the Closure algorithm and the CAL active learning algorithm).

With the above terminology in hand, we can now state our second abstract result.

Theorem 3 *Fix any $n \in \mathbb{N} \cup \{0\}$, let $\mathcal{A}$ be a collection of measurable subsets of $\mathcal{Z}$, and let $\psi_t : \mathcal{Z}^t \to \mathcal{A}$ (for $t \in \mathbb{N}$) be any consistent monotone sample compression rule of size n. Fix any $m \in \mathbb{N}$, $\delta \in (0,1)$, and any probability measure P on $\mathcal{Z}$. Letting $Z_1, \ldots, Z_m$ be independent P-distributed random variables, and denoting $A_m = \psi_m(Z_1, \ldots, Z_m)$, with probability at least $1-\delta$,*

$$P(A_m) \leq \frac{1}{m}\left(21n + 16\ln\left(\frac{3}{\delta}\right)\right). \tag{8}$$

Furthermore,

$$\mathbb{E}[P(A_m)] \leq \frac{21n + 34}{m}. \tag{9}$$

The proof of Theorem 3 relies on the following classic result due to Littlestone and Warmuth (1986); Floyd and Warmuth (1995) (see also Herbrich, 2002; Wiener, Hanneke, and El-Yaniv, 2015, for a clear and direct proof).

Lemma 4 *Fix any collection $\mathcal{A}$ of measurable subsets of $\mathcal{Z}$, any $m \in \mathbb{N}$ and $n \in \mathbb{N} \cup \{0\}$ with $n < m$, and any permutation-invariant functions $\phi_k : \mathcal{Z}^k \to \mathcal{A}$, $k \in [n] \cup \{0\}$. For any probability measure P on $\mathcal{Z}$, letting $Z_1, \ldots, Z_m$ be independent P-distributed random variables, for any $\delta \in (0,1)$, with probability at least $1-\delta$, every $k \in [n] \cup \{0\}$, and every distinct $i_1, \ldots, i_k \in [m]$ with $\phi_k(Z_{i_1}, \ldots, Z_{i_k}) \cap \{Z_1, \ldots, Z_m\} = \emptyset$ satisfy*

$$P\left(\phi_k(Z_{i_1}, \ldots, Z_{i_k})\right) \leq \frac{1}{m-n}\left(n\mathrm{Log}\left(\frac{em}{n}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

With this lemma in hand, we are ready for the proof of Theorem 3.

Proof of Theorem 3 The proof follows analogously to that of Theorem 1, but with several additional complications due to the form of Lemma 4 being somewhat different from that of Lemma 2. Let $\{(n_t, i_{t,1}, \ldots, i_{t,n_t}) : t \in \mathbb{N}\}$ and $\{\phi_{t,k} : t \in \mathbb{N}, k \in [\min\{n,t\}] \cup \{0\}\}$ be the compression function and reconstruction function of ψ_t, respectively. For convenience, also denote $\psi_0() = \mathcal{Z}$, and note that this extends the monotonicity property of ψ_t to $t \in \mathbb{N} \cup \{0\}$. Fix any probability measure P, let $Z_1, Z_2, \ldots$ be independent P-distributed random variables, and for each $m \in \mathbb{N}$ denote $A_m = \psi_m(Z_1, \ldots, Z_m)$.

We begin with the inequality in (8). The special case of $n = 0$ is directly implied by Lemma 4, so for the remainder of the proof of (8), we suppose $n \geq 1$. The proof proceeds by induction on m. Since $P(A) \leq 1$ for all $A \in \mathcal{A}$, and since $21 + 16\ln(3) > 38$, the stated bound is trivially satisfied for all $\delta \in (0,1)$ if $m \leq \max\{38, 21n\}$. Now, as an inductive hypothesis, fix any integer $m > \max\{38, 21n\}$ such that, $\forall m' \in [m-1]$, $\forall \delta \in (0,1)$, with probability at least $1-\delta$,

$$P(A_{m'}) \leq \frac{1}{m'}\left(21n + 16\ln\left(\frac{3}{\delta}\right)\right).$$

Fix any $\delta \in (0,1)$ and define

$$N = \left|\left\{Z_{\lfloor m/2\rfloor+1}, \ldots, Z_m\right\} \cap A_{\lfloor m/2\rfloor}\right|,$$

and enumerate the elements of $\{Z_{\lfloor m/2\rfloor+1}, \ldots, Z_m\} \cap A_{\lfloor m/2\rfloor}$ as $\hat{Z}_1, \ldots, \hat{Z}_N$. Also enumerate the elements of $\{Z_{\lfloor m/2\rfloor+1}, \ldots, Z_m\} \setminus A_{\lfloor m/2\rfloor}$ as $\hat{Z}'_1, \ldots, \hat{Z}'_{\lceil m/2\rceil - N}$. Now note that, by the monotonicity property of ψ_t, we have $A_m \subseteq A_{\lfloor m/2\rfloor}$. Furthermore, by permutation-invariance of ψ_t, we have that

$$A_m = \psi_m\left(\hat{Z}_1, \ldots, \hat{Z}_N, Z_1, \ldots, Z_{\lfloor m/2\rfloor}, \hat{Z}'_1, \ldots, \hat{Z}'_{\lceil m/2\rceil - N}\right).$$

Combined with the monotonicity property of ψ_t, this implies that $A_m \subseteq \psi_N\left(\hat{Z}_1, \ldots, \hat{Z}_N\right)$. Altogether, we have that

$$A_m \subseteq A_{\lfloor m/2\rfloor} \cap \psi_N\left(\hat{Z}_1, \ldots, \hat{Z}_N\right). \tag{10}$$

Note that $N = \sum_{t=\lfloor m/2\rfloor+1}^{m} \mathbb{1}_{A_{\lfloor m/2\rfloor}}(Z_t)$ is conditionally Binomial$(\lceil m/2\rceil, P(A_{\lfloor m/2\rfloor}))$-distributed given $Z_1, \ldots, Z_{\lfloor m/2\rfloor}$. In particular, with probability one, if $P(A_{\lfloor m/2\rfloor}) = 0$, then $N = 0 \le n$. Otherwise, if $P(A_{\lfloor m/2\rfloor}) > 0$, then note that $\hat{Z}_1, \ldots, \hat{Z}_N$ are conditionally independent and $P(\cdot|A_{\lfloor m/2\rfloor})$-distributed given N and $Z_1, \ldots, Z_{\lfloor m/2\rfloor}$. Since ψ_t is a consistent monotone rule, we have that $\psi_N(\hat{Z}_1, \ldots, \hat{Z}_N) \cap \{\hat{Z}_1, \ldots, \hat{Z}_N\} = \emptyset$. We also have, by definition of ψ_N, that $\psi_N(\hat{Z}_1, \ldots, \hat{Z}_N) = \phi_{N, n_N(\hat{Z}_{[N]})}\left(\hat{Z}_{i_{N,1}(\hat{Z}_{[N]})}, \ldots, \hat{Z}_{i_{N,n_N(\hat{Z}_{[N]})}(\hat{Z}_{[N]})}\right)$. Thus, applying Lemma 4 (under the conditional distribution given N and $Z_1, \ldots, Z_{\lfloor m/2\rfloor}$), combined with the law of total probability, we have that on an event E_1 of probability at least $1 - \delta/3$, if $N > n$, then

$$P\left(\psi_N\left(\hat{Z}_1, \ldots, \hat{Z}_N\right)\middle|A_{\lfloor m/2\rfloor}\right) \le \frac{1}{N-n}\left(n \ln\left(\frac{eN}{n}\right) + \ln\left(\frac{3}{\delta}\right)\right).$$

Combined with (10) and monotonicity of measures, this implies that on E_1, if $N > n$, then

$$\begin{aligned} P(A_m) \le P\left(A_{\lfloor m/2\rfloor} \cap \psi_N\left(\hat{Z}_1, \ldots, \hat{Z}_N\right)\right) &= P(A_{\lfloor m/2\rfloor}) P\left(A_{\lfloor m/2\rfloor} \cap \psi_N\left(\hat{Z}_1, \ldots, \hat{Z}_N\right)\middle|A_{\lfloor m/2\rfloor}\right) \\ &\le P(A_{\lfloor m/2\rfloor}) \frac{1}{N-n}\left(n \ln\left(\frac{eN}{n}\right) + \ln\left(\frac{3}{\delta}\right)\right). \end{aligned}$$

Additionally, again since N is conditionally Binomial$(\lceil m/2\rceil, P(A_{\lfloor m/2\rfloor}))$-distributed given $Z_1, \ldots, Z_{\lfloor m/2\rfloor}$, applying a Chernoff bound (under the conditional distribution given $Z_1, \ldots, Z_{\lfloor m/2\rfloor}$), combined with the law of total probability, we obtain that on an event E_2 of probability at least $1 - \delta/3$, if $P(A_{\lfloor m/2\rfloor}) \ge \frac{16}{m} \ln\left(\frac{3}{\delta}\right) \ge \frac{8}{\lceil m/2\rceil} \ln\left(\frac{3}{\delta}\right)$, then

$$N \ge P(A_{\lfloor m/2\rfloor}) \lceil m/2\rceil / 2 \ge P(A_{\lfloor m/2\rfloor}) m/4.$$

Also note that if $P(A_m) \ge \frac{1}{m}\left(21n + 16\ln\left(\frac{3}{\delta}\right)\right)$, then (10) and monotonicity of probability measures imply $P(A_{\lfloor m/2\rfloor}) \ge \frac{1}{m}\left(21n + 16\ln\left(\frac{3}{\delta}\right)\right)$ as well. In particular, if this occurs with

E_2, then we have $N \geq P(A_{\lfloor m/2 \rfloor})m/4 > 5n$. Thus, by monotonicity of $x \mapsto \mathrm{Log}(x)/x$ for $x > 0$, we have that on $E_1 \cap E_2$, if $P(A_m) \geq \frac{1}{m}\left(21n + 16\ln\left(\frac{3}{\delta}\right)\right)$, then

$$\begin{aligned} P(A_m) &< P(A_{\lfloor m/2 \rfloor})\frac{1}{N - (N/5)}\left(n\mathrm{Log}\left(\frac{eN}{n}\right) + \ln\left(\frac{3}{\delta}\right)\right) \\ &\leq \frac{5}{m}\left(n\mathrm{Log}\left(\frac{eP(A_{\lfloor m/2 \rfloor})m}{4n}\right) + \ln\left(\frac{3}{\delta}\right)\right). \end{aligned}$$

The inductive hypothesis implies that, on an event E_3 of probability at least $1 - \delta/3$,

$$P(A_{\lfloor m/2 \rfloor}) \leq \frac{1}{\lfloor m/2 \rfloor}\left(21n + 16\ln\left(\frac{9}{\delta}\right)\right).$$

Since $m \geq 39$, we have $\lfloor m/2 \rfloor \geq (m-2)/2 \geq (37/78)m$, so that the above implies

$$P(A_{\lfloor m/2 \rfloor}) \leq \frac{78}{37m}\left(21n + 16\ln\left(\frac{9}{\delta}\right)\right).$$

Thus, on $E_1 \cap E_2 \cap E_3$, if $P(A_m) \geq \frac{1}{m}\left(21n + 16\ln\left(\frac{3}{\delta}\right)\right)$, then

$$\begin{aligned} P(A_m) &< \frac{5}{m}\left(n\mathrm{Log}\left(\frac{78e}{4 \cdot 37}\left(21 + \frac{16}{n}\ln\left(\frac{9}{\delta}\right)\right)\right) + \ln\left(\frac{3}{\delta}\right)\right) \\ &\leq \frac{5}{m}\left(n\mathrm{Log}\left(\frac{78 \cdot 20}{37 \cdot 11}\left(\frac{21 \cdot 11e}{16 \cdot 5} + \frac{11e}{5}\ln(3) + \frac{11e}{5n}\ln\left(\frac{3}{\delta}\right)\right)\right) + \ln\left(\frac{3}{\delta}\right)\right). \end{aligned}$$

By Lemma 20 in Appendix A, this last expression is at most

$$\frac{5}{m}\left(n\mathrm{Log}\left(\frac{78 \cdot 20}{37 \cdot 11}\left(\frac{21 \cdot 11e}{16 \cdot 5} + \frac{11e}{5}\ln(3) + e\right)\right) + \frac{16}{5}\ln\left(\frac{3}{\delta}\right)\right) < \frac{1}{m}\left(21n + 16\ln\left(\frac{3}{\delta}\right)\right),$$

contradicting the condition $P(A_m) \geq \frac{1}{m}\left(21n + 16\ln\left(\frac{3}{\delta}\right)\right)$. Therefore, on $E_1 \cap E_2 \cap E_3$,

$$P(A_m) < \frac{1}{m}\left(21n + 16\ln\left(\frac{3}{\delta}\right)\right).$$

Noting that, by the union bound, the event $E_1 \cap E_2 \cap E_3$ has probability at least $1 - \delta$, this extends the inductive hypothesis to $m' = m$. By the principle of induction, this completes the proof of the first claim in Theorem 3.

For the bound on the expectation in (9), we note that (as in the proof of Theorem 1), letting $\varepsilon_m = \frac{1}{m}(21n + 16\ln(3))$, the result just established can be restated as: $\forall \varepsilon > \varepsilon_m$,

$$\mathbb{P}\left(P(A_m) > \varepsilon\right) \leq 3\exp\left\{(21/16)n - \varepsilon m/16\right\}.$$

Specifically, this is obtained by setting the bound in (8) equal to ε and solving for δ, the value of which is in $(0,1)$ for any $\varepsilon > \varepsilon_m$. Furthermore, for any $\varepsilon \leq \varepsilon_m$, we of course still have $\mathbb{P}\left(P(A_m) > \varepsilon\right) \leq 1$. Therefore, we have that

$$\begin{aligned} \mathbb{E}\left[P(A_m)\right] &= \int_0^\infty \mathbb{P}\left(P(A_m) > \varepsilon\right) \mathrm{d}\varepsilon \leq \varepsilon_m + \int_{\varepsilon_m}^\infty 3\exp\left\{(21/16)n - \varepsilon m/16\right\} \mathrm{d}\varepsilon \\ &= \varepsilon_m + \frac{3 \cdot 16}{m}\exp\left\{(21/16)n - \varepsilon_m m/16\right\} = \frac{1}{m}\left(21n + 16\ln(3)\right) + \frac{16}{m} \\ &= \frac{1}{m}\left(21n + 16\ln(3e)\right) \leq \frac{21n + 34}{m}. \end{aligned}$$

■

3. Application to the Closure Algorithm for Intersection-Closed Classes

One family of concept spaces studied in the learning theory literature, due to their interesting special properties, is the *intersection-closed* classes (Natarajan, 1987; Helmbold, Sloan, and Warmuth, 1990; Haussler, Littlestone, and Warmuth, 1994; Kuhlmann, 1999; Auer and Ortner, 2007). Specifically, the class $\mathbb{C}$ is called *intersection-closed* if the collection of sets $\{\{x : h(x) = +1\} : h \in \mathbb{C}\}$ is closed under intersections: that is, for every $h, g \in \mathbb{C}$, the classifier $x \mapsto 2\mathbb{1}[h(x) = g(x) = +1] - 1$ is also contained in $\mathbb{C}$. For instance, the class of conjunctions on $\{0,1\}^p$, the class of axis-aligned rectangles on $\mathbb{R}^p$, and the class $\{h : |\{x : h(x) = +1\}| \leq d\}$ of classifiers labeling at most d points positive, are all intersection-closed.

In the context of learning in the realizable case, there is a general learning strategy, called the *Closure* algorithm, designed for learning with intersection-closed concept spaces, which has been a subject of frequent study. Specifically, for any $m \in \mathbb{N} \cup \{0\}$, given any data set $L_m = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ with $\mathbb{C}[L_m] \neq \emptyset$, the Closure algorithm $\mathbb{A}(L_m)$ for $\mathbb{C}$ produces the classifier $\hat{h}_m : \mathcal{X} \to \mathcal{Y}$ with $\{x : \hat{h}_m(x) = +1\} = \bigcap_{h \in \mathbb{C}[L_m]} \{x : h(x) = +1\}$: that is, $\hat{h}_m(x) = +1$ if and only if every $h \in \mathbb{C}$ consistent with L_m (i.e., $\mathrm{er}_{L_m}(h) = 0$) has $h(x) = +1$.[1] Defining $\bar{\mathbb{C}}$ as the set of all classifiers $h : \mathcal{X} \to \mathcal{Y}$ for which there exists a nonempty $\mathcal{G} \subseteq \mathbb{C}$ with $\{x : h(x) = +1\} = \bigcap_{g \in \mathcal{G}} \{x : g(x) = +1\}$, Auer and Ortner (2007) have argued that $\bar{\mathbb{C}}$ is an intersection-closed concept space containing $\mathbb{C}$, with $\mathrm{vc}(\bar{\mathbb{C}}) = \mathrm{vc}(\mathbb{C})$. Thus, for $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$ (where $\mathbb{A}$ is the Closure algorithm), since $\hat{h}_m \in \bar{\mathbb{C}}[\mathcal{L}_m]$, Lemma 2 immediately implies that, for any $m \in \mathbb{N}$, with probability at least $1 - \delta$, $\mathrm{er}\left(\hat{h}_m\right) \lesssim \frac{1}{m}\left(d\mathrm{Log}(\frac{m}{d}) + \mathrm{Log}(\frac{1}{\delta})\right)$. However, by a more-specialized analysis, Auer and Ortner (2004, 2007) were able to show that, for intersection-closed classes $\mathbb{C}$, the Closure algorithm in fact achieves $\mathrm{er}\left(\hat{h}_m\right) \lesssim \frac{1}{m}\left(d\mathrm{Log}(d) + \mathrm{Log}(\frac{1}{\delta})\right)$ with probability at least $1 - \delta$, which is an improvement for large m. They also argued that, for a special subfamily of intersection-closed classes (namely, those with *homogeneous spans*), this bound can be further refined to $\frac{1}{m}\left(d + \mathrm{Log}(\frac{1}{\delta})\right)$, which matches (up to constant factors) the lower bound (2). However, they left open the question of whether this refinement is achievable for general intersection-closed concept spaces (by Closure, or any other algorithm).

In the following result, we prove that the Closure algorithm indeed always achieves the optimal bound (up to constant factors) for intersection-closed concept spaces, as a simple consequence of either Theorem 1 or Theorem 3. This fact was very recently also obtained by Darnstädt (2015) via a related direct approach; however, we note that the constant factors obtained here are significantly smaller (by roughly a factor of 15.5, for large d).

Theorem 5 *If $\mathbb{C}$ is intersection-closed and $\mathbb{A}$ is the Closure algorithm, then for any $m \in \mathbb{N}$ and $\delta \in (0,1)$, letting $\hat{h}_m = \mathbb{A}(\{(X_1, f^\star(X_1)), \ldots, (X_m, f^\star(X_m))\})$, with probability at least $1 - \delta$,*

$$\mathrm{er}\left(\hat{h}_m\right) \leq \frac{1}{m}\left(21d + 16\ln\left(\frac{3}{\delta}\right)\right).$$

1. For simplicity, we suppose $\mathbb{C}$ is such that this set $\bigcap_{h \in \mathbb{C}[L_m]} \{x : h(x) = +1\}$ is measurable for every L_m, which is the case for essentially all intersection-closed concept spaces of practical interest.

Furthermore,

$$\mathbb{E}\left[\mathrm{er}\left(\hat{h}_m\right)\right] \leq \frac{21d+34}{m}.$$

Proof For each $t \in \mathbb{N} \cup \{0\}$ and $x_1, \ldots, x_t \in \mathcal{X}$, define $\psi_t(x_1, \ldots, x_t) = \{x \in \mathcal{X} : \hat{h}_{x_{[t]}}(x) \neq f^\star(x)\}$, where $\hat{h}_{x_{[t]}} = \mathbb{A}(\{(x_1, f^\star(x_1)), \ldots, (x_t, f^\star(x_t))\})$. Fix any $x_1, x_2, \ldots \in \mathcal{X}$, let $L_t = \{(x_1, f^\star(x_1)), \ldots, (x_t, f^\star(x_t))\}$ for each $t \in \mathbb{N}$, and note that for any $t \in \mathbb{N}$, the classifier $\hat{h}_{x_{[t]}}$ produced by $\mathbb{A}(L_t)$ is consistent with L_t, which implies $\psi_t(x_1, \ldots, x_t) \cap \{x_1, \ldots, x_t\} = \emptyset$. Furthermore, since $f^\star \in \mathbb{C}[L_t]$, we have that $\{x : \hat{h}_{x_{[t]}}(x) = +1\} \subseteq \{x : f^\star(x) = +1\}$, which together with the definition of $\hat{h}_{x_{[t]}}$ implies

$$\begin{aligned}\psi_t(x_1, \ldots, x_t) &= \{x \in \mathcal{X} : \hat{h}_{x_{[t]}}(x) = -1, f^\star(x) = +1\} \\ &= \bigcup_{h \in \mathbb{C}[L_t]} \{x \in \mathcal{X} : h(x) = -1, f^\star(x) = +1\} \end{aligned} \tag{11}$$

for every $t \in \mathbb{N}$. Furthermore, for any $t \in \mathbb{N}$, $\mathbb{C}[L_{t+1}] \subseteq \mathbb{C}[L_t]$. Together with monotonicity of the union, these two observations imply

$$\begin{aligned}\psi_{t+1}(x_1, \ldots, x_{t+1}) &= \bigcup_{h \in \mathbb{C}[L_{t+1}]} \{x \in \mathcal{X} : h(x) = -1, f^\star(x) = +1\} \\ &\subseteq \bigcup_{h \in \mathbb{C}[L_t]} \{x \in \mathcal{X} : h(x) = -1, f^\star(x) = +1\} = \psi_t(x_1, \ldots, x_t).\end{aligned}$$

Thus, ψ_t defines a consistent monotone rule. Also, since $\mathbb{A}$ always produces a function in $\bar{\mathbb{C}}$, we have $\psi_t(x_1, \ldots, x_t) \in \{\{x \in \mathcal{X} : h(x) \neq f^\star(x)\} : h \in \bar{\mathbb{C}}\}$ for every $t \in \mathbb{N}$, and it is straightforward to show that the VC dimension of this collection of sets is exactly $\mathrm{vc}(\bar{\mathbb{C}})$ (see Vidyasagar, 2003, Lemma 4.12), which Auer and Ortner (2007) have argued equals d. From this, we can already infer a bound $\frac{4}{m}\left(17d + 4\ln\left(\frac{4}{\delta}\right)\right)$ via Theorem 1. However, we can refine the constant factors in this bound by noting that ψ_t can also be represented as a consistent monotone sample compression rule of size d, and invoking Theorem 3. The rest of this proof focuses on establishing this fact.

Fix any $t \in \mathbb{N}$. It is well known in the literature (see e.g., Auer and Ortner, 2007, Theorem 1) that there exist $k \in [d] \cup \{0\}$ and distinct $i_1, \ldots, i_k \in [t]$ such that $f^\star(x_{i_j}) = +1$ for all $j \in [k]$, and letting $L_{i_{[k]}} = \{(x_{i_1}, +1), \ldots, (x_{i_k}, +1)\}$, we have $\bigcap_{h \in \mathbb{C}[L_{i_{[k]}}]}\{x : h(x) = +1\} = \bigcap_{h \in \mathbb{C}[L_t]}\{x : h(x) = +1\}$; in particular, letting $\hat{h}_{x_{i_{[k]}}} = \mathbb{A}(L_{i_{[k]}})$, this implies $\hat{h}_{x_{i_{[k]}}} = \hat{h}_{x_{[t]}}$. This further implies $\psi_t(x_1, \ldots, x_t) = \psi_k(x_{i_1}, \ldots, x_{i_k})$, so that defining the compression function $(n_t(x_{[t]}), i_{t,1}(x_{[t]}), \ldots, i_{t,n_t(x_{[t]})}(x_{[t]})) = (k, i_1, \ldots, i_k)$ for k and $i_1, \ldots, i_k$ as above, for each $x_1, \ldots, x_t \in \mathcal{X}$, and defining the reconstruction function $\phi_{t,k'}(x'_1, \ldots, x'_{k'}) = \psi_{k'}(x'_1, \ldots, x'_{k'})$ for each $t \in \mathbb{N}$, $k' \in [d] \cup \{0\}$, and $x'_1, \ldots, x'_{k'} \in \mathcal{X}$, we have that $\psi_t(x_1, \ldots, x_t) = \phi_{t,n_t(x_{[t]})}(x_{i_{t,1}(x_{[t]})}, \ldots, x_{i_{t,n_t(x_{[t]})}(x_{[t]})})$ for all $t \in \mathbb{N}$ and $x_1, \ldots, x_t \in \mathcal{X}$. Furthermore, since $(x_1, \ldots, x_t) \mapsto \mathbb{C}[\{(x_1, f^\star(x_1)), \ldots, (x_t, f^\star(x_t))\}]$ is invariant to permutations of its arguments, it follows from (11) that ψ_t is permutation-invariant for every $t \in \mathbb{N}$; this also means that, for the choice of $\phi_{t,k'}$ above, the function $\phi_{t,k'}$ is also permutation-invariant. Altogether, we have that ψ_t is a consistent monotone sample compression rule of size d. Thus, since $\mathrm{er}\left(\hat{h}_m\right) = \mathcal{P}(\psi_m(X_1, \ldots, X_m))$ for $m \in \mathbb{N}$, the stated result follows directly from Theorem 3 (with $\mathcal{Z} = \mathcal{X}$, $P = \mathcal{P}$, and ψ_t defined as above). ■

4. Application to the CAL Active Learning Algorithm

As another interesting application of Theorem 3, we derive an improved bound on the label complexity of a well-studied active learning algorithm, usually referred to as CAL after its authors Cohn, Atlas, and Ladner (1994). Formally, in the active learning protocol, the learning algorithm $\mathbb{A}$ is given access to the *unlabeled* data sequence $X_1, X_2, \ldots$ (or some sufficiently-large finite initial segment thereof), and then sequentially requests to observe the labels: that is, it selects an index t_1 and requests to observe the label $f^\star(X_{t_1})$, at which time it is permitted access to $f^\star(X_{t_1})$; it may then select another index t_2 and request to observe the label $f^\star(X_{t_2})$, is then permitted access to $f^\star(X_{t_2})$, and so on. This continues until at most some given number n of labels have been requested (called the *label budget*), at which point the algorithm should halt and return a classifier $\hat{h}$; we denote this as $\hat{h} = \mathbb{A}(n)$ (leaving the dependence on the unlabeled data implicit, for simplicity). We are then interested in characterizing a sufficient size for the budget n so that, with probability at least $1 - \delta$, $\mathrm{er}(\hat{h}) \le \varepsilon$; this size is known as the *label complexity* of $\mathbb{A}$.

