Minimax Analysis of Active Learning

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Princeton, NJ 08542
Liu Yang LIUY@CS.CMU.EDU
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

Abstract

This work establishes distribution-free upper and lower bounds on the minimax label complexity of active learning with general hypothesis classes, under various noise models. The results reveal a number of surprising facts. In particular, under the noise model of Tsybakov (2004), the minimax label complexity of active learning with a VC class is always asymptotically smaller than that of passive learning, and is typically significantly smaller than the best previously-published upper bounds in the active learning literature. In high-noise regimes, it turns out that all active learning problems of a given VC dimension have roughly the same minimax label complexity, which contrasts with well-known results for bounded noise. In low-noise regimes, we find that the label complexity is well-characterized by a simple combinatorial complexity measure we call the *star number*. Interestingly, we find that almost all of the complexity measures previously explored in the active learning literature have worst-case values exactly equal to the star number. We also propose new active learning strategies that nearly achieve these minimax label complexities.

Keywords: Active Learning, Selective Sampling, Sequential Design, Adaptive Sampling, Statistical Learning Theory, Margin Condition, Tsybakov Noise, Sample Complexity, Minimax Analysis

1. Introduction

In many machine learning applications, in the process of training a high-accuracy classifier, the primary bottleneck in time and effort is often the annotation of the large quantities of data required for supervised learning. Active learning is a protocol designed to reduce this cost by allowing the learning algorithm to sequentially identify highly-informative data points to be annotated. In the specific model we study below, called *pool-based* active learning, the learning algorithm is initially given access to a large pool of unlabeled data points, which are considered inexpensive and abundant. It is then able to select any unlabeled data point from the pool and request to observe its label. Given the label of this data point, the algorithm can then select another unlabeled data point to be labeled, and so on. This interactive process continues until at most some prespecified number of rounds is reached, at which time the algorithm must halt and produce a classifier. This contrasts with *passive learning*, where the learning algorithm would be given access to a number of *random* labeled data points. The hope is that, by sequentially selecting the data points to be labeled, the active learning algorithm can direct the annotation effort toward only the highly-informative data points, given the information already gathered by the previously-labeled data points, and thereby reduce the total number of data points required to produce a classifier capable of predicting the labels of new instances with at most a desired error rate. This model of active learning has been successfully applied to a variety of learning problems, often with significant reductions in the number of label observations required to obtain a given error rate for the resulting classifier (see Settles, 2012, for a survey of several such applications).

This article studies the theoretical capabilities of active learning, regarding the number of label requests sufficient to learn a classifier to a desired error rate, known as the *label complexity*. There is now a substantial literature on this subject (see Hanneke, 2014, for a survey of known results), but on the important question of *optimal* performance in the general setting, the gaps present in the literature are quite substantial in some cases. In this work, we address this question by carefully studying the *minimax* performance. Specifically, we are interested in the *minimax label complexity*, defined as the smallest (over the choice of active learning algorithm) worst-case number of label requests sufficient for the active learning algorithm to produce a classifier of a specified error rate, in the context of various noise models (e.g., Tsybakov noise, bounded noise, agnostic noise, etc.). We derive upper and lower bounds on the minimax label complexity for several noise models, which reveal a variety of interesting and (in some cases) surprising observations. Furthermore, in establishing the upper bounds, we propose a novel active learning strategy, which often achieves significantly smaller label complexities than the active learning methods studied in the prior literature.

1.1 The Prior Literature on the Theory of Active Learning

Before getting into the technical details, we first review some background information about the prior literature on the theory of active learning. This will also allow us to introduce the key contributions of the present work.

The literature on the theory of active learning began with studies of the *realizable case*, a setting in which the labels are assumed to be consistent with some classifier in a known hypothesis class, and have no noise (Cohn, Atlas, and Ladner, 1994; Freund, Seung, Shamir, and Tishby, 1997; Dasgupta, 2004, 2005). In this simple setting, Dasgupta (2005) supplied the first general analysis of the label complexity of active learning, applicable to arbitrary hypothesis classes. However, Dasgupta (2005) found that there are a range of minimax label complexities, depending on the structure of the hypothesis class, so that even among hypothesis classes of roughly the same minimax sample complexities for passive learning, there can be widely varying minimax label complexities for active learning. In particular, he found that some hypothesis classes (e.g., interval classifiers) have minimax label complexity essentially no better than that of passive learning, while others have a minimax label complexity exponentially smaller than that of passive learning (e.g., threshold classifiers). Furthermore, most nontrivial hypothesis classes of interest in learning theory seem to fall into the former category, with minimax label complexities essentially no better than passive learning. Fortunately, Dasgupta (2005) also found that in some of these hard cases, it is still possible to show improvements over passive learning under restrictions on the data distribution.

Stemming from these observations, much of the literature on active learning in the realizable case has focused on describing various special conditions under which the label complexity of active learning is significantly better than that of passive learning: for instance, by placing restrictions on the marginal distribution of the unlabeled data (e.g., Dasgupta, Kalai, and Monteleoni, 2005; Balcan, Broder, and Zhang, 2007; El-Yaniv and Wiener, 2012; Balcan and Long, 2013; Hanneke, 2014), or abandoning the minimax approach by expressing the label complexity with an explicit dependence on the optimal classifier (e.g., Dasgupta, 2005; Balcan, Hanneke, and Vaughan, 2010; Hanneke, 2009b, 2012). In the general case, such results have been abstracted into various distribution-dependent (or sometimes data-dependent) *complexity measures*, such as the *splitting index* (Dasgupta, 2005), the *disagreement coefficient* (Hanneke, 2007b, 2009b), the *extended teaching dimension growth function* (Hanneke, 2007a), and the related *version space compression set size*

(El-Yaniv and Wiener, 2010, 2012; Wiener, Hanneke, and El-Yaniv, 2014). For each of these, there are general upper bounds (and in some cases, minimax lower bounds) on the label complexities achievable by active learning methods in the realizable case, expressed in terms of the complexity measure. By expressing bounds on the label complexity in terms of these quantities, the analysis of the label complexities achievable by active learning methods in the realizable case has been effectively reduced to the problem of bounding one of these complexity measures. In particular, these complexity measures are capable of exhibiting a range of behaviors, corresponding to the range of label complexities achievable by active learning. For certain values of the complexity measures, the resulting bounds reveal significant improvements over the minimax sample complexity of passive learning, while for other values, the resulting bounds are essentially no better than the minimax sample complexity of passive learning.

Moving beyond these initial studies of the realizable case, the more-recent literature has developed active learning algorithms that are provably robust to label noise. This advance was initiated by the seminal work of Balcan, Beygelzimer, and Langford (2006, 2009) on the A^2 (Agnostic Active) algorithm, and continued by a number of subsequent works (e.g., Dasgupta, Hsu, and Monteleoni, 2007; Balcan, Broder, and Zhang, 2007; Castro and Nowak, 2006, 2008; Hanneke, 2007a, 2009a,b, 2011, 2012; Minsker, 2012; Koltchinskii, 2010; Beygelzimer, Dasgupta, and Langford, 2009; Beygelzimer, Hsu, Langford, and Zhang, 2010; Hsu, 2010; Ailon, Begleiter, and Ezra, 2012; Hanneke and Yang, 2012). When moving into the analysis of label complexity in noisy settings, the literature continues to follow the same intuition from the realizable case: that is, that there should be some active learning problems that are inherently hard, sometimes no better than passive learning, while others are significantly easier, with significant savings compared to passive learning. As such, the general label complexity bounds proven in noisy settings have tended to follow similar patterns to those found in the realizable case. In some scenarios, the bounds reflect interesting savings compared to passive learning, while in other scenarios the bounds do not reflect any improvements at all. However, unlike the realizable case, these upper bounds on the label complexities of the various proposed methods for noisy settings lacked complementary minimax lower bounds showing that they were accurately describing the fundamental capabilities of active learning in these settings. For instance, in the setting of Tsybakov noise, there are essentially only two types of general lower bounds on the minimax label complexity in the prior literature: (1) lower bounds that hold for all nontrivial hypothesis classes of a given VC dimension, which therefore reflect a kind of best-case scenario (Hanneke, 2011, 2014), and (2) lower bounds inherited from the realizable case (which is a special case of Tsybakov noise). In particular, both of these lower bounds are always smaller than the minimax sample complexity of passive learning under Tsybakov noise. Thus, although the upper bounds on the label complexity of active learning in the literature are sometimes no better than the minimax sample complexity of passive learning, the existing lower bounds are unable to confirm that active learning truly cannot outperform passive learning in these scenarios. This gap in our understanding of active learning with noise has persisted for a number of years now, without really receiving a good explanation for why the gap exists and how it might be closed.

In the present work, we show that there is a very good reason for why better lower bounds have not been discovered in general for the noisy case. For certain ranges of the noise parameters (corresponding to the high-noise regime), these simple lower bounds are actually *tight* (up to certain constant and logarithmic factors): that is, the upper bounds can actually be reduced to nearly match these basic lower bounds. Proving this surprising fact requires the introduction of a new type of active learning strategy, which selects its queries based on both the structure of the hypothesis class and the estimated variances of the labels. In particular, in these high-noise regimes, we find that *all* hypothesis classes of the same VC dimension have essentially the same minimax label complexities (up to logarithmic factors), in stark contrast to the well-known differentiation of hypothesis classes observed in the realizable case by Dasgupta (2005).

For the remaining range of the noise parameters (the low-noise regime), we argue that the label complexity takes a value sometimes larger than this basic lower bound, yet still typically smaller than the known upper bounds. In this case, we further argue that the minimax label complexity is well-characterized by a simple combinatorial complexity measure, which we call the *star number*. In particular, these results reveal that for nonextremal parameter values, the minimax label complexity of active learning under Tsybakov noise with *any* VC class is *always* smaller than that of passive learning, a fact not implied by any results in the prior literature.

We further find that the star number can be used to characterize the minimax label complexities for a variety of other noise models. Interestingly, we also show that almost all of the distribution-dependent or data-dependent complexity measures from the prior literature on the label complexity of active learning are exactly *equal* to the star number when maximized over the choice of distribution or data set (including all of those mentioned above). Thus, the star number represents a unifying core concept within these disparate styles of analysis.

1.2 Our Contributions

Below, we summarize a few of the main contributions and interesting implications of this work.

- We develop a general noise-robust active learning strategy, which unlike previously-proposed general methods, selects its queries based on both the structure of the hypothesis class *and* the estimated variances of the labels.

- We obtain the first near-matching general distribution-free upper and lower bounds on the minimax label complexity of active learning, under a variety of noise models.

- In many cases, the upper bounds significantly improve over the best upper bounds implied by the prior literature.

- The upper bounds for Tsybakov noise *always* reflect improvements over the minimax sample complexity of passive learning (for non-extremal noise parameter values), a feat not previously known to be possible.

- In high-noise regimes of Tsybakov noise, our results imply that all hypothesis classes of a given VC dimension have roughly the same minimax label complexity (up to logarithmic factors), in contrast to well-known results for bounded noise. This fact is not implied by any results in the prior literature.

- We express our upper and lower bounds on the label complexity in terms of a simple combinatorial complexity measure, which we refer to as the *star number*.

- We show that for any hypothesis class, almost every complexity measure proposed to date in the active learning literature has worst-case value exactly *equal* to the star number, thus unifying the disparate styles of analysis in the active learning literature. We also prove that the doubling dimension is bounded if and only if the star number is finite.

- For most of the noise models studied here, we exhibit examples of hypothesis classes spanning the gaps between the upper and lower bounds, thus demonstrating that the gaps cannot generally be reduced (aside from logarithmic factors) without introducing additional complexity measures.

- We prove a separation result for Tsybakov noise vs the Bernstein class condition, establishing that the respective minimax label complexities can be significantly different. This contrasts with passive learning, where they are known to be equivalent up to a logarithmic factor.

The algorithmic techniques underlying the proofs of the most-interesting of our upper bounds involve a combination of the disagreement-based strategy of Cohn, Atlas, and Ladner (1994) (and the analysis thereof by Hanneke, 2011, and Wiener, Hanneke, and El-Yaniv, 2014), along with a repeated-querying technique of Kääriäinen (2006), modified to account for variations in label variances so that the algorithm does not waste too many queries determining the optimal classification of highly-noisy points; this modification represents the main algorithmic innovation in this work. In a supporting role, we also rely on auxiliary lemmas on the construction of ε-nets and ε-covers based on random samples, and the use of these to effectively discretize the instance space. The mathematical techniques underlying the proofs of the lower bounds are largely taken directly from the literature. Most of the lower bounds are established by a combination of a technique originating with Kääriäinen (2006) and refined by Beygelzimer, Dasgupta, and Langford (2009) and Hanneke (2011, 2014), and a technique of Raginsky and Rakhlin (2011) for incorporating a complexity measure into the lower bounds.

We note that, while the present work focuses on the distribution-free setting, in which the marginal distribution over the instance space is unrestricted, our results reveal that low-noise settings can still benefit from distribution-dependent analysis, as expected given the aforementioned observations by Dasgupta (2005) for the realizable case. For instance, under Tsybakov noise, it is often possible to obtain stronger upper bounds in low-noise regimes under assumptions restricting the distribution of the unlabeled data (see e.g., Balcan, Broder, and Zhang, 2007). We leave for future work the important problem of characterizing the minimax label complexity of active learning in the general case for an arbitrary fixed marginal distribution over the instance space.

1.3 Outline

The rest of this article is organized as follows. Section 2 introduces the formal setting and basic notation used throughout, followed in Section 3 with the introduction of the noise models studied in this work. Section 4 defines a combinatorial complexity measure – the star number – in terms of which we will express the label complexity bounds below. Section 5 provides statements of the main results of this work: upper and lower bounds on the minimax label complexities of active learning under each of the noise models defined in Section 3. That section also includes a discussion of the results, and a brief sketch of the arguments underlying the most-interesting among them. Section 6 compares the results from Section 5 to the known results on the minimax sample complexity of passive learning, revealing which scenarios yield improvements of active over passive. Next, in Section 7, we go through the various results on the label complexity of active learning from the literature, along with their corresponding complexity measures (most of which are distribution-dependent or data-dependent). We argue that all of these complexity measures are exactly equal to the star number when maximized over the choice of distribution or data set. This section also relates the star number to the well-known concept of *doubling dimension*, in particular showing that the doubling dimension is bounded if and only if the star number is finite.

We note that the article is written with the intention that it be read in-order; for instance, while Appendix B contains proofs of the results in Section 5, those proofs refer to quantities and results introduced in Sections 6 and 7 (which follow Section 5, but precede Appendix B).

2. Definitions

The rest of this paper makes use of the following formal definitions. There is a space $\mathcal{X}$, called the *instance space*. We suppose $\mathcal{X}$ is equipped with a σ-algebra $\mathcal{B}_{\mathcal{X}}$, and for simplicity we will assume $\{\{x\} : x \in \mathcal{X}\} \subseteq \mathcal{B}_{\mathcal{X}}$. There is also a set $\mathcal{Y} = \{-1, +1\}$, known as the *label space*. Any measurable function $h : \mathcal{X} \to \mathcal{Y}$ is called a *classifier*. There is an arbitrary set $\mathbb{C}$ of classifiers, known as the *hypothesis class*. To focus on nontrivial cases, we suppose $|\mathbb{C}| \geq 3$ throughout.

For any probability measure P over $\mathcal{X} \times \mathcal{Y}$ and any $x \in \mathcal{X}$, define $\eta(x; P) = \mathbb{P}(Y = +1|X = x)$ for $(X, Y) \sim P$, and let $f_P^\star(x) = \mathrm{sign}(2\eta(x; P)-1)$ denote the *Bayes optimal classifier*,[1] where $\mathrm{sign}(t) = +1$ if $t \geq 0$, and $\mathrm{sign}(t) = -1$ if $t < 0$. Define the *error rate* of a classifier h with respect to P as $\mathrm{er}_P(h) = P((x, y) : h(x) \neq y)$.

In the learning problem, there is a *target distribution* $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$, and a *data sequence* $(X_1, Y_1), (X_2, Y_2), \ldots$, which are independent $\mathcal{P}_{XY}$-distributed random variables. However, in the active learning protocol, the Y_i values are initially "hidden" until individually requested by the algorithm (see below). We refer to the sequence $X_1, X_2, \ldots$ as the *unlabeled data sequence*.[2] We will sometimes denote by $\mathcal{P}$ the marginal distribution of $\mathcal{P}_{XY}$ over $\mathcal{X}$: that is, $\mathcal{P}(\cdot) = \mathcal{P}_{XY}(\cdot \times \mathcal{Y})$.

In the *pool-based active learning* protocol,[3] we define an *active learning algorithm* $\mathcal{A}$ as an algorithm taking as input a budget $n \in \mathbb{N} \cup \{0\}$, and proceeding as follows. The algorithm initially has access to the unlabeled data sequence $X_1, X_2, \ldots$. If $n > 0$, the algorithm may then select an index $i_1 \in \mathbb{N}$ and request to observe the label Y_{i_1}. The algorithm may then observe the value of Y_{i_1}, and if $n \geq 2$, then based on both the unlabeled sequence and this new observation Y_{i_1}, it may select another index $i_2 \in \mathbb{N}$ and request to observe Y_{i_2}. This continues for a number of rounds at most n (i.e., it may request at most n labels), after which the algorithm must halt and produce a classifier $\hat{h}_n$. More formally, an active learning algorithm is defined by a random sequence $\{i_t\}_{t=1}^\infty$ in $\mathbb{N}$, a random variable N in $\mathbb{N}$, and a random classifier $\hat{h}_n$, satisfying the following properties. Each i_t is conditionally independent from $\{(X_i, Y_i)\}_{i=1}^\infty$ given $\{i_j\}_{j=1}^{t-1}$, $\{Y_{i_j}\}_{j=1}^{t-1}$, and $\{X_i\}_{i=1}^\infty$. The random variable N always has $N \leq n$, and for any $k \in \{0, \ldots, n\}$, $\mathbb{1}[N = k]$ is independent from

1. Since conditional probabilities are only defined up to probability zero differences, there can be multiple valid functions $\eta(\cdot; P)$ and $f_P^\star$, with any two such functions being equal with probability one. As such, we will interpret statements such as "$f_P^\star \in \mathbb{C}$" to mean that *there exists* a version of $f_P^\star$ contained in $\mathbb{C}$, and similarly for other claims and conditions for $f_P^\star$ and $\eta(\cdot; P)$.
2. Although, in practice, we would expect to have access to only a finite number of unlabeled samples, we expect this number would often be quite large (as unlabeled samples are considered inexpensive and abundant in many applications). For simplicity, and to focus the analysis purely on the number of *labels* required for learning, we approximate this scenario by supposing an *inexhaustible* source of unlabeled samples. We leave open the question of the number of unlabeled samples sufficient to obtain the minimax label complexity; in particular, we expect the number of such samples used by the methods obtaining our upper bounds to be quite large indeed.
3. Although technically we study the pool-based active learning protocol, all of our results apply equally well to the stream-based (selective sampling) model of active learning (in which the algorithm must decide whether or not to request the label Y_i before observing any X_j with $j > i$ or requesting any Y_j with $j > i$).

$\{(X_i, Y_i)\}_{i=1}^{\infty}$ given $\{i_j\}_{j=1}^{k}$, $\{Y_{i_j}\}_{j=1}^{k}$, and $\{X_i\}_{i=1}^{\infty}$. Finally, $\hat{h}_n$ is independent from $\{(X_i, Y_i)\}_{i=1}^{\infty}$ given N, $\{i_j\}_{j=1}^{N}$, $\{Y_{i_j}\}_{j=1}^{N}$, and $\{X_i\}_{i=1}^{\infty}$.

We are now ready for the definition of our primary quantity of study: the minimax label complexity. In the next section, we define several well-known noise models as specifications of the set $\mathbb{D}$ referenced in this definition.

Definition 1 *For a given set $\mathbb{D}$ of probability measures on $\mathcal{X} \times \mathcal{Y}$, $\forall \varepsilon \geq 0$, $\forall \delta \in [0,1]$, the* minimax label complexity *(of active learning) under $\mathbb{D}$ with respect to $\mathbb{C}$, denoted $\Lambda_{\mathbb{D}}(\varepsilon, \delta)$, is the smallest $n \in \mathbb{N} \cup \{0\}$ such that there exists an active learning algorithm $\mathcal{A}$ with the property that, for every $\mathcal{P}_{XY} \in \mathbb{D}$, the classifier $\hat{h}_n$ produced by $\mathcal{A}(n)$ based on the (independent $\mathcal{P}_{XY}$-distributed) data sequence $(X_1, Y_1), (X_2, Y_2), \ldots$ satisfies*

$$\mathbb{P}\left(\mathrm{er}_{\mathcal{P}_{XY}}\left(\hat{h}_n\right) - \inf_{h \in \mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) > \varepsilon\right) \leq \delta.$$

If no such n exists, we define $\Lambda_{\mathbb{D}}(\varepsilon, \delta) = \infty$.