The CAL active learning algorithm is based on a very elegant and natural principle: never request a label that can be deduced from information already obtained. CAL is defined solely by this principle, employing no additional criteria in its choice of queries. Specifically, the algorithm proceeds by considering randomly-sampled data points one at a time, and to each it applies the above principle, skipping over the labels that can be deduced, and requesting the labels that cannot be. In favorable scenarios, as the number of label requests grows, the frequency of encountering a sample whose label cannot be deduced should diminish. The key to bounding the label complexity of CAL is to characterize the rate at which this frequency shrinks. To further pursue this discussion with rigor, let us define the *region of disagreement* for any set $\mathcal{H}$ of classifiers:

$$\mathrm{DIS}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } h(x) \neq g(x)\}.$$

Then the CAL active learning algorithm is formally defined as follows.

Algorithm: CAL(n)
0. $m \leftarrow 0$, $t \leftarrow 0$, $V_0 \leftarrow \mathbb{C}$
1. While $t < n$ and $m < 2^n$
2. $m \leftarrow m + 1$
3.   If $X_m \in \mathrm{DIS}(V_{m-1})$
4.   Request label $Y_m = f^\star(X_m)$; let $V_m \leftarrow V_{m-1}[\{(X_m, Y_m)\}]$, $t \leftarrow t + 1$
5.   Else $V_m \leftarrow V_{m-1}$
6. Return any $\hat{h} \in V_m$

This algorithm has several attractive properties. One is that, since it only removes classifiers from V_m upon disagreement with $f^\star$, it maintains the invariant that $f^\star \in V_m$.

Another property is that, since it maintains $f^\star \in V_m$, and it only refrains from requesting a label if every classifier in V_m agrees on the label (and hence agrees with $f^\star$, so that requesting the label would not affect V_m anyway), it maintains the invariant that $V_m = \mathbb{C}[\mathcal{L}_m]$, where $\mathcal{L}_m = \{(X_1, f^\star(X_1)), \ldots, (X_m, f^\star(X_m))\}$.

This algorithm has been studied a great deal in the literature (Cohn, Atlas, and Ladner, 1994; Hanneke, 2009, 2011, 2012, 2014; El-Yaniv and Wiener, 2012; Wiener, Hanneke, and El-Yaniv, 2015), and has inspired an entire genre of active learning algorithms referred to as *disagreement-based* (or sometimes as *mellow*), including several methods possessing desirable properties such as robustness to classification noise (e.g., Balcan, Beygelzimer, and Langford, 2006, 2009; Dasgupta, Hsu, and Monteleoni, 2007; Koltchinskii, 2010; Hanneke and Yang, 2012; Hanneke, 2014). There is a substantial literature studying the label complexity of CAL and other disagreement-based active learning algorithms; the interested reader is referred to the recent survey article of Hanneke (2014) for a thorough discussion of this literature. Much of that literature discusses characterizations of the label complexity in terms of a quantity known as the *disagreement coefficient* (Hanneke, 2007b, 2009). However, Wiener, Hanneke, and El-Yaniv (2015) have recently discovered that a quantity known as the *version space compression set size* (a.k.a. *empirical teaching dimension*) can sometimes provide a smaller bound on the label complexity of CAL. Specifically, the following quantity was introduced in the works of El-Yaniv and Wiener (2010); Hanneke (2007a).

Definition 6 *For any $m \in \mathbb{N}$ and $\mathcal{L} \in (\mathcal{X} \times \mathcal{Y})^m$, the* version space compression set $\hat{C}_{\mathcal{L}}$ *is a smallest subset of $\mathcal{L}$ satisfying $\mathbb{C}[\hat{C}_{\mathcal{L}}] = \mathbb{C}[\mathcal{L}]$. We then define $\hat{n}(\mathcal{L}) = |\hat{C}_{\mathcal{L}}|$, the* version space compression set size*. In the special case $\mathcal{L} = \mathcal{L}_m$, we abbreviate $\hat{n}_m = \hat{n}(\mathcal{L}_m)$. Also define $\hat{n}_{1:m} = \max_{t \in [m]} \hat{n}_t$, and for any $\delta \in (0,1)$, define $\tilde{n}_m(\delta) = \min\{b \in [m] \cup \{0\} : \mathbb{P}(\hat{n}_m \leq b) \geq 1 - \delta\}$ and $\tilde{n}_{1:m}(\delta) = \min\{b \in [m] \cup \{0\} : \mathbb{P}(\hat{n}_{1:m} \leq b) \geq 1 - \delta\}$.*

The recent work of Wiener, Hanneke, and El-Yaniv (2015) studies this quantity for several concept spaces and distributions, and also identifies general relations between $\hat{n}_m$ and the more-commonly studied disagreement coefficient θ of (Hanneke, 2007b, 2009). Specifically, for any $r > 0$, define $\mathrm{B}(f^\star, r) = \{h \in \mathbb{C} : \mathcal{P}(x : h(x) \neq f^\star(x)) \leq r\}$. Then the disagreement coefficient is defined, for any $r_0 \geq 0$, as

$$\theta(r_0) = \sup_{r > r_0} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, r)))}{r} \vee 1.$$

Both $\tilde{n}_{1:m}(\delta)$ and $\theta(r_0)$ are complexity measures dependent on $f^\star$ and $\mathcal{P}$. Wiener, Hanneke, and El-Yaniv (2015) relate them by showing that

$$\theta(1/m) \lesssim \tilde{n}_{1:m}(1/20) \vee 1, \tag{12}$$

and for general $\delta \in (0,1)$,[2]

$$\tilde{n}_{1:m}(\delta) \lesssim \theta(d/m) \left(d\mathrm{Log}(\theta(d/m)) + \mathrm{Log}\left(\frac{\mathrm{Log}(m)}{\delta}\right)\right) \mathrm{Log}(m). \tag{13}$$

2. The original claim from Wiener, Hanneke, and El-Yaniv (2015) involved a maximum of minimal $(1-\delta)$-confidence bounds on $\hat{n}_t$ over $t \in [m]$, but the same proof can be used to establish this slightly stronger claim.

Wiener, Hanneke, and El-Yaniv (2015) prove that, for CAL(n) to produce $\hat{h}$ with $\operatorname{er}(\hat{h}) \leq \varepsilon$ with probability at least $1-\delta$, it suffices to take a budget n of size proportional to

$$\left(\max_{m \in [M(\varepsilon,\delta/2)]} \tilde{n}_m(\delta_m) \operatorname{Log}\left(\frac{m}{\tilde{n}_m(\delta_m)} \right) + \operatorname{Log}\left(\frac{\operatorname{Log}(M(\varepsilon,\delta/2))}{\delta} \right) \right) \operatorname{Log}(M(\varepsilon,\delta/2)), \quad (14)$$

where the values $\delta_m \in (0,1]$ are such that $\sum_{i=0}^{\lfloor \log_2(M(\varepsilon,\delta/2)) \rfloor} \delta_{2^i} \leq \delta/4$, and $M(\varepsilon,\delta/2)$ is the smallest $m \in \mathbb{N}$ for which $\mathbb{P}\left(\sup_{h \in \mathbb{C}[\mathcal{L}_m]} \operatorname{er}(h) \leq \varepsilon \right) \geq 1-\delta/2$; the quantity $M(\varepsilon,\delta)$ is discussed at length below in Section 5. They also argue that this is essentially a *tight* characterization of the label complexity of CAL, up to logarithmic factors.

The key to obtaining this result is establishing an upper bound on $\mathcal{P}(\operatorname{DIS}(V_m))$ as a function of m, where (as in CAL) $V_m = \mathbb{C}[\mathcal{L}_m]$. One basic observation indicating that $\mathcal{P}(\operatorname{DIS}(V_m))$ can be related to the version space compression set size is that, by exchangeability of the X_i random variables,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{P}(\operatorname{DIS}(V_m))\right] &= \mathbb{E}\left[\mathbb{1}\left[X_{m+1} \in \operatorname{DIS}(\mathbb{C}[\mathcal{L}_m])\right]\right] \\
&= \frac{1}{m+1} \sum_{i=1}^{m+1} \mathbb{E}\left[\mathbb{1}\left[X_i \in \operatorname{DIS}(\mathbb{C}[\mathcal{L}_{m+1} \setminus \{(X_i, f^\star(X_i))\}])\right]\right] \\
&\leq \frac{1}{m+1} \sum_{i=1}^{m+1} \mathbb{E}\left[\mathbb{1}\left[(X_i, f^\star(X_i)) \in \hat{C}_{\mathcal{L}_{m+1}}\right]\right] = \frac{\mathbb{E}\left[\hat{n}_{m+1}\right]}{m+1},
\end{aligned}$$

where the inequality is due to the observation that any $X_i \in \operatorname{DIS}(\mathbb{C}[\mathcal{L}_{m+1} \setminus \{(X_i, f^\star(X_i))\}])$ is necessarily in the version space compression set $\hat{C}_{\mathcal{L}_{m+1}}$, and the last equality is by linearity of the expectation. However, obtaining the bound (14) required a more-involved argument from Wiener, Hanneke, and El-Yaniv (2015), to establish a high-confidence bound on $\mathcal{P}(\operatorname{DIS}(V_m))$, rather than a bound on its expectation. Specifically, by combining a perspective introduced by El-Yaniv and Wiener (2010, 2012), with the observation that $\operatorname{DIS}(V_m)$ may be represented as a sample compression scheme of size $\hat{n}_m$, and invoking Lemma 4, Wiener, Hanneke, and El-Yaniv (2015) prove that, with probability at least $1-\delta$,

$$\mathcal{P}(\operatorname{DIS}(V_m)) \lesssim \frac{1}{m} \left(\hat{n}_m \operatorname{Log}\left(\frac{m}{\hat{n}_m} \right) + \operatorname{Log}\left(\frac{1}{\delta} \right) \right). \quad (15)$$

In the present work, we are able to entirely eliminate the factor $\operatorname{Log}\left(\frac{m}{\hat{n}_m} \right)$ from the first term, simply by observing that the region $\operatorname{DIS}(V_m)$ is *monotonic* in m. Specifically, by combining this monotonicity observation with the description of $\operatorname{DIS}(V_m)$ as a compression scheme from Wiener, Hanneke, and El-Yaniv (2015), the refined bound follows from arguments similar to the proof of Theorem 3. Formally, we have the following result.

Theorem 7 *For any $m \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathcal{P}(\operatorname{DIS}(V_m)) \leq \frac{16}{m} \left(2\hat{n}_{1:m} + \ln\left(\frac{3}{\delta} \right) \right).$$

We should note that, while Theorem 7 indeed eliminates a logarithmic factor compared to (15), this refinement is also accompanied by an increase in the complexity measure, replacing $\hat{n}_m$ with $\hat{n}_{1:m}$. This arises from our proof, since (as in the proof of Theorem 3) the argument relies on $\hat{n}_{1:m}$ being a sample compression set size, not just for the full sample, but also for any prefix of the sample. The effect of this increase is largely benign in this context, since the bound (14) on the label complexity of CAL, derived from (15), involves maximization over the sample size anyway.

Although Theorem 7 follows from the same principles as Theorem 3 (i.e., $\mathrm{DIS}(V_t)$ being a consistent monotone rule expressible as a sample compression scheme), it does not quite follow as an immediate consequence of Theorem 3, due fact that the size $\hat{n}_{1:m}$ of the sequence of sample compression schemes can vary based on the specific samples (including their *order*). For this reason, we provide a specialized proof of this result in Appendix B, which follows an argument nearly-identical to that of Theorem 3, with only a few minor changes to account for this variability of $\hat{n}_{1:m}$ using special properties of the sets $\mathrm{DIS}(V_t)$.

Based on this result, and following precisely the same arguments as Wiener, Hanneke, and El-Yaniv (2015),[3] we arrive at the following bound on the label complexity of CAL. For brevity, we omit the proof, referring the interested reader to the original exposition of Wiener, Hanneke, and El-Yaniv (2015) for the details.

Theorem 8 *There is a universal constant $c \in (0, \infty)$ such that, for any $\varepsilon, \delta \in (0, 1)$, for any $n \in \mathbb{N}$ with*

$$n \geq c \left(\tilde{n}_{1:M(\varepsilon,\delta/2)}(\delta/4) + \mathrm{Log}\left(\frac{\mathrm{Log}(M(\varepsilon, \delta/2))}{\delta} \right) \right) \mathrm{Log}(M(\varepsilon, \delta/2)),$$

with probability at least $1 - \delta$, the classifier $\hat{h}_n = \mathrm{CAL}(n)$ has $\mathrm{er}(\hat{h}_n) \leq \varepsilon$.

It is also possible to state a distribution-free variant of Theorem 7. Specifically, consider the following definition, from Hanneke and Yang (2015).

Definition 9 *The* star number $\mathfrak{s}$ *is the largest integer s such that there exist distinct points $x_1, \ldots, x_s \in \mathcal{X}$ and classifiers $h_0, h_1, \ldots, h_s \in \mathbb{C}$ with the property that $\forall i \in [s]$, $\mathrm{DIS}(\{h_0, h_i\}) \cap \{x_1, \ldots, x_s\} = \{x_i\}$; if no such largest integer exists, define $\mathfrak{s} = \infty$.*

The star number is a natural combinatorial complexity measure, corresponding to the largest possible degree in the data-induced one-inclusion graph. Hanneke and Yang (2015) provide several examples of concept spaces exhibiting a variety of values for the star number (though it should be noted that many commonly-used concept spaces have $\mathfrak{s} = \infty$: e.g., linear separators). As a basic relation, one can easily show that $\mathfrak{s} \geq d$. Hanneke and Yang (2015) also relate the star number to many other complexity measures arising in the learning theory literature, including $\hat{n}_m$. Specifically, they prove that, for every $m \in \mathbb{N}$ and

3. The only small twist is that we replace $\max_{m \leq M(\varepsilon,\delta/2)} \tilde{n}_m(\delta_m)$ from (14) with $\tilde{n}_{1:M(\varepsilon,\delta/2)}(\delta/4)$. As the purpose of these $\tilde{n}_m(\delta_m)$ values in the original proof is to provide bounds on their respective $\hat{n}_m$ values (which in our context, are $\hat{n}_{1:m}$ values), holding simultaneously for all $m = 2^i \in [M(\varepsilon, \delta/2)]$ with probability at least $1 - \delta/4$, the value $\tilde{n}_{1:M(\varepsilon,\delta/2)}(\delta/4)$ can clearly be used instead. If desired, by a union bound we can of course bound $\tilde{n}_{1:M(\varepsilon,\delta/2)}(\delta/4) \leq \max_{m \in [M(\varepsilon,\delta/2)]} \tilde{n}_m(\delta_m)$, for any sequence δ_m in $(0, 1]$ with $\sum_{m \in [M(\varepsilon,\delta/2)]} \delta_m \leq \delta/4$.

$\mathcal{L} \in (\mathcal{X} \times \mathcal{Y})^m$ with $\mathbb{C}[\mathcal{L}] \neq \emptyset$, $\hat{n}(\mathcal{L}) \leq \mathfrak{s}$, with equality in the worst case (over m and $\mathcal{L}$). Based on this fact, Theorem 3 implies the following result.

Theorem 10 *For any $m \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathcal{P}(\mathrm{DIS}(V_m)) \leq \frac{1}{m}\left(21\mathfrak{s} + 16 \ln\left(\frac{3}{\delta}\right)\right).$$

Proof For every $t \in \mathbb{N}$ and $x_1, \ldots, x_t \in \mathcal{X}$, define $\psi_t(x_1, \ldots, x_t) = \mathrm{DIS}(\mathbb{C}[\mathcal{L}_{x_{[t]}}])$, where $\mathcal{L}_{x_{[t]}} = \{(x_1, f^\star(x_1)), \ldots, (x_t, f^\star(x_t))\}$; ψ_t is clearly permutation-invariant, and satisfies $\psi_t(x_1, \ldots, x_t) \cap \{x_1, \ldots, x_t\} = \emptyset$ (since every $h \in \mathbb{C}[\mathcal{L}_{x_{[t]}}]$ agrees with $f^\star$ on $\{x_1, \ldots, x_t\}$). Furthermore, monotonicity of $\mathcal{L} \mapsto \mathbb{C}[\mathcal{L}]$ and $\mathcal{H} \mapsto \mathrm{DIS}(\mathcal{H})$ imply that any $t \in \mathbb{N}$ and $x_1, \ldots, x_{t+1} \in \mathcal{X}$ satisfy $\psi_{t+1}(x_1, \ldots, x_{t+1}) \subseteq \psi_t(x_1, \ldots, x_t)$, so that ψ_t is a consistent monotone rule. Also define $\phi_{t,k}(x_1, \ldots, x_k) = \psi_k(x_1, \ldots, x_k)$ for any $k \in [t]$ and $x_1, \ldots, x_k \in \mathcal{X}$, and $\phi_{t,0}() = \mathrm{DIS}(\mathbb{C})$. Since ψ_k is permutation-invariant for every $k \in [t]$, so is $\phi_{t,k}$. For any $x_1, \ldots, x_t \in \mathcal{X}$, from Definition 6, there exist distinct $i_{t,1}(x_{[t]}), \ldots, i_{t,\hat{n}(\mathcal{L}_{x_{[t]}})}(x_{[t]}) \in [t]$ such that $\hat{C}_{\mathcal{L}_{x_{[t]}}} = \{(x_{i_{t,j}(x_{[t]})}, f^\star(x_{i_{t,j}(x_{[t]})})) : j \in \{1, \ldots, \hat{n}(\mathcal{L}_{x_{[t]}})\}\}$, and since $\mathbb{C}[\hat{C}_{\mathcal{L}_{x_{[t]}}}] = \mathbb{C}[\mathcal{L}_{x_{[t]}}]$, it follows that $\phi_{t,\hat{n}(\mathcal{L}_{x_{[t]}})}(x_{i_{t,1}(x_{[t]})}, \ldots, x_{i_{t,\hat{n}(\mathcal{L}_{x_{[t]}})}(x_{[t]})}) = \psi_t(x_1, \ldots, x_t)$. Thus, since $\hat{n}(\mathcal{L}_{x_{[t]}}) \leq \mathfrak{s}$ for all $t \in \mathbb{N}$ (Hanneke and Yang, 2015), ψ_t is a consistent monotone sample compression rule of size $\mathfrak{s}$. The result immediately follows by applying Theorem 3 with $\mathcal{Z} = \mathcal{X}$, $P = \mathcal{P}$, and ψ_t as above. ■

As a final implication for CAL, we can also plug the inequality $\hat{n}(\mathcal{L}) \leq \mathfrak{s}$ into the bound from Theorem 8 to reveal that CAL achieves a label complexity upper-bounded by a value proportional to $\mathfrak{s}\mathrm{Log}(M(\varepsilon, \delta/2)) + \mathrm{Log}\left(\frac{\mathrm{Log}(M(\varepsilon,\delta/2))}{\delta}\right)\mathrm{Log}(M(\varepsilon, \delta/2))$.

Remark: In addition to the above applications to active learning, it is worth noting that, combined with the work of El-Yaniv and Wiener (2010), the above results also have implications for the setting of *selective classification*: that is, the setting in which, for each $t \in \mathbb{N}$, given access to $(X_1, f^\star(X_1)), \ldots, (X_{t-1}, f^\star(X_{t-1}))$ and X_t, a learning algorithm is required either to make a prediction $\hat{Y}_t$ for $f^\star(X_t)$, or to "abstain" from prediction; after each round t, the algorithm is permitted access to the value $f^\star(X_t)$. Then the error rate is the probability the prediction $\hat{Y}_t$ is incorrect (conditioned on $X_{[t-1]}$), given that the algorithm chooses to predict, and the *coverage* is the probability the algorithm chooses to make a prediction at time t (conditioned on $X_{[t-1]}$). El-Yaniv and Wiener (2010) explore an extreme variant, called *perfect selective classification*, in which the algorithm is required to only make predictions that will be correct with *certainty* (i.e., for any data sequence $x_1, x_2, \ldots$, the algorithm will never misclassify a point it chooses to predict for). El-Yaniv and Wiener (2010) find that a selective classification algorithm based on principles analogous to the CAL active learning algorithm obtains the optimal coverage among all perfect selective classification algorithms; the essential strategy is to predict only if $X_t \notin \mathrm{DIS}(V_{t-1})$, taking $\hat{Y}_t$ as the label agreed-upon by every $h \in V_{t-1}$. In particular, this implies that the optimal coverage rate in perfect selective classification, on round t, is $1 - \mathcal{P}(\mathrm{DIS}(V_{t-1}))$. Thus, combined with Theorem 7 or Theorem 10, we can immediately obtain bounds on the optimal coverage rate for perfect selective classification as well; in particular, this typically refines the bound of

El-Yaniv and Wiener (2010) (and a later refinement by Wiener, Hanneke, and El-Yaniv, 2015) by at least a logarithmic factor (though again, it is not a "pure" improvement, as Theorem 7 uses $\hat{n}_{1:m}$ in place of $\hat{n}_m$).

5. Application to General Consistent PAC Learners

In general, a *consistent learning algorithm* $\mathbb{A}$ is a learning algorithm such that, for any $m \in \mathbb{N}$ and $L \in (\mathcal{X} \times \mathcal{Y})^m$ with $\mathbb{C}[L] \neq \emptyset$, $\mathbb{A}(L)$ produces a classifier $\hat{h}$ consistent with L (i.e., $\hat{h} \in \mathbb{C}[L]$). In the context of learning in the realizable case, this is equivalent to $\mathbb{A}$ being an instance of the well-studied method of *empirical risk minimization*. The study of general consistent learning algorithms focuses on the quantity $\sup_{h \in V_m} \mathrm{er}(h)$, where $V_m = \mathbb{C}[\mathcal{L}_m]$, as above. It is clear that the error rate achieved by any consistent learning algorithm, given $\mathcal{L}_m$ as input, is at most $\sup_{h \in V_m} \mathrm{er}(h)$. Furthermore, it is not hard to see that, for any given $\mathcal{P}$ and $f^\star \in \mathbb{C}$, there exist consistent learning rules obtaining error rates arbitrarily close to $\sup_{h \in V_m} \mathrm{er}(h)$, so that obtaining guarantees on the error rate that hold *generally* for all consistent learning algorithms requires us to bound this value.

Based on Lemma 2 (taking $\mathcal{A} = \{\{x : h(x) \neq f^\star(x)\} : h \in \mathbb{C}\}$), one immediately obtains a classic result (due to Vapnik and Chervonenkis, 1974; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989), that with probability at least $1 - \delta$,

$$\sup_{h \in V_m} \mathrm{er}(h) \lesssim \frac{1}{m}\left(d\mathrm{Log}\left(\frac{m}{d}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

This has been refined by Giné and Koltchinskii (2006),[4] who argue that, with probability at least $1 - \delta$,

$$\sup_{h \in V_m} \mathrm{er}(h) \lesssim \frac{1}{m}\left(d\mathrm{Log}\left(\theta\left(\frac{d}{m}\right)\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right). \tag{16}$$

In the present work, by combining an argument of Hanneke (2009) with Theorem 7 above, we are able to obtain a new result, which replaces $\theta\left(\frac{d}{m}\right)$ in (16) with $\frac{\hat{n}_{1:m}}{d}$. Specifically, we have the following result.

Theorem 11 *For any $\delta \in (0,1)$ and $m \in \mathbb{N}$, with probability at least $1 - \delta$,*

$$\sup_{h \in V_m} \mathrm{er}(h) \leq \frac{8}{m}\left(d \ln\left(\frac{49e\hat{n}_{1:m}}{d} + 37\right) + 8\ln\left(\frac{6}{\delta}\right)\right).$$

The proof of Theorem 11 follows a similar strategy to the inductive step from the proofs of Theorems 1, 3, and 7. The details are included in Appendix C.

Additionally, since Hanneke and Yang (2015) prove that $\max_{t \in [m]} \max_{L \in (\mathcal{X}\times\mathcal{Y})^t} \hat{n}(L) = \min\{\mathfrak{s}, m\}$, where $\mathfrak{s}$ is the star number, the following new distribution-free bound immediately follows.[5]

4. See also Hanneke (2009), for a simple direct proof of this result.
5. The bound on the expectation follows by integrating the exponential bound on $\mathbb{P}(\sup_{h \in V_m} \mathrm{er}(h) > \varepsilon)$ implied by the first statement in the corollary, as was done, for instance, in the proofs of Theorems 1 and 3. We also note that, by using Theorem 10 in place of Theorem 7 in the proof of Theorem 11, one can obtain mildly better numerical constants in the logarithmic term in this corollary.

Corollary 12 *For any $m \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{h \in V_m} \mathrm{er}(h) \lesssim \frac{1}{m}\left(d\mathrm{Log}\left(\frac{\min\{\mathfrak{s}, m\}}{d}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

Furthermore,

$$\mathbb{E}\left[\sup_{h \in V_m} \mathrm{er}(h)\right] \lesssim \frac{d}{m}\mathrm{Log}\left(\frac{\min\{\mathfrak{s}, m\}}{d}\right).$$

Let us compare this result to (16). Since Hanneke and Yang (2015) prove that

$$\max_{\mathcal{P}} \max_{f^\star \in \mathbb{C}} \theta(r_0) = \min\left\{\mathfrak{s}, \frac{1}{r_0}\right\},$$

and also (as mentioned) that $\max_{t \in [m]} \max_{L \in (\mathcal{X} \times \mathcal{Y})^t} \hat{n}(L) = \min\{\mathfrak{s}, m\}$, we see that, at least in some scenarios (i.e., for some choices of $\mathcal{P}$ and $f^\star$), the new bound in Theorem 11 represents an improvement over (16). In particular, the best *distribution-free* bound obtainable from (16) is proportional to

$$\frac{1}{m}\left(d\mathrm{Log}\left(\frac{\min\{d\mathfrak{s}, m\}}{d}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right), \tag{17}$$

which is somewhat larger than the bound stated in Corollary 12 (which has $\mathfrak{s}$ in place of $d\mathfrak{s}$). Also, recalling that Wiener, Hanneke, and El-Yaniv (2015) established that $\theta(1/m) \lesssim \tilde{n}_{1:m}(\delta) \lesssim d\theta(d/m)\mathrm{polylog}(m, 1/\delta)$, we should expect that the bound in Theorem 11 is typically not much larger than (16) (and indeed will be smaller in many interesting cases).

5.1 Necessary and Sufficient Conditions for $1/m$ Rates for All Consistent Learners

Corollary 12 provides a sufficient condition for every consistent learning algorithm to achieve error rate with $O(1/m)$ asymptotic dependence on m: namely, $\mathfrak{s} < \infty$. Interestingly, we can show that this condition is in fact also *necessary* for every consistent learner to have a distribution-free bound on the error rate with $O(1/m)$ dependence on m. To be clear, in this context, we only consider m as the asymptotic variable: that is, $m \to \infty$ while δ and $\mathbb{C}$ (including d and $\mathfrak{s}$) are held fixed. This result is proven via the following theorem, establishing a worst-case lower bound on $\sup_{h \in V_m} \mathrm{er}(h)$.

Theorem 13 *For any $m \in \mathbb{N}$ and $\delta \in (0, 1/100)$, there exists a choice of $\mathcal{P}$ and $f^\star \in \mathbb{C}$ such that, with probability greater than δ,*

$$\sup_{h \in V_m} \mathrm{er}(h) \gtrsim \frac{d + \mathrm{Log}(\min\{\mathfrak{s}, m\}) + \mathrm{Log}\left(\frac{1}{\delta}\right)}{m} \wedge 1.$$

Furthermore,

$$\mathbb{E}\left[\sup_{h \in V_m} \mathrm{er}(h)\right] \gtrsim \frac{d + \mathrm{Log}(\min\{\mathfrak{s}, m\})}{m} \wedge 1.$$

Proof Since any $a, b, c \in \mathbb{R}$ have $a + b + c \leq 3 \max\{a, b, c\}$ and $a + b \leq 2 \max\{a, b\}$, it suffices to establish $\frac{d}{m} \wedge 1$, $\frac{\mathrm{Log}\left(\frac{1}{\delta}\right)}{m} \wedge 1$, and $\frac{\mathrm{Log}(\min\{\mathfrak{s}, m\})}{m}$ as lower bounds separately for the first bound, and $\frac{d}{m} \wedge 1$ and $\frac{\mathrm{Log}(\min\{\mathfrak{s}, m\})}{m}$ as lower bounds separately for the second bound. Lower bounds proportional to $\frac{d}{m} \wedge 1$ (in both bounds) and $\frac{\mathrm{Log}\left(\frac{1}{\delta}\right)}{m} \wedge 1$ (in the first bound) are known in the literature (Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Ehrenfeucht, Haussler, Kearns, and Valiant, 1989; Haussler, Littlestone, and Warmuth, 1994), and in fact hold as lower bounds on the error rate guarantees achievable by *any* learning algorithm.

For the remaining term, note that this term (with appropriately small constant factors) follows immediately from the others if $\mathfrak{s} \leq 56$, so suppose $\mathfrak{s} \geq 57$. Fix any $\varepsilon \in (0, 1/48)$, let $M_\varepsilon = \left\lfloor \frac{1+\varepsilon}{\varepsilon} \right\rfloor$, and let $x_1, \ldots, x_{\min\{\mathfrak{s}, M_\varepsilon\}} \in \mathcal{X}$ and $h_0, h_1, \ldots, h_{\min\{\mathfrak{s}, M_\varepsilon\}} \in \mathbb{C}$ be as in Definition 9. Choose the probability measure $\mathcal{P}$ such that $\mathcal{P}(\{x_i\}) = \varepsilon$ for every $i \in \{2, \ldots, \min\{\mathfrak{s}, M_\varepsilon\}\}$, and $\mathcal{P}(\{x_1\}) = 1 - (\min\{\mathfrak{s}, M_\varepsilon\} - 1)\varepsilon \geq 0$. Choose the target function $f^\star = h_0$. Then note that, for any $m \in \mathbb{N}$, if $\exists i \in \{2, \ldots, \min\{\mathfrak{s}, M_\varepsilon\}\}$ with $x_i \notin \{X_1, \ldots, X_m\}$, then $h_i \in V_m$, so that $\sup_{h \in V_m} \mathrm{er}(h) \geq \mathrm{er}(h_i) = \varepsilon$.

Characterizing the probability that $\{x_2, \ldots, x_{\min\{\mathfrak{s}, M_\varepsilon\}}\} \subseteq \{X_1, \ldots, X_m\}$ can be approached as an instance of the so-called *coupon collector's problem*. Specifically, let

$$\hat{M} = \min\left\{m \in \mathbb{N} : \{x_2, \ldots, x_{\min\{\mathfrak{s}, M_\varepsilon\}}\} \subseteq \{X_1, \ldots, X_m\}\right\}.$$

Note that $\hat{M}$ may be represented as a sum $\sum_{k=1}^{\min\{\mathfrak{s}, M_\varepsilon\}-1} G_k$ of independent geometric random variables $G_k \sim \mathrm{Geometric}(\varepsilon(\min\{\mathfrak{s}, M_\varepsilon\} - k))$, where G_k corresponds to the waiting time between encountering the $(k-1)^{\text{th}}$ and k^{th} distinct elements of $\{x_2, \ldots, x_{\min\{\mathfrak{s}, M_\varepsilon\}}\}$ in the X_t sequence. A simple calculation reveals that $\mathbb{E}[\hat{M}] = \frac{1}{\varepsilon} H_{\min\{\mathfrak{s}, M_\varepsilon\}-1}$, where H_t is the t^{th} harmonic number; in particular, $H_t \geq \ln(t)$. Another simple calculation with this sum of independent geometric random variables reveals $\mathrm{Var}(\hat{M}) < \frac{\pi^2}{6\varepsilon^2}$. Thus, Chebyshev's inequality implies that, with probability greater than $1/2$, $\hat{M} \geq \frac{1}{\varepsilon} \ln(\min\{\mathfrak{s}, M_\varepsilon\} - 1) - \frac{\pi}{\sqrt{3}\varepsilon}$. Since $\ln(\min\{\mathfrak{s}, M_\varepsilon\} - 1) \geq \ln(48) > 2\frac{\pi}{\sqrt{3}}$, the right hand side of this inequality is at least $\frac{1}{2\varepsilon} \ln(\min\{\mathfrak{s}, M_\varepsilon\} - 1) = \frac{1}{2\varepsilon} \ln\left(\min\left\{\mathfrak{s} - 1, \left\lfloor \frac{1}{\varepsilon} \right\rfloor\right\}\right)$. Altogether, we have that for any $m < \frac{1}{2\varepsilon} \ln\left(\min\left\{\mathfrak{s} - 1, \left\lfloor \frac{1}{\varepsilon} \right\rfloor\right\}\right)$, with probability greater than $1/2$, $\sup_{h \in V_m} \mathrm{er}(h) \geq \varepsilon$. By Markov's inequality, this further implies that, for any such m, $\mathbb{E}\left[\sup_{h \in V_m} \mathrm{er}(h)\right] > \varepsilon/2$.

For any $m \leq 47$, the $\frac{\mathrm{Log}(\min\{\mathfrak{s}, m\})}{m}$ term in both lower bounds (with appropriately small constant factors) follows from the lower bound proportional to $\frac{d}{m} \wedge 1$, so suppose $m \geq 48$. In particular, for any $c \in (4, \ln(56))$, letting $\varepsilon = \frac{\ln(\min\{\mathfrak{s}-1, m\})}{cm}$, one can easily verify that $0 < \varepsilon < 1/48$, and $m < \frac{1}{2\varepsilon} \ln\left(\min\left\{\mathfrak{s} - 1, \left\lfloor \frac{1}{\varepsilon} \right\rfloor\right\}\right)$. Therefore, with probability greater than $1/2 > \delta$,

$$\sup_{h \in V_m} \mathrm{er}(h) \geq \frac{\ln(\min\{\mathfrak{s} - 1, m\})}{cm},$$

and furthermore,

$$\mathbb{E}\left[\sup_{h \in V_m} \mathrm{er}(h)\right] > \frac{\ln(\min\{\mathfrak{s} - 1, m\})}{2cm}.$$

The result follows by noting $\ln(\min\{\mathfrak{s} - 1, m\}) \geq \ln(\min\{\mathfrak{s}, m\}/2) \geq \ln(\min\{\mathfrak{s}, m\})/2$ for $\mathfrak{s}, m \geq 4$. ■

Comparing Theorem 13 with Corollary 12, we see that the asymptotic dependences on m are identical, though they differ in their joint dependences on d and m. The precise dependence on both d and m from Corollary 12 can be included in the lower bound of Theorem 13 for certain types of concept spaces $\mathbb{C}$, but not all; the interested reader is referred to the recent article of Hanneke and Yang (2015) for discussions relevant to this type of gap, and constructions of concept spaces which (one can easily verify) span this gap: that is, for some spaces $\mathbb{C}$ the lower bound is tight, while for other spaces $\mathbb{C}$ the upper bound is tight, up to numerical constant factors.

An immediate corollary of Theorem 13 and Corollary 12 is that $\mathfrak{s} < \infty$ is *necessary and sufficient* for arbitrary consistent learners to achieve $O(1/m)$ rates. Formally, for any $\delta \in (0,1)$, let $R_m(\delta)$ denote the smallest value such that, for all $\mathcal{P}$ and all $f^\star \in \mathbb{C}$, with probability at least $1-\delta$, $\sup_{h \in V_m} \mathrm{er}(h) \leq R_m(\delta)$. Also let $\bar{R}_m$ denote the supremum value of $\mathbb{E}\left[\sup_{h \in V_m} \mathrm{er}(h)\right]$ over all $\mathcal{P}$ and all $f^\star \in \mathbb{C}$. We have the following corollary (which applies to any $\mathbb{C}$ with $0 < d < \infty$).

Corollary 14 *$\bar{R}_m = \Theta\left(\frac{1}{m}\right)$ if and only if $\mathfrak{s} < \infty$, and otherwise $\bar{R}_m = \Theta\left(\frac{\mathrm{Log}(m)}{m}\right)$. Likewise, $\forall \delta \in (0, 1/100)$, $R_m(\delta) = \Theta\left(\frac{1}{m}\right)$ if and only if $\mathfrak{s} < \infty$, and otherwise $R_m(\delta) = \Theta\left(\frac{\mathrm{Log}(m)}{m}\right)$.*

5.2 Using Subregions Smaller than the Region of Disagreement

In recent work, Zhang and Chaudhuri (2014) have proposed a general active learning strategy, which revises the CAL strategy so that the algorithm only requests a label if the corresponding X_m is in a well-chosen *subregion* of $\mathrm{DIS}(V_{m-1})$. This general idea was first explored in the more-specific context of learning linear separators under a uniform distribution by Balcan, Broder, and Zhang (2007) (see also Dasgupta, Kalai, and Monteleoni, 2005, for related arguments). Furthermore, following up on Balcan, Broder, and Zhang (2007), the work of Balcan and Long (2013) has also used this subregion idea to argue that any consistent learning algorithm achieves the optimal sample complexity (up to constants) for the problem of learning linear separators under isotropic log-concave distributions. In this section, we combine the abstract perspective of Zhang and Chaudhuri (2014) with our general bounding technique, to generalize the result of Balcan and Long (2013) by expressing a bound holding for arbitrary concept spaces $\mathbb{C}$, distributions $\mathcal{P}$, and target functions $f^\star \in \mathbb{C}$. First, we need to introduce the following complexity measure $\varphi_c(r_0)$ based on the work of Zhang and Chaudhuri (2014). As was true of $\theta(r_0)$ above, this complexity measure $\varphi_c(r_0)$ generally depends on both $\mathcal{P}$ and $f^\star$.

Definition 15 *For any nonempty set $\mathcal{H}$ of classifiers, and any $\eta \geq 0$, letting $X \sim \mathcal{P}$, define*

$$\Phi(\mathcal{H}, \eta) = \min \Bigg\{ \mathbb{E}[\gamma(X)] : \sup_{h \in \mathcal{H}} \mathbb{E}\left[\mathbb{1}[h(X) = +1]\zeta(X) + \mathbb{1}[h(X) = -1]\xi(X)\right] \leq \eta,$$
$$\text{where } \forall x \in \mathcal{X}, \gamma(x) + \zeta(x) + \xi(x) = 1 \text{ and } \gamma(x), \zeta(x), \xi(x) \in [0,1] \Bigg\}.$$

Then, for any $r_0 \in [0,1)$ *and* $c > 1$*, define*

$$\varphi_c(r_0) = \sup_{r_0 < r \le 1} \frac{\Phi(\mathrm{B}(f^\star, r), r/c)}{r} \vee 1.$$

One can easily observe that, for the optimal choices of γ, ζ, and ξ in the definition of Φ, we have $\gamma(x) = 0$ for (almost every) $x \notin \mathrm{DIS}(\mathcal{H})$. In the special case that γ is binary-valued, the aforementioned well-chosen "subregion" of $\mathrm{DIS}(\mathcal{H})$ corresponds to the set $\{x : \gamma(x) = 1\}$. In general, the definition also allows for $\gamma(x)$ values in between 0 and 1, in which case γ essentially re-weights the conditional distribution $\mathcal{P}(\cdot|\mathrm{DIS}(\mathcal{H}))$.[6] As an example where this quantity is informative, Zhang and Chaudhuri (2014) argue that, for $\mathbb{C}$ the class of homogeneous linear separators in $\mathbb{R}^k$ ($k \in \mathbb{N}$) and $\mathcal{P}$ any isotropic log-concave distribution, $\varphi_c(r_0) \lesssim \mathrm{Log}(c)$ (which follows readily from arguments of Balcan and Long, 2013). Furthermore, they observe that $\varphi_c(r_0) \le \theta(r_0)$ for any $c \in (1, \infty]$.

Zhang and Chaudhuri (2014) propose the above quantities for the purpose of proving a bound on the label complexity of a certain active learning algorithm, inspired both by the work of Balcan, Broder, and Zhang (2007) on active learning with linear separators, and by the connection between selective classification and active learning exposed by El-Yaniv and Wiener (2012). However, since the idea of using well-chosen subregions of $\mathrm{DIS}(V_m)$ in the analysis of consistent learning algorithms lead Balcan and Long (2013) to derive improved sample complexity bounds for these methods in the case of linear separators under isotropic log-concave distributions, and since the corresponding improvements for active learning are reflected in the general results of Zhang and Chaudhuri (2014), it is natural to ask whether the sample complexity improvements of Balcan and Long (2013) for that special scenario can also be extended to the general case by incorporating the complexity measure $\varphi_c(r_0)$. Here we provide such an extension. Specifically, following the same basic strategy from Theorem 7, with a few adjustments inspired by Zhang and Chaudhuri (2014) to allow us to consider only a *subregion* of $\mathrm{DIS}(V_m)$ in the argument (or more generally, a reweighting of the conditional distribution $\mathcal{P}(\cdot|\mathrm{DIS}(V_m))$), we arrive at the following result. The proof is included in Appendix D.

Theorem 16 *For any* $\delta \in (0,1)$ *and* $m \in \mathbb{N}$*, for* $c = 16$*, with probability at least* $1-\delta$*,*

$$\sup_{h \in V_m} \mathrm{er}(h) \le \frac{21}{m}\left(d \ln\left(83\varphi_c\left(\frac{d}{m}\right)\right) + 3\ln\left(\frac{4}{\delta}\right)\right).$$

In particular, in the special case of $\mathbb{C}$ the space of homogeneous linear separators on $\mathbb{R}^k$, and $\mathcal{P}$ an isotropic log-concave distribution, Theorem 16 recovers the bound of Balcan and Long (2013) proportional to $\frac{1}{m}(k + \mathrm{Log}(\frac{1}{\delta}))$ as a special case. Furthermore, one can easily construct scenarios (concept spaces $\mathbb{C}$, distributions $\mathcal{P}$, and target functions $f^\star \in \mathbb{C}$) where $\varphi_c\left(\frac{d}{m}\right)$ is bounded while $\frac{\hat{n}_{1:m}}{d} = \frac{m}{d}$ almost surely (e.g., $\mathbb{C} = \{x \mapsto 2\mathbb{1}_{\{t\}}(x) - 1 : t \in \mathbb{R}\}$ the class of impulse functions on $\mathbb{R}$, and $\mathcal{P}$ uniform on $(0,1)$), so that Theorem 16 sometimes reflects a significant improvement over Theorem 11.

6. Allowing these more-general values of $\gamma(x)$ typically does not affect the qualitative behavior of the minimal $\mathbb{E}[\gamma(X)]$ value; for instance, we argue in Lemma 24 of Appendix E that the minimal $\mathbb{E}[\gamma(X)]$ value achievable under the additional constraint that $\gamma(x) \in \{0,1\}$ is at most $2\Phi(\mathcal{H}, \eta/2)$. Thus, we do not lose much by thinking of $\Phi(\mathcal{H}, \eta)$ as describing the measure of a subregion of $\mathrm{DIS}(\mathcal{H})$.

One can easily show that we always have $\varphi_c(r_0) \leq \left(1 - \frac{1}{c}\right)\theta(r_0)$, so that Theorem 16 is never worse than the bound (16) of Giné and Koltchinskii (2006). However, we argue in Appendix D.1 that $\forall c \geq 2$, $\forall r_0 \in [0, 1)$,

$$\left(1 - \frac{1}{c}\right) \min\left\{\mathfrak{s}, \frac{1}{r_0} - \frac{1}{c-1}\right\} \leq \sup_{\mathcal{P}} \sup_{f^\star \in \mathbb{C}} \varphi_c(r_0) \leq \left(1 - \frac{1}{c}\right) \min\left\{\mathfrak{s}, \frac{1}{r_0}\right\}. \tag{18}$$

Thus, at least in some cases, the bound in Theorem 11 is smaller than that in Theorem 16 (as the former leads to Corollary 12 in the worst case, while the latter leads to (17) in the worst case). In fact, if we let $\varphi_c^{01}(r_0)$ be defined identically to $\varphi_c(r_0)$, except that γ is restricted to be $\{0, 1\}$-valued in Definition 15, then the same argument from Appendix D.1 reveals that, for any $c \geq 4$,

$$\sup_{\mathcal{P}} \sup_{f^\star \in \mathbb{C}} \varphi_c^{01}(r_0) = \min\left\{\mathfrak{s}, \frac{1}{r_0}\right\}.$$

Relation to the Doubling Dimension: To further put Theorem 16 in context, we also note that it is possible to relate $\varphi_c(r_0)$ to the *doubling dimension*. Specifically, the doubling dimension (also known as the local metric entropy) of $\mathbb{C}$ at $f^\star$ under $\mathcal{P}$, denoted $D(r_0)$, is defined as

$$D(r_0) = \max_{r \geq r_0} \log_2\left(\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P})\right),$$

for $r_0 > 0$, where $\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P})$ is the smallest $n \in \mathbb{N}$ such that there exist classifiers $h_1, \ldots, h_n$ for which $\sup_{h \in \mathrm{B}(f^\star, r)} \min_{1 \leq i \leq n} \mathcal{P}(x : h(x) \neq h_i(x)) \leq r/2$, known as the $(r/2)$-covering number for $\mathrm{B}(f^\star, r)$ under the $L_1(\mathcal{P})$ pseudo-metric. The notion of doubling dimension has been explored in a variety of contexts in the literature (e.g., LeCam, 1973; Yang and Barron, 1999; Gupta, Krauthgamer, and Lee, 2003; Bshouty, Li, and Long, 2009). We always have $D(r_0) \lesssim d\mathrm{Log}(1/r_0)$ (Haussler, 1995), though it can often be smaller than this, and in many interesting contexts, it can even be bounded by an r_0-invariant value (Bshouty, Li, and Long, 2009). Bshouty, Li, and Long (2009) construct a particular $\mathcal{P}$-dependent learning rule $\mathbb{A}$ such that, for any $\varepsilon, \delta \in (0, 1)$, and any

$$m \gtrsim \frac{1}{\varepsilon}\left(D(\varepsilon/c) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right), \tag{19}$$

where $c > 0$ is a specific constant, with probability at least $1 - \delta$, the classifier $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$ satisfies $\mathrm{er}(\hat{h}_m) \leq \varepsilon$. They also establish a weaker bound holding for all consistent learning rules: for any $\varepsilon > 0$, denoting $\varepsilon_0 = \varepsilon \exp\left\{-\sqrt{\ln(1/\varepsilon)}\right\}$, for any

$$m \gtrsim \frac{1}{\varepsilon}\left(\max\{d, D(\varepsilon_0)\}\sqrt{\mathrm{Log}\left(\frac{1}{\varepsilon}\right)} + \mathrm{Log}\left(\frac{1}{\delta}\right)\right), \tag{20}$$

with probability at least $1 - \delta$, $\sup_{h \in V_m} \mathrm{er}(h) \leq \varepsilon$.

Hanneke and Yang (2015) have proven that we always have $D(r_0) \lesssim d\mathrm{Log}(\theta(r_0))$, which immediately implies that (19) is never larger than the bound (16) for consistent learning rules (aside from constant factors), though (16) may often offer improvements over the weaker bound (20). Here we note that a related argument can be used to prove the following bound: for any $r_0 > 0$ and $c \geq 8$,

$$D(r_0) \leq 2d \log_2(96\varphi_c(r_0)). \tag{21}$$

In particular, this implies that the bound (19) is never larger than the bound in Theorem 16 for consistent learning rules (aside from constant factors), though again Theorem 16 may often offer improvements over the weaker bound (20). We also note that, combined with the above mentioned result of Zhang and Chaudhuri (2014) that $\varphi_c(r_0) \lesssim \mathrm{Log}(c)$ for $\mathbb{C}$ the class of homogeneous linear separators in $\mathbb{R}^k$ and $\mathcal{P}$ any isotropic log-concave distribution, (21) immediately implies a bound $D(r_0) \lesssim k$ for the doubling dimension in this scenario (recalling that $d = k$ for this class, from Cover, 1965), which appears to be new to the literature. The proof of (21) is included in Appendix D.2.

6. Learning with Noise

The previous sections demonstrate how variations on the basic technique of Hanneke (2009) lead to refined analyses of certain learning methods, in the *realizable case*, where $\exists f^\star \in \mathbb{C}$ with $\mathrm{er}(f^\star) = 0$. We can also apply this general technique in the more-general setting of learning with *classification noise*. Specifically, in this setting, there is a *joint* distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$, and the error rate of a classifier h is then defined as $\mathrm{er}(h) = \mathbb{P}(h(X) \neq Y)$ for $(X, Y) \sim \mathcal{P}_{XY}$. As above, we denote by $\mathcal{P}$ the marginal distribution $\mathcal{P}_{XY}(\cdot \times \mathcal{Y})$ on $\mathcal{X}$. We then let $(X_1, Y_1), (X_2, Y_2), \ldots$ denote a sequence of independent $\mathcal{P}_{XY}$-distributed random samples, and denoting $\mathcal{L}_m = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$, we are interested in obtaining bounds on $\mathrm{er}(\hat{h}_m) - \inf_{f \in \mathbb{C}} \mathrm{er}(f)$ (the *excess error rate*), where $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$ for some learning rule $\mathbb{A}$. This notation is consistent with the above, which represents the special case in which $\mathbb{P}(Y = f^\star(X)|X) = 1$ almost surely (i.e., the *realizable case*). While there are various noise models commonly studied in the literature, for our present discussion, we are primarily interested in two such models.

- For $\beta \in (0, 1/2)$, $\mathcal{P}_{XY}$ satisfies the *β-bounded noise* condition if $\exists h^\star \in \mathbb{C}$ such that $\mathbb{P}(Y \neq h^\star(X)|X) \leq \beta$ almost surely, where $(X, Y) \sim \mathcal{P}_{XY}$.
- For $a \in [1, \infty)$ and $\alpha \in [0, 1]$, $\mathcal{P}_{XY}$ satisfies the *(a, α)-Bernstein class* condition if, for $h^\star = \mathrm{argmin}_{h \in \mathbb{C}} \mathrm{er}(h)$,[7] we have $\forall h \in \mathbb{C}$, $\mathcal{P}(x : h(x) \neq h^\star(x)) \leq a\, (\mathrm{er}(h) - \mathrm{er}(h^\star))^\alpha$.

Note that β-bounded noise distributions also satisfy the Bernstein class condition, with $\alpha = 1$ and $a = \frac{1}{1-2\beta}$. These two conditions have been studied extensively in both the passive and active learning literatures (e.g., Mammen and Tsybakov, 1999; Tsybakov, 2004; Bartlett, Jordan, and McAuliffe, 2006; Massart and Nédélec, 2006; Koltchinskii, 2006; Bartlett and Mendelson, 2006; Giné and Koltchinskii, 2006; Hanneke, 2009, 2011, 2012, 2014; El-Yaniv and Wiener, 2011; Ailon, Begleiter, and Ezra, 2014; Zhang and Chaudhuri, 2014; Hanneke and Yang, 2015). In particular, for passive learning, much of this literature

7. For simplicity, we suppose the minimum error rate is achieved in $\mathbb{C}$. One can easily generalize the condition to the more-general case where the minimum is not necessarily achieved (see e.g., Koltchinskii, 2006), and the results below continue to hold with only minor technical adjustments to the proofs.

focuses on the analysis of *empirical risk minimization*. Specifically, for any $m \in \mathbb{N}$ and $L \in (\mathcal{X} \times \mathcal{Y})^m$, define $\mathrm{ERM}(\mathbb{C}, L) = \{h \in \mathbb{C} : \mathrm{er}_L(h) = \min_{g \in \mathbb{C}} \mathrm{er}_L(g)\}$, the set of *empirical risk minimizers*. Massart and Nédélec (2006) established that, for any $\mathcal{P}_{XY}$ satisfying the (a, α)-Bernstein class condition, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{h \in \mathrm{ERM}(\mathbb{C}, \mathcal{L}_m)} \mathrm{er}(h) - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \lesssim \left(\frac{a \left(d \mathrm{Log}\left(\frac{1}{a} \left(\frac{m}{ad} \right)^{\frac{\alpha}{2-\alpha}} \right) + \mathrm{Log}\left(\frac{1}{\delta} \right) \right)}{m} \right)^{\frac{1}{2-\alpha}}. \tag{22}$$

In the case of β-bounded noise, Giné and Koltchinskii (2006) showed that the logarithmic factor $\mathrm{Log}\left(\frac{m(1-2\beta)^2}{d}\right)$ implied by (22) can be replaced by $\mathrm{Log}\left(\theta\left(\frac{d}{m(1-2\beta)^2}\right)\right)$, where the disagreement coefficient $\theta(r_0)$ is defined as above, except with $h^\star$ in place of $f^\star$ in the definition. Furthermore, applying their arguments to the general case of the (a, α)-Bernstein class condition (see Hanneke and Yang, 2012, for an explicit derivation), one arrives at the fact that, with probability at least $1-\delta$,

$$\sup_{h \in \mathrm{ERM}(\mathbb{C}, \mathcal{L}_m)} \mathrm{er}(h) - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \lesssim \left(\frac{a \left(d \mathrm{Log}\left(\theta\left(a \left(\frac{ad}{m} \right)^{\frac{\alpha}{2-\alpha}} \right) \right) + \mathrm{Log}\left(\frac{1}{\delta} \right) \right)}{m} \right)^{\frac{1}{2-\alpha}}. \tag{23}$$

Since Hanneke and Yang (2015) have argued that $\theta(r_0) \leq \min\left\{\mathfrak{s}, \frac{1}{r_0}\right\}$ (with equality in the worst case), (23) further implies that, with probability at least $1-\delta$,

$$\sup_{h \in \mathrm{ERM}(\mathbb{C}, \mathcal{L}_m)} \mathrm{er}(h) - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \lesssim \left(\frac{a \left(d \mathrm{Log}\left(\min\left\{ \mathfrak{s}, \frac{1}{a} \left(\frac{m}{ad} \right)^{\frac{\alpha}{2-\alpha}} \right\} \right) + \mathrm{Log}\left(\frac{1}{\delta} \right) \right)}{m} \right)^{\frac{1}{2-\alpha}}. \tag{24}$$

Via the same integration argument used in Corollary 12, this further implies

$$\mathbb{E}\left[\sup_{h \in \mathrm{ERM}(\mathbb{C}, \mathcal{L}_m)} \mathrm{er}(h) \right] - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \lesssim \left(\frac{a d \mathrm{Log}\left(\min\left\{ \mathfrak{s}, \frac{1}{a} \left(\frac{m}{ad} \right)^{\frac{\alpha}{2-\alpha}} \right\} \right)}{m} \right)^{\frac{1}{2-\alpha}}. \tag{25}$$

It is worth noting that the bound (24) does not quite recover the bound of Corollary 12 in the realizable case (corresponding to $a = \alpha = 1$). Specifically, it contains a logarithmic factor $\mathrm{Log}\left(\frac{\min\{\mathfrak{s}d, m\}}{d}\right)$, rather than $\mathrm{Log}\left(\frac{\min\{\mathfrak{s}, m\}}{d}\right)$. I conjecture that this logarithmic factor in (24) can generally be improved so that, for any a and α, it is bounded by a numerical constant whenever $\mathfrak{s} \lesssim d$. This problem is intimately connected to a conjecture in active learning, proposed by Hanneke and Yang (2015), concerning the joint dependence on $\mathfrak{s}$ and d in the minimax label complexity of active learning under the Bernstein class condition.