Following Vapnik and Chervonenkis (1971); Anthony and Bartlett (1999), we say a collection of sets $\mathcal{T} \subseteq 2^{\mathcal{X}}$ *shatters* a sequence $S \in \mathcal{X}^k$ (for $k \in \mathbb{N}$) if $\{A \cap S : A \in \mathcal{T}\} = 2^S$. The *VC dimension* of $\mathcal{T}$ is then defined as the largest $k \in \mathbb{N} \cup \{0\}$ such that there exists $S \in \mathcal{X}^k$ shattered by $\mathcal{T}$; if no such largest k exists, the VC dimension is defined to be ∞. Overloading this terminology, the VC dimension of a set $\mathcal{H}$ of classifiers is defined as the VC dimension of the collection of sets $\{\{x : h(x) = +1\} : h \in \mathcal{H}\}$. Throughout this article, we denote by d the VC dimension of $\mathbb{C}$. We are particularly interested in the case $d < \infty$, in which case $\mathbb{C}$ is called a *VC class*.

For any set $\mathcal{H}$ of classifiers, define $\mathrm{DIS}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } h(x) \neq g(x)\}$, the *region of disagreement* of $\mathcal{H}$. Also, for any classifier h, any $r \geq 0$, and any probability measure P on $\mathcal{X}$, define $\mathrm{B}_P(h, r) = \{g \in \mathbb{C} : P(x : g(x) \neq h(x)) \leq r\}$, the *$r$-ball centered at h*.

Before proceeding, we introduce a few additional notational conventions that help to simplify the theorem statements and proofs. For any $\mathbb{R}$-valued functions f and g, we write $f(x) \lesssim g(x)$ (or equivalently $g(x) \gtrsim f(x)$) to express the fact that there is a *universal* finite numerical constant $c > 0$ such that $f(x) \leq c g(x)$. For any $x \in [0, \infty]$, we define $\mathrm{Log}(x) = \max\{\ln(x), 1\}$, where $\ln(0) = -\infty$ and $\ln(\infty) = \infty$. For simplicity, we define $\frac{\infty}{\mathrm{Log}(\infty)} = \infty$, but in any other context, we always define $0 \cdot \infty = 0$, and also define $\frac{a}{0} = \infty$ for any $a > 0$. For any function $\phi : \mathbb{R} \to \mathbb{R}$, we use the notation "$\lim_{\gamma \to 0} \phi(\gamma)$" to indicating taking the limit as γ approaches 0 *from above*: i.e., $\gamma \downarrow 0$. For $a, b \in \mathbb{R}$, we denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Finally, we remark that some of the claims below technically require additional qualifications to guarantee measurability of certain quantities (as is typically the case in empirical process theory); see Blumer, Ehrenfeucht, Haussler, and Warmuth (1989); van der Vaart and Wellner (1996, 2011) for some discussion of this issue. For simplicity, we do not mention these issues in the analysis below; rather, we implicitly qualify all of these results with the condition that $\mathbb{C}$ is such that all of the random variables and events arising in the proofs are measurable.

3. Noise Models

We now introduce the noise models under which we will study the minimax label complexity of active learning. These are defined as sets of probability measures on $\mathcal{X} \times \mathcal{Y}$, corresponding to specifications of the set $\mathbb{D}$ in Definition 1.

- (Realizable Case) Define RE as the collection of $\mathcal{P}_{XY}$ for which $f^{\star}_{\mathcal{P}_{XY}} \in \mathbb{C}$ and $2\eta(\cdot;\mathcal{P}_{XY}) - 1 = f^{\star}_{\mathcal{P}_{XY}}(\cdot)$ (almost everywhere w.r.t. $\mathcal{P}$).
- (Bounded Noise) For $\beta \in [0, 1/2)$, define $\mathrm{BN}(\beta)$ as the collection of joint distributions $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$ such that $f^{\star}_{\mathcal{P}_{XY}} \in \mathbb{C}$ and

$$\mathcal{P}\left(x : |\eta(x;\mathcal{P}_{XY}) - 1/2| \geq 1/2 - \beta\right) = 1.$$

- (Tsybakov Noise) For $a \in [1, \infty)$ and $\alpha \in (0, 1)$, define $\mathrm{TN}(a, \alpha)$ as the collection of joint distributions $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$ such that $f^{\star}_{\mathcal{P}_{XY}} \in \mathbb{C}$ and $\forall \gamma > 0$,

$$\mathcal{P}\left(x : |\eta(x;\mathcal{P}_{XY}) - 1/2| \leq \gamma\right) \leq a'\gamma^{\alpha/(1-\alpha)},$$

where $a' = (1-\alpha)(2\alpha)^{\alpha/(1-\alpha)} a^{1/(1-\alpha)}$.

- (Bernstein Class Condition) For $a \in [1, \infty)$ and $\alpha \in [0, 1]$, define $\mathrm{BC}(a, \alpha)$ as the collection of joint distributions $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$ such that, $\exists h_{\mathcal{P}_{XY}} \in \mathbb{C}$ for which $\forall h \in \mathbb{C}$,

$$\mathcal{P}(x : h(x) \neq h_{\mathcal{P}_{XY}}(x)) \leq a(\mathrm{er}_{\mathcal{P}_{XY}}(h) - \mathrm{er}_{\mathcal{P}_{XY}}(h_{\mathcal{P}_{XY}}))^{\alpha}.$$

- (Benign Noise) For $\nu \in [0, 1/2]$, define $\mathrm{BE}(\nu)$ as the collection of all joint distributions $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$ such that $f^{\star}_{\mathcal{P}_{XY}} \in \mathbb{C}$ and $\mathrm{er}_{\mathcal{P}_{XY}}(f^{\star}_{\mathcal{P}_{XY}}) \leq \nu$.
- (Agnostic Noise) For $\nu \in [0, 1]$, define $\mathrm{AG}(\nu)$ as the collection of all joint distributions $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$ such that $\inf_{h \in \mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) \leq \nu$.

It is known that $\mathrm{RE} \subseteq \mathrm{BN}(\beta) \subseteq \mathrm{BC}(1/(1-2\beta), 1)$, and also that $\mathrm{RE} \subseteq \mathrm{TN}(a, \alpha) \subseteq \mathrm{BC}(a, \alpha)$. Furthermore, $\mathrm{TN}(a, \alpha)$ is equivalent to the conditions in $\mathrm{BC}(a, \alpha)$ being satisfied for *all* classifiers h, rather than merely those in $\mathbb{C}$ (Mammen and Tsybakov, 1999; Tsybakov, 2004; Boucheron, Bousquet, and Lugosi, 2005). All of RE, $\mathrm{BN}(\beta)$, and $\mathrm{TN}(a, \alpha)$ are contained in $\bigcup_{\nu < 1/2} \mathrm{BE}(\nu)$, and in particular, $\mathrm{BN}(\beta) \subseteq \mathrm{BE}(\beta)$.

The realizable case is the simplest setting studied here, corresponding to the "optimistic case" of Vapnik (1998) or the PAC model of Valiant (1984). The bounded noise model has been studied under various names (e.g., Massart and Nédélec, 2006; Giné and Koltchinskii, 2006; Kääriäinen, 2006; Koltchinskii, 2010; Raginsky and Rakhlin, 2011); it is sometimes referred to as *Massart's noise condition*. The Tsybakov noise condition was introduced by Mammen and Tsybakov (1999) in a slightly stronger form (in the related context of discrimination analysis) and was distilled into the form stated above by Tsybakov (2004). There is now a substantial literature on the label complexity under this condition, both for passive learning and active learning (e.g., Mammen and Tsybakov, 1999; Tsybakov, 2004; Bartlett, Jordan, and McAuliffe, 2006; Koltchinskii, 2006; Balcan, Broder, and Zhang, 2007; Hanneke, 2011, 2012, 2014; Hanneke and Yang, 2012). However, in much of this literature, the results are in fact established under the weaker assumption given by the Bernstein class condition (Bartlett, Mendelson, and Philips, 2004), which is known to be implied by the Tsybakov noise condition (Mammen and Tsybakov, 1999; Tsybakov, 2004). For passive learning, it is known that the minimax sample complexities under Tsybakov noise and under the Bernstein class condition are equivalent up to a logarithmic factor. Interestingly, our results below imply that this is not the case for active learning. The benign noise condition (studied by Hanneke, 2009b) requires only that the Bayes optimal classifier be contained within the hypothesis class, and that the Bayes error rate be at most the value of the parameter ν. The agnostic noise condition (sometimes called *adversarial noise* in related contexts) is the weakest of the noise assumptions studied here, and admits any distribution for which the best error rate among classifiers in the hypothesis class is at most the value of the parameter ν. This model has been widely studied in the literature, for both passive and active learning (e.g., Vapnik and Chervonenkis, 1971; Vapnik, 1982, 1998; Kearns, Schapire, and Sellie, 1994; Kalai, Klivans, Mansour, and Servedio, 2005; Balcan, Beygelzimer, and Langford, 2006; Hanneke, 2007b,a; Awasthi, Balcan, and Long, 2014).

4. A Combinatorial Complexity Measure

There is presently a substantial literature on distribution-dependent bounds on the label complexities of various active learning algorithms. These bounds are expressed in terms of a variety of interesting complexity measures, designed to capture the behavior of each of these particular algorithms. These measures of complexity include the disagreement coefficient (Hanneke, 2007b), the reciprocal of the splitting index (Dasgupta, 2005), the extended teaching dimension growth function (Hanneke, 2007a), and the version space compression set size (El-Yaniv and Wiener, 2010, 2012). These quantities have been studied and bounded for a variety of learning problems (see Hanneke, 2014, for a summary). They each have many interesting properties, and in general can exhibit a wide variety of behaviors, as functions of the distribution over $\mathcal{X}$ (and in some cases, the distribution over $\mathcal{X} \times \mathcal{Y}$) and ε, or in some cases, the data itself. However, something remarkable happens when we maximize each of these complexity measures over the choice of distribution (or data set): they all become equal to a simple and easy-to-calculate combinatorial quantity (see Section 7 for proofs of these equivalences). Specifically, consider the following definition.[4]

Definition 2 *Define the* star number $\mathfrak{s}$ *as the largest integer s such that there exist distinct points $x_1, \ldots, x_s \in \mathcal{X}$ and classifiers $h_0, h_1, \ldots, h_s \in \mathbb{C}$ satisfying the property that $\forall i \in \{1, \ldots, s\}$,* $\mathrm{DIS}(\{h_0, h_i\}) \cap \{x_1, \ldots, x_s\} = \{x_i\}$*; if no such largest integer exists, define $\mathfrak{s} = \infty$.*

For any set $\mathcal{H}$ of functions $\mathcal{X} \to \mathcal{Y}$, any $t \in \mathbb{N}$, $x_1, \ldots, x_t \in \mathcal{X}$, and $h_0, h_1, \ldots, h_t \in \mathcal{H}$, we will say $\{x_1, \ldots, x_t\}$ is a *star set* for $\mathcal{H}$, *witnessed by* $\{h_0, h_1, \ldots, h_t\}$, if $\forall i \in \{1, \ldots, t\}$, $\mathrm{DIS}(\{h_0, h_i\}) \cap \{x_1, \ldots, x_t\} = \{x_i\}$. For brevity, in some instances below, we may simply say that $\{x_1, \ldots, x_t\}$ *is a star set for* $\mathcal{H}$, indicating that $\exists h_0, h_1, \ldots, h_t \in \mathcal{H}$ such that $\{x_1, \ldots, x_t\}$ is a star set for $\mathcal{H}$, witnessed by $\{h_0, h_1, \ldots, h_t\}$. We may also say that $\{x_1, \ldots, x_t\}$ *is a star set for* $\mathcal{H}$ *centered at* $h_0 \in \mathcal{H}$ if $\exists h_1, \ldots, h_t \in \mathcal{H}$ such that $\{x_1, \ldots, x_t\}$ is a star set for $\mathcal{H}$, witnessed by $\{h_0, h_1, \ldots, h_t\}$. For completeness, we also say that $\{\}$ (the empty sequence) is a star set for $\mathcal{H}$ (witnessed by $\{h_0\}$ for any $h_0 \in \mathcal{H}$), for any nonempty $\mathcal{H}$. In these terms, the star number of $\mathbb{C}$ is the maximum possible cardinality of a star set for $\mathbb{C}$, or ∞ if no such maximum exists.

Remark: The star number can equivalently be described as the maximum possible degree in the data-induced one-inclusion graph for $\mathbb{C}$ (see Haussler, Littlestone, and Warmuth, 1994), where the

4. A similar notion previously appeared in a lower-bound argument of Dasgupta (2005), including a kind of distribution-dependent version of the "star set" idea. Indeed, we explore these connections formally in Section 7, where we additionally prove this definition is exactly equivalent to a quantity studied by Hanneke (2007a) (namely, the distribution-free version of the extended teaching dimension growth function), and has connections to several other complexity measures in the literature.

maximum is over all possible data sets and nodes in the graph.[5] To relate this to the VC dimension, one can show that the VC dimension is the maximum possible degree of a *hypercube* in the data-induced one-inclusion graph for $\mathbb{C}$ (maximized over all possible data sets). From this, it is clear that $\mathfrak{s} \geq d$. Indeed, any set $\{x_1, \ldots, x_k\}$ shatterable by $\mathbb{C}$ is also a star set for $\mathbb{C}$, since some $h_0 \in \mathbb{C}$ classifies all k points -1, and for each x_i, some $h_i \in \mathbb{C}$ has $h_i(x_i) = +1$ while $h_i(x_j) = -1$ for every $j \neq i$ (where h_i is guaranteed to exist by shatterability of the set). On the other hand, there is no general upper bound on $\mathfrak{s}$ in terms of d, and the gap between $\mathfrak{s}$ and d can generally be infinite.

Examples: Before continuing, we briefly go through a few simple example calculations of the star number. For the class of *threshold* classifiers on $\mathbb{R}$ (i.e., $\mathbb{C} = \{x \mapsto 2\mathbb{1}_{[t,\infty)}(x) - 1 : t \in \mathbb{R}\}$), we have $\mathfrak{s} = 2$, as $\{x_1, x_2\}$ is a star set for $\mathbb{C}$ centered at $2\mathbb{1}_{[t,\infty)} - 1$ if and only if $x_1 < t \leq x_2$, and any set $\{x_1, x_2, x_3\}$ cannot be a star set for $\mathbb{C}$ centered at any given $2\mathbb{1}_{[t,\infty)} - 1$ since, of the (at least) two of these points on the same side of t, any threshold classifier disagreeing with $2\mathbb{1}_{[t,\infty)} - 1$ on the one further from t must also disagree with $2\mathbb{1}_{[t,\infty)} - 1$ on the one closer to t. In contrast, for the class of *interval* classifiers on $\mathbb{R}$ (i.e., $\mathbb{C} = \{x \mapsto 2\mathbb{1}_{[a,b]}(x) - 1 : -\infty < a \leq b < \infty\}$), we have $\mathfrak{s} = \infty$, since for *any* distinct points $x_0, x_1, \ldots, x_s \in \mathbb{R}$, $\{x_1, \ldots, x_s\}$ is a star set for $\mathbb{C}$ witnessed by $\{2\mathbb{1}_{[x_0,x_0]} - 1, 2\mathbb{1}_{[x_1,x_1]} - 1, \ldots, 2\mathbb{1}_{[x_s,x_s]} - 1\}$. It is an easy exercise to verify that we also have $\mathfrak{s} = \infty$ for the classes of *linear separators* on $\mathbb{R}^k$ ($k \geq 2$) and axis-aligned rectangles on $\mathbb{R}^k$ ($k \geq 1$), since the above construction for interval classifiers can be embedded into these spaces, with the star set lying within a lower-dimensional manifold in $\mathbb{R}^k$ (see Dasgupta, 2004, 2005; Hanneke, 2014).

As an intermediate case, where $\mathfrak{s}$ has a range of values, consider the class of *intervals of width at least* $w \in (0, 1)$ (i.e., $\mathbb{C} = \{x \mapsto 2\mathbb{1}_{[a,b]}(x) - 1 : -\infty < a \leq b < \infty, b - a \geq w\}$), for the space $\mathcal{X} = [0, 1]$. In this case, we can show that $\lfloor 2/w \rfloor \leq \mathfrak{s} \leq \lfloor 2/w \rfloor + 2$, as follows. We may note that letting $k = \lfloor 2/(w + \varepsilon) \rfloor + 1$ (for $\varepsilon > 0$), and taking $x_i = (w + \varepsilon)(i - 1)/2$ for $1 \leq i \leq k$, we have that $\{x_1, \ldots, x_k\}$ is a star set for $\mathbb{C}$, witnessed by $\{2\mathbb{1}_{[-2w,-w]} - 1, 2\mathbb{1}_{[x_1-w/2,x_1+w/2]} - 1, \ldots, 2\mathbb{1}_{[x_k-w/2,x_k+w/2]} - 1\}$. Thus, taking $\varepsilon \to 0$ reveals that $\mathfrak{s} \geq \lfloor 2/w \rfloor$. On the other hand, for any $k' \in \mathbb{N}$ with $k' > 2$, and points $x_1, \ldots, x_{k'} \in [0, 1]$, suppose $\{x_1, \ldots, x_{k'}\}$ is a star set for $\mathbb{C}$ witnessed by $\{h_0, h_1, \ldots, h_{k'}\}$. Without loss of generality, suppose $x_1 \leq x_2 \leq \cdots \leq x_{k'}$. First suppose h_0 classifies all of these points -1. Note that, for any $i \in \{3, \ldots, k'\}$, since the interval corresponding to h_{i-1} has width at least w and contains x_{i-1} but not x_{i-2} or x_i, we have $x_i - x_{i-1} > \max\{0, w - (x_{i-1} - x_{i-2})\}$. Thus, $1 \geq \sum_{i=2}^{k'} x_i - x_{i-1} > x_2 - x_1 + \sum_{i=3}^{k'} \max\{0, w - (x_{i-1} - x_{i-2})\} \geq (k'-2)w - \sum_{i=3}^{k'-1} x_i - x_{i-1} = (k'-2)w - (x_{k'-1} - x_2)$, so that $x_{k'-1} - x_2 > (k'-2)w - 1$. But $x_{k'-1} - x_2 \leq 1$, so that $k' < 2/w + 2$. Since k' is an integer, this implies $k' \leq \lfloor 2/w \rfloor + 2$. For the remaining case, if h_0 classifies some x_i as $+1$, then let $x_{i_0} = \min\{x_i : h_0(x_i) = +1\}$ and $x_{i_1} = \max\{x_i : h_0(x_i) = +1\}$. Note that, if $i_0 > 1$, then for any $x < x_{i_0-1}$, any $h \in \mathbb{C}$ with $h(x_{i_0}) = h(x) = +1 \neq h_0(x)$ must have $h(x_{i_0-1}) = +1 \neq h_0(x_{i_0-1})$, so that $\{x, x_{i_0-1}\} \subseteq \mathrm{DIS}(\{h, h_0\})$. Therefore, $\nexists x_i < x_{i_0-1}$ (since otherwise $\mathrm{DIS}(\{h_i, h_0\}) \cap \{x_1, \ldots, x_{k'}\} = \{x_i\}$ would be violated), so that $i_0 \leq 2$. Symmetric reasoning implies $i_1 \geq k' - 1$. Similarly, if $\exists x \in (x_{i_0}, x_{i_1})$, then any $h \in \mathbb{C}$ with $h(x) = -1 \neq h_0(x)$ must have either $h(x_{i_0}) = -1 \neq h_0(x_{i_0})$ or $h(x_{i_1}) = -1 \neq h_0(x_{i_1})$, so that either $\{x, x_{i_0}\} \subseteq \mathrm{DIS}(\{h, h_0\})$ or $\{x, x_{i_1}\} \subseteq \mathrm{DIS}(\{h, h_0\})$. Therefore, $\nexists x_i \in (x_{i_0}, x_{i_1})$ (since again, $\mathrm{DIS}(\{h_i, h_0\}) \cap \{x_1, \ldots, x_{k'}\} = \{x_i\}$ would be violated),

5. The maximum degree in the one-inclusion graph was recently studied in the context of teaching complexity by Fan (2012). However, using the data-induced one-inclusion graph of Haussler, Littlestone, and Warmuth (1994) (rather than the graph based on the full space $\mathcal{X}$) can substantially increase the maximum degree by omitting certain highly-informative points.

so that $i_1 \in \{i_0, i_0+1\}$. Combined, these facts imply $k' \le i_1 + 1 \le i_0 + 2 \le 4 \le \lfloor 2/w \rfloor + 2$. Altogether, we have $\mathfrak{s} \le \lfloor 2/w \rfloor + 2$.