6.1 Necessary and Sufficient Conditions for $1/m$ Minimax Rates under Bounded Noise

In the case of bounded noise (where $a = \frac{1}{1-2\beta}$ and $\alpha = 1$), Massart and Nédélec (2006) have shown that for some concept spaces $\mathbb{C}$, the factor $\mathrm{Log}\left(\frac{m(1-2\beta)^2}{d}\right)$ is present even in a *lower* bound on the minimax excess error rate, so that it cannot generally be removed. Raginsky and Rakhlin (2011) further discuss a range of lower bounds on the minimax excess error rate for various spaces $\mathbb{C}$ they construct, where the appropriate factor ranges between $\mathrm{Log}\left(\frac{m(1-2\beta)^2}{d}\right)$ at the highest, to a constant factor at the lowest. The bound in (24) provides a *sufficient* condition for all empirical risk minimization algorithms to achieve excess error rate with $O(1/m)$ asymptotic dependence on m under β-bounded noise: namely $\mathfrak{s} < \infty$. Recall that this condition was both sufficient *and necessary* for $O(1/m)$ error rates to be achievable by every algorithm of this type for all distributions in the realizable case (Corollary 14). It is therefore natural to wonder whether this remains the case for bounded noise as well. In this section, we find this is indeed the case. In fact, following a generalization of the technique of Raginsky and Rakhlin (2011) explored by Hanneke and Yang (2015) for active learning, we are here able to provide a general lower bound on the *minimax* excess error rate of passive learning, expressed in terms of $\mathfrak{s}$. This immediately implies a corollary that $\mathfrak{s} < \infty$ is both *necessary* and *sufficient* for the *minimax optimal* bound on the excess error rate to have dependence on m of $\Theta(1/m)$ under bounded noise, and otherwise the minimax optimal bound is $\Theta(\mathrm{Log}(m)/m)$. Note that this is a stronger type of result than that given by Corollary 14, as the lower bounds here apply to *all* learning rules. Formally, we have the following theorem. The proof is included in Appendix E.1.

Theorem 17 *For any $\beta \in (0, 1/2)$, $m \in \mathbb{N}$, and $\delta \in (0, 1/24]$, for any (passive) learning rule $\mathbb{A}$, there exists a choice of $\mathcal{P}_{XY}$ satisfying the β-bounded noise condition such that, denoting $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$, with probability greater than δ,*

$$\mathrm{er}(\hat{h}_m) - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \gtrsim \frac{d + \beta \mathrm{Log}\left(\min\left\{\mathfrak{s}, (1-2\beta)^2 m\right\}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)}{(1-2\beta)m} \wedge (1-2\beta).$$

Furthermore,

$$\mathbb{E}\left[\mathrm{er}(\hat{h}_m)\right] - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \gtrsim \frac{d + \beta \mathrm{Log}\left(\min\left\{\mathfrak{s}, (1-2\beta)^2 m\right\}\right)}{(1-2\beta)m} \wedge (1-2\beta).$$

As was the case in Theorem 13, the joint dependence on d and m in this lower bound does not match that in (24) in the case $\mathfrak{s} = \infty$. One can show that the dependence in this lower bound can be made to nearly match that in (24) for certain specially-constructed spaces $\mathbb{C}$ under bounded noise (Massart and Nédélec, 2006; Raginsky and Rakhlin, 2011; Hanneke and Yang, 2015) (the only gap being that $\mathfrak{s}$ is replaced by $\mathfrak{s}/d$ in (24) to obtain the lower bound); however, there also exist spaces $\mathbb{C}$ where these lower bounds are nearly tight (for β bounded away from 0), so that they cannot be improved in the general case (see Hanneke and Yang, 2015, for construction of spaces $\mathbb{C}$ with arbitrary d and $\mathfrak{s}$, for which one can show this is the case).

As mentioned above, an immediate corollary of Theorem 17, in combination with (24), is that $\mathfrak{s} < \infty$ is necessary and sufficient for the minimax excess error rate to have $O(1/m)$ dependence on m for bounded noise. Formally, for $m \in \mathbb{N}$, $\beta \in [0, 1/2)$, and $\delta \in (0, 1)$, let $R_m(\delta, \beta)$ denote the smallest value such that there exists a learning rule $\mathbb{A}$ for which, for all $\mathcal{P}_{XY}$ satisfying the β-bounded noise condition, with probability at least $1 - \delta$, $\mathrm{er}(\mathbb{A}(\mathcal{L}_m)) - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \leq R_m(\delta, \beta)$. Also let $\bar{R}_m(\beta)$ denote the smallest value such that there exists a learning rule $\mathbb{A}$ for which, for all $\mathcal{P}_{XY}$ satisfying the β-bounded noise condition, $\mathbb{E}[\mathrm{er}(\mathbb{A}(\mathcal{L}_m))] - \inf_{h\in\mathbb{C}} \mathrm{er}(h) \leq \bar{R}_m(\beta)$. We have the following corollary (which applies to any $\mathbb{C}$ with $0 < d < \infty$).

Corollary 18 *Fix any $\beta \in (0, 1/2)$. $\bar{R}_m(\beta) = \Theta\left(\frac{1}{m}\right)$ if and only if $\mathfrak{s} < \infty$, and otherwise $\bar{R}_m(\beta) = \Theta\left(\frac{\mathrm{Log}(m)}{m}\right)$. Likewise, $\forall \delta \in (0, 1/24]$, $R_m(\delta, \beta) = \Theta\left(\frac{1}{m}\right)$ if and only if $\mathfrak{s} < \infty$, and otherwise $R_m(\delta, \beta) = \Theta\left(\frac{\mathrm{Log}(m)}{m}\right)$.*

Again, note that this is a stronger type of result than Corollary 14 above, which only found $\mathfrak{s} < \infty$ as necessary and sufficient for a *particular family* of learning rules to obtain $O(1/m)$ rates. In contrast, this result applies even to the *minimax optimal* learning rule.

We conclude this section by noting that the technique leading to Theorem 17 appears not to straightforwardly extend to the general (a, α)-Bernstein class condition. Indeed, though one can certainly exhibit specific spaces $\mathbb{C}$ for which the minimax excess risk has $\Theta\left(\left(\frac{\mathrm{Log}(m)}{m}\right)^{\frac{1}{2-\alpha}}\right)$ dependence on m (e.g., impulse functions on $\mathbb{R}$; see Hanneke and Yang, 2015, for related discussions), it appears a much more challenging problem to construct general lower bounds describing the range of possible dependences on m. Thus, the more general question of establishing necessary and sufficient conditions for $O\left(1/m^{\frac{1}{2-\alpha}}\right)$ excess error rates under the (a, α)-Bernstein class condition remains open.

6.2 Using Subregions to Achieve Improved Excess Error Bounds

In general, note that plugging into (23) the parameters $a = \alpha = 1$ admitted by the realizable case, (23) recovers the bound (16). Recalling that we were able to refine the bound (16) via techniques from the subregion-based analysis of Zhang and Chaudhuri (2014), yielding Theorem 16 above, it is natural to consider whether we might be able to refine (23) in a similar way. We find that this is indeed the case, though we establish this refinement for a different learning rule (described in Appendix E.2). Letting $c = 128$, for any $r_0 \in [0, 1)$, $a \geq 1$ and $\alpha \in (0, 1]$, define

$$\hat{\varphi}_{a,\alpha}(r_0) = \sup_{h\in\mathbb{C}} \sup_{r>r_0} \frac{\Phi\left(\mathrm{B}(h, r), (r/a)^{1/\alpha}/c\right)}{r} \vee 1.$$

For completeness, also define $\hat{\varphi}_{a,\alpha}(r_0) = 1$ for any $r_0 \geq 1$, $a \geq 1$, and $\alpha \in [0, 1]$. We have the following theorem.

Theorem 19 *For any $a \geq 1$ and $\alpha \in (0, 1]$, for any probability measure $\mathcal{P}$ over $\mathcal{X}$, for any $\delta \in (0, 1)$, there exists a learning rule $\mathbb{A}$ such that, for any $\mathcal{P}_{XY}$ satisfying the (a, α)-Bernstein class condition with marginal distribution $\mathcal{P}$ over $\mathcal{X}$, for any $m \in \mathbb{N}$, letting $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$, with probability at least $1 - \delta$,*

$$\mathrm{er}(\hat{h}_m) - \inf_{h\in\mathbb{C}} \mathrm{er}(h) \lesssim \left(\frac{a\left(d\mathrm{Log}\left(\hat{\varphi}_{a,\alpha}\left(a\left(\frac{ad}{m}\right)^{\frac{\alpha}{2-\alpha}}\right)\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)}{m}\right)^{\frac{1}{2-\alpha}}.$$

The proof is included in Appendix E.2. We should emphasize that the bound in Theorem 19 is established for a particular learning method (described in Appendix E.2), *not* for empirical risk minimization. Thus, whether or not this bound can be established for the general family of empirical risk minimization rules remains an open question. We should also note that $\hat{\varphi}_{a,\alpha}(r_0)$ involves a supremum over $h \in \mathbb{C}$ only so that we may allow the algorithm to explicitly depend on $\hat{\varphi}_{a,\alpha}(r_0)$ (noting that, as stated, Theorem 19 allows $\mathcal{P}$-dependence in the algorithm). It is conceivable that this dependence on $\hat{\varphi}_{a,\alpha}(r_0)$ in $\mathbb{A}$ can be removed, for instance via a stratification and model selection technique (see e.g., Koltchinskii, 2006), in which case this supremum over h would be replaced by fixing $h = h^\star$.

We conclude this section with some basic observations about the bound in Theorem 19. First, in the special case of $\mathbb{C}$ the class of homogeneous linear separators on $\mathbb{R}^k$ and $\mathcal{P}$ any isotropic log-concave distribution, Theorem 19 recovers a bound of Balcan and Long (2013) (established for a closely related method), since a result of Zhang and Chaudhuri (2014) implies $\hat{\varphi}_{a,\alpha}(a\varepsilon^\alpha) \lesssim \mathrm{Log}\left(a\varepsilon^{\alpha-1}\right)$ in that case. Additionally, we note that a result similar to (24) also generally holds for the method $\mathbb{A}$ from Theorem 19, since (18) implies we always have

$$\frac{\Phi(\mathrm{B}(h, a\varepsilon^\alpha), \varepsilon/c)}{a\varepsilon^\alpha} \le \left(1 - \frac{1}{ca}\varepsilon^{1-\alpha}\right) \min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^\alpha}\right\}.$$

Appendix A. A Technical Lemma

The following lemma is useful in the proofs of several of the main results of this paper.[8]

Lemma 20 *For any $a, b, c_1 \in [1, \infty)$ and $c_2 \in [0, \infty)$,*

$$a \ln\left(c_1\left(c_2 + \frac{b}{a}\right)\right) \le a \ln\left(c_1(c_2 + e)\right) + \frac{1}{e}b.$$

Proof By subtracting $a\ln(c_1)$ from both sides, we see that it suffices to verify that $a\ln\left(c_2 + \frac{b}{a}\right) \le a\ln(c_2 + e) + \frac{1}{e}b$. If $\frac{b}{a} \le e$, then monotonicity of $\ln(\cdot)$ implies

$$a \ln\left(c_2 + \frac{b}{a}\right) \le a\ln(c_2 + e),$$

which is clearly no greater than $a\ln(c_2 + e) + \frac{1}{e}b$. On the other hand, if $\frac{b}{a} > e$, then

$$a \ln\left(c_2 + \frac{b}{a}\right) \le a \ln\left(\max\{c_2, 2\}\frac{b}{a}\right) = a \ln\left(\max\{c_2, 2\}\right) + a \ln\left(\frac{b}{a}\right).$$

The first term in the rightmost expression is at most $a\ln(c_2 + 2) \le a\ln(c_2 + e)$. The second term in the rightmost expression can be rewritten as $b\frac{\ln(b/a)}{b/a}$. Since $x \mapsto \ln(x)/x$ is nonincreasing on (e, ∞), in the case $\frac{b}{a} > e$ this is at most $\frac{1}{e}b$. Together, we have that

$$a \ln\left(c_2 + \frac{b}{a}\right) \le a\ln(c_2 + e) + \frac{1}{e}b$$

in this case as well. ■

8. This lemma and proof also appear in a sibling paper (Hanneke, 2016).

Appendix B. Proof of Theorem 7

Here we present the proof of Theorem 7.

Proof of Theorem 7 The structure of the proof is nearly identical to that of Theorem 3, with only a few small changes to account for the fact that $\hat{n}_{1:m}$ depends on the specific samples, and in particular, on the order of the samples.

The proof proceeds by induction on m. Since $\mathcal{P}(\mathrm{DIS}(V_m)) \leq 1$ always, the stated bound is trivially satisfied for all $\delta \in (0,1)$ if $m \leq 16$. Now, as an inductive hypothesis, fix any integer $m \geq 17$ such that, $\forall \delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor})) \leq \frac{16}{\lfloor m/2 \rfloor}\left(2\hat{n}_{1:\lfloor m/2 \rfloor} + \ln\left(\frac{3}{\delta}\right)\right).$$

Fix any $\delta \in (0,1)$. Define

$$N = \left|\left\{X_{\lfloor m/2 \rfloor+1}, \ldots, X_m\right\} \cap \mathrm{DIS}(V_{\lfloor m/2 \rfloor})\right|,$$

and enumerate the elements of $\{X_{\lfloor m/2 \rfloor+1}, \ldots, X_m\} \cap \mathrm{DIS}(V_{\lfloor m/2 \rfloor})$ as $\hat{X}_1, \ldots, \hat{X}_N$. Let $\mathcal{L}_t = \{(X_1, f^\star(X_1)), \ldots, (X_t, f^\star(X_t))\}$ for every $t \in [m]$, and $\hat{n}'_m = \left|\hat{C}_{\mathcal{L}_m} \setminus \mathcal{L}_{\lfloor m/2 \rfloor}\right|$, and enumerate as $i'_1, \ldots, i'_{\hat{n}'_m}$ the indices $i \in \{\lfloor m/2 \rfloor+1, \ldots, m\}$ with $(X_i, f^\star(X_i)) \in \hat{C}_{\mathcal{L}_m} \setminus \mathcal{L}_{\lfloor m/2 \rfloor}$. In particular, note that $\hat{n}'_m \leq \hat{n}_m$, and $\hat{C}_{\mathcal{L}_m} \subseteq \mathcal{L}_{\lfloor m/2 \rfloor} \cup \{(X_{i'_1}, f^\star(X_{i'_1})), \ldots, (X_{i'_{\hat{n}'_m}}, f^\star(X_{i'_{\hat{n}'_m}}))\}$, so that

$$\mathbb{C}\left[\mathcal{L}_{\lfloor m/2 \rfloor} \cup \{(X_{i'_1}, f^\star(X_{i'_1})), \ldots, (X_{i'_{\hat{n}'_m}}, f^\star(X_{i'_{\hat{n}'_m}}))\}\right] = V_m.$$

Next, let $\hat{n}''_m = \left|\{j \in [\hat{n}'_m] : X_{i'_j} \in \mathrm{DIS}(V_{\lfloor m/2 \rfloor})\}\right|$, and enumerate as $i''_1, \ldots, i''_{\hat{n}''_m}$ the indices $i \in [N]$ such that $(\hat{X}_i, f^\star(\hat{X}_i)) \in \{(X_{i'_1}, f^\star(X_{i'_1})), \ldots, (X_{i'_{\hat{n}'_m}}, f^\star(X_{i'_{\hat{n}'_m}}))\}$. Note that, since every $j \in [\hat{n}'_m]$ with $X_{i'_j} \notin \mathrm{DIS}(V_{\lfloor m/2 \rfloor})$ has $h(X_{i'_j}) = f^\star(X_{i'_j})$ for every $h \in \mathbb{C}[\mathcal{L}_{\lfloor m/2 \rfloor} \cup \{(\hat{X}_{i''_1}, f^\star(\hat{X}_{i''_1})), \ldots, (\hat{X}_{i''_{\hat{n}''_m}}, f^\star(\hat{X}_{i''_{\hat{n}''_m}}))]$ (by definition of DIS and monotonicity of $\mathcal{L} \mapsto \mathbb{C}[\mathcal{L}]$), we have

$$\begin{aligned}&\mathbb{C}[\mathcal{L}_{\lfloor m/2 \rfloor} \cup \{(\hat{X}_{i''_1}, f^\star(\hat{X}_{i''_1})), \ldots, (\hat{X}_{i''_{\hat{n}''_m}}, f^\star(\hat{X}_{i''_{\hat{n}''_m}}))] \\ &\qquad = \mathbb{C}\left[\mathcal{L}_{\lfloor m/2 \rfloor} \cup \{(X_{i'_1}, f^\star(X_{i'_1})), \ldots, (X_{i'_{\hat{n}'_m}}, f^\star(X_{i'_{\hat{n}'_m}}))\}\right] = V_m,\end{aligned}$$

so that $\mathrm{DIS}(V_m)$ may be expressed as a fixed function of $X_1, \ldots, X_{\lfloor m/2 \rfloor}$ and $\hat{X}_{i''_1}, \ldots, \hat{X}_{i''_{\hat{n}''_m}}$. Furthermore, note that the set $\mathrm{DIS}\left(\mathbb{C}[\mathcal{L}_{\lfloor m/2 \rfloor} \cup \{(\hat{X}_{i''_1}, f^\star(\hat{X}_{i''_1})), \ldots, (\hat{X}_{i''_{\hat{n}''_m}}, f^\star(\hat{X}_{i''_{\hat{n}''_m}}))]\right)$ is invariant to permutations of the $i''_1, \ldots, i''_{\hat{n}''_m}$ indices.

Now note that N is conditionally Binomial$(\lceil m/2 \rceil, \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor})))$-distributed given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$. In particular, with probability one, if $P(\mathrm{DIS}(V_{\lfloor m/2 \rfloor})) = 0$, then $N = 0$. Otherwise, if $\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor})) > 0$, then note that $\hat{X}_1, \ldots, \hat{X}_N$ are conditionally independent and $\mathcal{P}(\cdot|\mathrm{DIS}(V_{\lfloor m/2 \rfloor}))$-distributed given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$ and N. Thus, since $\mathrm{DIS}(V_m) \cap \{\hat{X}_1, \ldots, \hat{X}_N\} = \emptyset$ (since every $h \in V_m$ agrees with $f^\star$ on $X_1, \ldots, X_m$), combining the above with Lemma 4 (applied under the conditional distribution given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$

and N), combined with the law of total probability, implies that for every $n \in [m] \cup \{0\}$, with probability at least $1 - \delta/(n+3)^2$, if $\hat{n}''_m = n$ and $N > n$, then

$$\mathcal{P}\left(\mathrm{DIS}(V_m) \middle| \mathrm{DIS}(V_{\lfloor m/2 \rfloor})\right) \leq \frac{1}{N-n}\left(n\mathrm{Log}\left(\frac{eN}{n}\right) + \mathrm{Log}\left(\frac{(n+3)^2}{\delta}\right)\right).$$

By a union bound, this holds simultaneously for all $n \in [m] \cup \{0\}$ on an event E_1 of probability at least $1 - \sum_{i=0}^{m} \frac{\delta}{(i+3)^2} > 1 - \frac{2}{5}\delta$. In particular, since the right hand side of the above inequality is nondecreasing in n, and $\hat{n}''_m \leq \hat{n}_m$, and since $\mathrm{DIS}(V_m) \subseteq \mathrm{DIS}(V_{\lfloor m/2 \rfloor})$, we have that on E_1, if $N > \hat{n}_m$, then

$$\mathcal{P}(\mathrm{DIS}(V_m)) \leq \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor}))\frac{1}{N-\hat{n}_m}\left(\hat{n}_m\mathrm{Log}\left(\frac{eN}{\hat{n}_m}\right) + \mathrm{Log}\left(\frac{(\hat{n}_m+3)^2}{\delta}\right)\right).$$

Next, again since N is conditionally Binomial$(\lceil m/2 \rceil, \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor})))$-distributed given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$, by a Chernoff bound (applied under the conditional distribution given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$), combined with the law of total probability, we obtain that on an event E_2 of probability at least $1 - \delta/3$, if $\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor})) \geq \frac{16}{m}\ln\left(\frac{3}{\delta}\right) \geq \frac{8}{\lceil m/2 \rceil}\ln\left(\frac{3}{\delta}\right)$, then

$$N \geq \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor}))\lceil m/2 \rceil/2 \geq \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor}))m/4.$$

Also note that if $\mathcal{P}(\mathrm{DIS}(V_m)) \geq \frac{16}{m}\left(2\hat{n}_m + \ln\left(\frac{3}{\delta}\right)\right)$, then monotonicity of $t \mapsto \mathrm{DIS}(V_t)$ and monotonicity of probability measures imply $\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor})) \geq \frac{16}{m}\left(2\hat{n}_m + \ln\left(\frac{3}{\delta}\right)\right)$ as well. In particular, if this occurs with E_2, then we have $N \geq \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor}))m/4 > 8\hat{n}_m$. Thus, by monotonicity of $x \mapsto \mathrm{Log}(x)/x$ for $x > 0$, we have that on $E_1 \cap E_2$, if $\mathcal{P}(\mathrm{DIS}(V_m)) \geq \frac{16}{m}\left(2\hat{n}_m + \ln\left(\frac{3}{\delta}\right)\right)$, then

$$\begin{aligned}\mathcal{P}(\mathrm{DIS}(V_m)) &< \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor}))\frac{8}{7N}\left(\hat{n}_m\mathrm{Log}\left(\frac{eN}{\hat{n}_m}\right) + \ln\left(\frac{(\hat{n}_m+3)^2}{\delta}\right)\right) \\ &\leq \frac{32}{7m}\left(\hat{n}_m\mathrm{Log}\left(\frac{e\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor}))m}{4\hat{n}_m}\right) + \ln\left(\frac{(\hat{n}_m+3)^2}{\delta}\right)\right).\end{aligned}$$

The inductive hypothesis implies that, on an event E_3 of probability at least $1 - \delta/4$,

$$\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor})) \leq \frac{16}{\lfloor m/2 \rfloor}\left(2\hat{n}_{1:\lfloor m/2 \rfloor} + \ln\left(\frac{12}{\delta}\right)\right).$$

Since $m \geq 17$, we have $\lfloor m/2 \rfloor \geq (m-2)/2 \geq (15/34)m$, so that the above implies

$$\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2 \rfloor})) \leq \frac{544}{15m}\left(2\hat{n}_{1:\lfloor m/2 \rfloor} + \ln\left(\frac{12}{\delta}\right)\right).$$

Thus, on $E_1 \cap E_2 \cap E_3$, if $\mathcal{P}(\mathrm{DIS}(V_m)) \geq \frac{16}{m}\left(2\hat{n}_m + \ln\left(\frac{3}{\delta}\right)\right)$, then

$$\begin{aligned}\mathcal{P}(\mathrm{DIS}(V_m)) &< \frac{32}{7m}\left(\hat{n}_m\mathrm{Log}\left(\frac{136e}{15}\left(2\frac{\hat{n}_{1:\lfloor m/2 \rfloor}}{\hat{n}_m} + \frac{1}{\hat{n}_m}\ln\left(\frac{12}{\delta}\right)\right)\right) + \ln\left(\frac{(\hat{n}_m+3)^2}{\delta}\right)\right) \\ &\leq \frac{32}{7m}\left(\hat{n}_{1:m}\mathrm{Log}\left(\frac{136e}{15}\left(2 + \frac{1}{\hat{n}_{1:m}}\ln(4) + \frac{1}{\hat{n}_{1:m}}\ln\left(\frac{3}{\delta}\right)\right)\right) + \ln\left(\frac{(\hat{n}_{1:m}+3)^2}{\delta}\right)\right). \quad (26)\end{aligned}$$

By straightforward calculus, one can easily verify that, when $\hat{n}_{1:m} \in \{0,1\}$, the right hand side of (26) is at most $\frac{16}{m}\left(2\hat{n}_{1:m} + \ln\left(\frac{3}{\delta}\right)\right)$ (recalling our conventions that $1/0 = \infty$ and $0\mathrm{Log}(\infty) = 0$). Otherwise, supposing $\hat{n}_{1:m} \geq 2$, Lemma 20 in Appendix A (applied with $b = \frac{5e}{2}\ln(3/\delta)$) implies the right hand side of (26) is at most

$$\frac{32}{7m}\left(\hat{n}_{1:m}\mathrm{Log}\left(\frac{136e}{15}\left(2+\ln(4)+\frac{2}{5}\right)\right) + 2\ln(\hat{n}_{1:m}+3) + \frac{7}{2}\ln\left(\frac{3}{\delta}\right)\right)$$
$$\leq \frac{32}{7m}\left(5\hat{n}_{1:m} + 2\ln(\hat{n}_{1:m}+3) + \frac{7}{2}\ln\left(\frac{3}{\delta}\right)\right).$$

Since $5x + 2\ln(x+3) < 7x$ for any $x \geq 2$, the above is at most

$$\frac{32}{7m}\left(7\hat{n}_{1:m} + \frac{7}{2}\ln\left(\frac{3}{\delta}\right)\right) = \frac{16}{m}\left(2\hat{n}_{1:m} + \ln\left(\frac{3}{\delta}\right)\right).$$

Thus, since $\frac{16}{m}\left(2\hat{n}_m + \ln\left(\frac{3}{\delta}\right)\right) \leq \frac{16}{m}\left(2\hat{n}_{1:m} + \ln\left(\frac{3}{\delta}\right)\right)$ as well, in either case we have that, on $E_1 \cap E_2 \cap E_3$,

$$\mathcal{P}(\mathrm{DIS}(V_m)) \leq \frac{16}{m}\left(2\hat{n}_{1:m} + \ln\left(\frac{3}{\delta}\right)\right).$$

Noting that, by a union bound, the event $E_1 \cap E_2 \cap E_3$ has probability at least $1 - \frac{2}{5}\delta - \frac{1}{3}\delta - \frac{1}{4}\delta > 1 - \delta$, this extends the result to m. By the principle of induction, this completes the proof of Theorem 7. ■

Appendix C. Proof of Theorem 11

We now present the proof of Theorem 11.

Proof of Theorem 11 The result trivially holds for $m \leq \lfloor 8(\ln(37) + 8\ln(6))\rfloor = 143$, so suppose $m \geq 144$. Let $N = |\{X_{\lfloor m/2\rfloor+1}, \ldots, X_m\} \cap \mathrm{DIS}(V_{\lfloor m/2\rfloor})|$ and enumerate the elements of $\{X_{\lfloor m/2\rfloor+1}, \ldots, X_m\} \cap \mathrm{DIS}(V_{\lfloor m/2\rfloor})$ as $\hat{X}_1, \ldots, \hat{X}_N$. Note that N is conditionally $\mathrm{Binomial}(\lceil m/2\rceil, \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor})))$ -distributed given $X_1, \ldots, X_{\lfloor m/2\rfloor}$. In particular, with probability one, if $\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor})) = 0$, then $N = 0$. Otherwise, if $\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor})) > 0$, then note that $\hat{X}_1, \ldots, \hat{X}_N$ are conditionally independent $\mathcal{P}(\cdot|\mathrm{DIS}(V_{\lfloor m/2\rfloor}))$-distributed random variables, given $X_1, \ldots, X_{\lfloor m/2\rfloor}$ and N. Also, note that (one can easily show) $\mathrm{vc}(\{\{x : h(x) \neq f^\star(x)\} : h \in \mathbb{C}\}) = d$. Together with Lemma 2 (applied under the conditional distribution given $X_1, \ldots, X_{\lfloor m/2\rfloor}$ and N), combined with the law of total probability, these observations imply that there is an event H_1 of probability at least $1 - \delta/3$, on which, if $N > 0$, then $\forall h \in V_m$,

$$\mathcal{P}(\mathrm{DIS}(\{h, f^\star\})|\mathrm{DIS}(V_{\lfloor m/2\rfloor})) \leq \frac{2}{N}\left(d\mathrm{Log}_2\left(\frac{2eN}{d}\right) + \log_2\left(\frac{6}{\delta}\right)\right).$$

In particular, noting that $\forall h \in V_m$, since $f^\star \in V_m$ as well, $\mathrm{DIS}(\{h, f^\star\}) \subseteq \mathrm{DIS}(V_m) \subseteq \mathrm{DIS}(V_{\lfloor m/2\rfloor})$, we have that on H_1, $\forall h \in V_m$,

$$\mathrm{er}(h) = \mathcal{P}(\mathrm{DIS}(\{h, f^\star\})) = \mathcal{P}(\mathrm{DIS}(\{h, f^\star\})|\mathrm{DIS}(V_{\lfloor m/2\rfloor}))\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor}))$$
$$\leq \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor}))\frac{2}{N}\left(d\mathrm{Log}_2\left(\frac{2eN}{d}\right) + \log_2\left(\frac{6}{\delta}\right)\right).$$

Next, again since N is conditionally Binomial$(\lceil m/2\rceil, \mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor})))$-distributed given $X_1,\ldots,X_{\lfloor m/2\rfloor}$, by a Chernoff bound (applied under the conditional distribution given $X_1,\ldots,X_{\lfloor m/2\rfloor}$), combined with the law of total probability, there is an event H_2 of probability at least $1-\delta/3$, on which, if $\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor})) \geq \frac{32}{\lceil m/2\rceil}\ln\left(\frac{3}{\delta}\right)$, then

$$N \geq (3/4)\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor}))\lceil m/2\rceil \geq (3/8)\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor}))m,$$

which (by $\mathrm{Log}_2(x) \leq \mathrm{Log}(x)/\ln(2)$ and monotonicity of $x \mapsto \mathrm{Log}(x)/x$ for $x>0$) implies

$$\begin{aligned}&\frac{2}{N}\left(d\mathrm{Log}_2\left(\frac{2eN}{d}\right)+\log_2\left(\frac{6}{\delta}\right)\right)\\ &\leq \frac{16}{3\ln(2)\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor}))m}\left(d\ln\left(\frac{3e\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor}))m}{4d}\right)+\ln\left(\frac{6}{\delta}\right)\right).\end{aligned}$$

Also, by Theorem 7, on an event H_3 of probability at least $1-\delta/3$,

$$\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor})) \leq \frac{16}{\lfloor m/2\rfloor}\left(2\hat{n}_{1:\lfloor m/2\rfloor}+\ln\left(\frac{9}{\delta}\right)\right).$$

Together with the facts that $\frac{16}{3\ln(2)} < 8$ and $\lfloor m/2\rfloor \geq \frac{m-2}{m}\frac{m}{2} \geq \frac{142}{144}\frac{m}{2}$, we have that, on $H_1 \cap H_2 \cap H_3$, if $\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor})) \geq \frac{32}{\lceil m/2\rceil}\ln\left(\frac{3}{\delta}\right)$, then

$$\begin{aligned}\sup_{h\in V_m}\mathrm{er}(h) &\leq \frac{8}{m}\left(d\ln\left(\frac{e24\cdot 144(2\hat{n}_{1:\lfloor m/2\rfloor}+\ln(9/\delta))}{142d}\right)+\ln\left(\frac{6}{\delta}\right)\right)\\ &= \frac{8}{m}\left(d\ln\left(\frac{24\cdot 144}{7\cdot 142}\left(\frac{14e\hat{n}_{1:\lfloor m/2\rfloor}+7e\ln(3/2)}{d}+\frac{7e\ln(6/\delta)}{d}\right)\right)+\ln\left(\frac{6}{\delta}\right)\right).\end{aligned}$$

By Lemma 20 in Appendix A, this last expression is at most

$$\begin{aligned}&\frac{8}{m}\left(d\ln\left(\frac{24\cdot 144}{7\cdot 142}\left(\frac{14e\hat{n}_{1:\lfloor m/2\rfloor}+7e\ln(3/2)}{d}+e\right)\right)+8\ln\left(\frac{6}{\delta}\right)\right)\\ &\leq \frac{8}{m}\left(d\ln\left(\left(\frac{49e\hat{n}_{1:\lfloor m/2\rfloor}}{d}+37\right)\right)+8\ln\left(\frac{6}{\delta}\right)\right).\end{aligned}$$

Furthermore, since $\mathrm{DIS}(\{h,f^\star\}) \subseteq \mathrm{DIS}(V_{\lfloor m/2\rfloor})$ for every $h \in V_m$, if $\mathcal{P}(\mathrm{DIS}(V_{\lfloor m/2\rfloor})) < \frac{32}{\lceil m/2\rceil}\ln\left(\frac{3}{\delta}\right) \leq \frac{64}{m}\ln\left(\frac{3}{\delta}\right)$, then

$$\sup_{h\in V_m}\mathrm{er}(h) < \frac{64}{m}\ln\left(\frac{3}{\delta}\right) < \frac{8}{m}\left(d\mathrm{Log}\left(\frac{49e\hat{n}_{1:\lfloor m/2\rfloor}}{d}+37\right)+8\ln\left(\frac{6}{\delta}\right)\right).$$

Thus, in either case, we have that, on $H_1 \cap H_2 \cap H_3$,

$$\sup_{h\in V_m}\mathrm{er}(h) \leq \frac{8}{m}\left(d\mathrm{Log}\left(\frac{49e\hat{n}_{1:\lfloor m/2\rfloor}}{d}+37\right)+8\ln\left(\frac{6}{\delta}\right)\right).$$

The proof is completed by noting that $\hat{n}_{1:\lfloor m/2\rfloor} \leq \hat{n}_{1:m}$, and that, by the union bound, the event $H_1 \cap H_2 \cap H_3$ has probability at least $1-\delta$. ∎

Appendix D. Proof of Theorem 16

We now present the proof of Theorem 16.

Proof of Theorem 16 The proof essentially combines the argument of Hanneke (2009) (which proves (16)) with the subsample-based ideas of Zhang and Chaudhuri (2014). Fix $c = 16$. The proof proceeds by induction on m. Since $\sup_{h \in \mathbb{C}} \mathrm{er}(h) \leq 1$, the result trivially holds for $m < 21(d \ln(83) + 3\ln(4))$. Now, as an inductive hypothesis, fix any $m \geq 21(d\ln(83) + 3\ln(4))$ such that $\forall m' \in [m-1]$, $\forall \delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{h \in V_{m'}} \mathrm{er}(h) \leq \frac{21}{m'} \left(d\mathrm{Log}\left(83\varphi_c\left(\frac{d}{m'}\right)\right) + 3\mathrm{Log}\left(\frac{4}{\delta}\right)\right).$$

Fix any $\delta \in (0,1)$ and $\eta \in [0,1]$. Let γ^*, ζ^*, ξ^* be the functions γ, ζ, and ξ from Definition 15 (each mapping $\mathcal{X} \to [0,1]$) with $\gamma^*(x) + \zeta^*(x) + \xi^*(x) = 1$ for all $x \in \mathcal{X}$, and $\mathbb{E}\left[\gamma^*(X) \middle| X_1, \ldots, X_{\lfloor m/2 \rfloor}\right]$ minimal subject to

$$\sup_{h \in V_{\lfloor m/2 \rfloor}} \mathbb{E}\left[\mathbb{1}[h(X) = +1]\zeta^*(X) + \mathbb{1}[h(X) = -1]\xi^*(X) \middle| X_1, \ldots, X_{\lfloor m/2 \rfloor}\right] \leq \eta,$$

where $X \sim \mathcal{P}$ is independent of $X_1, X_2, \ldots$.[9] Note that these functions are themselves random, having dependence on $X_1, \ldots, X_{\lfloor m/2 \rfloor}$. In particular, $\mathbb{E}\left[\gamma^*(X) \middle| X_1, \ldots, X_{\lfloor m/2 \rfloor}\right] = \Phi(V_{\lfloor m/2 \rfloor}, \eta)$.

Let $\Gamma_{\lfloor m/2 \rfloor + 1}, \ldots, \Gamma_m$ be conditionally independent random variables given $X_1, \ldots, X_m$, with Γ_i having conditional distribution Bernoulli$(\gamma^*(X_i))$ given $X_1, \ldots, X_m$, for each $i \in \{\lfloor m/2 \rfloor + 1, \ldots, m\}$. Let $N = |\{i \in \{\lfloor m/2 \rfloor + 1, \ldots, m\} : \Gamma_i = 1\}|$, and enumerate the elements of $\{X_i : i \in \{\lfloor m/2 \rfloor + 1, \ldots, m\}, \Gamma_i = 1\}$ as $\hat{X}_1, \ldots, \hat{X}_N$ (retaining their original order). For $X \sim \mathcal{P}$ independent of $X_1, X_2, \ldots$, let $\Gamma(X)$ denote a random variable that is conditionally Bernoulli$(\gamma^*(X))$ given X and $X_1, \ldots, X_{\lfloor m/2 \rfloor}$. Also define a (random) probability measure $P_{\lfloor m/2 \rfloor}$ such that, given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$, $P_{\lfloor m/2 \rfloor}(A) = \mathbb{P}(X \in A | \Gamma(X) = 1, X_1, \ldots, X_{\lfloor m/2 \rfloor})$ for all measurable $A \subseteq \mathcal{X}$.

Note that $N = \sum_{t=\lfloor m/2 \rfloor + 1}^{m} \Gamma_i$ is conditionally Binomial$\left(\lceil m/2 \rceil, \Phi(V_{\lfloor m/2 \rfloor}, \eta)\right)$ given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$. In particular, with probability one, if $\Phi(V_{\lfloor m/2 \rfloor}, \eta) = 0$, then $N = 0$. Otherwise, if $\Phi(V_{\lfloor m/2 \rfloor}, \eta) > 0$, then $\hat{X}_1, \ldots, \hat{X}_N$ are conditionally i.i.d. given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$ and N, each with conditional distribution $P_{\lfloor m/2 \rfloor}$ given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$ and N. Thus, since every $h \in V_m$ has $\{x : h(x) \neq f^\star(x)\} \cap \{\hat{X}_1, \ldots, \hat{X}_N\} \subseteq \{x : h(x) \neq f^\star(x)\} \cap \{X_1, \ldots, X_m\} = \emptyset$, and (one can easily show) $\mathrm{vc}(\{\{x : h(x) \neq f^\star(x)\} : h \in \mathbb{C}\}) = d$, applying Lemma 2 (under the conditional distribution given N and $X_1, \ldots, X_{\lfloor m/2 \rfloor}$), combined with the law of total probability, we have that on an event E_1 of probability at least $1 - \delta/2$, if $N > 0$, then

$$\sup_{h \in V_m} P_{\lfloor m/2 \rfloor}(x : h(x) \neq f^\star(x)) \leq \frac{2}{N}\left(d\mathrm{Log}_2\left(\frac{2eN}{d}\right) + \log_2\left(\frac{4}{\delta}\right)\right).$$

Next, since N is conditionally Binomial$\left(\lceil m/2 \rceil, \Phi(V_{\lfloor m/2 \rfloor}, \eta)\right)$ given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$, applying a Chernoff bound (under the conditional distribution given $X_1, \ldots, X_{\lfloor m/2 \rfloor}$), combined with the law of total probability, we obtain that on an event E_2 of probability at least

9. Note that the minimum is actually achieved here, since the objective function is continuous and convex, and the feasible region is nonempty, closed, bounded, and convex (see Bowers and Kalton, 2014, Proposition 5.50).

$1-\delta/4$, if $\Phi(V_{\lfloor m/2\rfloor},\eta) \geq \frac{18}{\lceil m/2\rceil}\ln\left(\frac{4}{\delta}\right)$, then

$$N \geq (2/3)\Phi(V_{\lfloor m/2\rfloor},\eta)\lceil m/2\rceil \geq \Phi(V_{\lfloor m/2\rfloor},\eta)m/3.$$

In particular, if $\Phi(V_{\lfloor m/2\rfloor},\eta) \geq \frac{18}{\lceil m/2\rceil}\ln\left(\frac{4}{\delta}\right)$, then the right hand side is strictly greater than 0, so that if this occurs with E_2, then we have $N > 0$. Thus, by the fact that $\mathrm{Log}_2(x) \leq \mathrm{Log}(x)/\ln(2)$, combined with monotonicity of $x \mapsto \mathrm{Log}(x)/x$ for $x > 0$, we have that on $E_1 \cap E_2$, if $\Phi(V_{\lfloor m/2\rfloor},\eta) \geq \frac{18}{\lceil m/2\rceil}\ln\left(\frac{4}{\delta}\right)$, then

$$\sup_{h\in V_m} P_{\lfloor m/2\rfloor}(x : h(x) \neq f^\star(x)) \leq \frac{6/\ln(2)}{\Phi(V_{\lfloor m/2\rfloor},\eta)m}\left(d\mathrm{Log}\left(\frac{2e\Phi(V_{\lfloor m/2\rfloor},\eta)m}{3d}\right)+\ln\left(\frac{4}{\delta}\right)\right).$$

Next (following an argument of Zhang and Chaudhuri, 2014), note that $\forall h \in V_m$,

$$\begin{aligned}
\mathrm{er}(h) &= \mathbb{E}\left[\mathbb{1}[h(X)\neq f^\star(X)]\left(\gamma^*(X)+\zeta^*(X)+\xi^*(X)\right)\middle|X_1,\ldots,X_{\lfloor m/2\rfloor}\right] \\
&= P_{\lfloor m/2\rfloor}(x : h(x)\neq f^\star(x))\mathbb{P}(\Gamma(X)=1|X_1,\ldots,X_{\lfloor m/2\rfloor}) \\
&\quad + \mathbb{E}\Big[\big(\mathbb{1}[h(X)=+1]\mathbb{1}[f^\star(X)=-1] \\
&\qquad\qquad + \mathbb{1}[h(X)=-1]\mathbb{1}[f^\star(X)=+1]\big)\left(\zeta^*(X)+\xi^*(X)\right)\Big|X_1,\ldots,X_{\lfloor m/2\rfloor}\Big] \\
&\leq P_{\lfloor m/2\rfloor}(x : h(x)\neq f^\star(x))\Phi(V_{\lfloor m/2\rfloor},\eta) \\
&\quad + \mathbb{E}\left[\mathbb{1}[h(X)=+1]\zeta^*(X)+\mathbb{1}[h(X)=-1]\xi^*(X)\middle|X_1,\ldots,X_{\lfloor m/2\rfloor}\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}[f^\star(X)=+1]\zeta^*(X)+\mathbb{1}[f^\star(X)=-1]\xi^*(X)\middle|X_1,\ldots,X_{\lfloor m/2\rfloor}\right].
\end{aligned}$$

Since $h, f^\star \in V_{\lfloor m/2\rfloor}$, the definition of ζ^* and ξ^* implies this last expression is at most

$$P_{\lfloor m/2\rfloor}(x : h(x)\neq f^\star(x))\Phi(V_{\lfloor m/2\rfloor},\eta)+2\eta.$$

Therefore, on $E_1 \cap E_2$, if $\Phi(V_{\lfloor m/2\rfloor},\eta) \geq \frac{18}{\lceil m/2\rceil}\ln\left(\frac{4}{\delta}\right)$, then

$$\sup_{h\in V_m}\mathrm{er}(h) \leq 2\eta + \frac{6/\ln(2)}{m}\left(d\mathrm{Log}\left(\frac{2e\Phi(V_{\lfloor m/2\rfloor},\eta)m}{3d}\right)+\ln\left(\frac{4}{\delta}\right)\right).$$

The inductive hypothesis implies that, on an event E_3 of probability at least $1-\delta/4$,

$$\sup_{h\in V_{\lfloor m/2\rfloor}}\mathrm{er}(h) \leq \frac{21}{\lfloor m/2\rfloor}\left(d\mathrm{Log}\left(83\varphi_c\left(\frac{d}{\lfloor m/2\rfloor}\right)\right)+3\mathrm{Log}\left(\frac{16}{\delta}\right)\right).$$

Since $m \geq \lceil 21(d\ln(83)+3\ln(4))\rceil \geq 181$, we have $\lfloor m/2\rfloor \geq (m-2)/2 \geq (179/362)m$, so that (together with monotonicity of $\varphi_c(\cdot)$) the above implies $V_{\lfloor m/2\rfloor} \subseteq \mathrm{B}(f^\star, r_{\lfloor m/2\rfloor})$, where

$$r_{\lfloor m/2\rfloor} = \frac{21\cdot 362}{179m}\left(d\ln\left(83\varphi_c\left(\frac{d}{m}\right)\right)+3\ln\left(\frac{16}{\delta}\right)\right).$$

Altogether, plugging in $\eta = (r_{\lfloor m/2\rfloor}/c)\wedge 1$, and noting that $\mathcal{H} \mapsto \Phi(\mathcal{H},\eta)$ is nondecreasing in $\mathcal{H}$, and that $d/m \leq r_{\lfloor m/2\rfloor}$, we have that on $E_1 \cap E_2 \cap E_3$, if $\Phi(V_{\lfloor m/2\rfloor},(r_{\lfloor m/2\rfloor}/c)\wedge 1) \geq$

$\frac{18}{\lceil m/2 \rceil} \ln\left(\frac{4}{\delta}\right)$, then

$$\sup_{h \in V_m} \mathrm{er}(h) \le \frac{2r_{\lfloor m/2 \rfloor}}{c} + \frac{6/\ln(2)}{m}\left(d\mathrm{Log}\left(\frac{2e\Phi(\mathrm{B}(f^{\star}, r_{\lfloor m/2 \rfloor}), (r_{\lfloor m/2 \rfloor}/c) \wedge 1)m}{3d}\right) + \ln\left(\frac{4}{\delta}\right)\right)$$
$$\le \frac{2r_{\lfloor m/2 \rfloor}}{c} + \frac{6/\ln(2)}{m}\left(d\ln\left(\frac{2e\varphi_c(d/m) r_{\lfloor m/2 \rfloor} m}{3d}\right) + \ln\left(\frac{4}{\delta}\right)\right). \tag{27}$$

The second term in this last expression equals

$$\frac{6/\ln(2)}{m}\left(d\ln\left(\frac{14 \cdot 362}{179}\varphi_c\left(\frac{d}{m}\right)\left(e\ln\left(83\varphi_c\left(\frac{d}{m}\right)\right) + \frac{3e}{d}\ln\left(\frac{16}{\delta}\right)\right)\right) + \ln\left(\frac{4}{\delta}\right)\right)$$
$$\le \frac{6/\ln(2)}{m}\left(d\ln\left(\frac{14 \cdot 362 \cdot 6}{179 \cdot 7}\varphi_c\left(\frac{d}{m}\right)\left(\frac{7e}{6}\ln\left(64 \cdot 83\varphi_c\left(\frac{d}{m}\right)\right) + \frac{7e}{2d}\ln\left(\frac{4}{\delta}\right)\right)\right) + \ln\left(\frac{4}{\delta}\right)\right).$$

Applying Lemma 20 (with $b = (7e/2)\ln(4/\delta)$), this is at most

$$\frac{6/\ln(2)}{m}\left(d\ln\left(\frac{14 \cdot 362 \cdot 6}{179 \cdot 7}\varphi_c\left(\frac{d}{m}\right)\left(\frac{7e}{6}\ln\left(64 \cdot 83\varphi_c\left(\frac{d}{m}\right)\right) + e\right)\right) + \frac{9}{2}\ln\left(\frac{4}{\delta}\right)\right),$$

and a simple relaxation of the expression in the logarithm reveals this is at most

$$\frac{6/\ln(2)}{m}\left(\frac{3}{2}d\ln\left(83\varphi_c\left(\frac{d}{m}\right)\right) + \frac{9}{2}\ln\left(\frac{4}{\delta}\right)\right) \le \frac{13}{m}\left(d\ln\left(83\varphi_c\left(\frac{d}{m}\right)\right) + 3\ln\left(\frac{4}{\delta}\right)\right).$$

Additionally, some straightforward reasoning about numerical constants reveals that

$$\frac{2r_{\lfloor m/2 \rfloor}}{c} \le \frac{8}{m}\left(d\ln\left(83\varphi_c\left(\frac{d}{m}\right)\right) + 3\ln\left(\frac{4}{\delta}\right)\right).$$

Plugging these two facts back into (27), we have that on $E_1 \cap E_2 \cap E_3$, if $\Phi(V_{\lfloor m/2 \rfloor}, (r_{\lfloor m/2 \rfloor}/c) \wedge 1) \ge \frac{18}{\lceil m/2 \rceil}\ln\left(\frac{4}{\delta}\right)$, then

$$\sup_{h \in V_m} \mathrm{er}(h) \le \frac{21}{m}\left(d\ln\left(83\varphi_c\left(\frac{d}{m}\right)\right) + 3\ln\left(\frac{4}{\delta}\right)\right). \tag{28}$$

On the other hand, if $\Phi(V_{\lfloor m/2 \rfloor}, (r_{\lfloor m/2 \rfloor}/c) \wedge 1) < \frac{18}{\lceil m/2 \rceil}\ln\left(\frac{4}{\delta}\right)$, then recalling that (as established above) $\sup_{h \in V_m} \mathrm{er}(h) \le 2\eta + \sup_{h \in V_m} P_{\lfloor m/2 \rfloor}(x : h(x) \ne f^{\star}(x))\Phi(V_{\lfloor m/2 \rfloor}, \eta)$, plugging in $\eta = (r_{\lfloor m/2 \rfloor}/c) \wedge 1$ and noting that $P_{\lfloor m/2 \rfloor}(x : h(x) \ne f^{\star}(x)) \le 1$, we have

$$\sup_{h \in V_m} \mathrm{er}(h) \le \frac{2r_{\lfloor m/2 \rfloor}}{c} + \Phi(V_{\lfloor m/2 \rfloor}, (r_{\lfloor m/2 \rfloor}/c) \wedge 1)$$
$$< \frac{8}{m}\left(d\ln\left(83\varphi_c\left(\frac{d}{m}\right)\right) + 3\ln\left(\frac{4}{\delta}\right)\right) + \frac{18}{\lceil m/2 \rceil}\ln\left(\frac{4}{\delta}\right)$$
$$\le \frac{21}{m}\left(d\ln\left(83\varphi_c\left(\frac{d}{m}\right)\right) + 3\ln\left(\frac{4}{\delta}\right)\right).$$

Thus, in either case, on $E_1 \cap E_2 \cap E_3$, (28) holds. Noting that, by the union bound, the event $E_1 \cap E_2 \cap E_3$ has probability at least $1 - \delta$, this extends the inductive hypothesis to m. The result then follows by the principle of induction. ■

D.1 The Worst-Case Value of φ_c

Next, we prove (18). Fix any $c \geq 2$. First, suppose $r_0 \in (0,1)$, and let $m = \min\left\{\mathfrak{s}, \left\lceil \frac{1}{r_0} \right\rceil\right\}$; note that our assumption that $|\mathbb{C}| \geq 3$ implies $\mathfrak{s} \geq 2$, so that $m \geq 2$ here. Let $x_1, \ldots, x_m \in \mathcal{X}$ and $h_0, h_1, \ldots, h_m \in \mathbb{C}$ be as in Definition 9. Let $\mathcal{P}(\{x_i\}) = 1/m$ for each $i \in [m]$, and take $f^\star = h_0$.