5. Main Results

We are now ready to state the main results of this article: upper and lower bounds on the minimax label complexities under the above noise models. For the sake of making the theorem statements more concise, we abstract the dependence on logarithmic factors in several of the upper bounds into a simple "polylog(x)" factor, meaning a value $\lesssim \mathrm{Log}^k(x)$, for some $k \in [1, \infty)$ (in fact, all of these results hold with values of $k \le 4$); the reader is referred to the proofs for a description of the actual logarithmic factors this polylog function represents, along with tighter expressions of the upper bounds. The formal proofs of all of these results are included in Appendix B.

Theorem 3 *For any $\varepsilon \in (0, 1/9)$, $\delta \in (0, 1/3)$,*

$$\max\left\{\min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}, d, \mathrm{Log}\left(\min\left\{\frac{1}{\varepsilon}, |\mathbb{C}|\right\}\right)\right\} \lesssim \Lambda_{\mathrm{RE}}(\varepsilon,\delta) \lesssim \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}, \frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})}\right\} \mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

Theorem 4 *For any $\beta \in [0, 1/2)$, $\varepsilon \in (0, (1-2\beta)/24)$, $\delta \in (0, 1/24]$,*

$$\frac{1}{(1-2\beta)^2}\max\left\{\min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon}\right\}\beta \mathrm{Log}\left(\frac{1}{\delta}\right), d\right\}$$
$$\lesssim \Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \lesssim \frac{1}{(1-2\beta)^2}\min\left\{\mathfrak{s}, \frac{(1-2\beta)d}{\varepsilon}\right\}\mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

Theorem 5 *For any $a \in [4, \infty)$, $\alpha \in (0,1)$, $\varepsilon \in (0, 1/(24a^{1/\alpha}))$, and $\delta \in (0, 1/24]$,*
if $0 < \alpha \le 1/2$,

$$a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \lesssim \Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta) \lesssim a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha} d\cdot \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right)$$

and if $1/2 < \alpha < 1$,

$$a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\max\left\{\min\left\{\mathfrak{s}, \frac{1}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1}\mathrm{Log}\left(\frac{1}{\delta}\right), d\right\}$$
$$\lesssim \Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta) \lesssim a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\min\left\{\frac{\mathfrak{s}}{d}, \frac{1}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1} d\cdot\mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

Theorem 6 *For any $a \in [4, \infty)$, $\alpha \in (0,1)$, $\varepsilon \in (0, 1/(24a^{1/\alpha}))$, and $\delta \in (0, 1/24]$,*
if $0 \le \alpha \le 1/2$,

$$a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \lesssim \Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta) \lesssim a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^{\alpha}}\right\} d\cdot\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right),$$

and if $1/2 < \alpha \le 1$,

$$a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\max\left\{\min\left\{\mathfrak{s}, \frac{1}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1}\mathrm{Log}\left(\frac{1}{\delta}\right), d\right\}$$
$$\lesssim \Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta) \lesssim a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^{\alpha}}\right\} d\cdot\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right).$$

Theorem 7 *For any* $\nu \in [0, 1/2)$, $\varepsilon \in (0, (1-2\nu)/24)$, *and* $\delta \in (0, 1/24]$,

$$\frac{\nu^2}{\varepsilon^2}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) + \min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\} \lesssim \Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta) \lesssim \left(\frac{\nu^2}{\varepsilon^2}d + \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}\right) \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

Theorem 8 *For any* $\nu \in [0, 1/2)$, $\varepsilon \in (0, (1-2\nu)/24)$, *and* $\delta \in (0, 1/24]$,

$$\frac{\nu^2}{\varepsilon^2}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) + \min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}$$
$$\lesssim \Lambda_{\mathrm{AG}(\nu)}(\varepsilon, \delta) \lesssim \min\left\{\mathfrak{s}, \frac{1}{\nu+\varepsilon}\right\}\left(\frac{\nu^2}{\varepsilon^2} + 1\right) d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right).$$

Here, we mention a few noteworthy observations and comments regarding the above theorems. We sketch the main innovations underlying the active learning algorithm achieving these upper bounds in Section 5.1. Sections 6 and 7 include further detailed and thorough comparisons of each of these results to those in the prior literature on passive and active learning.

Comparison to the previous best known results: Aside from Theorems 6 and 8, each of the above results offers some kind of refinement over the previous best known results on the label complexity of active learning. Some of these refinements are relatively mild, such as those for the realizable case and bounded noise. However, our refinements under Tsybakov noise and benign noise are far more significant. In particular, perhaps the most surprising and interesting of the above results are the upper bounds in Theorem 5, which can be considered the primary contribution of this work.

As discussed above, the prior literature on noise-robust active learning is largely rooted in the intuitions and techniques developed for the realizable case. As indicated by Theorem 3, there is a wide spread of label complexities for active learning problems in the realizable case, depending on the structure of the hypothesis class. In particular, when $\mathfrak{s} < \infty$, we have $O(\mathrm{Log}(1/\varepsilon))$ label complexity in the realizable case, representing a nearly-exponential improvement over passive learning, which has $\tilde{\Theta}(1/\varepsilon)$ dependence on ε. On the other hand, when $\mathfrak{s} = \infty$, we have $\Omega(1/\varepsilon)$ minimax label complexity for active learning, which is the same dependence on ε as known for passive learning (see Section 6). Thus, for active learning in the realizable case, some hypothesis classes are "easy" (such as threshold classifiers), offering strong improvements over passive learning, while others are "hard" (such as interval classifiers), offering almost no improvements over passive.

With the realizable case as inspiration, the results in the prior literature on general noise-robust active learning have all continued to reflect these distinctions, and the label complexity bounds in those works continue to exhibit this wide spread. In the case of Tsybakov noise, the best general results in the prior literature (from Hanneke and Yang, 2012; Hanneke, 2014) correspond to an upper bound of roughly $a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^\alpha}\right\} d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$ (after converting those complexity measures into the star number via the results in Section 7 below). When $\mathfrak{s} < \infty$, this has dependence $\tilde{\Theta}(\varepsilon^{2\alpha-2})$ on ε, which reflects a strong improvement over the $\tilde{\Theta}(\varepsilon^{\alpha-2})$ minimax sample complexity of passive learning for this problem (see Section 6). On the other hand, when $\mathfrak{s} = \infty$, this bound is $\tilde{\Theta}(\varepsilon^{\alpha-2})$, so that as in the realizable case, the bound is no better than that of passive learning for these hypothesis classes. Thus, the prior results in the literature continue the trend observed in the realizable case, in which the "easy" hypothesis classes admit strong improvements over passive learning, while the "hard" hypothesis classes have a bound that is no better than the sample complexity of passive learning.

With this as background, it comes as quite a surprise that the upper bounds in Theorem 5 are *always* smaller than the corresponding minimax sample complexities of passive learning, in terms of their asymptotic dependence on ε for $0 < \alpha < 1$. Specifically, these upper bounds reveal a label complexity $\tilde{O}(\varepsilon^{2\alpha-2})$ when $\mathfrak{s} < \infty$, and $\tilde{O}(\varepsilon^{2\alpha-2} \vee (1/\varepsilon))$ when $\mathfrak{s} = \infty$. Comparing to the $\tilde{\Theta}(\varepsilon^{\alpha-2})$ minimax sample complexity of passive learning, the improvement for active learning is by a factor of $\tilde{\Theta}(\varepsilon^{-\alpha})$ when $\mathfrak{s} < \infty$, and by a factor of $\tilde{\Theta}(\varepsilon^{-\min\{\alpha,1-\alpha\}})$ when $\mathfrak{s} = \infty$. As a further surprise, when $0 < \alpha \leq 1/2$ (the high-noise regime), we see that the distinctions between active learning problems of a given VC dimension essentially *vanish* (up to logarithmic factors), so that the familiar spread of label complexities from the realizable case is no longer present. Indeed, in this latter case, *all* hypothesis classes with finite VC dimension exhibit the strong improvements over passive learning, previously only known to hold for the "easy" hypothesis classes (such as threshold classifiers): that is, $\tilde{O}(\varepsilon^{2\alpha-2})$ label complexity.

Further examining these upper bounds, we see that the spread of label complexities between "easy" and "hard" hypothesis classes increasingly re-emerges as α approaches 1, beginning with $\alpha = 1/2$. This transition point is quite sensible, since this is precisely the point at which the label complexity has dependence on ε of $\tilde{\Theta}(1/\varepsilon)$, which is roughly the same as the minimax label complexity of the "hard" hypothesis classes in the realizable case, which is, after all, included in $\mathrm{TN}(a, \alpha)$. Thus, as α increases above $1/2$, the "easy" hypothesis classes (with $\mathfrak{s} < \infty$) exhibit stronger improvements over passive learning, while the "hard" hypothesis classes (with $\mathfrak{s} = \infty$) continue to exhibit precisely this $\tilde{\Theta}\left(\frac{1}{\varepsilon}\right)$ behavior. In either case, the label complexity exhibits an improvement in dependence on ε compared to passive learning for the same α value. But since the label complexity of passive learning decreases to $\tilde{\Theta}\left(\frac{1}{\varepsilon}\right)$ as $\alpha \to 1$, we naturally have that for the "hard" hypothesis classes, the gap between the passive and active label complexities shrinks as α approaches 1. In contrast, the "easy" hypothesis classes exhibit a gap between passive and active label complexities that becomes more pronounced as α approaches 1 (with a near-exponential improvement over passive learning exhibited in the limiting case, corresponding to bounded noise).

This same pattern is present, though to a lesser extent, in the benign noise case. In this case, the best general results in the prior literature (from Dasgupta, Hsu, and Monteleoni, 2007; Hanneke, 2007a, 2014) correspond to an upper bound of roughly $\min\left\{\mathfrak{s}, \frac{1}{\nu+\varepsilon}\right\}\left(\frac{\nu^2}{\varepsilon^2}+1\right) d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$ (again, after converting those complexity measures into the star number via the results in Section 7 below). When $\mathfrak{s} < \infty$, the dependence on ν and ε is roughly $\tilde{\Theta}\left(\frac{\nu^2}{\varepsilon^2}\right)$ (aside from logarithmic factors and constants, and for $\nu > \varepsilon$). However, when $\mathfrak{s} = \infty$, this dependence becomes roughly $\tilde{\Theta}\left(\frac{\nu}{\varepsilon^2}\right)$, which is the same as in the minimax sample complexity of passive learning (see Section 6). Thus, for these results in the prior literature, we again see that the "easy" hypothesis classes have a bound reflecting improvements over passive learning, while the bound for the "hard" hypothesis classes fail to reflect any improvements over passive learning at all.

In contrast, consider the upper bound in Theorem 7. In this case, when $\nu \geq \sqrt{\varepsilon}$ (again, the high-noise regime), for *all* hypothesis classes with finite VC dimension, the dependence on ν and ε is roughly $\tilde{\Theta}\left(\frac{\nu^2}{\varepsilon^2}\right)$ (aside from logarithmic factors and constants). Again, this makes almost no distinction between "easy" hypothesis classes (with $\mathfrak{s} < \infty$) and "hard" hypothesis classes (with $\mathfrak{s} = \infty$), and instead always exhibits the strongest possible improvements (up to logarithmic factors), previously only known to hold for the "easy" classes (such as threshold classifiers): namely, reduction in label complexity by roughly a factor of $1/\nu$ compared to passive learning. The improvements in this case are typically milder than we found in Theorem 5, but noteworthy nonetheless.

Again, as ν decreases below $\sqrt{\varepsilon}$, the distinction between "easy" and "hard" hypothesis classes begins to re-emerge, with the harder classes maintaining a $\tilde{\Theta}\left(\frac{1}{\varepsilon}\right)$ dependence (which is equivalent to the realizable-case label complexity for these classes, up to logarithmic factors), while the easier classes continue to exhibit the $\tilde{\Theta}\left(\frac{\nu^2}{\varepsilon^2}\right)$ behavior, approaching $O\left(\mathrm{polylog}\left(\frac{1}{\varepsilon}\right)\right)$ as ν shrinks.

The dependence on $\boldsymbol{\delta}$: One remarkable fact about $\Lambda_{\mathrm{RE}}(\varepsilon, \delta)$ is that there is *no* significant dependence on δ in the optimal label complexity for the given range of δ.[6] Note that this is not the case in noisy settings, where the lower bounds have an explicit dependence on δ. In the proofs, this dependence on δ is introduced via randomness of the labels. However, as argued by Kääriäinen (2006), a dependence on δ is sometimes still required in $\Lambda_{\mathbb{D}}(\varepsilon, \delta)$, even if we restrict $\mathbb{D}$ to those $\mathcal{P}_{XY} \in \mathrm{AG}(\nu)$ inducing *deterministic* labels: that is, $\eta(x; \mathcal{P}_{XY}) \in \{0, 1\}$ for all x.

Spanning the gaps: All of these results have gaps between the lower and upper bounds. It is interesting to note that one can construct examples of hypothesis classes spanning these gaps, for Theorems 3, 4, 5, and 7 (up to logarithmic factors). For instance, for sufficiently large d and $\mathfrak{s}$ and sufficiently small ε and δ, these upper bounds are tight (up to logarithmic factors) in the case where $\mathbb{C} = \{x \mapsto 2\mathbb{1}_S(x) - 1 : S \subseteq \{1, \ldots, \mathfrak{s}\}, |S| \le d\}$, for $\mathcal{X} = \mathbb{N}$ (taking inspiration from a suggested modification by Hanneke, 2014, of the proof of a related result of Raginsky and Rakhlin, 2011). Likewise, these lower bounds are tight (up to logarithmic factors) in the case that $\mathcal{X} = \mathbb{N}$ and $\mathbb{C} = \{x \mapsto 2\mathbb{1}_S(x) - 1 : S \in 2^{\{1,\ldots,d\}} \cup \{\{i\} : d+1 \le i \le \mathfrak{s}\}\}$.[7] Thus, these upper and lower bounds cannot be significantly refined (without loss of generality) without introducing additional complexity measures to distinguish these cases. For completeness, we include proofs of these claims in Appendix D. It immediately follows from this (and monotonicity of the respective noise models in $\mathbb{C}$) that the upper and lower bounds in Theorems 3, 4, 5, and 7 are each sometimes tight in the case $\mathfrak{s} = \infty$, as limiting cases of the above constructions: that is, the upper bounds are tight (up to logarithmic factors) for $\mathbb{C} = \{x \mapsto 2\mathbb{1}_S(x) - 1 : S \subseteq \mathbb{N}, |S| \le d\}$, and the lower bounds are tight (up to logarithmic factors) for $\mathbb{C} = \{x \mapsto 2\mathbb{1}_S(x) - 1 : S \in 2^{\{1,\ldots,d\}} \cup \{\{i\} : d+1 \le i < \infty\}\}$. It is interesting to note that the above space $\mathbb{C}$ for which the upper bounds are tight can be embedded in a variety of hypothesis classes in common use in machine learning (while maintaining VC dimension $\lesssim d$ and star number $\lesssim \mathfrak{s}$): for instance, in the case of $\mathfrak{s} = \infty$, this is true of linear separators in $\mathbb{R}^{3d}$ and axis-aligned rectangles in $\mathbb{R}^{2d}$. It follows that the upper bounds in these theorems are tight (up to logarithmic factors) for each of these hypothesis classes.

Separation of $\mathbf{TN}(\boldsymbol{a}, \boldsymbol{\alpha})$ and $\mathbf{BC}(\boldsymbol{a}, \boldsymbol{\alpha})$: Another interesting implication of these results is a separation between the noise models $\mathrm{TN}(a, \alpha)$ and $\mathrm{BC}(a, \alpha)$ not previously noted in the literature. Specifically, if we consider any class $\mathbb{C}$ comprised of only the $\mathfrak{s} + 1$ classifiers in Definition 2, then one can show[8] that (for $\mathfrak{s} \ge 3$), for any $\alpha \in (0, 1]$, $a \in [4, \infty)$, $\varepsilon \in (0, 1/(4a^{1/\alpha}))$, and $\delta \in (0, 1/16]$,

$$\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon, \delta) \gtrsim a^2 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^\alpha}\right\} \mathrm{Log}\left(\frac{1}{\delta}\right).$$

6. We should expect a more significant dependence on δ near 1, since one case easily prove that $\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \to 0$ as $\delta \to 1$.
7. Technically, for Theorems 4 and 7, we require slightly stronger versions of the lower bound to establish tightness for β or ν near 0: namely, adding the lower bound from Theorem 3 to these lower bounds. The validity of this stronger lower bound follows immediately from the facts that $\mathrm{RE} \subseteq \mathrm{BN}(\beta)$ and $\mathrm{RE} \subseteq \mathrm{BE}(\nu)$.
8. Specifically, this follows by taking $\zeta = \frac{a}{2}(4\varepsilon)^\alpha$, $\beta = \frac{1}{2} - \frac{2}{a4^\alpha}\varepsilon^{1-\alpha}$, and $k = \min\{\mathfrak{s} - 1, \lfloor 1/\zeta \rfloor\}$ in Lemma 26 of Appendix A.2, and noting that the resulting set of distributions $\mathrm{RR}(k, \zeta, \beta)$ is contained in $\mathrm{BC}(a, \alpha)$ for this $\mathbb{C}$.

In particular, when $\mathfrak{s} > \frac{1}{a\varepsilon^{\alpha}}$, we have $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta) \gtrsim a\varepsilon^{\alpha-2}\mathrm{Log}(1/\delta)$, which is larger than the upper bound on $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta)$. Furthermore, when $\mathfrak{s} = \infty$, this lower bound has asymptotic dependence on ε that is $\Omega(\varepsilon^{\alpha-2})$, which is the same dependence found in the sample complexity of passive learning, up to a logarithmic factor (see Section 6 below). Comparing this to the upper bounds in Theorem 5, which exhibit asymptotic dependence on ε as $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta) = \tilde{O}(\varepsilon^{\min\{2\alpha-1,0\}-1})$ when $\mathfrak{s} = \infty$, we see that for this class, any $\alpha \in (0,1)$ has $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta) \ll \Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta)$. One reason this separation is interesting is that most of the existing literature on active learning under $\mathrm{TN}(a,\alpha)$ makes use of the noise condition via the fact that it implies $\mathcal{P}(x : h(x) \neq f^{\star}_{\mathcal{P}_{XY}}(x)) \leq a(\mathrm{er}_{\mathcal{P}_{XY}}(h) - \mathrm{er}_{\mathcal{P}_{XY}}(f^{\star}_{\mathcal{P}_{XY}}))^{\alpha}$ for all $h \in \mathbb{C}$: that is, $\mathrm{TN}(a,\alpha) \subseteq \mathrm{BC}(a,\alpha)$. This separation indicates that, to achieve the optimal performance under $\mathrm{TN}(a,\alpha)$, one needs to consider more-specific properties of this noise model, beyond those satisfied by $\mathrm{BC}(a,\alpha)$. Another reason this separation is quite interesting is that it contrasts with the known results for *passive* learning, where (as we discuss in Section 6 below) the sample complexities under these two noise models are *equivalent* (up to an unresolved logarithmic factor).

Gaps in Theorems 6 and 8, and related open problems: We conjecture that the dependence on d and $\mathfrak{s}$ in the upper bounds of Theorem 6 can be refined in general (where presently it is linear in $\mathfrak{s}d$). More specifically, we conjecture that the upper bound can be improved to

$$\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta) \lesssim a^2 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \min\left\{\mathfrak{s}, \frac{d}{a\varepsilon^{\alpha}}\right\} \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right),$$

though it is unclear at this time as to how this might be achieved. The above example (separating $\mathrm{BC}(a,\alpha)$ from $\mathrm{TN}(a,\alpha)$) indicates that we generally cannot hope to reduce the upper bound on the label complexity for $\mathrm{BC}(a,\alpha)$ much beyond this.