Let r_1 be any value satisfying $\max\{1/m, r_0\} < r_1 \leq 1$ chosen sufficiently close to $\max\{1/m, r_0\}$ so that $\frac{m r_1}{c} < 1$. Consider now the definition of $\Phi(\mathrm{B}(f^\star, r_1), r_1/c)$ from Definition 15. For any functions $\chi_0, \chi_1 : \mathcal{X} \to [0,1]$, let $\zeta(x) = \mathbb{1}[h_0(x) = -1]\chi_0(x) + \mathbb{1}[h_0(x) = +1]\chi_1(x)$ and $\xi(x) = \mathbb{1}[h_0(x) = -1]\chi_1(x) + \mathbb{1}[h_0(x) = +1]\chi_0(x)$. In particular, note that it is possible to specify any functions $\zeta, \xi : \mathcal{X} \to [0,1]$ by choosing appropriate χ_0, χ_1 values (namely, $\chi_0(x) = \mathbb{1}[h_0(x) = -1]\zeta(x) + \mathbb{1}[h_0(x) = +1]\xi(x)$ and $\chi_1(x) = \mathbb{1}[h_0(x) = -1]\xi(x) + \mathbb{1}[h_0(x) = +1]\zeta(x)$). Noting that, for any classifier h and any $x \in \mathcal{X}$, $\mathbb{1}[h(x) = +1]\zeta(x) + \mathbb{1}[h(x) = -1]\xi(x) = \mathbb{1}[h(x) \neq h_0(x)]\chi_0(x) + \mathbb{1}[h(x) = h_0(x)]\chi_1(x)$, and $\zeta(x) + \xi(x) = \chi_0(x) + \chi_1(x)$, we may re-express the constraints in the optimization problem defining $\Phi(\mathrm{B}(f^\star, r_1), r_1/c)$ in Definition 15 as $\sup_{h \in \mathrm{B}(f^\star, r_1)} \mathbb{E}[\mathbb{1}[h(X) \neq h_0(X)]\chi_0(X) + \mathbb{1}[h(X) = h_0(X)]\chi_1(X)] \leq r_1/c$ and $\forall x \in \mathcal{X}$, $\gamma(x) + \chi_0(x) + \chi_1(x) = 1$ while $\gamma(x), \chi_0(x), \chi_1(x) \in [0,1]$. We may further simplify the problem by noting that $\gamma(x) = 1 - \chi_0(x) - \chi_1(x)$, so that these last two constraints become $\chi_0(x) + \chi_1(x) \leq 1$ while $\chi_0(x), \chi_1(x) \geq 0$, and the value $\Phi(\mathrm{B}(f^\star, r_1), r_1/c)$ is the minimum achievable value of $\mathbb{E}[1 - \chi_0(X) - \chi_1(X)]$ subject to these constraints. Furthermore, noting that $h_i \in \mathrm{B}(f^\star, r_1)$ for every $i \in [m]$, we have that

$$
\begin{aligned}
&\Phi(\mathrm{B}(f^\star, r_1), r_1/c) \\
&\geq \min\Bigg\{\mathbb{E}[1 - \chi_0(X) - \chi_1(X)] : \\
&\qquad\qquad \max_{i \in [m]} \mathbb{E}\left[\mathbb{1}[h_i(X) \neq h_0(X)]\chi_0(X) + \mathbb{1}[h_i(X) = h_0(X)]\chi_1(X)\right] \leq \frac{r_1}{c}, \\
&\qquad\qquad \text{where } \forall x \in \mathcal{X}, \chi_0(x) + \chi_1(x) \leq 1 \text{ and } \chi_0(x), \chi_1(x) \geq 0\Bigg\} \\
&= \min\Bigg\{\sum_{i=1}^{m} \frac{1}{m}(1 - \chi_0(x_i) - \chi_1(x_i)) : \\
&\qquad \forall i \in [m], \chi_0(x_i) + \sum_{j \neq i} \chi_1(x_j) \leq \frac{m r_1}{c}, \chi_0(x_i) + \chi_1(x_i) \leq 1, \chi_0(x_i), \chi_1(x_i) \geq 0\Bigg\}.
\end{aligned}
$$

This is a simple linear program with linear inequality constraints. We can explicitly solve this problem to find an optimal solution with $\chi_1(x_i) = 0$ and $\chi_0(x_i) = \frac{m r_1}{c}$ for all $i \in [m]$, at which the value of the objective function $\sum_{i=1}^{m} \frac{1}{m}(1 - \chi_0(x_i) - \chi_1(x_i))$ is $1 - \frac{m r_1}{c}$. One can easily verify that this choice of χ_0 and χ_1 satisfies the constraints above. To see that this is an optimal choice, we note that the objective function can be re-expressed as $\sum_{i=1}^{m} \frac{1}{m}(1 - \chi_0(x_i) - \chi_1(x_{\sigma(i)}))$, where $\sigma(i) = i + 1$ for $i \in [m-1]$, and $\sigma(m) = 1$. In particular, since $m \geq 2$, we have $\sigma(i) \neq i$ for each $i \in [m]$. Thus, for any χ_0 and χ_1 satisfying the constraints above, we have $\chi_0(x_i) + \chi_1(x_{\sigma(i)}) \leq \chi_0(x_i) + \sum_{j \neq i} \chi_1(x_j) \leq \frac{m r_1}{c}$

for each $i \in [m]$, so that $\sum_{i=1}^{m} \frac{1}{m}(1 - \chi_0(x_i) - \chi_1(x_{\sigma(i)})) \geq 1 - \frac{mr_1}{c}$, which is precisely the value obtained with the above choices of χ_0 and χ_1.

Thus, since the above argument holds for any choice of $r_1 > \max\{1/m, r_0\}$ sufficiently close to $\max\{1/m, r_0\}$, we have

$$\varphi_c(r_0) = \sup_{r_0 < r \leq 1} \frac{\Phi(\mathrm{B}(f^\star, r), r/c)}{r} \vee 1 \geq \lim_{r_1 \searrow \max\{1/m, r_0\}} \frac{1 - \frac{mr_1}{c}}{r_1} = \frac{1 - \frac{1}{c}\max\{1, mr_0\}}{\max\{1/m, r_0\}}.$$

If $\mathfrak{s} < \frac{1}{r_0}$, then $m = \mathfrak{s}$, and the rightmost expression above equals $(1 - 1/c)\mathfrak{s}$. Otherwise, if $s \geq \frac{1}{r_0}$, then $m = \left\lceil \frac{1}{r_0} \right\rceil$, and the rightmost expression above equals

$$\left(1 - \frac{1}{c}\left\lceil \frac{1}{r_0} \right\rceil r_0\right) \frac{1}{r_0} \geq \left(1 - \frac{1 + r_0}{c}\right) \frac{1}{r_0} = \left(1 - \frac{1}{c}\right)\left(\frac{1}{r_0} - \frac{1}{c-1}\right).$$

Either way, we have

$$\varphi_c(r_0) \geq \left(1 - \frac{1}{c}\right) \min\left\{\mathfrak{s}, \frac{1}{r_0} - \frac{1}{c-1}\right\}.$$

For the case $r_0 = 0$, we note that $\forall \varepsilon > 0$, any $c \geq 2$ has

$$\sup_{\mathcal{P}} \sup_{f^\star \in \mathbb{C}} \varphi_c(0) \geq \sup_{\mathcal{P}} \sup_{f^\star \in \mathbb{C}} \varphi_c(\varepsilon) \geq \left(1 - \frac{1}{c}\right) \min\left\{\mathfrak{s}, \frac{1}{\varepsilon} - \frac{1}{c-1}\right\}.$$

Taking the limit $\varepsilon \to 0$ yields $\sup_{\mathcal{P}} \sup_{f^\star \in \mathbb{C}} \varphi_c(0) \geq \left(1 - \frac{1}{c}\right)\mathfrak{s} = \left(1 - \frac{1}{c}\right)\min\left\{\mathfrak{s}, \frac{1}{r_0} - \frac{1}{c-1}\right\}$.

For the upper bound, we clearly have $\varphi_c(r_0) \leq (1 - 1/c)\theta(r_0)$ for every $c > 1$. To see this, take $\zeta(x) = (1/c)\mathbb{1}[x \in \mathrm{DIS}(\mathrm{B}(f^\star, r))]\mathbb{1}[f^\star(x) = -1] + \mathbb{1}[x \notin \mathrm{DIS}(\mathrm{B}(f^\star, r))]\mathbb{1}[f^\star(x) = -1]$ and $\xi(x) = (1/c)\mathbb{1}[x \in \mathrm{DIS}(\mathrm{B}(f^\star, r))]\mathbb{1}[f^\star(x) = +1] + \mathbb{1}[x \notin \mathrm{DIS}(\mathrm{B}(f^\star, r))]\mathbb{1}[f^\star(x) = +1]$ in the optimization problem defining $\Phi(\mathrm{B}(f^\star, r), r/c)$ in Definition 15. With these choices of ζ and ξ, we have $\mathbb{E}[\gamma(X)] = (1 - 1/c)\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f^\star, r)))$; also, for any $h \in \mathrm{B}(f^\star, r)$, since $\mathrm{DIS}(\{h, f^\star\}) \subseteq \mathrm{DIS}(\mathrm{B}(f^\star, r))$, we have $\mathbb{E}[\mathbb{1}[h(X) = +1]\zeta(X) + \mathbb{1}[h(X) = -1]\xi(X)] = \mathbb{E}[(1/c)\mathbb{1}[h(X) \neq f^\star(X)]] = (1/c)\mathcal{P}(x : h(x) \neq f^\star(x)) \leq r/c$; one can easily verify that the remaining constraints are also satisfied. Thus, since Hanneke and Yang (2015) prove $\sup_{\mathcal{P}} \sup_{f^\star \in \mathbb{C}} \theta(r_0) = \min\left\{\mathfrak{s}, \frac{1}{r_0}\right\}$, we have $\sup_{\mathcal{P}} \sup_{f^\star \in \mathbb{C}} \varphi_c(r_0) \leq (1 - 1/c)\min\left\{\mathfrak{s}, \frac{1}{r_0}\right\}$.

We also note that, if we define $\varphi_c^{01}(r_0)$ identically to $\varphi_c(r_0)$ except that γ is restricted to have *binary* values (i.e., in $\{0, 1\}$), then for $c \geq 4$, this same construction giving the lower bound above must have $\gamma(x_i) = 1$ for every $i \in [m]$, which implies $\varphi_c^{01}(r_0) \geq \min\left\{\mathfrak{s}, \frac{1}{r_0}\right\}$ in this case. To see this, consider any $r_1 > \max\{1/m, r_0\}$ sufficiently small so that $\frac{mr_1}{c} < \frac{1}{2}$; then to satisfy the constraints $\chi_0(x_i) + \sum_{j \neq i} \chi_1(x_j) \leq \frac{mr_1}{c} < \frac{1}{2}$ for every $i \in [m]$, while $\chi_0(x_i), \chi_1(x_i) \geq 0$, we must have every $\chi_0(x_i)$ and $\chi_1(x_i)$ strictly less than $\frac{1}{2}$, so that $\gamma(x_i) = 1 - \chi_0(x_i) - \chi_1(x_i) > 0$ (and hence, $\gamma(x_i) = 1$, due to the constraint to binary values). As we always have $\varphi_c^{01}(r_0) \leq \theta(r_0)$, and Hanneke and Yang (2015) have shown $\sup_{\mathcal{P}} \sup_{f^\star \in \mathbb{C}} \theta(r_0) = \min\left\{\mathfrak{s}, \frac{1}{r_0}\right\}$, this implies $\sup_{\mathcal{P}} \sup_{f^\star \in \mathbb{C}} \varphi_c^{01}(r_0) = \min\left\{\mathfrak{s}, \frac{1}{r_0}\right\}$ as well.

D.2 Relation of $\varphi_c(r_0)$ to the Doubling Dimension

Here we present the proof of (21), via a modification of an argument of Hanneke and Yang (2015). We in fact prove the following slightly stronger inequality: for any $c \geq 8$ and $r > 0$,

$$\log_2\left(\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P})\right) \leq 2d \log_2\left(96\left(\frac{\Phi(\mathrm{B}(f^\star, r), r/c)}{r} \vee 1\right)\right), \tag{29}$$

which will immediately imply (21) by taking the supremum of both sides over $r > r_0$ (with some careful consideration of the special case $r = r_0$; see below).

Fix any $c > 4$ and $r \in (0, 1]$. Let G_r denote any maximal $(r/2)$-packing of $\mathrm{B}(f^\star, r)$: that is, G_r is a subset of $\mathrm{B}(f^\star, r)$ of maximal cardinality such that $\min_{h,g \in G_r : h \neq g} \mathcal{P}(x : h(x) \neq g(x)) > r/2$. It is known that any such set G_r satisfies

$$\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P}) \leq |G_r| \leq \mathcal{N}(r/4, \mathrm{B}(f^\star, r), \mathcal{P}) \tag{30}$$

(see e.g., Kolmogorov and Tikhomirov, 1959, 1961; Vidyasagar, 2003). In particular, since we have assumed $d < \infty$, in our case this further implies $|G_r| < \infty$ (Haussler, 1995). Also, this implies that if $|G_r| = 1$, then (29) trivially holds, so let us suppose $|G_r| \geq 2$.

Now fix any measurable functions γ, ζ, ξ mapping $\mathcal{X} \to [0, 1]$ satisfying the constraint $\sup_{h \in \mathrm{B}(f^\star, r)} \mathbb{E}[\mathbb{1}[h(X) = +1]\zeta(X) + \mathbb{1}[h(X) = -1]\xi(X)] \leq r/c$, where $X \sim \mathcal{P}$, and $\forall x \in \mathcal{X}$, $\gamma(x) + \zeta(x) + \xi(x) = 1$; for simplicity, also suppose $\mathbb{E}[\gamma(X)] \geq r$. As above, for $m \in \mathbb{N}$, let $X_1, \ldots, X_m$ be independent $\mathcal{P}$-distributed random variables. Then let $\Gamma_1, \ldots, \Gamma_m$ be conditionally independent given $X_1, \ldots, X_m$, with the conditional distribution of each Γ_i as Bernoulli$(\gamma(X_i))$ given $X_1, \ldots, X_m$. Let $N_m = |\{i \in [m] : \Gamma_i = 1\}|$, and let $\hat{X}_1, \ldots, \hat{X}_{N_m}$ denote the subsequence of $X_1, \ldots, X_m$ for which the respective $\Gamma_i = 1$.

By two applications of the Chernoff bound, combined with the union bound, the event $E_1 = \{m\mathbb{E}[\gamma(X)]/2 \leq N_m \leq 2m\mathbb{E}[\gamma(X)]\}$ has probability at least $1 - 2\exp\{-m\mathbb{E}[\gamma(X)]/8\}$. Additionally, $\forall f, g \in G_r$ with $f \neq g$, $\forall i \in [m]$,

$$\begin{aligned}
&\mathbb{P}(f(X_i) \neq g(X_i) \text{ and } \Gamma_i = 0) \\
&= \mathbb{E}[\mathbb{1}[f(X) \neq g(X)](1 - \gamma(X))] = \mathbb{E}[\mathbb{1}[f(X) \neq g(X)](\zeta(X) + \xi(X))] \\
&= \mathbb{E}\left[(\mathbb{1}[f(X) = +1]\mathbb{1}[g(X) = -1] + \mathbb{1}[f(X) = -1]\mathbb{1}[g(X) = +1])\,(\zeta(X) + \xi(X))\right] \\
&\leq \mathbb{E}\left[\mathbb{1}[f(X) = +1]\zeta(X) + \mathbb{1}[f(X) = -1]\xi(X)\right] + \mathbb{E}\left[\mathbb{1}[g(X) = -1]\xi(X) + \mathbb{1}[g(X) = +1]\zeta(X)\right] \\
&\leq \frac{2r}{c},
\end{aligned}$$

so that

$$\mathbb{P}(f(X_i) \neq g(X_i) \text{ and } \Gamma_i = 1) = \mathbb{P}(f(X_i) \neq g(X_i)) - \mathbb{P}(f(X_i) \neq g(X_i) \text{ and } \Gamma_i = 0) > \frac{r}{2} - \frac{2r}{c}.$$

In particular, this implies

$$\mathbb{P}\left(f(X_i) \neq g(X_i) | \Gamma_i = 1\right) \geq \left(\frac{1}{2} - \frac{2}{c}\right)\frac{r}{\mathbb{E}[\gamma(X)]}.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\exists i \in [N_m] : f(\hat{X}_i) \neq g(\hat{X}_i) \middle| N_m\right) = 1 - (1 - \mathbb{P}(f(X_1) \neq g(X_1) | \Gamma_1 = 1))^{N_m} \\
&\geq 1 - \left(1 - \left(\frac{1}{2} - \frac{2}{c}\right)\frac{r}{\mathbb{E}[\gamma(X)]}\right)^{N_m} \geq 1 - \exp\left\{-\left(\frac{1}{2} - \frac{2}{c}\right)\frac{r}{\mathbb{E}[\gamma(X)]}N_m\right\}.
\end{aligned}$$

On the event E_1, this is at least $1 - \exp\left\{-\left(\frac{1}{4} - \frac{1}{c}\right) rm\right\}$. Altogether, we have that

$$
\begin{aligned}
&\mathbb{P}\left(E_1 \text{ and } \exists i \in [N_m] : f(\hat{X}_i) \neq g(\hat{X}_i)\right) = \mathbb{E}\left[\mathbb{1}_{E_1} \cdot \mathbb{P}\left(\exists i \in [N_m] : f(\hat{X}_i) \neq g(\hat{X}_i) \middle| N_m\right)\right] \\
&\geq \left(1 - \exp\left\{-\left(\frac{1}{4} - \frac{1}{c}\right) rm\right\}\right) \mathbb{P}(E_1) \\
&\geq 1 - \exp\left\{-\left(\frac{1}{4} - \frac{1}{c}\right) rm\right\} - 2\exp\{-m\mathbb{E}[\gamma(X)]/8\} \\
&\geq 1 - \exp\left\{-\left(\frac{c-4}{4c}\right) rm\right\} - 2\exp\{-mr/8\}.
\end{aligned}
$$

In particular, choosing

$$
m = \left\lceil \frac{1}{r}\left(\frac{4c}{c-4} \vee 8\right) \ln\left(2|G_r|^2\right) \right\rceil,
$$

we have that $\mathbb{P}\left(E_1 \text{ and } \exists i \in [N_m] : f(\hat{X}_i) \neq g(\hat{X}_i)\right) \geq 1 - \frac{2}{|G_r|^2}$. By a union bound, this implies that with probability at least $1 - \frac{2}{|G_r|^2}\binom{|G_r|}{2} = \frac{1}{|G_r|} > 0$, E_1 holds and, for every $f, g \in G_r$ with $f \neq g$, $\exists i \in [N_m]$ for which $f(\hat{X}_i) \neq g(\hat{X}_i)$: that is, every $f \in G_r$ classifies $\hat{X}_1, \ldots, \hat{X}_{N_m}$ distinctly. But for this to be the case, $|G_r|$ can be at most the number of distinct classifications of a sequence of N_m points in $\mathcal{X}$ realizable by classifiers in $\mathbb{C}$, where (since E_1 also holds) $N_m \leq 2m\mathbb{E}[\gamma(X)]$. Together with the VC-Sauer lemma (Vapnik and Chervonenkis, 1971; Sauer, 1972), this implies that

$$
\begin{aligned}
\log_2(|G_r|) &\leq d \log_2\left(\frac{2em\mathbb{E}[\gamma(X)]}{d} \vee 2\right) \\
&\leq d \log_2\left(\frac{35 \cdot 4e}{33}\left(\frac{4c}{c-4} \vee 8\right)\frac{\mathbb{E}[\gamma(X)]}{r}\frac{1}{d}\left(\ln(\sqrt{2}) + \ln(|G_r|)\right) \vee 2\right) \\
&= d \log_2\left(\frac{35 \cdot 4e}{33\log_2(e)}\left(\frac{4c}{c-4} \vee 8\right)\frac{\mathbb{E}[\gamma(X)]}{r}\frac{1}{d}\left((1/2) + \log_2(|G_r|)\right) \vee 2\right),
\end{aligned}
$$

where the second inequality follows from the fact that $8\ln(2|G_r|^2) > 16.5$ (since $|G_r| \geq 2$), so that $m \leq \frac{17.5}{16.5}\frac{1}{r}\left(\frac{4c}{c-4} \vee 8\right)\ln(2|G_r|^2) = \frac{35}{33}\frac{1}{r}\left(\frac{4c}{c-4} \vee 8\right)\ln(2|G_r|^2)$.

If $\log_2(|G_r|) \leq d$, then together with (30), the inequality (29) trivially holds. Otherwise, if $\log_2(|G_r|) > d$, then letting $K = \frac{1}{d}\log_2(|G_r|) \geq 1$, the above implies

$$
\begin{aligned}
K &\leq \log_2\left(\frac{35 \cdot 4e}{33\log_2(e)}\left(\frac{4c}{c-4} \vee 8\right)\frac{\mathbb{E}[\gamma(X)]}{r}\frac{3}{2}K\right) \\
&= \log_2\left(\frac{35 \cdot 4e}{22\log_2(e)}\left(\frac{4c}{c-4} \vee 8\right)\frac{\mathbb{E}[\gamma(X)]}{r}\right) + \log_2(K).
\end{aligned}
$$

Via some simple calculus (see e.g., Vidyasagar, 2003, Lemma 4.6), this implies

$$
K \leq 2\log_2\left(\frac{35 \cdot 4e}{22\log_2(e)}\left(\frac{4c}{c-4} \vee 8\right)\frac{\mathbb{E}[\gamma(X)]}{r}\right).
$$

Noting that $\frac{35 \cdot 4e}{22 \log_2(e)} < 12$, together with (30), we have that

$$\log_2(\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P})) \leq 2d \log_2\left(12\left(\frac{4c}{c-4} \vee 8\right)\frac{\mathbb{E}[\gamma(X)]}{r}\right). \tag{31}$$

This inequality holds for any choice of γ, ζ, ξ satisfying the constraints in the definition of $\Phi(\mathrm{B}(f^\star, r), r/c)$ from Definition 15, with the additional constraint that $\mathbb{E}[\gamma(X)] \geq r$. Thus, if $\Phi(\mathrm{B}(f^\star, r), r/c) \geq r$, then by minimizing the right hand side of (31) over the choice of γ, ζ, ξ, it follows that

$$\log_2(\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P})) \leq 2d \log_2\left(12\left(\frac{4c}{c-4} \vee 8\right)\frac{\Phi(\mathrm{B}(f^\star, r), r/c)}{r}\right).$$

Otherwise, if $\Phi(\mathrm{B}(f^\star, r), r/c) < r$, then we note that, for any functions γ^*, ζ^*, ξ^* satisfying the constraints from the definition of $\Phi(\mathrm{B}(f^\star, r), r/c)$ such that $\mathbb{E}[\gamma^*(X)] = \Phi(\mathrm{B}(f^\star, r), r/c)$, there exists functions γ, ζ, ξ satisfying the constraints from the definition of $\Phi(\mathrm{B}(f^\star, r), r/c)$ for which $\mathbb{E}[\gamma(X)] = r$. For instance, we can take γ based on a convex combination of γ^* and 1: $\gamma(x) = \frac{1-r}{1-\mathbb{E}[\gamma^*(X)]}\gamma^*(x) + \frac{r-\mathbb{E}[\gamma^*(X)]}{1-\mathbb{E}[\gamma^*(X)]}$, $\zeta(x) = (\zeta^*(x) - (\gamma(x) - \gamma^*(x))) \vee 0$, $\xi(x) = 1 - \gamma(x) - \zeta(x)$; one can easily verify that, since $0 \leq \zeta(x) \leq \zeta^*(x)$ and $0 \leq \xi(x) \leq \xi^*(x)$, this choice of γ, ζ, ξ still satisfy the requirements for γ, ζ, ξ above, and that furthermore, $\mathbb{E}[\gamma(X)] = r$. Therefore, (31) implies $\log_2(\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P})) \leq 2d \log_2\left(12\left(\frac{4c}{c-4} \vee 8\right)\right)$. Thus, either way, we have established that

$$\log_2(\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P})) \leq 2d \log_2\left(12\left(\frac{4c}{c-4} \vee 8\right)\left(\frac{\Phi(\mathrm{B}(f^\star, r), r/c)}{r} \vee 1\right)\right). \tag{32}$$

Noting that, for any $c \geq 8$, $\frac{4c}{c-4} \leq 8$, this establishes (29) for any $c \geq 8$ and $r \in (0, 1]$.

In the case of $r > 1$, a result of Haussler (1995) implies that

$$\begin{aligned}
&\log_2(\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P})) \leq \log_2(\mathcal{N}(1/2, \mathbb{C}, \mathcal{P})) \leq d \log_2(4e) + \log_2(e(d+1)) \\
&\leq d \log_2(4e) + d + \log_2(e) \leq d \log_2(8e^2) \leq d \log_2(96) \leq 2d \log_2\left(96\left(\frac{\Phi(\mathrm{B}(f^\star, r), r/c)}{r} \vee 1\right)\right),
\end{aligned}$$

so that both (29) and (32) are also valid for $r > 1$. This completes the proof of (29).

As a final step in the proof of (21), we note that there is a slight complication to be resolved, since the defintion of $D(r_0)$ includes r_0 in the range of r, while the definition of $\varphi_c(r_0)$ does not. However, we note that, for any $c > 4$, any $r_0 > 0$, and any $r > r_0$ sufficiently close to r_0, we have $c > cr_0/r > 4$, so that (32) would imply

$$\begin{aligned}
\log_2(\mathcal{N}(r_0/2, \mathrm{B}(f^\star, r_0), \mathcal{P})) &\leq 2d \log_2\left(12\left(\frac{4(cr_0/r)}{(cr_0/r) - 4} \vee 8\right)\left(\frac{\Phi(\mathrm{B}(f^\star, r_0), r_0/(cr_0/r))}{r_0} \vee 1\right)\right) \\
&\leq 2d \log_2\left(12\left(\frac{4c}{(cr_0/r) - 4} \vee \frac{8r}{r_0}\right)\left(\frac{\Phi(\mathrm{B}(f^\star, r), r/c)}{r} \vee 1\right)\right).
\end{aligned}$$

Then taking the limit as $r \searrow r_0$ implies

$$\begin{aligned}
\log_2\left(\mathcal{N}(r_0/2, \mathrm{B}(f^\star, r_0), \mathcal{P})\right) &\leq 2d \log_2\left(12\left(\frac{4c}{c-4} \vee 8\right)\lim_{r \searrow r_0}\left(\frac{\Phi(\mathrm{B}(f^\star, r), r/c)}{r} \vee 1\right)\right) \\
&\leq 2d \log_2\left(12\left(\frac{4c}{c-4} \vee 8\right)\varphi_c(r_0)\right).
\end{aligned}$$

In particular, for any $c \geq 8$, $\frac{4c}{c-4} \leq 8$, so that

$$\log_2\left(\mathcal{N}(r_0/2, \mathrm{B}(f^\star, r_0), \mathcal{P})\right) \leq 2d \log_2\left(96\varphi_c(r_0)\right).$$

Together with the above, we therefore have that, for any $c \geq 8$ and $r_0 > 0$,

$$\begin{aligned}
D(r_0) &= \max\left\{\log_2(\mathcal{N}(r_0/2, \mathrm{B}(f^\star, r_0), \mathcal{P})), \sup_{r > r_0} \log_2(\mathcal{N}(r/2, \mathrm{B}(f^\star, r), \mathcal{P}))\right\} \\
&\leq \max\left\{2d \log_2\left(96\varphi_c(r_0)\right), \sup_{r > r_0} 2d \log_2\left(96\left(\frac{\Phi(\mathrm{B}(f^\star, r), r/c)}{r} \vee 1\right)\right)\right\} \\
&= 2d \log_2\left(96\varphi_c(r_0)\right).
\end{aligned}$$

Thus, we have established (21).

Appendix E. Proofs of Results on Learning with Noise

This appendix includes the proofs of results in Section 6: namely, Theorems 17 and 19.

E.1 Proof of Theorem 17

We begin with the proof of Theorem 17. The proof follows a technique of Hanneke and Yang (2015), which identifies a subset of classifiers in $\mathbb{C}$, corresponding to a certain concept space for which Raginsky and Rakhlin (2011) have established lower bounds. Specifically, the following setup is taken directly from Hanneke and Yang (2015). Fix $\zeta \in (0,1]$, $\beta \in [0, 1/2)$, and $k \in \mathbb{N}$ with $k \leq \min\{1/\zeta, |\mathcal{X}| - 1\}$. Let $\mathcal{X}_k = \{x_1, \ldots, x_{k+1}\}$ be a set of $k+1$ distinct elements of $\mathcal{X}$, and define $\mathbb{C}_k = \{x \mapsto 2\mathbb{1}_{\{x_i\}}(x) - 1 : i \in [k]\}$. Let $\mathcal{P}_{k,\zeta}$ be a probability measure over $\mathcal{X}$ with $\mathcal{P}(\{x_i\}) = \zeta$ for each $i \in [k]$, and $\mathcal{P}_{k,\zeta}(\{x_{k+1}\}) = 1 - \zeta k$. For each $t \in [k]$, let $P'_{k,\zeta,t}$ be a probability measure over $\mathcal{X} \times \mathcal{Y}$ with marginal distribution $\mathcal{P}_{k,\zeta}$ over $\mathcal{X}$, such that for $(X, Y) \sim P'_{k,\zeta,t}$, every $i \in [k]$ has $\mathbb{P}(Y = 2\mathbb{1}_{\{x_t\}}(X) - 1 | X = x_i) = 1 - \beta$, and $\mathbb{P}(Y = -1 | X = x_{k+1}) = 1$. Raginsky and Rakhlin (2011) prove the following result (see the proof of their Theorem 1).[10]

Lemma 21 *For k, ζ, β as above, with $k \geq 2$, for any $\delta \in (0, 1/4)$, for any (passive) learning rule $\mathbb{A}$, and any $m \in \mathbb{N}$ with*

$$m < \max\left\{\frac{\beta \ln\left(\frac{1}{4\delta}\right)}{2\zeta(1-2\beta)^2}, \frac{3\beta \ln\left(\frac{k}{96}\right)}{16\zeta(1-2\beta)^2}\right\},$$

if $\mathbb{C}_k \subseteq \mathbb{C}$, then there exists a $t \in [k]$ such that, if $\mathcal{P}_{XY} = P'_{k,\zeta,t}$, then denoting $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$, with probability greater than δ,

$$\mathrm{er}(\hat{h}_m) - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \geq (\zeta/2)(1 - 2\beta).$$

10. As noted by Hanneke and Yang (2015), although technically the proof of this result by Raginsky and Rakhlin (2011) relies on a lemma (their Lemma 4) that imposes additional restrictions on k and a parameter "d", one can easily verify that the conclusions of that lemma continue to hold in the special case considered here (corresponding to $d = 1$ and arbitrary $k \in \mathbb{N}$) by defining $\mathcal{M}_{k,1} = \{0,1\}^k_1$ in their construction.

Continuing to follow Hanneke and Yang (2015), we embed the above scenario into the general case, so that Lemma 21 provides a lower bound. Fix any $\zeta \in (0,1]$, $\beta \in [0,1/2)$, and $k \in \mathbb{N}$ with $k \leq \min\{\mathfrak{s}-1, \lfloor 1/\zeta \rfloor\}$, and let $x_1, \ldots, x_{k+1}$ and $h_0, h_1, \ldots, h_k$ be as in Definition 9. Let $\mathcal{P}_{k,\zeta}$ be as above (for this choice of $x_1, \ldots, x_{k+1}$), and for each $t \in [k]$, let $P_{k,\zeta,t}$ denote a probability measure over $\mathcal{X} \times \mathcal{Y}$ with marginal distribution $\mathcal{P}_{k,\zeta}$ over $\mathcal{X}$ such that, for $(X,Y) \sim P_{k,\zeta,t}$, $\mathbb{P}(Y = h_t(X) | X = x_i) = 1 - \beta$ for every $i \in [k]$, while $\mathbb{P}(Y = h_t(X) | X = x_{k+1}) = 1$.

Lemma 22 *For k, ζ, β as above, with $k \geq 96e$, for any $\delta \in (0,1/4)$, for any (passive) learning rule $\mathbb{A}$, and any $m \in \mathbb{N}$ with*

$$m < \frac{3\beta \ln\left(\frac{k}{96}\right)}{16\zeta(1-2\beta)^2},$$

there exists a $t \in [k]$ such that, if $\mathcal{P}_{XY} = P_{k,\zeta,t}$, then denoting $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$, with probability greater than δ,

$$\mathrm{er}(\hat{h}_m) - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \geq (\zeta/2)(1-2\beta).$$

The proof of Lemma 22 is essentially identical to the proof of Hanneke and Yang (2015, Lemma 26), except that the algorithm $\mathbb{A}$ here is restricted to be a passive learning rule so that Lemma 21 can be applied (in place of Lemma 25 there). As such, we omit the details here for brevity.

We are now ready for the proof of Theorem 17.

Proof of Theorem 17 Fix any $\beta \in (0,1/2)$, $\delta \in (0,1/24)$, $m \in \mathbb{N}$, and any (passive) learning rule $\mathbb{A}$. First consider the case of $\mathfrak{s} \geq 97e$. Fix $\varepsilon \in (0, (1-2\beta)/(384e^2)]$, and let $\zeta = \frac{2\varepsilon}{1-2\beta}$ and $k = \min\{\mathfrak{s}-1, \lfloor 1/\zeta \rfloor\}$. Then, noting that the distributions $P_{k,\zeta,t}$ above satisfy the β-bounded noise condition, Lemma 22 implies that if

$$m < \frac{3\beta \ln\left(\frac{k}{96}\right)}{32\varepsilon(1-2\beta)}, \tag{33}$$

then there exists a choice of $\mathcal{P}_{XY}$ satisfying the β-bounded noise condition such that, with probability greater than δ, the classifier $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$ has

$$\mathrm{er}(\hat{h}_m) - \inf_{h \in \mathbb{C}} \mathrm{er}(h) \geq \varepsilon.$$

Note that for any $m \in \mathbb{N}$ and $\varepsilon \in (0, (1-2\beta)/(384e^2)]$, it holds that (see e.g., Vidyasagar, 2003, Corollary 4.1)

$$\begin{aligned} & m \leq \frac{3\beta}{64\varepsilon(1-2\beta)} \ln\left(\frac{(1-2\beta)^2 m}{18\beta}\right) \\ \implies & m < \frac{3\beta \ln\left(\frac{1-2\beta}{384\varepsilon}\right)}{32\varepsilon(1-2\beta)} \leq \frac{3\beta \ln\left(\frac{\lfloor 1/\zeta \rfloor}{96}\right)}{32\varepsilon(1-2\beta)}. \end{aligned}$$

Thus, the inequality in (33) is satisfied if both

$$m < \frac{3\beta \ln\left(\frac{\mathfrak{s}-1}{96}\right)}{32\varepsilon(1-2\beta)}$$

and

$$m \leq \frac{3\beta}{64\varepsilon(1-2\beta)} \ln\left(\frac{(1-2\beta)^2 m}{18\beta}\right).$$

Solving for a value $\varepsilon \in (0, (1-2\beta)/(384e^2)]$ that satisfies both of these, we have that for any $m \in \mathbb{N}$ with $m \geq \frac{18e\beta}{(1-2\beta)^2}$, there is a choice of $\mathcal{P}_{XY}$ satisfying the β-bounded noise condition such that, with probability greater than δ,

$$\begin{aligned}\operatorname{er}(\hat{h}_m) - \inf_{h \in \mathbb{C}} \operatorname{er}(h) &\geq \frac{3\beta \ln\left(\min\left\{\frac{\mathfrak{s}-1}{96}, \frac{(1-2\beta)^2 m}{18\beta}\right\}\right)}{64(1-2\beta)m} \wedge \frac{1-2\beta}{384e^2} \\ &\gtrsim \frac{\beta \mathrm{Log}\left(\min\left\{\mathfrak{s}, (1-2\beta)^2 m\right\}\right)}{(1-2\beta)m} \wedge (1-2\beta).\end{aligned}$$

Furthermore, for $m < \frac{18e\beta}{(1-2\beta)^2}$, we may also think of $\hat{h}_m$ as the output of $\mathbb{A}'(\mathcal{L}_{m'})$ for $m' = \left\lceil \frac{18e\beta}{(1-2\beta)^2} \right\rceil > m$, for a learning rule $\mathbb{A}'$ which simply discards the last $m'-m$ samples and runs $\mathbb{A}(\mathcal{L}_m)$ to produce its return classifier. Thus, the above result implies that for $m < \frac{18e\beta}{(1-2\beta)^2}$, with probability greater than δ,

$$\operatorname{er}(\hat{h}_m) - \inf_{h \in \mathbb{C}} \operatorname{er}(h) \geq \frac{3\beta \ln\left(\min\left\{\frac{\mathfrak{s}-1}{96}, \frac{(1-2\beta)^2 m'}{18\beta}\right\}\right)}{64(1-2\beta)m'} \wedge \frac{1-2\beta}{384e^2}.$$

Since $m, m' \in \mathbb{N}$ and $m' > m$, we know that $m' \geq 2$, so that $\frac{18e\beta}{(1-2\beta)^2} \leq m' \leq \frac{36e\beta}{(1-2\beta)^2}$. Therefore,

$$\frac{3\beta \ln\left(\min\left\{\frac{\mathfrak{s}-1}{96}, \frac{(1-2\beta)^2 m'}{18\beta}\right\}\right)}{64(1-2\beta)m'} \geq \frac{3\beta}{64(1-2\beta)m'} \geq \frac{3(1-2\beta)}{64 \cdot 36e} > \frac{(1-2\beta)}{384e^2}.$$

Thus, in this case, we have that with probability greater than δ,

$$\operatorname{er}(\hat{h}_m) - \inf_{h \in \mathbb{C}} \operatorname{er}(h) \geq \frac{(1-2\beta)}{384e^2} \gtrsim (1-2\beta) \geq \frac{\beta \mathrm{Log}\left(\min\left\{\mathfrak{s}, (1-2\beta)^2 m\right\}\right)}{(1-2\beta)m} \wedge (1-2\beta).$$

Next, we return to the general case of arbitrary $\mathfrak{s} \in \mathbb{N} \cup \{\infty\}$. In particular, since any $\mathfrak{s} < 97e$ has $\frac{\beta \mathrm{Log}\left(\min\left\{\mathfrak{s}, (1-2\beta)^2 m\right\}\right)}{(1-2\beta)m} \lesssim \frac{d}{(1-2\beta)m}$, to complete the proof it suffices to establish a lower bound

$$\operatorname{er}(\hat{h}_m) - \inf_{h \in \mathbb{C}} \operatorname{er}(h) \gtrsim \frac{1}{(1-2\beta)m}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \wedge (1-2\beta),$$

holding with probability greater than δ. This lower bound is already known, and frequently referred to in the literature; it follows from well-known constructions (see e.g., Anthony and Bartlett, 1999; Massart and Nédélec, 2006; Hanneke, 2011, 2014). The case $\beta < 3/8$ is covered by the classic minimax lower bound of Ehrenfeucht, Haussler, Kearns, and Valiant (1989) for the realizable case, while the case $\beta \geq 3/8$ is addressed by Hanneke (2014, Theorem 3.5). However, it seems an explicit proof of this latter result has not actually appeared in the literature. As such, for completeness, we include a brief sketch of the argument here.

Suppose $\beta \geq 3/8$. We begin with the term $\frac{1}{(1-2\beta)m}\mathrm{Log}\left(\frac{1}{\delta}\right)$. Since we have assumed $|\mathbb{C}| \geq 3$, there must exist $x_0, x_1 \in \mathcal{X}$ and $h_0, h_1 \in \mathbb{C}$ such that $h_0(x_0) = h_1(x_0)$ while $h_0(x_1) \neq h_1(x_1)$. Now fix $\varepsilon = \frac{3}{8(1-2\beta)m}\ln\left(\frac{1}{5\delta}\right) \wedge (1-2\beta)$, let $\mathcal{P}(\{x_1\}) = \frac{\varepsilon}{1-2\beta}$, and let $\mathcal{P}(\{x_0\}) = 1-\mathcal{P}(\{x_1\})$. Then, for $b \in \{0,1\}$, we let P_b be a distribution on $\mathcal{X}\times\mathcal{Y}$ with marginal $\mathcal{P}$ over $\mathcal{X}$, and with $P_b(\{(x_0, h_0(x_0))\}|\{x_0\}\times\mathcal{Y}) = 1$ and $P_b(\{(x_1, h_b(x_1))\}|\{x_1\}\times\mathcal{Y}) = 1-\beta$. Then one can easily check that, for $\mathcal{P}_{XY} = P_b$, any classifier h with $h(x_1) \neq h_b(x_1)$ has $\mathrm{er}(h) - \inf_{g\in\mathbb{C}}\mathrm{er}(g) \geq \varepsilon$. But since $\mathrm{KL}(P_0^m \| P_1^m) = m\mathrm{KL}(P_0\|P_1) = m\varepsilon\ln\left(\frac{1-\beta}{\beta}\right)$, and $\ln\left(\frac{1-\beta}{\beta}\right) \leq \frac{1-\beta}{\beta} - 1 = \frac{1-2\beta}{\beta} \leq \frac{8}{3}(1-2\beta)$ (since $\beta \geq 3/8$), classic hypothesis testing lower bounds (see Tsybakov, 2009, Theorem 2.2) imply that there exists a choice of $b \in \{0,1\}$ such that, with $\mathcal{P}_{XY} = P_b$ and $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$, $\mathbb{P}(\hat{h}_m(x_1) \neq h_b(x_1)) \geq \frac{1}{4}\exp\left\{-m\varepsilon\frac{8}{3}(1-2\beta)\right\} \geq (5/4)\delta > \delta$. Thus, with probability greater than δ, $\mathrm{er}(\hat{h}_m) - \inf_{g\in\mathbb{C}}\mathrm{er}(g) \geq \varepsilon \gtrsim \frac{1}{(1-2\beta)m}\mathrm{Log}\left(\frac{1}{\delta}\right)$.

Next, we present a proof for the term $\frac{d}{(1-2\beta)m}$, again for $\beta \geq 3/8$. This term is trivially implied by the term $\frac{1}{(1-2\beta)m}\mathrm{Log}\left(\frac{1}{\delta}\right)$ in the case $d = 1$, so suppose $d \geq 2$. This time, we let $\{x_0, \ldots, x_{d-1}\}$ denote a subset of $\mathcal{X}$ shatterable by $\mathbb{C}$, fix $\varepsilon = \frac{3(d-1)}{64e(1-2\beta)m} \wedge \frac{1-2\beta}{8e}$, and let $\mathcal{P}(\{x_i\}) = \frac{8e\varepsilon}{(d-1)(1-2\beta)}$ for $i \in \{1, \ldots, d-1\}$, and $\mathcal{P}(\{x_0\}) = 1 - \frac{8e\varepsilon}{1-2\beta}$. Now for each $\bar{b} = (b_1, \ldots, b_{d-1}) \in \{0,1\}^{d-1}$, let $P_{\bar{b}}$ denote a probability measure on $\mathcal{X}\times\mathcal{Y}$ with marginal $\mathcal{P}$ over $\mathcal{X}$, and with $P_{\bar{b}}(\{(x_i, 2b_i - 1)\}|\{x_i\}\times\mathcal{Y}) = 1-\beta$ for every $i \in \{1, \ldots, d-1\}$, and $P_{\bar{b}}(\{(x_0, -1)\}|\{x_0\}\times\mathcal{Y}) = 1$. In particular, note that any $\bar{b}, \bar{b}' \in \{0,1\}^{d-1}$ with Hamming distance $\|\bar{b} - \bar{b}'\|_1 = 1$ have $\mathrm{KL}(P_{\bar{b}}^m \| P_{\bar{b}'}^m) = m\mathrm{KL}(P_{\bar{b}}\|P_{\bar{b}'}) = m\frac{8e\varepsilon}{d-1}\ln\left(\frac{1-\beta}{\beta}\right)$, and as above, $\ln\left(\frac{1-\beta}{\beta}\right) \leq \frac{8}{3}(1-2\beta)$. Now Assouad's lemma (see Tsybakov, 2009, Theorem 2.12) implies that there exists a $\bar{b} \in \{0,1\}^{d-1}$ such that, with $\mathcal{P}_{XY} = P_{\bar{b}}$ and $\hat{h}_m = \mathbb{A}(\mathcal{L}_m)$, denoting $\hat{b} = ((1+\hat{h}_m(x_1))/2, \ldots, (1+\hat{h}_m(x_{d-1}))/2)$, we have $\mathbb{E}\left[\|\hat{b} - \bar{b}\|_1\right] \geq \frac{d-1}{4}\exp\left\{-m\frac{8e\varepsilon}{d-1}\frac{8}{3}(1-2\beta)\right\} \geq \frac{d-1}{4e}$. Noting that $0 \leq \|\hat{b} - \bar{b}\|_1 \leq d-1$, this further implies that $\mathbb{P}\left(\|\hat{b} - \bar{b}\|_1 \geq \frac{d-1}{8e}\right) \geq \frac{1}{8e}$. Furthermore, note that $\mathrm{er}(\hat{h}_m) - \inf_{g\in\mathbb{C}}\mathrm{er}(g) \geq \|\hat{b} - \bar{b}\|_1\frac{8e\varepsilon}{d-1}$. Thus, $\mathbb{P}\left(\mathrm{er}(\hat{h}_m) - \inf_{g\in\mathbb{C}}\mathrm{er}(g) \geq \varepsilon\right) \geq \frac{1}{8e} > \delta$. Finally, note that $\varepsilon \gtrsim \frac{d}{(1-2\beta)m} \wedge (1-2\beta)$.

Altogether, by choosing which ever of these lower bounds is greatest, we have that for any $m \in \mathbb{N}$, there exists a choice of $\mathcal{P}_{XY}$ satisfying the β-bounded noise condition such that, with probability greater than δ,

$$\mathrm{er}(\hat{h}_m) - \inf_{h\in\mathbb{C}}\mathrm{er}(h) \gtrsim \frac{\max\left\{d, \beta\mathrm{Log}\left(\min\left\{\mathfrak{s}, (1-2\beta)^2 m\right\}\right), \mathrm{Log}\left(\frac{1}{\delta}\right)\right\}}{(1-2\beta)m} \wedge (1-2\beta).$$

Applying the relaxation $\max\{a,b,c\} \geq (1/3)(a+b+c)$ (for nonnegative values a,b,c) then completes the proof of the first lower bound stated in the theorem.

For the second inequality, note that by taking $\delta = 1/24$, the inequality proven above implies that there exists a distribution $\mathcal{P}_{XY}$ satisfying the β-bounded noise condition such that, with probability greater than $1/24$,

$$\mathrm{er}(\hat{h}_m) - \inf_{h\in\mathbb{C}}\mathrm{er}(h) \gtrsim \frac{d + \beta\mathrm{Log}\left(\min\left\{\mathfrak{s}, (1-2\beta)^2 m\right\}\right)}{(1-2\beta)m} \wedge (1-2\beta).$$

Furthermore, since bounded noise distributions have $\inf_{h\in\mathbb{C}} \mathrm{er}(h)$ equal the Bayes risk, $\mathrm{er}(\hat{h}_m) - \inf_{h\in\mathbb{C}} \mathrm{er}(h)$ is always nonnegative. We therefore have

$$\begin{aligned}\mathbb{E}\left[\mathrm{er}(\hat{h}_m) - \inf_{h\in\mathbb{C}} \mathrm{er}(h)\right] &\gtrsim \frac{23}{24}0 + \frac{1}{24}\frac{d+\beta\mathrm{Log}\left(\min\left\{\mathfrak{s},(1-2\beta)^2 m\right\}\right)}{(1-2\beta)m} \wedge (1-2\beta)\\ &\gtrsim \frac{d+\beta\mathrm{Log}\left(\min\left\{\mathfrak{s},(1-2\beta)^2 m\right\}\right)}{(1-2\beta)m} \wedge (1-2\beta).\end{aligned}$$

Finally, since $\inf\limits_{h\in\mathbb{C}} \mathrm{er}(h)$ is nonrandom, $\mathbb{E}\left[\mathrm{er}(\hat{h}_m)\right] - \inf\limits_{h\in\mathbb{C}} \mathrm{er}(h) = \mathbb{E}\left[\mathrm{er}(\hat{h}_m) - \inf\limits_{h\in\mathbb{C}} \mathrm{er}(h)\right]$. ■

E.2 Proof of Theorem 19

Next, we present the proof of Theorem 19. We begin by stating a classic result, due to Giné and Koltchinskii (2006) (see also van der Vaart and Wellner, 2011; Hanneke and Yang, 2012). For any set $\mathcal{H}$ of classifiers, denote $\mathrm{diam}_{\mathcal{P}}(\mathcal{H}) = \sup_{h,g\in\mathcal{H}} \mathcal{P}(x : h(x) \neq g(x))$.

Lemma 23 *There is a universal constant $c_0 \in (1,\infty)$ such that, for any set $\mathcal{H}$ of classifiers, for any $\delta \in (0,1)$ and $m \in \mathbb{N}$, defining*

$$U(\mathcal{H},m,\delta;R) = 1 \wedge \inf_{r>\mathrm{diam}_{\mathcal{P}}(\mathcal{H})} c_0\sqrt{r\frac{\mathrm{vc}(\mathcal{H})\mathrm{Log}\left(\frac{\mathcal{P}(R)}{r}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{m}} + c_0\frac{\mathrm{vc}(\mathcal{H})\mathrm{Log}\left(\frac{\mathcal{P}(R)}{r}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{m}$$

for every measurable $R \subseteq \mathcal{X}$, with probability at least $1-\delta$, $\forall h \in \mathcal{H}$,

$$\mathrm{er}(h) - \inf_{g\in\mathcal{H}} \mathrm{er}(g) \leq \max\left\{2\left(\mathrm{er}_{\mathcal{L}_m}(h) - \min_{g\in\mathcal{H}} \mathrm{er}_{\mathcal{L}_m}(g)\right), U(\mathcal{H},m,\delta;\mathrm{DIS}(\mathcal{H}))\right\},$$
$$\mathrm{er}_{\mathcal{L}_m}(h) - \min_{g\in\mathcal{H}} \mathrm{er}_{\mathcal{L}_m}(g) \leq \max\left\{2\left(\mathrm{er}(h) - \inf_{g\in\mathcal{H}} \mathrm{er}(g)\right), U(\mathcal{H},m,\delta;\mathrm{DIS}(\mathcal{H}))\right\}.$$

Next, we note that we lose very little by requiring the γ function in Definition 15 to be binary. This allows us to simplify certain parts of the proof of Theorem 19 below.