As for whether the form of the upper bound on $\Lambda_{\mathrm{AG}(\nu)}(\varepsilon,\delta)$ in Theorem 8 can generally be improved to match the form of the upper bound for $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta)$, this remains a fascinating open question. We conjecture that at least the dependence on d and $\mathfrak{s}$ can be improved to some extent (where presently it is linear in $d\mathfrak{s}$).

Minutiae: We note that the restrictions to the ranges of ε and δ in the above results are required only for the lower bounds (aside from $\delta \in (0,1]$, $\varepsilon > 0$), as are the restrictions to the ranges of the parameters a, α, and ν, aside from the constraints in the definitions in Section 3; the upper bounds are proven without any such restrictions in Appendix B. Also, several of the upper bounds above (e.g., Theorems 5 and 7) are slightly looser (by logarithmic factors) than those actually proven in Appendix B, which are typically stated in a different form (e.g., with factors of $d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)$, rather than simply $d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$). We state the weaker results here purely to simplify the theorem statements, referring the interested reader to the proofs for the refined versions. However, aside from Theorem 3, we believe it is possible to further optimize the logarithmic factors in all of these upper bounds.

We additionally note that we can also obtain results by the subset relations between the noise models. For instance, since $\mathrm{RE} \subseteq \mathrm{BN}(\beta) \subseteq \mathrm{BE}(\beta) \subseteq \mathrm{AG}(\beta)$, in the case β is close to 0 we can increase the lower bounds in Theorems 4, 7, and 8 based on the lower bound in Theorem 3: that is, for $\nu \geq \beta \geq 0$,

$$\Lambda_{\mathrm{AG}(\nu)}(\varepsilon,\delta) \geq \Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \geq \Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \geq \Lambda_{\mathrm{RE}}(\varepsilon,\delta) \gtrsim \max\left\{\min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}, d\right\}.$$

Similarly, since RE is contained in all of the noise models studied here, $\mathrm{Log}\left(\min\left\{\frac{1}{\varepsilon}, |\mathbb{C}|\right\}\right)$ can also be included as a lower bound in each of these results. Likewise, in the cases that a is very large or α is very close to 0, we can get a more informative upper bound in Theorem 5 via Theorem 7, since $\mathrm{TN}(a, \alpha) \subseteq \mathrm{BE}(1/2)$. For simplicity, in many cases we have not explicitly included the various compositions of the above results that can be obtained in this way (with only a few exceptions).

5.1 The Strategy behind Theorems 5 and 7

The upper bounds in Theorems 5 and 7 represent the main results of this work, and along with the upper bound in Theorem 4, are based on a general argument with essentially three main components. The first component is a more-sophisticated variant of a basic approach introduced to the active learning literature by Kääriäinen (2006): namely, reduction to the realizable case via repeatedly querying for the label at a point in $\mathcal{X}$ until its Bayes optimal classification can be determined (based on a sequential probability ratio test, as studied by Wald, 1945, 1947). Of course, in the present model of active learning, repeatedly requesting a label Y_i yields no new information beyond requesting Y_i once, since we are not able to resample from the distribution of Y_i given X_i (as Kääriäinen, 2006, does). To resolve this, we argue that it is possible to partition the space $\mathcal{X}$ into cells, in a way such that $f^{\star}_{\mathcal{P}_{XY}}$ is nearly constant in the vast majority of cells (without direct knowledge of $f^{\star}_{\mathcal{P}_{XY}}$ or $\mathcal{P}$); this is essentially a data-dependent approximation to the recently-discovered finite approximability property of VC classes (Adams and Nobel, 2012). Given this partition, for a given point X_i, we can find many other points X_j in the same cell of the partition with X_i, and request labels for these points until we can determine what the majority label for the cell is. We show that, with high probability, this value will equal $f^{\star}_{\mathcal{P}_{XY}}(X_i)$, so that we can effectively use these majority labels in an active learning algorithm for the realizable case.

However, we note that in the case of $\mathrm{TN}(a, \alpha)$, if we simply apply this repeated querying strategy to random $\mathcal{P}$-distributed samples, the resulting label complexity would be too large, and we would sometimes expect to exhaust most of the queries determining the optimal labels in very *noisy* regions (i.e., in cells of the partition where $\eta(\cdot; \mathcal{P}_{XY})$ is close to $1/2$ on average). This is because Tsybakov's condition allows that such regions can have non-negligible probability, and the number of samples required to determine the majority value of a ± 1 random variable becomes unbounded as its mean approaches zero. However, we can note that it is also less important for the final classifier $\hat{h}$ to agree with $f^{\star}_{\mathcal{P}_{XY}}$ on these high-noise points than it is for low-noise points, since classifying them opposite from $f^{\star}_{\mathcal{P}_{XY}}$ has less impact on the excess error rate $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) - \mathrm{er}_{\mathcal{P}_{XY}}(f^{\star}_{\mathcal{P}_{XY}})$. Therefore, as the second main component of our active learning strategy, we take a tiered approach to learning, effectively shifting the distribution $\mathcal{P}$ to favor points in cells with average $\eta(\cdot; \mathcal{P}_{XY})$ value further from $1/2$. We achieve this by discarding a point X_i if the number of queries exhausted toward determining the majority label in its cell of the partition becomes excessively large, and we gradually decrease this threshold as the data set grows, so that the points making it through this filter have progressively less and less noisy labels. By choosing $\hat{h}$ to agree with the inferred $f^{\star}_{\mathcal{P}_{XY}}$ classification of every point passing this filter, and combining this with the standard analysis of learning in the realizable case (Vapnik, 1982, 1998; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989), this allows us to provide a bound on the fraction of points in $\mathcal{X}$ at a given level of noisiness (i.e., $|\eta(\cdot; \mathcal{P}_{XY}) - 1/2|$) on which the produced classifier $\hat{h}$ disagrees with $f^{\star}_{\mathcal{P}_{XY}}$, such that this bound decreases as the noisiness decreases (i.e., as $|\eta(\cdot; \mathcal{P}_{XY}) - 1/2|$ increases). Furthermore, by discarding many of the points in high-noise regions without exhausting too many label requests trying to determine their $f^\star_{\mathcal{P}_{XY}}$ classifications, we are able to reduce the total number of label requests needed to obtain ε excess error rate.

Already these two components comprise the essential strategy that achieves these upper bounds in the case of $\mathfrak{s} = \infty$. However, to obtain the stated dependence on $\mathfrak{s}$ in these bounds when $\mathfrak{s} < \infty$, we need to introduce a third component: namely, using the inferred values of $f^\star_{\mathcal{P}_{XY}}(X_i)$ in the context of an active learning algorithm for the realizable case. For this, we specifically use the disagreement-based strategy of Cohn, Atlas, and Ladner (1994) (known as CAL), which processes the unlabeled data in sequence, and requests to observe the classification $f^\star_{\mathcal{P}_{XY}}(X_i)$ if and only if X_i is in the region of disagreement of the set of classifiers in $\mathbb{C}$ consistent with all previously-observed $f^\star_{\mathcal{P}_{XY}}(X_j)$ values. Using a modification of a recent analysis of this algorithm by Wiener, Hanneke, and El-Yaniv (2014) (applied to each tier of label-noise separately), combined with the results below (in Section 7.3) relating the complexity measure used in that analysis to the star number, we obtain the dependence on $\mathfrak{s}$ stated in the above results.

6. Comparison to Passive Learning

The natural baseline for comparison in active learning is the *passive learning* protocol, in which the labeled data are i.i.d. samples with common distribution $\mathcal{P}_{XY}$: that is, the input to the passive learning algorithm is $(X_1, Y_1), \ldots, (X_n, Y_n)$. In this context, the minimax sample complexity of passive learning, denoted $\mathcal{M}_{\mathbb{D}}(\varepsilon, \delta)$, is defined as the smallest $n \in \mathbb{N} \cup \{0\}$ for which there exists a passive learning rule mapping $(X_1, Y_1), \ldots, (X_n, Y_n)$ to a classifier $\hat{h} : \mathcal{X} \to \mathcal{Y}$ such that, for any $\mathcal{P}_{XY} \in \mathbb{D}$, with probability at least $1 - \delta$, $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) - \inf_{h \in \mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) \leq \varepsilon$.

Clearly $\Lambda_{\mathbb{D}}(\varepsilon, \delta) \leq \mathcal{M}_{\mathbb{D}}(\varepsilon, \delta)$ for any $\mathbb{D}$, since for every passive learning algorithm $\mathcal{A}$, there is an active learning algorithm that requests $Y_1, \ldots, Y_n$ and then runs $\mathcal{A}$ with $(X_1, Y_1), \ldots, (X_n, Y_n)$ to determine the returned classifier. One of the main interests in the theory of active learning is determining the size of the gap between these two complexities, for various sets $\mathbb{D}$. For the purpose of this comparison, we now review several results known to hold for $\mathcal{M}_{\mathbb{D}}(\varepsilon, \delta)$, for various sets $\mathbb{D}$. Specifically, the following bounds are known to hold for any choice of hypothesis class $\mathbb{C}$, and for β, a, α, ν, ε, and δ as in the respective theorems from Section 5 (Vapnik and Chervonenkis, 1971; Vapnik, 1982, 1998; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Ehrenfeucht, Haussler, Kearns, and Valiant, 1989; Haussler, Littlestone, and Warmuth, 1994; Massart and Nédélec, 2006; Hanneke, 2014).

- $\frac{1}{\varepsilon}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \lesssim \mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta) \lesssim \frac{1}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\max\{\varepsilon,\delta\}}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$.
- $\frac{1}{(1-2\beta)\varepsilon}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \lesssim \mathcal{M}_{\mathrm{BN}(\beta)}(\varepsilon, \delta) \lesssim \frac{1}{(1-2\beta)\varepsilon}\left(d\mathrm{Log}\left(\frac{1-2\beta}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$.
- $\frac{a}{\varepsilon^{2-\alpha}}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \lesssim \mathcal{M}_{\mathrm{TN}(a,\alpha)}(\varepsilon, \delta) \leq \mathcal{M}_{\mathrm{BC}(a,\alpha)} \lesssim \frac{a}{\varepsilon^{2-\alpha}}\left(d\mathrm{Log}\left(\frac{1}{a\varepsilon^{\alpha}}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$.
- $\frac{\nu+\varepsilon}{\varepsilon^2}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \lesssim \mathcal{M}_{\mathrm{BE}(\nu)}(\varepsilon, \delta) \leq \mathcal{M}_{\mathrm{AG}(\nu)}(\varepsilon, \delta) \lesssim \frac{\nu+\varepsilon}{\varepsilon^2}\left(d\mathrm{Log}\left(\frac{1}{\nu+\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$.

Let us compare these to the results for active learning in Section 5 on a case-by-case basis. In the realizable case, we observe clear improvements of active learning over passive learning in the case $\mathfrak{s} \ll \frac{d}{\varepsilon}$ (aside from logarithmic factors). In particular, based on the upper and lower bounds for both passive and active learning, we may conclude that $\mathfrak{s} < \infty$ is necessary and sufficient for the asymptotic dependence on ε to satisfy $\Lambda_{\mathrm{RE}}(\varepsilon, \cdot) = o(\mathcal{M}_{\mathrm{RE}}(\varepsilon, \cdot))$; specifically, when $\mathfrak{s} < \infty$, $\Lambda_{\mathrm{RE}}(\varepsilon, \cdot) =$

$O(\mathrm{Log}(\mathcal{M}_{\mathrm{RE}}(\varepsilon, \cdot)))$, and when $\mathfrak{s} = \infty$, $\Lambda_{\mathrm{RE}}(\varepsilon, \cdot) = \Theta(\mathcal{M}_{\mathrm{RE}}(\varepsilon, \cdot))$. For bounded noise, we have a similar asymptotic behavior. When $\mathfrak{s} < \infty$, again $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \cdot) = O(\mathrm{polylog}(\mathcal{M}_{\mathrm{BN}(\beta)}(\varepsilon, \cdot)))$, and when $\mathfrak{s} = \infty$, $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \cdot) = \tilde{\Theta}(\mathcal{M}_{\mathrm{BN}(\beta)}(\varepsilon, \cdot))$. In terms of the constants, to obtain improvements over passive learning (aside from the effects of logarithmic factors), it suffices to have $\mathfrak{s} \ll \frac{(1-2\beta)d}{\varepsilon}$, which is somewhat smaller (depending on β) than was sufficient in the realizable case.

Under Tsybakov's noise condition, every $\alpha \in (0, 1/2]$ shows an improvement in the upper bounds for active learning over the lower bound for passive learning by a factor of roughly $\frac{1}{a\varepsilon^{\alpha}}$ (aside from logarithmic factors). On the other hand, when $\alpha \in (1/2, 1)$, if $\mathfrak{s} < \frac{d}{a^{1/\alpha}\varepsilon}$, the improvement of active upper bounds over the passive lower bound is by a factor of roughly $\frac{1}{a\varepsilon^{\alpha}}\left(\frac{d}{\mathfrak{s}}\right)^{2\alpha-1}$, while for $\mathfrak{s} \geq \frac{d}{a^{1/\alpha}\varepsilon}$, the improvement is by a factor of roughly $\frac{1}{a^{\frac{1-\alpha}{\alpha}}\varepsilon^{1-\alpha}}$ (again, ignoring logarithmic factors in both cases). In particular, for *any* $\alpha \in (0, 1)$, when $\mathfrak{s} < \infty$, the asymptotic dependence on ε satisfies $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon, \cdot) = \tilde{\Theta}\left(\varepsilon^{\alpha}\mathcal{M}_{\mathrm{TN}(a,\alpha)}(\varepsilon, \cdot)\right)$, and when $\mathfrak{s} = \infty$, the asymptotic dependence on ε satisfies $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon, \cdot) = \tilde{\Theta}\left(\varepsilon^{\min\{\alpha, 1-\alpha\}}\mathcal{M}_{\mathrm{TN}(a,\alpha)}(\varepsilon, \cdot)\right)$. In either case, we have that for any $\alpha \in (0, 1)$, $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon, \cdot) = o(\mathcal{M}_{\mathrm{TN}(a,\alpha)}(\varepsilon, \cdot))$.

For the Bernstein class condition, the gaps in the upper and lower bounds of Theorem 6 render unclear the necessary and sufficient conditions for $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon, \cdot) = o(\mathcal{M}_{\mathrm{BC}(a,\alpha)}(\varepsilon, \cdot))$. Certainly $\mathfrak{s} < \infty$ is a sufficient condition for this, in which case the improvements are by a factor of roughly $\frac{1}{a\varepsilon^{\alpha}}$. However, in the case of $\mathfrak{s} = \infty$, the upper bounds do not reveal any improvements over those given above for $\mathcal{M}_{\mathrm{BC}(a,\alpha)}(\varepsilon, \delta)$. Indeed, the example given above in Section 5 reveals that, in some nontrivial cases, $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon, \delta) \gtrsim \mathcal{M}_{\mathrm{BC}(a,\alpha)}(\varepsilon, \delta)/\mathrm{Log}(1/\varepsilon)$, in which case any improvements would be, at best, in the constant and logarithmic factors. Note that this example also presents an interesting contrast between active and passive learning, since it indicates that in some cases $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon, \delta)$ and $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon, \delta)$ are quite different, while the above bounds for passive learning reveal that $\mathcal{M}_{\mathrm{BC}(a,\alpha)}(\varepsilon, \delta)$ is equivalent to $\mathcal{M}_{\mathrm{TN}(a,\alpha)}(\varepsilon, \delta)$ up to constant and logarithmic factors.

In the case of benign noise, comparing the above bounds for passive learning to Theorem 7, we see that (aside from logarithmic factors) the upper bound for active learning improves over the lower bound for passive learning by a factor of roughly $\frac{1}{\nu}$ when $\nu \geq \sqrt{\varepsilon}$. When $\nu < \sqrt{\varepsilon}$, if $\mathfrak{s} > \frac{d}{\varepsilon}$, the improvements are by a factor of roughly $\frac{\nu+\varepsilon}{\varepsilon}$, and if $\mathfrak{s} \leq \frac{d}{\varepsilon}$, the improvements are by roughly a factor of $\min\left\{\frac{1}{\nu}, \frac{(\nu+\varepsilon)d}{\varepsilon^2\mathfrak{s}}\right\}$ (again, ignoring logarithmic factors). However, as has been known for this noise model for some time (Kääriäinen, 2006), there are no gains in terms of the asymptotic dependence on ε for fixed ν. However, if we consider ν_{ε} such that $\varepsilon \leq \nu_{\varepsilon} = o(1)$, then for $\mathfrak{s} < \infty$ we have $\Lambda_{\mathrm{BE}(\nu_{\varepsilon})}(\varepsilon, \cdot) = \tilde{\Theta}(\nu_{\varepsilon}\mathcal{M}_{\mathrm{BE}(\nu_{\varepsilon})}(\varepsilon, \cdot))$, and for $\mathfrak{s} = \infty$ we have $\Lambda_{\mathrm{BE}(\nu_{\varepsilon})}(\varepsilon, \cdot) = \tilde{O}\left(\max\left\{\nu_{\varepsilon}, \frac{\varepsilon}{\nu_{\varepsilon}}\right\}\mathcal{M}_{\mathrm{BE}(\nu_{\varepsilon})}(\varepsilon, \cdot)\right)$.

Finally, in the case of agnostic noise, similarly to the Bernstein class condition, the gaps between the upper and lower bounds in Theorem 8 render unclear precisely what types of improvements we can expect when $\mathfrak{s} > \frac{1}{\nu+\varepsilon}$, ranging from the lower bound, which has the behavior described above for $\Lambda_{\mathrm{BE}(\nu)}$, to the upper bound, which reflects no improvements over passive learning in this case. When $\mathfrak{s} < \frac{1}{\nu+\varepsilon}$, the upper bound for active learning reflects an improvement over the lower bound for passive learning by roughly a factor of $\frac{1}{(\nu+\varepsilon)\mathfrak{s}}$ (aside from logarithmic factors). It remains an interesting open problem to determine whether the stronger improvements observed for benign noise generally also hold for agnostic noise.

Technique	Source	Relation to $\mathfrak{s}$
disagreement coefficient	(Hanneke, 2007b)	$\sup_P \theta_P(\varepsilon) = \mathfrak{s} \wedge \frac{1}{\varepsilon}$
splitting index	(Dasgupta, 2005)	$\sup_{h,P} \lim_{\tau \to 0} \left\lfloor \frac{1}{\rho_{h,P}(\varepsilon;\tau)} \right\rfloor = \mathfrak{s} \wedge \left\lfloor \frac{1}{\varepsilon} \right\rfloor$
teaching dimension	(Hanneke, 2007a)	$\mathrm{XTD}(\mathbb{C}, m) = \mathfrak{s} \wedge m$
version space compression	(El-Yaniv and Wiener, 2010)	$\max_{h \in \mathbb{C}} \max_{\mathcal{U} \in \mathcal{X}^m} \hat{n}_h(\mathcal{U}) = \mathfrak{s} \wedge m$
doubling dimension	(Li and Long, 2007)	$\sup_{h,P} D_{h,P}(\varepsilon) \in [1, O(d)] \log\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right)$

Table 1: Many of the complexity measures from the literature are related to the star number.

A remark on logarithmic factors: It is known that the terms of the form "$d\mathrm{Log}(x)$" in each of the above upper bounds for passive learning can be refined to replace x with the maximum of the disagreement coefficient (see Section 7.1 below) over the distributions in $\mathbb{D}$ (Giné and Koltchinskii, 2006; Hanneke and Yang, 2012; Hanneke, 2014). Therefore, based on the results in Section 7.1 relating the disagreement coefficient to the star number, we can replace these "$d\mathrm{Log}(x)$" terms with "$d\mathrm{Log}(\mathfrak{s} \wedge x)$". In the case of $\mathrm{BN}(\beta)$, Massart and Nédélec (2006) and Raginsky and Rakhlin (2011) have argued that, at least in some cases, this logarithmic factor can also be included in the lower bounds. It is presently not known whether this is the case for the other noise models studied here.

7. Connections to the Prior Literature on Active Learning

As mentioned, there is already a substantial literature bounding the label complexities of various active learning algorithms under various noise models. It is natural to ask how the results in the prior literature compare to those stated above. However, as most of the prior results are $\mathcal{P}_{XY}$-dependent, the appropriate comparison is to the worst-case values of those results: that is, maximizing the bounds over $\mathcal{P}_{XY}$ in the respective noise model. This section makes this comparison. In particular, we will see that the label complexity upper bounds above for RE, $\mathrm{BN}(\beta)$, $\mathrm{TN}(a, \alpha)$, and $\mathrm{BE}(\nu)$ all show some improvements over the known results, with the last two of these showing the strongest improvements.

The general results in the prior literature each express their label complexity bounds in terms of some kind of complexity measure. There are now several such complexity measures in use, each appropriate for studying some family of active learning algorithms under certain noise models. Most of these quantities are dependent on the distribution $\mathcal{P}_{XY}$ or the data, and their definitions are quite diverse. For some pairs of them, there are known inequalities loosely relating them, while other pairs have defied attempts to formally relate the quantities. The dependence on $\mathcal{P}_{XY}$ in the general results in the prior literature is typically isolated to the various complexity measures they are expressed in terms of. Thus, the natural first step is to characterize the worst-case values of these complexity measures, for any given hypothesis class $\mathbb{C}$. Plugging these worst-case values into the original bounds then allows us to compare to the results stated above.

In the process of studying the worst-case behaviors of these complexity measures, we also identify a *very* interesting fact that has heretofore gone unnoticed: namely, that almost all of the complexity measures in the relevant prior literature on the label complexity of active learning are in fact *equal* to the star number when maximized over the choice of distribution or data set. In some sense, this fact is quite surprising, as this seemingly-eclectic collection of complexity measures includes disparate definitions and interpretations, corresponding to entirely distinct approaches to the analysis of the respective algorithms these quantities are used to bound the label complexities of. Thus, this equivalence is interesting in its own right; additionally, it plays an important role in our proofs of the main results above, since it allows us to build on these diverse techniques from the prior literature when establishing these results.

Each subsection below is devoted to a particular complexity measure from the prior literature on active learning, each representing an established technique for obtaining label complexity bounds. Together, they represent a summary of the best-known general results from the prior literature relevant to our present discussion. In each case, we show the equivalence of the worst-case value of the complexity measure to the star number, and then combine this fact with the known results to obtain the corresponding bounds on the minimax label complexities implicit in the prior literature. In each case, we then compare this result to those obtained above.

We additionally study the *doubling dimension*, a quantity which has been used to bound the sample complexity of passive learning, and can be used to provide a loose bound on the label complexity of certain active learning algorithms. Below we argue that, when maximized over the choice of distribution, the doubling dimension can be upper and lower bounded in terms of the star number. One immediate implication of these bounds is that the doubling dimension is bounded if and only if the star number is finite.

Our findings on the relations of these various complexity measures to the star number are summarized in Table 1.

7.1 The Disagreement Coefficient

We begin with, what is perhaps the most well-studied complexity measure in the active learning literature: the *disagreement coefficient* (Hanneke, 2007b, 2009b).

Definition 9 *For any $r_0 \geq 0$, any classifier h, and any probability measure $\mathcal{P}$ over $\mathcal{X}$, the disagreement coefficient of h with respect to $\mathbb{C}$ under $\mathcal{P}$ is defined as*

$$\theta_{h,\mathcal{P}}(r_0) = \sup_{r > r_0} \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}_{\mathcal{P}}\left(h, r\right)\right)\right)}{r} \vee 1.$$

Also, for any probability measure $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$, letting $\mathcal{P}$ denote the marginal distribution of $\mathcal{P}_{XY}$ over $\mathcal{X}$, and letting $h^*_{\mathcal{P}_{XY}}$ denote a classifier with $\mathrm{er}_{\mathcal{P}_{XY}}(h^*_{\mathcal{P}_{XY}}) = \inf_{h \in \mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h)$ and $\inf_{h \in \mathbb{C}} \mathcal{P}(x : h(x) \neq h^*_{\mathcal{P}_{XY}}(x)) = 0$,[9] define the disagreement coefficient of the class $\mathbb{C}$ with respect to $\mathcal{P}_{XY}$ as $\theta_{\mathcal{P}_{XY}}(r_0) = \theta_{h^*_{\mathcal{P}_{XY}},\mathcal{P}}(r_0)$.

The disagreement coefficient is used to bound the label complexities of a family of active learning algorithms, described as *disagreement-based*. This line of work was initiated by Cohn, Atlas, and Ladner (1994), who propose an algorithm effective in the realizable case. That method was extended to be robust to label noise by Balcan, Beygelzimer, and Langford (2006, 2009), which then inspired a slew of papers studying variants of this idea; the interested reader is referred to Hanneke (2014) for a thorough survey of this literature. The general-case label complexity analysis of disagreement-based active learning (in terms of the disagreement coefficient) was initiated in the

9. See Hanneke (2012) for a proof that such a classifier always exists (though not necessarily in $\mathbb{C}$).

work of Hanneke (2007b, 2009b), and followed up by many papers since then (e.g., Dasgupta, Hsu, and Monteleoni, 2007; Hanneke, 2009a, 2011, 2012; Koltchinskii, 2010; Hanneke and Yang, 2012), as well as many works characterizing the value of the disagreement coefficient under various conditions (e.g., Hanneke, 2007b; Friedman, 2009; Balcan, Hanneke, and Vaughan, 2010; Wang, 2011; Balcan and Long, 2013; Hanneke, 2014); again, see Hanneke (2014) for a thorough survey of the known results on the disagreement coefficient.

To study the worst-case values of the label complexity bounds expressed in terms of the disagreement coefficient, let us define

$$\hat{\hat{\theta}}(\varepsilon) = \sup_{\mathcal{P}_{XY}} \theta_{\mathcal{P}_{XY}}(\varepsilon).$$

In fact, a result of Hanneke (2014, Theorem 7.4) implies that $\hat{\hat{\theta}}(\varepsilon) = \sup_{\mathcal{P}} \sup_{h \in \mathbb{C}} \theta_{h,\mathcal{P}}(\varepsilon)$, so that this would be an equivalent way to define $\hat{\hat{\theta}}(\varepsilon)$, which can sometimes be simpler to work with. We can now express the bounds on the minimax label complexity implied by the best general results to date in the prior literature on disagreement-based active learning (namely, the results of Hanneke, 2011; Dasgupta, Hsu, and Monteleoni, 2007; Koltchinskii, 2010; Hanneke and Yang, 2012; Hanneke, 2014), summarized as follows (see the survey of Hanneke, 2014, for detailed descriptions of the best-known logarithmic factors in these results).

- $\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \lesssim \hat{\hat{\theta}}(\varepsilon) d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$.
- $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \delta) \lesssim \frac{1}{(1-2\beta)^2} \hat{\hat{\theta}}(\varepsilon/(1-2\beta)) d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$.
- $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon, \delta) \lesssim a^2 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \hat{\hat{\theta}}(a\varepsilon^{\alpha}) d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$.
- $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon, \delta) \lesssim a^2 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \hat{\hat{\theta}}(a\varepsilon^{\alpha}) d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$.
- $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta) \lesssim \left(\frac{\nu^2}{\varepsilon^2} + 1\right) \hat{\hat{\theta}}(\nu + \varepsilon) d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$.
- $\Lambda_{\mathrm{AG}(\nu)}(\varepsilon, \delta) \lesssim \left(\frac{\nu^2}{\varepsilon^2} + 1\right) \hat{\hat{\theta}}(\nu + \varepsilon) d \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$.

In particular, these bounds on $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon, \delta)$, $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon, \delta)$, $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta)$, and $\Lambda_{\mathrm{AG}(\nu)}(\varepsilon, \delta)$ are the best general-case bounds on the label complexity of active learning in the prior literature (up to logarithmic factors), so that any improvements over these should be considered an interesting advance in our understanding of the capabilities of active learning methods. To compare these results to those stated in Section 5, we need to relate $\hat{\hat{\theta}}(\varepsilon)$ to the star number. Interestingly, we find that these quantities are *equal* (for $\varepsilon = 0$). Specifically, the following result describes the relation between these two quantities; its proof is included in Appendix C.1. This connection also plays a role in the proofs of some of our results from Section 5.

Theorem 10 $\forall \varepsilon \in (0, 1]$, $\hat{\hat{\theta}}(\varepsilon) = \mathfrak{s} \wedge \frac{1}{\varepsilon}$ *and* $\hat{\hat{\theta}}(0) = \mathfrak{s}$.

With this result in hand, we immediately observe that several of the upper bounds from Section 5 offer refinements over those stated in terms of $\hat{\hat{\theta}}(\cdot)$ above. For simplicity, we do not discuss differences in the logarithmic factors here. Specifically, the upper bound on $\Lambda_{\mathrm{RE}}(\varepsilon, \delta)$ in Theorem 3 refines that stated here by replacing the factor $\hat{\hat{\theta}}(\varepsilon)d = \min\left\{\mathfrak{s}d, \frac{d}{\varepsilon}\right\}$ with the sometimes-smaller factor $\min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}$. Likewise, the upper bound on $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta)$ in Theorem 4 refines the result stated here, again by replacing the factor $\hat{\hat{\theta}}(\varepsilon/(1-2\beta))d = \min\left\{\mathfrak{s}d, \frac{(1-2\beta)d}{\varepsilon}\right\}$ with the sometimes-smaller factor $\min\left\{\mathfrak{s}, \frac{(1-2\beta)d}{\varepsilon}\right\}$. On the other hand, Theorem 5 offers a much stronger refinement over the result stated above. Specifically, in the case $\alpha \le 1/2$, the upper bound in Theorem 5 completely *eliminates* the factor of $\hat{\hat{\theta}}(a\varepsilon^{\alpha})$ from the upper bound on $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta)$ stated here (i.e., replacing it with a universal constant). For the case $\alpha > 1/2$, the upper bound on $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta)$ in Theorem 5 replaces this factor of $\hat{\hat{\theta}}(a\varepsilon^{\alpha}) = \min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^{\alpha}}\right\}$ with the factor $\min\left\{\frac{\mathfrak{s}}{d}, \frac{1}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1}$, which is always smaller (for small ε and large d). The upper bounds on $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta)$ and $\Lambda_{\mathrm{AG}(\nu)}(\varepsilon,\delta)$ in Theorems 6 and 8 are equivalent to those stated here; indeed, this is precisely how these results are obtained in Appendix B. We have conjectured above that at least the dependence on d and $\mathfrak{s}$ can be refined, analogous to the refinements for the realizable case and bounded noise noted above. However, we *do* obtain refinements for the bound on $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta)$ in Theorem 7, replacing the factor of $\left(\frac{\nu^2}{\varepsilon^2}+1\right)\hat{\theta}(\nu+\varepsilon)d = \left(\frac{\nu^2}{\varepsilon^2}+1\right)\min\left\{\mathfrak{s}d, \frac{d}{\nu+\varepsilon}\right\}$ in the upper bound here with a factor $\frac{\nu^2}{\varepsilon^2}d + \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}$, which is sometimes significantly smaller (for $\varepsilon \ll \nu \ll 1$ and large d).

7.2 The Splitting Index

Another, very different, approach to the design and analysis of active learning algorithms was proposed by Dasgupta (2005): namely, the *splitting* approach. In particular, this technique has the desirable property that it yields distribution-dependent label complexity bounds for the realizable case which, even when the marginal distribution $\mathcal{P}$ is held fixed, (almost) imply near-minimax performance. The intuition behind this technique is that the objective in the realizable case (achieving error rate at most ε) is typically well-approximated by the related objective of reducing the *diameter* of the version space (set of classifiers consistent with the observed labels) to size at most ε. From this perspective, at any given time, the impediments to achieving this objective are clearly identifiable: pairs of classifiers $\{h,g\}$ in $\mathbb{C}$ consistent with all labels observed thus far, yet with $\mathcal{P}(x : h(x) \neq g(x)) > \varepsilon$. Supposing we have only a finite number of such classifiers (which can be obtained if we first replace $\mathbb{C}$ by a fine-grained finite *cover* of $\mathbb{C}$), we can then estimate the *usefulness* of a given point X_i by the number of these pairs it would be guaranteed to eliminate if we were to request its label (supposing the worse of the two possible labels); by "eliminate," we mean that at least one of the two classifiers will be inconsistent with the observed label. If we always request labels of points guaranteed to eliminate a large fraction of the surviving ε-separated pairs, we will quickly arrive at a version space of diameter ε, and can then return any surviving classifier. Dasgupta (2005) further applies this strategy in tiers, first eliminating at least one classifier from every $\frac{1}{2}$-separated pair, then repeating this for the remaining $\frac{1}{4}$-separated pairs, and so on. This allows the label complexity to be *localized*, in the sense that the surviving Δ-separated pairs we need to eliminate will be composed of classifiers within distance 2Δ of $f^{\star}_{\mathcal{P}_{XY}}$ (or the representative thereof in the initial finite cover of $\mathbb{C}$). The analysis of this method naturally leads to the following definition from Dasgupta (2005).

For any finite set $Q \subseteq \{\{h, g\} : h, g \in \mathbb{C}\}$ of unordered pairs of classifiers in $\mathbb{C}$, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, let $Q_x^y = \{\{h, g\} \in Q : h(x) = g(x) = y\}$, and define

$$\mathrm{Split}(Q, x) = |Q| - \max_{y \in \mathcal{Y}} |Q_x^y|.$$

This represents the number of pairs guaranteed to be eliminated (as described above) by requesting the label at a point x. The splitting index is then defined as follows.

Definition 11 *For any $\rho, \Delta, \tau \in [0, 1]$, a set $\mathcal{H} \subseteq \mathbb{C}$ is said to be (ρ, Δ, τ)-splittable under a probability measure $\mathcal{P}$ over $\mathcal{X}$ if, for all finite $Q \subseteq \{\{h, g\} \subseteq \mathcal{H} : \mathcal{P}(x : h(x) \neq g(x)) \geq \Delta\}$,*

$$\mathcal{P}(x : \mathrm{Split}(Q, x) \geq \rho|Q|) \geq \tau.$$

For any classifier $h : \mathcal{X} \to \mathcal{Y}$, any probability measure $\mathcal{P}$ over $\mathcal{X}$, and any $\varepsilon, \tau \in [0, 1]$, the splitting index *is defined as*

$$\rho_{h,\mathcal{P}}(\varepsilon; \tau) = \sup \left\{\rho \in [0, 1] : \forall \Delta \geq \varepsilon, \mathrm{B}_{\mathcal{P}}(h, 4\Delta) \textit{ is } (\rho, \Delta, \tau)\textit{-splittable under } \mathcal{P}\right\}.$$

Dasgupta (2005) proves a bound on the label complexity of a general active learning algorithm based on the above strategy, in the realizable case, expressed in terms of the splitting index. Specifically, for any $\tau > 0$, letting $\rho = \rho_{f^{\star}_{\mathcal{P}_{XY}},\mathcal{P}}(\varepsilon/4; \tau)$, Dasgupta (2005) finds that for that algorithm to achieve error rate at most ε with probability at least $1 - \delta$, it suffices to use a number of label requests

$$\frac{d}{\rho} \mathrm{polylog}\left(\frac{d}{\varepsilon\delta\tau\rho}\right). \tag{1}$$

The τ argument to $\rho_{h,\mathcal{P}}(\varepsilon; \tau)$ captures the trade-off between the number of label requests and the number of unlabeled samples available, with smaller τ corresponding to the scenario where more unlabeled data are available, and a larger value of $\rho_{h,\mathcal{P}}(\varepsilon; \tau)$. Specifically, Dasgupta (2005) argues that $\tilde{O}\left(\frac{d}{\tau\rho}\right)$ unlabeled samples suffice to achieve the above result. In our present model, we suppose an abundance of unlabeled data, and as such, we are interested in the behavior for very small τ. However, note that the logarithmic factors in the above bound have an inverse dependence on τ, so that taking τ too small can potentially increase the value of the bound. It is not presently known whether or not this is necessary (though intuitively it seems not to be). However, for the purpose of comparison to our results in Section 5, we will ignore this logarithmic dependence on $1/\tau$, and focus on the leading factor. In this case, we are interested in the value $\lim_{\tau \to 0} \rho_{h,\mathcal{P}}(\varepsilon; \tau)$. Additionally, to convert (1) into a distribution-free bound for the purpose of comparison to the results in Section 5, we should minimize this value over the choice of $\mathcal{P}$ and $h \in \mathbb{C}$. Formally, we are interested in the following quantity, defined for any $\varepsilon \in [0, 1]$.

$$\hat{\hat{\rho}}(\varepsilon) = \inf_{P} \inf_{h \in \mathbb{C}} \lim_{\tau \to 0} \rho_{h,P}(\varepsilon; \tau).$$

In particular, in terms of this quantity, the maximum possible value of the bound (1) for a given hypothesis class $\mathbb{C}$ is at least

$$\frac{d}{\hat{\hat{\rho}}(\varepsilon/4)} \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

To compare this to the upper bound in Theorem 3, we need to relate $\frac{1}{\hat{\hat{\rho}}(\varepsilon)}$ to the star number. Again, we find that these quantities are essentially *equal* (as $\varepsilon \to 0$), as stated in the following theorem.

Theorem 12 $\forall \varepsilon \in (0,1]$, $\left\lfloor \frac{1}{\hat{\hat{\rho}}(\varepsilon)} \right\rfloor = \mathfrak{s} \wedge \left\lfloor \frac{1}{\varepsilon} \right\rfloor$.

The proof of this result is included in Appendix C.2. We note that the inequalities $\mathfrak{s} \wedge \left\lfloor \frac{1}{\varepsilon} \right\rfloor \leq \left\lfloor \frac{1}{\hat{\hat{\rho}}(\varepsilon)} \right\rfloor \leq \left\lfloor \frac{1}{\varepsilon} \right\rfloor$ were already implicit in the original work of Dasgupta (2005, Corollary 3 and Lemma 1). For completeness (and to make the connection explicit), we include these arguments in the proof given in Appendix C.2, along with our proof that $\left\lfloor \frac{1}{\hat{\hat{\rho}}(\varepsilon)} \right\rfloor \leq \mathfrak{s}$ (which was heretofore unknown).

Plugging this into the above bound, we see that the maximum possible value of the bound (1) for a given hypothesis class $\mathbb{C}$ is at least

$$\min\left\{\mathfrak{s}d, \frac{d}{\varepsilon}\right\} \operatorname{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

Note that the upper bound in Theorem 3 refines this by reducing the first term in the "$\min$" from $\mathfrak{s}d$ to simply $\mathfrak{s}$.

Dasgupta (2005) also argues for a kind of lower bound in terms of the splitting index, which was reformulated as a lower bound on the minimax label complexity (for a fixed $\mathcal{P}$) in the realizable case by Balcan and Hanneke (2012); Hanneke (2014). In our present distribution-free style of analysis, the implication of that result is the following lower bound.

$$\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \gtrsim \frac{1}{\hat{\hat{\rho}}(4\varepsilon)}.$$

Based on Theorem 12, we see that the $\min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}$ term in the lower bound of Theorem 3 follows immediately from this lower bound. For completeness, in Appendix B, we directly prove this term in the lower bound, based on a more-direct argument than that used to establish the above lower bound. We note, however, that Dasgupta (2005, Corollary 3) also describes a technique for obtaining lower bounds, which is essentially equivalent to that used in Appendix B to obtain this term (and furthermore, makes use of a distribution-dependent version of the "star" idea).

The upper bounds of Dasgupta (2005) have also been extended to the bounded noise setting. In particular, Balcan and Hanneke (2012) and Hanneke (2014) have proposed variants of the splitting approach, which are robust to bounded noise. They have additionally bounded the label complexities of these methods in terms of the splitting index. Similarly to the above discussion of the realizable case, the worst-case values of these bounds for any given hypothesis class $\mathbb{C}$ are larger than those stated in Theorem 4 by factors related to the VC dimension (logarithmic factors aside). We refer the interested readers to these sources for the details of those bounds.

7.3 The Teaching Dimension

Another quantity that has been used to bound the label complexity of certain active learning methods is the *extended teaching dimension growth function*. This quantity was introduced by Hanneke (2007a), inspired by analogous notions used to tightly-characterize the query complexity of *Exact* learning with membership queries (Hegedüs, 1995; Hellerstein, Pillaipakkamnatt, Raghavan, and Wilkins, 1996). The term *teaching dimension* takes its name from the literature on Exact teaching (Goldman and Kearns, 1995), where the teaching dimension characterizes the minimum number of well-chosen labeled data points sufficient to guarantee that the only classifier in $\mathbb{C}$ consistent with these labels is the target function. Hegedüs (1995) extends this to target functions not contained in

$\mathbb{C}$, in which case the objective is simply to leave at most one consistent classifier in $\mathbb{C}$; he refers to the minimum number of points sufficient to achieve this as the *extended teaching dimension*, and argues that this quantity can be used to characterize the minimum number of *membership queries* by a learning algorithm sufficient to guarantee that the only classifier in $\mathbb{C}$ consistent with the returned labels is the target function (which is the objective in the *Exact* learning model).