Lemma 24 *For any set $\mathcal{H}$ of classifiers, and any $\eta \in [0,1]$, for $X \sim \mathcal{P}$, letting*

$$\begin{aligned}\Phi_{\{0,1\}}(\mathcal{H},\eta) = \inf\Bigg\{&\mathbb{E}[\gamma(X)] : \sup_{h\in\mathcal{H}} \mathbb{E}\left[\mathbb{1}[h(X)=+1]\zeta(X) + \mathbb{1}[h(X)=-1]\xi(X)\right] \leq \eta,\\ &\text{where } \forall x \in \mathcal{X}, \gamma(x)+\zeta(x)+\xi(x) = 1 \text{ and } \zeta(x),\xi(x) \in [0,1], \gamma(x) \in \{0,1\}\Bigg\},\end{aligned}$$

we have that

$$\Phi(\mathcal{H},\eta) \leq \Phi_{\{0,1\}}(\mathcal{H},\eta) \leq 2\Phi(\mathcal{H},\eta/2).$$

Proof The left inequality is clear from the definitions. For the right inequality, let γ^*, ζ^*, ξ^* be the functions at the optimal solution achieving $\Phi(\mathcal{H},\eta/2)$ in Definition 15. For every

$x \in \mathcal{X}$, if $\gamma^*(x) \geq 1/2$, define $\gamma(x) = 1$ and $\zeta(x) = \xi(x) = 0$, and otherwise define $\gamma(x) = 0$, $\zeta(x) = \zeta^*(x)/(\zeta^*(x) + \xi^*(x))$, and $\xi(x) = \xi^*(x)/(\zeta^*(x) + \xi^*(x))$. By design, we have that $\gamma(x) \in \{0,1\}$, $\zeta(x), \xi(x) \in [0,1]$, and $\gamma(x) + \zeta(x) + \xi(x) = 1$ for every $x \in \mathcal{X}$. Since every $x \in \mathcal{X}$ has $\gamma(x) \leq 2\gamma^*(x)$, we have $\mathbb{E}[\gamma(X)] \leq 2\mathbb{E}[\gamma^*(X)] = 2\Phi(\mathcal{H}, \eta/2)$. Furthermore, for every $x \in \mathcal{X}$, we either have $\zeta(x) = 0 \leq 2\zeta^*(x)$ and $\xi(x) = 0 \leq 2\xi^*(x)$, or else $\gamma^*(x) < 1/2$, in which case $\zeta^*(x)+\xi^*(x) = 1-\gamma^*(x) > 1/2$, so that $\zeta(x) = \zeta^*(x)/(\zeta^*(x)+\xi^*(x)) \leq 2\zeta^*(x)$ and $\xi(x) = \xi^*(x)/(\zeta^*(x) + \xi^*(x)) \leq 2\xi^*(x)$. Therefore,

$$
\begin{aligned}
&\sup_{h \in \mathcal{H}} \mathbb{E}\left[\mathbb{1}[h(X) = +1]\zeta(X) + \mathbb{1}[h(X) = -1]\xi(X)\right] \\
&\leq 2 \sup_{h \in \mathcal{H}} \mathbb{E}\left[\mathbb{1}[h(X) = +1]\zeta^*(X) + \mathbb{1}[h(X) = -1]\xi^*(X)\right] \leq \eta.
\end{aligned}
$$

Thus, γ, ζ, ξ are functions in the feasible region of the optimization problem defining $\Phi_{\{0,1\}}(\mathcal{H}, \eta)$, so that $\Phi_{\{0,1\}}(\mathcal{H}, \eta) \leq \mathbb{E}[\gamma(X)] \leq 2\Phi(\mathcal{H}, \eta/2)$. ■

We will establish the claim in Theorem 19 for the following algorithm (which has the data set $\mathcal{L}_m$ as input). For simplicity, this algorithm is stated in a way that makes it $\mathcal{P}$-dependent (which is consistent with the statement of Theorem 19). It may be possible to remove this dependence by replacing the $\mathcal{P}$-dependent quantities with empirical estimates, but we leave this task to future work (e.g., see the work of Koltchinskii, 2006, for discussion of empirical estimation of $U(\mathcal{H}, m, \delta; R)$; Zhang and Chaudhuri, 2014, additionally discuss estimating the minimizing function γ from the definition of Φ, though some refinement to their concentration arguments would be needed for our purposes). For any $k \in \{0, 1, \ldots, \lfloor \log_2(m) \rfloor - 1\}$, define $\delta_k = \frac{\delta}{(\log_2(2m)-k)^2}$, and fix a value $\eta_k \geq 0$ (to be specified in the proof below).

Algorithm 1:
0. $\mathcal{G}_0 \leftarrow \mathbb{C}$
1. For $k = 0, 1, \ldots, \lfloor \log_2(m) \rfloor - 1$
2. Let γ_k be the function γ at the solution defining $\Phi_{\{0,1\}}(\mathcal{G}_k, \eta_k)$
3. $R_k \leftarrow \{x \in \mathcal{X} : \gamma_k(x) = 1\}$
4. $D_k \leftarrow \{(X_i, Y_i) : 2^k + 1 \leq i \leq 2^{k+1}, X_i \in R_k\}$
5. $\mathcal{G}_{k+1} \leftarrow \left\{ h \in \mathcal{G}_k : 2^{-k}|D_k| \left(\mathrm{er}_{D_k}(h) - \min_{g \in \mathcal{G}_k} \mathrm{er}_{D_k}(g) \right) \leq \max\{4\eta_k, U(\mathcal{G}_k, 2^k, \delta_k; R_k)\} \right\}$
6. Return any $\hat{h} \in \mathcal{G}_{\lfloor \log_2(m) \rfloor}$

For simplicity, we suppose the function γ_k in Step 2 actually minimizes $\mathbb{E}[\gamma_k(X)]$ subject to the constraints in the definition of $\Phi_{\{0,1\}}(\mathcal{G}_k, \eta_k)$. However, the proof below would remain valid for any γ_k satisfying these constraints, with $\mathbb{E}[\gamma_k(X)] \leq 2\Phi(\mathcal{G}_k, \eta_k/2)$: for instance, the proof of Lemma 24 reveals this would be satisfied by $\gamma_k(x) = \mathbb{1}[\gamma^*(x) \geq 1/2]$ for the γ^* achieving the minimum value of $\mathbb{E}[\gamma^*(X)]$ in the definition of $\Phi(\mathcal{G}_k, \eta_k/2)$. Indeed, it would even suffice to choose γ_k satisfying the constraints of $\Phi_{\{0,1\}}(\mathcal{G}_k, \eta_k)$ with $\mathbb{E}[\gamma_k(X)] \leq c'\Phi(\mathcal{G}_k, \eta_k/2)$, for any finite numerical constant c', as this would only affect the numerical constant factors in Theorem 19.

We are now ready for the proof of Theorem 19.

Proof of Theorem 19 The proof is similar to those given above (e.g., that of Theorem 16), except that the stronger form of Lemma 23 (compared to Lemma 2) affords us a simplification that avoids the step in which we lower-bound the sample size under the conditional distribution given $\Gamma_i = 1$.

Fix any $a \geq 1$ and $\alpha \in (0,1]$, and fix $c = 128$. We establish the claim for Algorithm 1, described above. Define $\eta_0 = 2/c$ and $\tilde{U}_0 = 1$, and for each $k \in \{1, \ldots, \lfloor \log_2(m) \rfloor\}$, inductively define

$$\tilde{U}_k = \min\left\{1, 2\eta_{k-1} + \max\left\{8\eta_{k-1}, 2U(\mathcal{G}_{k-1}, 2^{k-1}, \delta_{k-1}; R_{k-1})\right\}\right\},$$

$$r_k = ac_1\left(a2^{1-k}\left(d\mathrm{Log}\left(\hat{\varphi}_{a,\alpha}\left(a\left(ad2^{1-k}\right)^{\frac{\alpha}{2-\alpha}}\right)\right) + \mathrm{Log}\left(\frac{1}{\delta_{k-1}}\right)\right)\right)^{\frac{\alpha}{2-\alpha}},$$

$$\eta_k = \frac{2}{c}\left(\frac{r_k}{a}\right)^{1/\alpha},$$

where $c_1 = (32c_0)^{\frac{2\alpha}{2-\alpha}}$. We proceed by induction on k in the algorithm. Suppose that, for some $k \in \{0, 1, \ldots, \lfloor \log_2(m) \rfloor - 1\}$, there is an event E_k of probability at least $1 - \sum_{k'=0}^{k-1} \delta_{k'}$ (or probability 1 if $k = 0$), on which $h^\star \in \mathcal{G}_k$, and for some universal constant $c_1 \in (1, \infty)$, every $k' \in \{0, \ldots, k\}$ has

$$\tilde{U}_{k'} \leq (c/2)\eta_{k'},$$

and

$$\mathcal{G}_{k'} \subseteq \left\{h \in \mathbb{C} : \mathrm{er}(h) - \mathrm{er}(h^\star) \leq \tilde{U}_{k'}\right\}.$$

In particular, these conditions are trivially satisfied for $k = 0$, so this may serve as a base case for this inductive argument. Next we must extend these conditions to $k+1$.

For each $h \in \mathcal{G}_k$, define $h_{R_k}(x) = h(x)\mathbb{1}[x \in R_k] + h^\star(x)\mathbb{1}[x \notin R_k]$, and denote $\mathcal{H}_k = \{h_{R_k} : h \in \mathcal{G}_k\}$. Noting that $R_k \supseteq \mathrm{DIS}(\mathcal{H}_k)$, and that this implies $U\big(\mathcal{H}_k, 2^k, \delta_k; R_k\big) \geq U\big(\mathcal{H}_k, 2^k, \delta_k; \mathrm{DIS}(\mathcal{H}_k)\big)$, Lemma 23 (applied under the conditional distribution given $\mathcal{G}_k$) and the law of total probability imply that there exists an event E'_{k+1} of probability at least $1 - \delta_k$, on which, $\forall h_{R_k} \in \mathcal{H}_k$, denoting $\tilde{\mathcal{L}}_k = \{(X_i, Y_i) : 2^k + 1 \leq i \leq 2^{k+1}\}$ (which is distributionally equivalent to $\mathcal{L}_{2^k}$ but independent of $\mathcal{G}_k$),

$$\mathrm{er}(h_{R_k}) - \inf_{g_{R_k} \in \mathcal{H}_k} \mathrm{er}(g_{R_k}) \leq \max\left\{2\left(\mathrm{er}_{\tilde{\mathcal{L}}_k}(h_{R_k}) - \min_{g_{R_k} \in \mathcal{H}_k} \mathrm{er}_{\tilde{\mathcal{L}}_k}(g_{R_k})\right), U\left(\mathcal{H}_k, 2^k, \delta_k; R_k\right)\right\},$$

$$\mathrm{er}_{\tilde{\mathcal{L}}_k}(h_{R_k}) - \min_{g_{R_k} \in \mathcal{H}_k} \mathrm{er}_{\tilde{\mathcal{L}}_k}(g_{R_k}) \leq \max\left\{2\left(\mathrm{er}(h_{R_k}) - \inf_{g_{R_k} \in \mathcal{H}_k} \mathrm{er}(g_{R_k})\right), U(\mathcal{H}_k, 2^k, \delta_k; R_k)\right\}.$$

First we note that, since every h_{R_k} and g_{R_k} in $\mathcal{H}_k$ agree on the labels of all samples in $\tilde{\mathcal{L}}_k \setminus D_k$, and they each agree with their respective classifiers h and g in $\mathcal{G}_k$ on D_k, we have that

$$\mathrm{er}_{\tilde{\mathcal{L}}_k}(h_{R_k}) - \min_{g_{R_k} \in \mathcal{H}_k} \mathrm{er}_{\tilde{\mathcal{L}}_k}(g_{R_k}) = 2^{-k}|D_k|\left(\mathrm{er}_{D_k}(h) - \min_{g \in \mathcal{G}_k} \mathrm{er}_{D_k}(g)\right).$$

Next, let ζ_k and ξ_k denote the functions ζ and ξ from the definition of $\Phi_{\{0,1\}}(\mathcal{G}_k, \eta_k)$ at the solution with γ equal γ_k. Note that ζ_k and ξ_k are themselves random, but are competely determined by $\mathcal{G}_k$. The definition of R_k guarantees that for every $h, g \in \mathcal{G}_k$, for $X \sim \mathcal{P}$ (independent from $\mathcal{L}_m$)

$$\begin{aligned}
&\mathcal{P}(x \notin R_k : h(x) \neq g(x)) = \mathbb{E}\left[\mathbb{1}[h(X) \neq g(X)](\zeta_k(X) + \xi_k(X)) | \mathcal{G}_k\right] \\
&= \mathbb{E}\left[(\mathbb{1}[h(X) = +1]\mathbb{1}[g(X) = -1] + \mathbb{1}[h(X) = -1]\mathbb{1}[g(X) = +1])\left(\zeta_k(X) + \xi_k(X)\right) | \mathcal{G}_k\right] \\
&\leq \mathbb{E}\left[\mathbb{1}[h(X) = +1]\zeta_k(X) + \mathbb{1}[h(X) = -1]\xi_k(X) | \mathcal{G}_k\right] \\
&\qquad + \mathbb{E}\left[\mathbb{1}[g(X) = +1]\zeta_k(X) + \mathbb{1}[g(X) = -1]\xi_k(X) | \mathcal{G}_k\right] \leq 2\eta_k.
\end{aligned}$$

Therefore,

$$\operatorname{er}(h_{R_k}) - \operatorname{er}(g_{R_k}) \leq \operatorname{er}(h) - \operatorname{er}(g) + \mathcal{P}(x \notin R_k : h(x) \neq g(x)) \leq \operatorname{er}(h) - \operatorname{er}(g) + 2\eta_k,$$

and similarly

$$\operatorname{er}(h_{R_k}) - \operatorname{er}(g_{R_k}) \geq \operatorname{er}(h) - \operatorname{er}(g) - \mathcal{P}(x \notin R_k : h(x) \neq g(x)) \geq \operatorname{er}(h) - \operatorname{er}(g) - 2\eta_k.$$

In particular, noting that $\operatorname{er}(h_{R_k}) - \inf_{g_{R_k} \in \mathcal{H}_k} \operatorname{er}(g_{R_k}) = \sup_{g \in \mathcal{G}_k} \operatorname{er}(h_{R_k}) - \operatorname{er}(g_{R_k})$ and $\sup_{g \in \mathcal{G}_k} \operatorname{er}(h) - \operatorname{er}(g) = \operatorname{er}(h) - \inf_{g \in \mathcal{G}_k} \operatorname{er}(g)$, this implies

$$\operatorname{er}(h) - \inf_{g \in \mathcal{G}_k} \operatorname{er}(g) - 2\eta_k \leq \operatorname{er}(h_{R_k}) - \inf_{g_{R_k} \in \mathcal{H}_k} \operatorname{er}(g_{R_k}) \leq \operatorname{er}(h) - \inf_{g \in \mathcal{G}_k} \operatorname{er}(g) + 2\eta_k.$$

We also note that $\operatorname{vc}(\mathcal{H}_k) \leq \operatorname{vc}(\mathcal{G}_k)$ and $\operatorname{diam}_{\mathcal{P}}(\mathcal{H}_k) \leq \operatorname{diam}_{\mathcal{P}}(\mathcal{G}_k)$, which together imply $U(\mathcal{H}_k, 2^k, \delta_k; R_k) \leq U(\mathcal{G}_k, 2^k, \delta_k; R_k)$. Altogether, we have that on E'_{k+1}, $\forall h \in \mathcal{G}_k$,

$$\operatorname{er}(h) - \inf_{g \in \mathcal{G}_k} \operatorname{er}(g) \leq 2\eta_k + \max\left\{2^{1-k}|D_k|\left(\operatorname{er}_{D_k}(h) - \min_{g \in \mathcal{G}_k} er_{D_k}(g)\right), U(\mathcal{G}_k, 2^k, \delta_k; R_k)\right\},$$

$$2^{-k}|D_k|\left(\operatorname{er}_{D_k}(h) - \min_{g \in \mathcal{G}_k} \operatorname{er}_{D_k}(g)\right) \leq \max\left\{2\left(\operatorname{er}(h) - \inf_{g \in \mathcal{G}_k} \operatorname{er}(g) + 2\eta_k\right), U(\mathcal{G}_k, 2^k, \delta_k; R_k)\right\}.$$

In particular, defining $E_{k+1} = E'_{k+1} \cap E_k$, we have that on E_{k+1}, $h^\star \in \mathcal{G}_k$, and

$$2^{-k}|D_k|\left(\operatorname{er}_{D_k}(h^\star) - \min_{g \in \mathcal{G}_k} \operatorname{er}_{D_k}(g)\right) \leq \max\left\{4\eta_k, U(\mathcal{G}_k, 2^k, \delta_k; R_k)\right\},$$

so that $h^\star \in \mathcal{G}_{k+1}$ as well. Furthermore, combined with the definition of $\mathcal{G}_{k+1}$, this further implies that on E_{k+1},

$$\begin{aligned}
\mathcal{G}_{k+1} &\subseteq \left\{h \in \mathbb{C} : \operatorname{er}(h) - \operatorname{er}(h^\star) \leq 2\eta_k + \max\left\{8\eta_k, 2U\left(\mathcal{G}_k, 2^k, \delta_k; R_k\right)\right\}\right\} \\
&= \left\{h \in \mathbb{C} : \operatorname{er}(h) - \operatorname{er}(h^\star) \leq \tilde{U}_{k+1}\right\}.
\end{aligned}$$

It remains only to establish the bound on $\tilde{U}_{k+1}$. For this, we first note that, combining the inductive hypothesis with the (a, α)-Bernstein class condition, on E_{k+1} we have

$$\mathcal{G}_k \subseteq \mathrm{B}\left(h^\star, a\tilde{U}_k^\alpha\right) \subseteq \mathrm{B}\left(h^\star, r_k\right).$$

Combining this with Lemma 24 and monotonicity of $\Phi(\cdot, \eta_k/2)$, we have that

$$\mathcal{P}(R_k) \leq 2\Phi\left(\mathrm{B}\left(h^\star, r_k\right), \eta_k/2\right) = 2\Phi\left(\mathrm{B}\left(h^\star, r_k\right), (r_k/a)^{1/\alpha}/c\right) \leq 2\hat{\varphi}_{a,\alpha}(r_k) r_k.$$

The above also implies that $\mathrm{diam}_{\mathcal{P}}(\mathcal{G}_k) \leq 2r_k$ on E_{k+1}. Together with the fact that $\mathrm{vc}(\mathcal{G}_k) \leq d$, we have that on E_{k+1},

$$U(\mathcal{G}_k, 2^k, \delta_k; R_k) \leq c_0 \sqrt{2r_k 2^{-k}\left(d\mathrm{Log}\left(\hat{\varphi}_{a,\alpha}(r_k)\right) + \mathrm{Log}\left(\frac{1}{\delta_k}\right)\right)} + c_0 2^{-k}\left(d\mathrm{Log}\left(\hat{\varphi}_{a,\alpha}(r_k)\right) + \mathrm{Log}\left(\frac{1}{\delta_k}\right)\right). \quad (34)$$

Furthermore, monotonicity of $\hat{\varphi}_{a,\alpha}(\cdot)$ implies $\hat{\varphi}_{a,\alpha}(r_k) \leq \hat{\varphi}_{a,\alpha}\left(a(ad2^{-k})^{\frac{\alpha}{2-\alpha}}\right)$. Plugging the definition of r_k into (34) along with this relaxation of $\hat{\varphi}_{a,\alpha}(r_k)$ and simplifying, the minimum of 1 and the right hand side of (34) is at most

$$8c_0\sqrt{c_1}\left(a2^{-k}\left(d\mathrm{Log}\left(\hat{\varphi}_{a,\alpha}\left(a\left(ad2^{-k}\right)^{\frac{\alpha}{2-\alpha}}\right)\right) + \mathrm{Log}\left(\frac{1}{\delta_k}\right)\right)\right)^{\frac{1}{2-\alpha}} = 8c_0\sqrt{c_1}\left(\frac{r_{k+1}}{c_1 a}\right)^{1/\alpha} = \frac{4c_0 c}{c_1^{\frac{2-\alpha}{2\alpha}}}\eta_{k+1} = \frac{c}{8}\eta_{k+1}.$$

We may also observe that

$$\eta_k \leq 4^{\frac{1}{2-\alpha}}\eta_{k+1} \leq 4\eta_{k+1}.$$

Combining the above with the definition of $\tilde{U}_{k+1}$, we have that on E_{k+1},

$$\tilde{U}_{k+1} \leq 8\eta_{k+1} + \max\left\{32\eta_{k+1}, \frac{c}{4}\eta_{k+1}\right\} = 40\eta_{k+1} \leq 64\eta_{k+1} = \frac{c}{2}\eta_{k+1}.$$

Finally, noting that the union bound implies E_{k+1} has probability at least $1 - \sum_{k'=0}^{k}\delta_{k'}$ completes the inductive step.

By the principle of induction, we have thus established that, on an event $E_{\lfloor \log_2(m)\rfloor}$ of probability at least $1 - \sum_{k=0}^{\lfloor \log_2(m)\rfloor - 1}\delta_k > 1 - \delta\sum_{i=2}^{\infty}\frac{1}{i^2} > 1 - \delta$,

$$h^\star \in \mathcal{G}_{\lfloor \log_2(m)\rfloor} \subseteq \left\{h \in \mathbb{C} : \mathrm{er}(h) - \mathrm{er}(h^\star) \leq \frac{c}{2}\eta_{\lfloor \log_2(m)\rfloor}\right\}.$$

In particular, this implies that $\hat{h}$ exists in Step 6, and satisfies $\mathrm{er}(\hat{h}) - \inf_{g\in\mathbb{C}}\mathrm{er}(g) = \mathrm{er}(\hat{h}) - \mathrm{er}(h^\star) \leq \frac{c}{2}\eta_{\lfloor \log_2(m)\rfloor}$. Noting that

$$\frac{c}{2}\eta_{\lfloor \log_2(m)\rfloor} \leq c_1^{1/\alpha}\left(\frac{4a\left(d\mathrm{Log}\left(\hat{\varphi}_{a,\alpha}\left(a\left(\frac{ad}{m}\right)^{\frac{\alpha}{2-\alpha}}\right)\right) + \mathrm{Log}\left(\frac{4}{\delta}\right)\right)}{m}\right)^{\frac{1}{2-\alpha}} \leq 6(32c_0)^2\left(\frac{a\left(d\mathrm{Log}\left(\hat{\varphi}_{a,\alpha}\left(a\left(\frac{ad}{m}\right)^{\frac{\alpha}{2-\alpha}}\right)\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)}{m}\right)^{\frac{1}{2-\alpha}}$$

completes the proof. ■

References

N. Ailon, R. Begleiter, and E. Ezra. Active learning using smooth relative regret approximations with applications. *Journal of Machine Learning Research*, 15(3):885–920, 2014. 6

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. E.1

P. Auer and R. Ortner. A new PAC bound for intersection-closed concept classes. In *Proceedings of the* 17^{th} *Conference on Learning Theory*, 2004. 1.2, 3

P. Auer and R. Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2-3):151–163, 2007. 1, 1.2, 1.2, 3, 3

M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the* 26^{th} *Conference on Learning Theory*, 2013. 1.2, 5.2, 5.2, 5.2, 6.2

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the* 23^{rd} *International Conference on Machine Learning*, 2006. 4

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the* 20^{th} *Conference on Learning Theory*, 2007. 1.2, 5.2, 5.2

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009. 4

P. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. 6

P. L. Bartlett and S. Mendelson. Discussion: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2657–2663, 2006. 6

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4): 929–965, 1989. 1.2, 2, 5, 5.1

A. Bowers and N. J. Kalton. *An Introductory Course in Functional Analysis*. Springer, 2014. 9

N. H. Bshouty, Y. Li, and P. M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *Journal of Computer and System Sciences*, 75(6): 323–335, 2009. 5.2

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994. 1.2, 4

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14 (3):326–334, 1965. 5.2

M. Darnstädt. The optimal PAC bound for intersection-closed concept classes. *Information Processing Letters*, 115(4):458–461, 2015. 1.2, 3

S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the* 18^{th} *Conference on Learning Theory*, 2005. 5.2

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems* 20, 2007. 4

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989. 1.2, 1.2, 5.1, E.1

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5):1605–1641, 2010. 4, 4, 4

R. El-Yaniv and Y. Wiener. Agnostic selective classification. In *Advances in Neural Information Processing Systems* 24, 2011. 6

R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2):255–279, 2012. 1.2, 4, 4, 5.2

S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995. 1.2, 2, 2

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006. 1.2, 1.2, 5, 5.2, 6, 6, E.2

A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *Proceedings of the* 44^{th} *Annual IEEE Symposium on Foundations of Computer Science*, 2003. 5.2

S. Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of the* 20^{th} *Conference on Learning Theory*, 2007a. 4

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the* 24^{th} *International Conference on Machine Learning*, 2007b. 4, 4

S. Hanneke. *Theoretical Foundations of Active Learning.* PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009. 1, 1.2, 1.2, 1.2, 2, 4, 4, 5, 4, 6, D

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011. 1.2, 4, 6, E.1

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13(5):1469–1587, 2012. 4, 6

S. Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2–3):131–309, 2014. 1.2, 4, 6, E.1

S. Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016. 1, 1.2, 1.2, 8

S. Hanneke and L. Yang. Surrogate losses in passive and active learning. *arXiv:1207.3772*, 2012. 1.2, 4, 6, E.2

S. Hanneke and L. Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(12):3487–3602, 2015. 1.2, 1.2, 4, 4, 4, 5, 5, 5.1, 5.2, 6, 6, 6, 6.1, 6.1, 6.1, D.1, D.2, E.1, 10, E.1, E.1

D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory A*, 69(2):217–232, 1995. 5.2, D.2, D.2

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994. 1.2, 1.2, 3, 5.1

D. Helmbold, R. Sloan, and M. Warmuth. Learning nested differences of intersection-closed concept classes. *Machine Learning*, 5(2):165–196, 1990. 3

R. Herbrich. *Learning Kernel Classifiers*. The MIT Press. Cambridge, MA, 2002. 2

A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959. D.2

A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961. D.2

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. 6, 7, 6.2, E.2

V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11(9):2457–2485, 2010. 4

C. Kuhlmann. On teaching and learning intersection-closed concept classes. In *Proceedings of the* 12^{th} *Conference on Learning Theory*, 1999. 3

L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973. 5.2

Y. Li, P. M. Long, and A. Srinivasan. The one-inclusion graph algorithm is near-optimal for the prediction model of learning. *IEEE Transactions on Information Theory*, 47(3): 1257–1261, 2001. 1.2

N. Littlestone and M. Warmuth. Relating data compression and learnability. *Unpublished manuscript*, 1986. 1.2, 2, 2

P. M. Long. An upper bound on the sample complexity of PAC learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–234, 2003. 1.2

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999. 6

P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006. 1.2, 6, 6.1, 6.1, E.1

B. K. Natarajan. On learning Boolean functions. In *Proceedings of the* 19^{th} *Annual ACM Symposium on Theory of Computing*, 1987. 1.2, 3

M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems* 24, 2011. 1.2, 6.1, 6.1, E.1, 10

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13(1): 145–147, 1972. D.2

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. 6

A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009. E.1

A. van der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. 1.1, E.2

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer, 1996. 1.1

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971. 1.1, D.2

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition.* Nauka, Moscow, 1974. 1.2, 2, 5

M. Vidyasagar. *Learning and Generalization with Applications to Neural Networks.* Springer-Verlag, 2^{nd} edition, 2003. 3, D.2, E.1

M. Warmuth. The optimal PAC algorithm. In *Proceedings of the* 17^{th} *Conference on Learning Theory*, 2004. 1.2

Y. Wiener, S. Hanneke, and R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *Journal of Machine Learning Research*, 16(4):713–745, 2015. 1.2, 2, 4, 4, 2, 4, 4, 4, 4, 4, 5

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. 5.2

C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems* 27, 2014. 1.2, 5.2, 5.2, 5.2, 6, 6.2, 6.2, D, E.2