Hanneke (2007a) transfers this strategy to the statistical setting studied here (where the objective is only to obtain excess error rate ε with probability $1-\delta$, rather than exactly identifying a target function). That work introduces empirical versions of the teaching dimension and extended teaching dimension, and defines distribution-dependent bounds on these quantities. It then proves upper and lower bounds on the label complexity in terms of these quantities. For our present purposes, we will be most-interested in a particular distribution-free upper bound on these quantities, called the *extended teaching dimension growth function*, also introduced by Hanneke (2006, 2007a). Since both this quantity and the star number are distribution-free, they can be directly compared.

We introduce these quantities formally as follows. For any $m \in \mathbb{N} \cup \{0\}$ and $S \in \mathcal{X}^m$, and for any $h : \mathcal{X} \to \mathcal{Y}$, define the *version space* $V_{S,h} = \{g \in \mathbb{C} : \forall x \in S, g(x) = h(x)\}$ (Mitchell, 1977). For any $m \in \mathbb{N}$ and $\mathcal{U} \in \mathcal{X}^m$, let $\mathbb{C}[\mathcal{U}]$ denote an arbitrary subset of classifiers in $\mathbb{C}$ such that, $\forall h \in \mathbb{C}$, $|\mathbb{C}[\mathcal{U}] \cap V_{\mathcal{U},h}| = 1$: that is, $\mathbb{C}[\mathcal{U}]$ contains exactly one classifier from each equivalence class in $\mathbb{C}$ induced by the classifications of $\mathcal{U}$. For any classifier $h : \mathcal{X} \to \mathcal{Y}$, define

$$\mathrm{TD}(h, \mathbb{C}[\mathcal{U}], \mathcal{U}) = \min\{t \in \mathbb{N} \cup \{0\} : \exists S \in \mathcal{U}^t \text{ s.t. } |V_{S,h} \cap \mathbb{C}[\mathcal{U}]| \leq 1\},$$

the *empirical teaching dimension* of h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$. Any $S \in \bigcup_t \mathcal{U}^t$ with $|V_{S,h} \cap \mathbb{C}[\mathcal{U}]| \leq 1$ is called a *specifying set* for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$; thus, $\mathrm{TD}(h, \mathbb{C}[\mathcal{U}], \mathcal{U})$ is the size of a *minimal specifying set* for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$. Equivalently, $S \in \bigcup_t \mathcal{U}^t$ is a specifying set for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$ if and only if $\mathrm{DIS}(V_{S,h}) \cap \mathcal{U} = \emptyset$. Also define $\mathrm{TD}(h, \mathbb{C}, m) = \max\limits_{\mathcal{U} \in \mathcal{X}^m} \mathrm{TD}(h, \mathbb{C}[\mathcal{U}], \mathcal{U})$, $\mathrm{TD}(\mathbb{C}, m) = \max\limits_{h \in \mathbb{C}} \mathrm{TD}(h, \mathbb{C}, m)$ (the *teaching dimension growth function*), and $\mathrm{XTD}(\mathbb{C}, m) = \max\limits_{h : \mathcal{X} \to \mathcal{Y}} \mathrm{TD}(h, \mathbb{C}, m)$ (the *extended teaching dimension growth function*).

Hanneke (2007a) proves two upper bounds on the label complexity of active learning relevant to our present discussion. They are summarized as follows (see the original source for the precise logarithmic factors).[10]

- $\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \lesssim \mathrm{XTD}\left(\mathbb{C}, \left\lceil \frac{1}{\varepsilon} \right\rceil\right) d \cdot \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right)$.
- $\Lambda_{\mathrm{AG}(\nu)}(\varepsilon, \delta) \lesssim \left(\frac{\nu^2}{\varepsilon^2} + 1\right) \mathrm{XTD}\left(\mathbb{C}, \left\lceil \frac{1}{\nu+\varepsilon} \right\rceil\right) d \cdot \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right)$.

Since $\mathrm{BE}(\nu) \subseteq \mathrm{AG}(\nu)$, we have the further implication that

$$\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta) \lesssim \left(\frac{\nu^2}{\varepsilon^2} + 1\right) \mathrm{XTD}\left(\mathbb{C}, \left\lceil \frac{1}{\nu+\varepsilon} \right\rceil\right) d \cdot \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

Additionally, by a refined argument of Hegedüs (1995), the ideas of Hanneke (2007a) can be applied (see Hanneke, 2006, 2009b) to show that

$$\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \lesssim \frac{\mathrm{XTD}(\mathbb{C}, \lceil d/\varepsilon \rceil)}{\log_2(\mathrm{XTD}(\mathbb{C}, \lceil d/\varepsilon \rceil))} d \cdot \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

10. Here we have simplified the arguments m to the $\mathrm{XTD}(\mathbb{C}, m)$ instances compared to those of Hanneke (2007a), using monotonicity of $m \mapsto \mathrm{XTD}(\mathbb{C}, m)$, combined with the basic observation that $\mathrm{XTD}(\mathbb{C}, mk) \leq \mathrm{XTD}(\mathbb{C}, m)k$ for any integer $k \geq 1$.

To compare these bounds to the results stated in Section 5, we will need to relate the quantity $\mathrm{XTD}(\mathbb{C}, m)$ to the star number. Although it may not be obvious from a superficial reading of the definitions, we find that these quantities are *exactly equal* (as $m \to \infty$). Thus, the extended teaching dimension growth function is simply an alternative way of referring to the star number (and vice versa), as they define the same quantity.[11] This equivalence is stated formally in the following theorem, the proof of which is included in Appendix C.3.

Theorem 13 $\forall m \in \mathbb{N}$, $\mathrm{XTD}(\mathbb{C}, m) = \mathrm{TD}(\mathbb{C}, m) = \min\{\mathfrak{s}, m\}$.

We note that the inequalities $\min\{\mathfrak{s}, m\} \leq \mathrm{TD}(\mathbb{C}, m) \leq \mathrm{XTD}(\mathbb{C}, m) \leq m$ follow readily from previously-established facts about the teaching dimension. For instance, Fan (2012) notes that the teaching dimension of any class is at least the maximum degree of its one-inclusion graph; applying this fact to $\mathbb{C}[\mathcal{U}]$ and maximizing over the choice of $\mathcal{U} \in \mathcal{X}^m$, this maximum degree becomes $\min\{\mathfrak{s}, m\}$ (by definition of $\mathfrak{s}$). However, the inequality $\mathrm{XTD}(\mathbb{C}, m) \leq \mathfrak{s}$ and the resulting fact that $\mathrm{XTD}(\mathbb{C}, m) = \mathrm{TD}(\mathbb{C}, m)$ are apparently new.

In fact, in the process of proving this theorem, we establish another remarkable fact: that *every* minimal specifying set is a star set. This is stated formally in the following lemma, the proof of which is also included in Appendix C.3.

Lemma 14 *For any $h : \mathcal{X} \to \mathcal{Y}$, $m \in \mathbb{N}$, and $\mathcal{U} \in \mathcal{X}^m$, every minimal specifying set for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$ is a star set for $\mathbb{C} \cup \{h\}$ centered at h.*

Using Theorem 13, we can now compare the results above to those in Section 5. For simplicity, we will not discuss the differences in logarithmic factors here. Specifically, Theorem 3 refines the results here on $\Lambda_{\mathrm{RE}}(\varepsilon, \delta)$, replacing a factor of $\min\left\{\mathrm{XTD}(\mathbb{C}, \lceil 1/\varepsilon \rceil)d, \frac{\mathrm{XTD}(\mathbb{C}, \lceil d/\varepsilon \rceil)d}{\log(\mathrm{XTD}(\mathbb{C}, \lceil d/\varepsilon \rceil))}\right\} \approx \min\left\{\mathfrak{s}d, \frac{d}{\varepsilon}, \frac{\mathfrak{s}d}{\log(\mathfrak{s})}, \frac{d^2}{\varepsilon \log(d/\varepsilon)}\right\}$ implied by the above results with a factor of $\min\left\{\mathfrak{s}, \frac{d}{\varepsilon}, \frac{\mathfrak{s}d}{\log(\mathfrak{s})}\right\}$, thus reducing the first term in the "min" by a factor of d (though see below, as Wiener, Hanneke, and El-Yaniv, 2014, have already shown this to be possible, directly in terms of $\mathrm{XTD}(\mathbb{C}, m)$). Theorem 13 further reveals that the above bound on $\Lambda_{\mathrm{AG}(\nu)}(\varepsilon, \delta)$ is equivalent (up to logarithmic factors) to that stated in Theorem 8. However, the bound on $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta)$ in Theorem 7 refines that implied above, replacing a factor $\left(\frac{\nu^2}{\varepsilon^2} + 1\right) \mathrm{XTD}\left(\mathbb{C}, \left\lceil \frac{1}{\nu+\varepsilon} \right\rceil\right) d \approx \left(\frac{\nu^2}{\varepsilon^2} + 1\right) \min\left\{\mathfrak{s}d, \frac{d}{\nu+\varepsilon}\right\}$ with a factor $\frac{\nu^2}{\varepsilon^2}d + \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}$, which can be significantly smaller for $\varepsilon \ll \nu \ll 1$ and large d.

Hanneke (2006, 2007a) also proves a *lower bound* on the label complexity of active learning in the realizable case, based on the following modification of the extended teaching dimension. For any set $\mathcal{H} \subseteq \mathbb{C}$, classifier $h : \mathcal{X} \to \mathcal{Y}$, $m \in \mathbb{N}$, $\mathcal{U} \in \mathcal{X}^m$, and $\delta \in [0, 1]$, define the *partial teaching dimension* as

$$\mathrm{XPTD}(h, \mathcal{H}[\mathcal{U}], \mathcal{U}, \delta) = \min\{t \in \mathbb{N} \cup \{0\} : \exists S \in \mathcal{U}^t \text{ s.t. } |V_{S,h} \cap \mathcal{H}[\mathcal{U}]| \leq \delta |\mathcal{H}[\mathcal{U}]| + 1\},$$

and let $\mathrm{XPTD}(\mathcal{H}, m, \delta) = \max_{h : \mathcal{X} \to \mathcal{Y}} \max_{\mathcal{U} \in \mathcal{X}^m} \mathrm{XPTD}(h, \mathcal{H}[\mathcal{U}], \mathcal{U}, \delta)$. Hanneke (2006, 2007a) proves that

$$\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \geq \max_{\mathcal{H} \subseteq \mathbb{C}} \mathrm{XPTD}\left(\mathcal{H}, \left\lceil \frac{1 - \varepsilon}{\varepsilon} \right\rceil, \delta\right).$$

The following result relates this quantity to the star number.

11. In this sense, the star number is not really a *new* quantity to the active learning literature, but rather a simplified definition for the already-familiar extended teaching dimension growth function.

Theorem 15 $\forall m \in \mathbb{N}$, $\forall \delta \in [0, 1/2]$,

$$\lceil (1-2\delta) \min\{\mathfrak{s}, m\} \rceil \leq \max_{\mathcal{H} \subseteq \mathbb{C}} \mathrm{XPTD}(\mathcal{H}, m, \delta) \leq \left\lceil \left(1 - \frac{\delta}{1+\delta}\right) \min\{\mathfrak{s}, m\} \right\rceil.$$

The proof is in Appendix C.3. Note that, combined with the lower bound of Hanneke (2006, 2007a), this immediately implies the part of the lower bound in Theorem 3 involving $\mathfrak{s}$. In Appendix B, we provide a direct proof for this term in the lower bound, based on an argument similar to that of Hanneke (2007a).

7.3.1 THE VERSION SPACE COMPRESSION SET SIZE

More-recently, El-Yaniv and Wiener (2010, 2012); Wiener, Hanneke, and El-Yaniv (2014) have studied a quantity $\hat{n}_h(\mathcal{U})$ (for a sequence $\mathcal{U} \in \bigcup_m \mathcal{X}^m$ and classifier h), termed the minimal *version space compression set size*, defined as the size of the smallest subsequence $S \subseteq \mathcal{U}$ for which $V_{S,h} = V_{\mathcal{U},h}$.[12]

It is easy to see that, when $h \in \mathbb{C}$, the version space compression set size is equivalent to the empirical teaching dimension: that is, $\forall h \in \mathbb{C}$,

$$\hat{n}_h(\mathcal{U}) = \mathrm{TD}(h, \mathbb{C}[\mathcal{U}], \mathcal{U}).$$

To see this, note that since $|V_{\mathcal{U},h} \cap \mathbb{C}[\mathcal{U}]| = 1$, any $S \subseteq \mathcal{U}$ with $V_{S,h} = V_{\mathcal{U},h}$ has $|V_{S,h} \cap \mathbb{C}[\mathcal{U}]| = 1$, and hence is a specifying set for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$. On the other hand, for any $S \subseteq \mathcal{U}$, we (always) have $V_{S,h} \supseteq V_{\mathcal{U},h}$, so that if $|V_{S,h} \cap \mathbb{C}[\mathcal{U}]| \leq 1$, then $V_{S,h} \cap \mathbb{C}[\mathcal{U}] \supseteq V_{\mathcal{U},h} \cap \mathbb{C}[\mathcal{U}]$ and $|V_{S,h} \cap \mathbb{C}[\mathcal{U}]| \geq |V_{\mathcal{U},h} \cap \mathbb{C}[\mathcal{U}]| = 1 \geq |V_{S,h} \cap \mathbb{C}[\mathcal{U}]|$, which together imply $V_{S,h} \cap \mathbb{C}[\mathcal{U}] = V_{\mathcal{U},h} \cap \mathbb{C}[\mathcal{U}]$; thus, $V_{S,h} \subseteq \{g \in \mathbb{C} : \forall x \in \mathcal{U}, g(x) = h(x)\} = V_{\mathcal{U},h} \subseteq V_{S,h}$, so that $V_{S,h} = V_{\mathcal{U},h}$: that is, S is a version space compression set. Thus, in the case $h \in \mathbb{C}$, any version space compression set S is a specifying set for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$ and vice versa. That $\hat{n}_h(\mathcal{U}) = \mathrm{TD}(h, \mathbb{C}[\mathcal{U}], \mathcal{U})$ $\forall h \in \mathbb{C}$ follows immediately from this equivalence.

In particular, combined with Theorem 13, this implies that $\forall m \in \mathbb{N}$,

$$\max_{\mathcal{U} \in \mathcal{X}^m} \max_{h \in \mathbb{C}} \hat{n}_h(\mathcal{U}) = \mathrm{TD}(\mathbb{C}, m) = \min\{\mathfrak{s}, m\}. \tag{2}$$

Letting $\hat{n}_m = \hat{n}_{f^\star_{\mathcal{P}_{XY}}}(\{X_1, \ldots, X_m\})$, Wiener, Hanneke, and El-Yaniv (2014) have shown that, in the realizable case, for the CAL active learning algorithm (proposed by Cohn, Atlas, and Ladner, 1994) to achieve error rate at most ε with probability at least $1 - \delta$, it suffices to use a budget n of any size at least

$$\max_{1 \leq m \leq M_{\varepsilon,\delta}} \hat{n}_m \cdot \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right),$$

where $M_{\varepsilon,\delta} \lesssim \frac{1}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$ is a bound on the sample complexity of passive learning by returning an arbitrary classifier in the version space (Vapnik, 1982, 1998; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989). They further provide a distribution-dependent bound (to remove the dependence on the data here) based on confidence bounds on $\hat{n}_m$ (analogous to the aforementioned distribution-dependent bounds on the empirical teaching dimension studied by Hanneke, 2007a).

12. The quantity studied there is defined slightly differently, but is easily seen to be equivalent to this definition.

For our purposes (distribution-free, data-independent bounds), we can simply take the maximum over possible data sets and possible $f^{\star}_{\mathcal{P}_{XY}}$ functions, so that the above bound becomes

$$\max_{x_1,x_2,\ldots\in\mathcal{X}} \max_{h\in\mathbb{C}} \max_{1\leq m\leq M_{\varepsilon,\delta}} \hat{n}_h(\{x_1,\ldots,x_m\})\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$$
$$= \mathrm{TD}\left(\mathbb{C}, M_{\varepsilon,\delta}\right)\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right) \lesssim \mathrm{TD}\left(\mathbb{C}, \left\lfloor\frac{d}{\varepsilon}\right\rfloor\right)\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right).$$

Combining this with (2), we find that the label complexity of CAL in the realizable case is at most

$$\min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right),$$

which matches the upper bound on the minimax label complexity from Theorem 3 up to logarithmic factors.

7.4 The Doubling Dimension

Another quantity of interest in the learning theory literature is the *doubling dimension*, also known as the *local metric entropy* (LeCam, 1973; Yang and Barron, 1999; Gupta, Krauthgamer, and Lee, 2003; Bshouty, Li, and Long, 2009). Specifically, for any set $\mathcal{H}$ of classifiers, a set of classifiers $\mathcal{G}$ is an ε-cover of $\mathcal{H}$ (with respect to the $\mathcal{P}(\mathrm{DIS}(\{\cdot,\cdot\}))$ pseudometric) if

$$\sup_{h\in\mathcal{H}} \inf_{g\in\mathcal{G}} \mathcal{P}(x : g(x) \neq h(x)) \leq \varepsilon.$$

Let $\mathcal{N}(\varepsilon,\mathcal{H},\mathcal{P})$ denote the minimum cardinality $|\mathcal{G}|$ over all ε-covers $\mathcal{G}$ of $\mathcal{H}$, or else $\mathcal{N}(\varepsilon,\mathcal{H},\mathcal{P}) = \infty$ if no finite ε-cover of $\mathcal{H}$ exists. The doubling dimension (at h) is defined as follows.

Definition 16 *For any $\varepsilon \in (0,1]$, any probability measure P over $\mathcal{X}$, and any classifier h, define*

$$D_{h,P}(\varepsilon) = \max_{r\geq\varepsilon} \log_2\left(\mathcal{N}\left(r/2, \mathrm{B}_P(h,r), P\right)\right).$$

The quantity $D_\varepsilon = D_{f^{\star}_{\mathcal{P}_{XY}},\mathcal{P}}(\varepsilon)$ is known to be useful in bounding the sample complexity of passive learning. Specifically, Li and Long (2007); Bshouty, Li, and Long (2009) have shown that there is a passive learning algorithm achieving sample complexity $\lesssim \frac{D_{\varepsilon/4}}{\varepsilon} + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)$ for $\mathcal{P}_{XY} \in$ RE. Furthermore, though we do not go into the details here, by a combination of the ideas from Dasgupta (2005), Balcan, Beygelzimer, and Langford (2009), and Hanneke (2007b), it is possible to show that a certain active learning algorithm achieves a label complexity $\lesssim 4^{D_\varepsilon} D_\varepsilon \cdot \mathrm{polylog}(\frac{1}{\varepsilon\delta})$ for $\mathcal{P}_{XY} \in$ RE, though this is typically a very loose upper bound.

To our knowledge, the question of the worst-case value of the doubling dimension for a given hypothesis class $\mathbb{C}$ has not previously been explored in the literature (though there is an obvious $O(d\log(1/\varepsilon))$ upper bound). Here we obtain upper and lower bounds on this worst-case value, expressed in terms of the star number. While this relation generally has a wide range (roughly a factor of d), it does have the interesting implication that the doubling dimension is *bounded* if and only if $\mathfrak{s} < \infty$. Specifically, we have the following theorem, the proof of which is included in Appendix C.4.

Theorem 17 $\forall \varepsilon \in (0, 1/4]$, $\max\left\{d, \mathrm{Log}\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right)\right\} \lesssim \sup_{P} \sup_{h \in \mathbb{C}} D_{h,P}(\varepsilon) \lesssim d\mathrm{Log}\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right)$.

One can show that the gap between the upper and lower bounds on $\sup_P \sup_{h \in \mathbb{C}} D_{h,P}(\varepsilon)$ in this result cannot generally be improved by much without sacrificing generality or introducing additional quantities. Specifically, for the class $\mathbb{C}$ discussed in Appendix D.2, we have $\sup_P \sup_{h \in \mathbb{C}} D_{h,P}(\varepsilon) \leq \sup_P \log_2(\mathcal{N}(\varepsilon/2, \mathbb{C}, P)) \lesssim \max\left\{d, \mathrm{Log}\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right)\right\}$, so that the lower bound above is sometimes tight to within a universal constant factor. For the class $\mathbb{C}$ discussed in Appendix D.1, based on a result of Raginsky and Rakhlin (2011, Lemma 4), one can show $\sup_P \sup_{h \in \mathbb{C}} D_{h,P}(\varepsilon) \gtrsim d\mathrm{Log}\left(\frac{\mathfrak{s}}{d} \wedge \frac{1}{\varepsilon}\right)$, so that the above upper bound is sometimes tight, aside from a small difference in the logarithmic factor (dividing $\mathfrak{s}$ by d).

Interestingly, in the process of proving the upper bound in Theorem 17, we also establish the following inequality relating the doubling dimension and the disagreement coefficient, holding for any classifier h, any probability measure $\mathcal{P}$ over $\mathcal{X}$, and any $\varepsilon \in (0, 1]$.

$$D_{h,\mathcal{P}}(\varepsilon) \leq 2d \log_2\left(22e^2 \theta_{h,\mathcal{P}}(\varepsilon)\right).$$

This inequality may be of independent interest, as it enables comparisons between results in the literature expressed in terms of these quantities. For instance, it implies that in the realizable case, the passive learning sample complexity bound of Bshouty, Li, and Long (2009) is no larger than that of Giné and Koltchinskii (2006) (aside from constant factors).

8. Conclusions

In this work, we derived upper and lower bounds on the minimax label complexity of active learning under several noise models. In most cases, these new bounds offer refinements over the best results in the prior literature. Furthermore, in the case of Tsybakov noise, we discovered the heretofore-unknown fact that the minimax label complexity of active learning with VC classes is *always* smaller than that of passive learning. We expressed each of these bounds in terms of a simple combinatorial complexity measure, termed the *star number*. We further found that almost all of the distribution-dependent and sample-dependent complexity measures in the prior active learning literature are exactly equal to the star number when maximized over the choice of distribution or data set.

The bounds derived here are all distribution-free, in the sense that they are expressed without dependence or restrictions on the marginal distribution $\mathcal{P}$ over $\mathcal{X}$. They are also worst-case bounds, in the sense that they express the maximum of the label complexity over the distributions in the noise model $\mathbb{D}$, rather than expressing a bound on the label complexity achieved by a given algorithm as a function of $\mathcal{P}_{XY}$. As observed by Dasgupta (2005), there are some cases in which smaller label complexities can be achieved under restrictions on the marginal distribution $\mathcal{P}$, and some cases in which there are achievable label complexities which exhibit a range of values depending on $\mathcal{P}_{XY}$ (see also Balcan, Hanneke, and Vaughan, 2010; Hanneke, 2012, for further exploration of this). Our results reveal that in some cases, such as Tsybakov noise with $\alpha \leq 1/2$, these issues might typically not be of much significance (aside from logarithmic factors). However, in other cases, particularly when $\mathfrak{s} = \infty$, the issue of expressing distribution-dependent bounds on the label complexity is clearly an important one. In particular, the question of the minimax label complexity of active learning under the restrictions of the above noise models that explicitly fix the marginal distribution $\mathcal{P}$ remains an important and challenging open problem. In deriving such bounds, the present work should be considered a kind of guide, in that we should restrict our focus to deriving distribution-dependent label complexity bounds with worst-case values that are never worse than the distribution-free bounds proven here.

Appendix A. Preliminary Lemmas

Before presenting the proofs of the main results above, we begin by introducing some basic lemmas, which will be useful in the main proofs below.

A.1 ε-nets and ε-covers

For a collection $\mathcal{T}$ of measurable subsets of $\mathcal{X}$, a value $\varepsilon \geq 0$, and a probability measure $\mathcal{P}$ on $\mathcal{X}$, we say a set $N \subseteq \mathcal{X}$ is an ε-net of $\mathcal{P}$ for $\mathcal{T}$ if $N \cap A \neq \emptyset$ for every $A \in \mathcal{T}$ with $\mathcal{P}(A) > \varepsilon$ (Haussler and Welzl, 1987). Also, a finite set $\mathcal{H}$ of classifiers is called an ε-cover of $\mathbb{C}$ (under the $\mathcal{P}(\mathrm{DIS}(\{\cdot,\cdot\}))$ pseudometric) if $\sup_{g \in \mathbb{C}} \min_{h \in \mathcal{H}} \mathcal{P}(x : h(x) \neq g(x)) \leq \varepsilon$.

The following lemma bounds the probabilities and empirical probabilities of sets in a collection in terms of each other. This result is based on the work of Vapnik and Chervonenkis (1974) (see also Vapnik, 1982, Theorem A.3); this version is taken from Bousquet, Boucheron, and Lugosi (2004, Theorem 7), in combination with the VC-Sauer Lemma (Vapnik and Chervonenkis, 1971; Sauer, 1972) and a union bound.

Lemma 18 *For any collection $\mathcal{T}$ of measurable subsets of $\mathcal{X}$, letting k denote the VC dimension of $\mathcal{T}$, for any $\delta \in (0,1)$, for any integer $m > k$, for any probability measure $\mathcal{P}$ over $\mathcal{X}$, if $X'_1, \ldots, X'_m$ are independent $\mathcal{P}$-distributed random variables, then with probability at least $1-\delta$, it holds that $\forall A \in \mathcal{T}$, letting $\hat{\mathcal{P}}(A) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_A(X'_i)$,*

$$\mathcal{P}(A) \leq \hat{\mathcal{P}}(A) + 2\sqrt{\mathcal{P}(A)\frac{k\mathrm{Log}\left(\frac{2em}{k}\right) + \mathrm{Log}\left(\frac{8}{\delta}\right)}{m}}$$

$$\text{and } \hat{\mathcal{P}}(A) \leq \mathcal{P}(A) + 2\sqrt{\hat{\mathcal{P}}(A)\frac{k\mathrm{Log}\left(\frac{2em}{k}\right) + \mathrm{Log}\left(\frac{8}{\delta}\right)}{m}}.$$

In particular, with a bit of algebra, this implies the following corollary.

Corollary 19 *There exists a finite universal constant $c_0 \geq 1$ such that, for any collection $\mathcal{T}$ of measurable subsets of $\mathcal{X}$, letting k denote the VC dimension of $\mathcal{T}$, for any $\varepsilon, \delta \in (0,1)$, for any integer $m \geq \frac{c_0}{\varepsilon}\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$, for any probability measure $\mathcal{P}$ over $\mathcal{X}$, if $X'_1, \ldots, X'_m$ are independent $\mathcal{P}$-distributed random variables, then with probability at least $1-\delta$, it holds that $\forall A \in \mathcal{T}$, letting $\hat{\mathcal{P}}(A) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_A(X'_i)$,*

- $\hat{\mathcal{P}}(A) \leq \frac{3}{4}\varepsilon \implies \mathcal{P}(A) < \varepsilon$.
- $\mathcal{P}(A) \leq \frac{1}{2}\varepsilon \implies \hat{\mathcal{P}}(A) < \frac{3}{4}\varepsilon$.

Proof Let $\mathcal{E}(m) = 4\frac{k\mathrm{Log}\left(\frac{2em}{k}\right) + \mathrm{Log}\left(\frac{8}{\delta}\right)}{m}$, and note that for $m \geq \frac{c_0}{\varepsilon}\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$,

$$\mathcal{E}(m) \leq \frac{4\varepsilon}{c_0}\frac{k\mathrm{Log}\left(\frac{2ec_0}{k\varepsilon}\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right) + \mathrm{Log}\left(\frac{8}{\delta}\right)}{k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)} \tag{3}$$

If $k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) \geq \mathrm{Log}\left(\frac{1}{\delta}\right)$, then

$$\begin{aligned}
&k\mathrm{Log}\left(\frac{2ec_0}{k\varepsilon}\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right) + \mathrm{Log}\left(\frac{8}{\delta}\right) \\
&\leq k\mathrm{Log}\left(\frac{4ec_0}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right) + \mathrm{Log}\left(\frac{8}{\delta}\right) \leq k\mathrm{Log}\left(\frac{4ec_0}{\varepsilon^2}\right) + \mathrm{Log}\left(\frac{8}{\delta}\right) \\
&\leq 2k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + k\mathrm{Log}(4ec_0) + \mathrm{Log}(8) + \mathrm{Log}\left(\frac{1}{\delta}\right) \leq \mathrm{Log}\left(32e^3c_0\right)\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).
\end{aligned}$$

Otherwise, if $k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) < \mathrm{Log}\left(\frac{1}{\delta}\right)$, then

$$\begin{aligned}
&k\mathrm{Log}\left(\frac{2ec_0}{k\varepsilon}\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right) + \mathrm{Log}\left(\frac{8}{\delta}\right) \\
&\leq k\mathrm{Log}\left(\frac{4ec_0}{k\varepsilon}\mathrm{Log}\left(\frac{1}{\delta}\right)\right) + \mathrm{Log}\left(\frac{8}{\delta}\right) \leq k\mathrm{Log}\left(\frac{4ec_0}{\varepsilon}\right) + k\mathrm{Log}\left(\frac{1}{k}\mathrm{Log}\left(\frac{1}{\delta}\right)\right) + \mathrm{Log}\left(\frac{8}{\delta}\right),
\end{aligned}$$

and since $x \mapsto x\mathrm{Log}\left(\frac{1}{x}\mathrm{Log}\left(\frac{1}{\delta}\right)\right)$ is nondecreasing for $x > 0$, and $k \leq k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) \leq \mathrm{Log}\left(\frac{1}{\delta}\right)$, the above is at most

$$\begin{aligned}
&k\mathrm{Log}\left(\frac{4ec_0}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right) + \mathrm{Log}\left(\frac{8}{\delta}\right) \\
&\leq k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + k\mathrm{Log}(4ec_0) + \mathrm{Log}(8) + 2\mathrm{Log}\left(\frac{1}{\delta}\right) \leq \mathrm{Log}\left(32e^2c_0\right)\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).
\end{aligned}$$

In either case, we have that the right hand side of (3) is at most $\frac{4\varepsilon}{c_0}\mathrm{Log}\left(32e^3c_0\right)$. In particular, taking $c_0 = 2^{14}$ suffices to make $\frac{4}{c_0}\mathrm{Log}\left(32e^3c_0\right) \leq \frac{1}{64}$, so that (3) implies $\mathcal{E}(m) \leq \frac{\varepsilon}{64}$.

Lemma 18 implies that with probability at least $1-\delta$, every $A \in \mathcal{T}$ has

$$\mathcal{P}(A) \leq \hat{\mathcal{P}}(A) + \sqrt{\mathcal{P}(A)\mathcal{E}(m)}$$

and

$$\hat{\mathcal{P}}(A) \leq \mathcal{P}(A) + \sqrt{\hat{\mathcal{P}}(A)\mathcal{E}(m)}.$$

Solving these quadratic expressions in $\sqrt{\mathcal{P}(A)}$ and $\sqrt{\hat{\mathcal{P}}(A)}$, respectively, we have

$$\mathcal{P}(A) \leq \hat{\mathcal{P}}(A) + \frac{1}{2}\mathcal{E}(m) + \frac{1}{2}\sqrt{\mathcal{E}(m)^2 + 4\mathcal{E}(m)\hat{\mathcal{P}}(A)} \tag{4}$$

and

$$\hat{\mathcal{P}}(A) \leq \mathcal{P}(A) + \frac{1}{2}\mathcal{E}(m) + \frac{1}{2}\sqrt{\mathcal{E}(m)^2 + 4\mathcal{E}(m)\mathcal{P}(A)}. \tag{5}$$

Therefore, if $\hat{\mathcal{P}}(A) \leq \frac{3}{4}\varepsilon$, then (4) implies

$$\begin{aligned}
\mathcal{P}(A) &\leq \frac{3}{4}\varepsilon + \frac{1}{2}\mathcal{E}(m) + \frac{1}{2}\sqrt{\mathcal{E}(m)^2 + 3\mathcal{E}(m)\varepsilon} \\
&\leq \left(\frac{3}{4} + \frac{1}{128} + \frac{1}{2}\sqrt{\frac{1}{64^2} + \frac{3}{64}}\right)\varepsilon < \left(\frac{3}{4} + \frac{1}{128} + \frac{1}{8}\right)\varepsilon < \varepsilon,
\end{aligned}$$

and likewise, if $\mathcal{P}(A) \leq \frac{1}{2}\varepsilon$, then (5) implies

$$\begin{aligned}\hat{\mathcal{P}}(A) &\leq \frac{1}{2}\varepsilon + \frac{1}{2}\mathcal{E}(m) + \frac{1}{2}\sqrt{\mathcal{E}(m)^2 + 2\mathcal{E}(m)\varepsilon} \\ &\leq \left(\frac{1}{2} + \frac{1}{128} + \frac{1}{2}\sqrt{\frac{1}{64^2} + \frac{1}{32}}\right)\varepsilon < \left(\frac{1}{2} + \frac{1}{128} + \frac{1}{8}\right)\varepsilon < \frac{3}{4}\varepsilon.\end{aligned}$$

■

We will be interested in applying these results to the collection of sets $\{\mathrm{DIS}(\{h,g\}) : h, g \in \mathbb{C}\}$. For this, the following lemma of Vidyasagar (2003, Theorem 4.5) will be useful.

Lemma 20 *The VC dimension of the collection* $\{\mathrm{DIS}(\{h,g\}) : h, g \in \mathbb{C}\}$ *is at most* $10d$.

Together, these results imply the following lemma (see also Vapnik and Chervonenkis, 1974; Vapnik, 1982; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Haussler and Welzl, 1987).

Lemma 21 *There exists a finite universal constant* $c \geq 1$ *such that, for any* $\varepsilon, \delta \in (0,1)$*, for any integer* $m \geq \frac{c}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$*, for any probability measure* $\mathcal{P}$ *over* $\mathcal{X}$*, if* $X'_1, \ldots, X'_m$ *are independent* $\mathcal{P}$*-distributed random variables, then with probability at least* $1-\delta$*, it holds that* $\forall h, g \in \mathbb{C}$*, if* $(g(X'_1), \ldots, g(X'_m)) = (h(X'_1), \ldots, h(X'_m))$*, then* $\mathcal{P}(x : g(x) \neq h(x)) \leq \varepsilon$*.*
In particular, this implies that with probability at least $1-\delta$*, letting* $\mathbb{C}[(X'_1, \ldots, X'_m)]$ *be as in Section 7.3,* $\mathbb{C}[(X'_1, \ldots, X'_m)]$ *is an* ε*-cover of* $\mathbb{C}$ *(under the* $\mathcal{P}(\mathrm{DIS}(\{\cdot,\cdot\}))$ *pseudometric), and* $\{X'_1, \ldots, X'_m\}$ *is an* ε*-net of* $\mathcal{P}$ *for* $\{\mathrm{DIS}(\{h,g\}) : h, g \in \mathbb{C}\}$*.*

Proof Let c_0 be as in Corollary 19, and let k denote the VC dimension of $\{\mathrm{DIS}(\{h,g\}) : h, g \in \mathbb{C}\}$. Corollary 19 implies that, if $m \geq \frac{c_0}{\varepsilon}\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$, then there is an event E of probability at least $1-\delta$, on which every $h, g \in \mathbb{C}$ with $\sum_{t=1}^{m} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_t) = 0$ satisfy $\mathcal{P}(\mathrm{DIS}(\{h,g\})) < \varepsilon$; in particular, this proves that on the event E, $\{X'_1, \ldots, X'_m\}$ is an ε-net of $\mathcal{P}$ for $\{\mathrm{DIS}(\{h,g\}) : h, g \in \mathbb{C}\}$. Furthermore, by definition of $\mathbb{C}[(X'_1, \ldots, X'_m)]$, for every $h \in \mathbb{C}$, $\exists g \in \mathbb{C}[(X'_1, \ldots, X'_m)]$ with $\sum_{t=1}^{m} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_t) = 0$, which (on the event E) therefore also satisfies $\mathcal{P}(\mathrm{DIS}(\{h,g\})) < \varepsilon$. Thus, on the event E, $\mathbb{C}[(X'_1, \ldots, X'_m)]$ is an ε-cover of $\mathbb{C}$ (under the $\mathcal{P}(\mathrm{DIS}(\{\cdot,\cdot\}))$ pseudometric). To complete the proof, we note that Lemma 20 implies $k \leq 10d$, so that by choosing $c = 10c_0$, the condition $m \geq \frac{c_0}{\varepsilon}\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$ will be satisfied for any $m \geq \frac{c}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$. ■

Based on this result, it is straightforward to construct an ε-net of $\mathcal{P}$ for $\{\mathrm{DIS}(\{h,g\}) : h, g \in \mathbb{C}\}$ of size $\lesssim \frac{d}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$, based on a relatively small number of random samples. Specifically, we have the following lemma.

Lemma 22 *There exists a finite universal constant* $c' \geq 1$ *such that, for any probability measure* $\mathcal{P}$ *on* $\mathcal{X}$*, if* $X'_1, X'_2, \ldots$ *are independent* $\mathcal{P}$*-distributed random variables, then* $\forall \varepsilon, \delta \in (0,1)$*, for any integers* $m \geq \frac{c'd}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ *and* $\ell \geq \frac{c'}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$*, defining* $N_i = \{X'_{m(i-1)+1}, \ldots, X'_{mi}\}$

for each $i \in \{1, \ldots, \lceil \log_2(2/\delta) \rceil\}$*, letting*

$$\hat{i} = \underset{i \in \{1, \ldots, \lceil \log_2(2/\delta) \rceil\}}{\operatorname{argmin}} \max \left\{ \sum_{j=m\lceil \log_2(2/\delta)\rceil+1}^{m\lceil \log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) : \right. \\ \left. h, g \in \mathbb{C}, \sum_{j=m(i-1)+1}^{mi} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) = 0 \right\},$$

and $\hat{N} = N_{\hat{i}}$*, with probability at least* $1 - \delta$*,* $\hat{N}$ *is an* ε*-net of* $\mathcal{P}$ *for* $\{\mathrm{DIS}(\{h, g\}) : h, g \in \mathbb{C}\}$.

Proof Let k denote the VC dimension of the collection of sets $\{\mathrm{DIS}(\{h, g\}) : h, g \in \mathbb{C}\}$. Letting c_0 be as in Corollary 19, taking $c' \geq 10c_0$, we have $\ell \geq \frac{c_0}{\varepsilon}\left(10d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{2}{\delta}\right)\right)$, which is at least $\frac{c_0}{\varepsilon}\left(k\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{2}{\delta}\right)\right)$ by Lemma 20. Therefore, Corollary 19 implies there exists an event E' of probability at least $1 - \delta/2$ such that, on E', $\forall h, g \in \mathbb{C}$,

$$\sum_{m\lceil \log_2(2/\delta)\rceil+1}^{m\lceil \log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) \leq \frac{3}{4}\varepsilon\ell \implies \mathcal{P}(\mathrm{DIS}(\{h, g\})) \leq \varepsilon, \tag{6}$$

$$\mathcal{P}(\mathrm{DIS}(\{h, g\})) \leq \frac{\varepsilon}{2} \implies \sum_{m\lceil \log_2(2/\delta)\rceil+1}^{m\lceil \log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) \leq \frac{3}{4}\varepsilon\ell. \tag{7}$$

Let c be as in Lemma 21. Taking $c' \geq 6c$, we have $m \geq \frac{2c}{\varepsilon}\left(d\mathrm{Log}\left(\frac{2}{\varepsilon}\right) + \mathrm{Log}\left(2\right)\right)$, so that Lemma 21 implies that, for each $i \in \{1, \ldots, \lceil \log_2(2/\delta) \rceil\}$, N_i is an $\frac{\varepsilon}{2}$-net of $\mathcal{P}$ for $\{\mathrm{DIS}(\{h, g\}) : h, g \in \mathbb{C}\}$ with probability at least $1/2$. Since the N_i sets are independent, there is an event E of probability at least $1 - (1 - 1/2)^{\lceil \log_2(2/\delta) \rceil} \geq 1 - \delta/2$, on which $\exists i^* \in \{1, \ldots, \lceil \log_2(2/\delta) \rceil\}$ such that N_{i^*} is an $\frac{\varepsilon}{2}$-net of $\mathcal{P}$ for $\{\mathrm{DIS}(\{h, g\}) : h, g \in \mathbb{C}\}$. In particular, this implies that on E,

$$\sup \left\{ \mathcal{P}(\mathrm{DIS}(\{h, g\})) : h, g \in \mathbb{C}, \sum_{j=m(i^*-1)+1}^{mi^*} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) = 0 \right\} \leq \frac{\varepsilon}{2}. \tag{8}$$

Therefore, on the event $E' \cap E$, we have

$$\max \left\{ \sum_{j=m\lceil \log_2(2/\delta)\rceil+1}^{m\lceil \log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) : h, g \in \mathbb{C}, \sum_{j=m(\hat{i}-1)+1}^{m\hat{i}} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) = 0 \right\}$$
$$\leq \max \left\{ \sum_{j=m\lceil \log_2(2/\delta)\rceil+1}^{m\lceil \log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) : h, g \in \mathbb{C}, \sum_{j=m(i^*-1)+1}^{mi^*} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) = 0 \right\} \leq \frac{3}{4}\varepsilon\ell,$$

where the first inequality is by definition of $\hat{i}$, and the second inequality is by a combination of (8) with (7). Therefore, by (6), on the event $E' \cap E$, we have

$$\max \left\{ \mathcal{P}(\mathrm{DIS}(\{h, g\})) : h, g \in \mathbb{C}, \sum_{j=m(\hat{i}-1)+1}^{m\hat{i}} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_j) = 0 \right\} \leq \varepsilon,$$

or equivalently, $N_{\hat{i}}$ is an ε-net of $\mathcal{P}$ for $\{\mathrm{DIS}(\{h,g\}) : h, g \in \mathbb{C}\}$. To complete the proof, we take $c' = \max\{10c_0, 6c\}$, and note that the event $E' \cap E$ has probability at least $1-\delta$ by a union bound. ■

There are also variants of the above two lemmas applicable to sample compression schemes. Specifically, the next lemma is due to Littlestone and Warmuth (1986); Floyd and Warmuth (1995).

Lemma 23 *There exists a finite universal constant $\tilde{c} \geq 1$ such that, for any collection $\mathcal{T}$ of measurable subsets of $\mathcal{X}$, any $n \in \mathbb{N} \cup \{0\}$, and any function $\phi_n : \mathcal{X}^n \to \mathcal{T}$, for any $\varepsilon, \delta \in (0,1)$, for any integer $m \geq \frac{\tilde{c}}{\varepsilon}\left(n\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$, for any probability measure $\mathcal{P}$ over $\mathcal{X}$, if $X'_1, \ldots, X'_m$ are independent $\mathcal{P}$-distributed random variables, then with probability at least $1-\delta$, it holds that every $i_1, \ldots, i_n \in \{1, \ldots, m\}$ with $i_1 \leq \cdots \leq i_n$ and $\{X'_1, \ldots, X'_m\} \cap \phi_n(X'_{i_1}, \ldots, X'_{i_n}) = \emptyset$ has $\mathcal{P}\left(\phi_n(X'_{i_1}, \ldots, X'_{i_n})\right) \leq \varepsilon$: that is, $\{X'_1, \ldots, X'_m\}$ is an ε-net of $\mathcal{P}$ for $\{\phi_n(X'_{i_1}, \ldots, X'_{i_n}) : i_1, \ldots, i_n \in \{1, \ldots, m\}, i_1 \leq \cdots \leq i_n\}$.*

This implies the following result.

Lemma 24 *There exists a finite universal constant $\tilde{c}' \geq 1$ such that, for any collection $\mathcal{T}$ of measurable subsets of $\mathcal{X}$, any $n \in \mathbb{N}$, and any function $\phi_n : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{T}$, for any probability measure $\mathcal{P}$ on $\mathcal{X}$, if $X'_1, X'_2, \ldots$ are independent $\mathcal{P}$-distributed random variables, then for any $\varepsilon, \delta \in (0,1)$, for any integers $m \geq \frac{\tilde{c}' n}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ and $\ell \geq \frac{\tilde{c}'}{\varepsilon}\left(n\mathrm{Log}\left(\frac{m}{n}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$, defining $N_i = \{X'_{m(i-1)+1}, \ldots, X'_{mi}\}$ for each $i \in \{1, \ldots, \lceil \log_2(2/\delta) \rceil\}$, letting*

$$\hat{i} = \underset{i \in \{1, \ldots, \lceil \log_2(2/\delta) \rceil\}}{\operatorname{argmin}} \max \left\{ \sum_{j=m\lceil \log_2(2/\delta)\rceil+1}^{m\lceil \log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n)}(X'_j) : y_1, \ldots, y_n \in \mathcal{Y}, \right.$$
$$\left. m(i-1) < i_1 \leq \cdots \leq i_n \leq mi, \sum_{j=m(i-1)+1}^{mi} \mathbb{1}_{\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n)}(X'_j) = 0 \right\} \cup \{0\},$$

and $\hat{N} = N_{\hat{i}}$, with probability at least $1-\delta$, $\hat{N}$ is an ε-net of $\mathcal{P}$ for $\{\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n) : m(\hat{i}-1) < i_1 \leq \cdots \leq i_n \leq m\hat{i}, y_1, \ldots, y_n \in \mathcal{Y}\}$.

Proof Let $\tilde{c}$ be as in Lemma 23, define $\tilde{c}' = \max\{8\tilde{c}, 128\}$, and let m and ℓ be as described in the lemma statement. Noting that $\frac{2\tilde{c}}{\varepsilon}\left(n\mathrm{Log}\left(\frac{2}{\varepsilon}\right) + \mathrm{Log}\left(2^{n+1}\right)\right) \leq \frac{8\tilde{c}n}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$, we have that $m \geq \frac{2\tilde{c}}{\varepsilon}\left(n\mathrm{Log}\left(\frac{2}{\varepsilon}\right) + \mathrm{Log}\left(2^{n+1}\right)\right)$. Thus, by Lemma 23, for each $i \in \{1, \ldots, \lceil \log_2(2/\delta) \rceil\}$ and $y_1, \ldots, y_n \in \mathcal{Y}$, with probability at least $1 - 2^{-n-1}$, $\left\{X'_{m(i-1)+1}, \ldots, X'_{mi}\right\}$ is an $\frac{\varepsilon}{2}$-net of $\mathcal{P}$ for $\left\{\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n) : m(i-1) < i_1 \leq \cdots \leq i_n \leq mi\right\}$. By a union bound, this holds simultaneously for all $y_1, \ldots, y_n \in \mathcal{Y}$ with probability at least $\frac{1}{2}$. In particular, since the $\left\{X'_{m(i-1)+1}, \ldots, X'_{mi}\right\}$ subsequences are independent over values of i, we have that there is an event E of probability at least $1 - \left(\frac{1}{2}\right)^{\lceil \log_2(2/\delta) \rceil} \geq 1 - \frac{\delta}{2}$, on which $\exists i^* \in \{1, \ldots, \lceil \log_2(2/\delta) \rceil\}$ such that $\left\{X'_{m(i^*-1)+1}, \ldots, X'_{mi^*}\right\}$ is an $\frac{\varepsilon}{2}$-net of $\mathcal{P}$ for $\left\{\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n) : m(i^*-1) < i_1 \leq \cdots \leq i_n \leq mi^*, y_1, \ldots, y_n \in \mathcal{Y}\right\}$.

For any $i \in \{1, \ldots, \lceil \log_2(2/\delta) \rceil\}$, any $i_1, \ldots, i_n \in \{m(i-1)+1, \ldots, mi\}$ with $i_1 \leq \cdots \leq i_n$, and any $y_1, \ldots, y_n \in \mathcal{Y}$, Chernoff bounds (applied under the conditional distribution given

$X'_{i_1}, \ldots, X'_{i_n}$) and the law of total probability imply that, with probability at least $1-\exp\{-\varepsilon\ell/32\}$, if $\mathcal{P}\left(\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n)\right) \leq \frac{\varepsilon}{2}$, then

$$\sum_{j=m\lceil\log_2(2/\delta)\rceil+1}^{m\lceil\log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\phi_n(X'_{i_1},\ldots,X'_{i_n},y_1,\ldots,y_n)}(X'_j) \leq \frac{3}{4}\varepsilon\ell,$$

while if $\mathcal{P}\left(\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n)\right) > \varepsilon$, then

$$\sum_{j=m\lceil\log_2(2/\delta)\rceil+1}^{m\lceil\log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\phi_n(X'_{i_1},\ldots,X'_{i_n},y_1,\ldots,y_n)}(X'_j) > \frac{3}{4}\varepsilon\ell.$$

The number of distinct nondecreasing sequences $(i_1, \ldots, i_n) \in \{m(i-1)+1, \ldots, mi\}^n$ is $\binom{n+m-1}{n}$ $\leq \left(\frac{2em}{n}\right)^n$. Therefore, by a union bound, there exists an event E' of probability at least

$$1 - 2^n \left(\frac{2em}{n}\right)^n \lceil\log_2(2/\delta)\rceil \exp\{-\varepsilon\ell/32\},$$

on which this holds for every such $y_1, \ldots, y_n, i, i_1, \ldots, i_n$. Noting that

$$\frac{32}{\varepsilon}\mathrm{Log}\left(2^n\lceil\log_2(2/\delta)\rceil\left(\frac{2em}{n}\right)^n\frac{2}{\delta}\right) \leq \frac{128}{\varepsilon}\left(n\mathrm{Log}\left(\frac{m}{n}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \leq \ell,$$

we have that E' has probability at least $1 - \frac{\delta}{2}$.

In particular, defining for each $i \in \{1, \ldots, \lceil\log_2(2/\delta)\rceil\}$,

$$\hat{p}_i = \max\left\{\sum_{j=m\lceil\log_2(2/\delta)\rceil+1}^{m\lceil\log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\phi_n(X'_{i_1},\ldots,X'_{i_n},y_1,\ldots,y_n)}(X'_j) : y_1, \ldots, y_n \in \mathcal{Y},\right.$$
$$\left. m(i-1) < i_1 \leq \cdots \leq i_n \leq mi, \sum_{j=m(i-1)+1}^{mi} \mathbb{1}_{\phi_n(X'_{i_1},\ldots,X'_{i_n},y_1,\ldots,y_n)}(X'_j) = 0\right\} \cup \{0\},$$

we have that, on $E \cap E'$, $\hat{p}_{i^*} \leq \frac{3}{4}\varepsilon\ell$. Furthermore, for every $i \in \{1, \ldots, \lceil\log_2(2/\delta)\rceil\}$ for which $\left\{X'_{m(i-1)+1}, \ldots, X'_{mi}\right\}$ is *not* an ε-net of $\mathcal{P}$ for $\left\{\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n) : m(i-1) < i_1 \leq \cdots \leq i_n \leq mi, y_1, \ldots, y_n \in \mathcal{Y}\right\}$, by definition $\exists i_1, \ldots, i_n \in \{m(i-1)+1, \ldots, mi\}$ with $i_1 \leq \cdots \leq i_n$, and $y_1, \ldots, y_n \in \mathcal{Y}$, such that $\mathcal{P}\left(\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n)\right) > \varepsilon$ while $\sum_{j=m(i-1)+1}^{mi} \mathbb{1}_{\phi_n(X'_{i_1},\ldots,X'_{i_n},y_1,\ldots,y_n)}(X'_j) = 0$; thus, on the event E',

$$\sum_{j=m\lceil\log_2(2/\delta)\rceil+1}^{m\lceil\log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\phi_n(X'_{i_1},\ldots,X'_{i_n},y_1,\ldots,y_n)}(X'_j) > \frac{3}{4}\varepsilon\ell$$

for this choice of $i_1, \ldots, i_n, y_1, \ldots, y_n$. In particular, this implies that $\hat{p}_i > \frac{3}{4}\varepsilon\ell$. Altogether, we have that on the event $E \cap E'$, any such i has $\hat{p}_{\hat{i}} \leq \hat{p}_{i^*} \leq \frac{3}{4}\varepsilon\ell < \hat{p}_i$, so that $\hat{i} \neq i$. Therefore, on the event $E \cap E'$, $\left\{X'_{m(\hat{i}-1)+1}, \ldots, X'_{m\hat{i}}\right\}$ is an ε-net of $\mathcal{P}$ for $\{\phi_n(X'_{i_1}, \ldots, X'_{i_n}, y_1, \ldots, y_n) : m(\hat{i}-1) < i_1 \leq \cdots \leq i_n \leq m\hat{i}, y_1, \ldots, y_n \in \mathcal{Y}\}$.

To complete the proof, we note that the event $E \cap E'$ has probability at least $1 - \delta$ by a union bound. ■

A.2 Lower Bound Constructions for Noisy Settings

Fix any $\zeta \in (0, 1]$, $\beta \in [0, 1/2)$, and $k \in \mathbb{N}$ with $k \leq 1/\zeta$. Let $\mathcal{X}_k = \{x_1, \ldots, x_{k+1}\}$ be any $k + 1$ distinct elements of $\mathcal{X}$ (assuming $|\mathcal{X}| \geq k + 1$), and let $\mathbb{C}_k = \{x \mapsto 2\mathbb{1}_{\{x_i\}}(x) - 1 : i \in \{1, \ldots, k\}\}$, a set of functions mapping $\mathcal{X}$ to $\{-1, +1\}$. Let $\mathcal{P}_{k,\zeta}$ be a probability measure over $\mathcal{X}$ with $\mathcal{P}_{k,\zeta}(\{x_i\}) = \zeta$ for each $i \in \{1, \ldots, k\}$, and $\mathcal{P}_{k,\zeta}(\{x_{k+1}\}) = 1 - \zeta k$. For each $t \in \{1, \ldots, k\}$, let $P'_{k,\zeta,t}$ denote the probability measure over $\mathcal{X} \times \mathcal{Y}$ having marginal distribution $\mathcal{P}_{k,\zeta}$ over $\mathcal{X}$, such that if $(X, Y) \sim P'_{k,\zeta,t}$, then every $i \in \{1, \ldots, k\}$ has $\mathbb{P}(Y = 2\mathbb{1}_{\{x_t\}}(X) - 1 | X = x_i) = 1 - \beta$, and furthermore $\mathbb{P}(Y = -1 | X = x_{k+1}) = 1$. Finally, define $\mathrm{RR}'(k, \zeta, \beta) = \left\{P'_{k,\zeta,t} : t \in \{1, \ldots, k\}\right\}$. Raginsky and Rakhlin (2011) prove the following result (see the proof of their Theorem 2).[13]

Lemma 25 *For ζ, β, k as above, if $k \geq 2$ and $\mathbb{C}_k \subseteq \mathbb{C}$, then for any $\delta \in (0, 1/4)$,*

$$\Lambda_{\mathrm{RR}'(k,\zeta,\beta)}((\zeta/2)(1 - 2\beta), \delta) \geq \frac{\beta k \ln\left(\frac{1}{4\delta}\right)}{3(1 - 2\beta)^2}.$$

This has the following immediate implication for general $\mathcal{X}$ and $\mathbb{C}$. Fix any $\zeta \in (0, 1]$ and $\beta \in [0, 1/2)$, let $k \in \mathbb{N} \cup \{0\}$ satisfy $k \leq \min\{\mathfrak{s} - 1, \lfloor 1/\zeta \rfloor\}$, and let $x_1, \ldots, x_{k+1}$ and $h_0, h_1, \ldots, h_k$ be as in Definition 2. Let $\mathcal{P}_{k,\zeta}$ be as above (for this choice of $x_1, \ldots, x_{k+1}$), and for each $t \in \{1, \ldots, k\}$, let $P_{k,\zeta,t}$ denote the probability measure over $\mathcal{X} \times \mathcal{Y}$ having marginal distribution $\mathcal{P}_{k,\zeta}$ over $\mathcal{X}$, such that if $(X, Y) \sim P_{k,\zeta,t}$, then every $i \in \{1, \ldots, k\}$ has $\mathbb{P}(Y = h_t(X) | X = x_i) = 1 - \beta$, and furthermore $\mathbb{P}(Y = h_t(X) | X = x_{k+1}) = 1$. Define $\mathrm{RR}(k, \zeta, \beta) = \{P_{k,\zeta,t} : t \in \{1, \ldots, k\}\}$. We have the following result.

Lemma 26 *For k, ζ, β as above, for any $\delta \in (0, 1/4)$,*

$$\Lambda_{\mathrm{RR}(k,\zeta,\beta)}((\zeta/2)(1 - 2\beta), \delta) \geq \frac{\beta(k-1) \ln\left(\frac{1}{4\delta}\right)}{3(1 - 2\beta)^2}.$$

Proof First note that if $k \leq 1$, then the lemma trivially holds (since $\Lambda_{\mathrm{RR}(k,\zeta,\beta)}(\cdot, \cdot) \geq 0$). For this same reason, the result also trivially holds if $\beta = 0$. Otherwise, suppose $k \geq 2$ and $\beta > 0$, and fix any n less than the right hand side of the above inequality. Let $\mathcal{A}$ be any active learning algorithm, and consider the following modification $\mathcal{A}'$ of $\mathcal{A}$. For any given sequence $X_1, X_2, \ldots$ of unlabeled data, $\mathcal{A}'(n)$ simulates the execution of $\mathcal{A}(n)$, except that when $\mathcal{A}(n)$ would request the label Y_i of a point X_i in the sequence, $\mathcal{A}'(n)$ requests the label Y_i, but proceeds as $\mathcal{A}(n)$ would if the label value

13. Technically, the proof of Raginsky and Rakhlin (2011, Theorem 2) relies on a lemma (their Lemma 4), with various conditions on both k and a parameter "d" in their construction. However, one can easily verify that the conclusions of that lemma continue to hold (in fact, with improved constants) in our special case (corresponding to $d = 1$ and arbitrary $k \in \mathbb{N}$) by defining $\mathcal{M}_{k,1} = \{0, 1\}^k_1$ in their construction.

had been $-Y_i h_0(X_i)$ instead of Y_i. When the simulation of $\mathcal{A}(n)$ concludes, if $\hat{h}$ is its return value, $\mathcal{A}'(n)$ instead returns the function $x \mapsto \hat{h}'(x) = -\hat{h}(x)h_0(x)$.

Now fix a $P'_{k,\zeta,t} \in \mathrm{RR}'(k,\zeta,\beta)$ minimizing the probability that $\mathrm{er}_{P'_{k,\zeta,t}}(\hat{h}') - \inf_{h\in\mathbb{C}_k} \mathrm{er}_{P'_{k,\zeta,t}}(h) \leq (\zeta/2)(1-2\beta)$ when $\mathcal{A}'$ is run with $\mathcal{P}_{XY} = P'_{k,\zeta,t}$, and let $(X,Y) \sim P'_{k,\zeta,t}$. Note that the marginal distribution of $P'_{k,\zeta,t}$ over $\mathcal{X}$ is $\mathcal{P}_{k,\zeta}$, that for any $i \in \{1,\ldots,k\}$, $\mathbb{P}(-Yh_0(X) = h_t(X)|X = x_i) = \mathbb{P}(Y = 2\mathbb{1}_{\{x_t\}}(X) - 1|X = x_i) = 1-\beta$, and that $\mathbb{P}(-Yh_0(X) = h_t(X)|X = x_{k+1}) = \mathbb{P}(Y = -1|X = x_{k+1}) = 1$. In particular, this implies $(X, -Yh_0(X)) \sim P_{k,\zeta,t}$. Therefore, running the active learning algorithm $\mathcal{A}'(n)$ with a sequence $(X_1,Y_1),(X_2,Y_2),\ldots$ of independent $P'_{k,\zeta,t}$-distributed samples, the algorithm behaves as $\mathcal{A}(n)$ would under $P_{k,\zeta,t}$, except that its returned classifier is $\hat{h}'$ instead of $\hat{h}$. Next, note that

$$\begin{aligned}
\mathrm{er}_{P'_{k,\zeta,t}}(\hat{h}') &= \mathbb{P}(-\hat{h}(X)h_0(X) \neq Y) \\
&= \mathbb{E}[\mathbb{P}(\hat{h}(X) \neq -Y|X)\mathbb{1}[h_0(X) = 1] + \mathbb{P}(\hat{h}(X) \neq Y|X)\mathbb{1}[h_0(X) = -1]] \\
&= \mathbb{P}(\hat{h}(X) \neq -Yh_0(X)) = \mathrm{er}_{P_{k,\zeta,t}}(\hat{h}),
\end{aligned}$$

and furthermore

$$\inf_{h\in\mathbb{C}_k} \mathrm{er}_{P'_{k,\zeta,t}}(h) = \mathrm{er}_{P'_{k,\zeta,t}}(2\mathbb{1}_{\{x_t\}} - 1) = \beta\zeta k = \mathrm{er}_{P_{k,\zeta,t}}(h_t) = \inf_{h\in\mathbb{C}} \mathrm{er}_{P_{k,\zeta,t}}(h).$$

Thus, if $\mathrm{er}_{P_{k,\zeta,t}}(\hat{h}) - \inf_{h\in\mathbb{C}} \mathrm{er}_{P_{k,\zeta,t}}(h) \leq (\zeta/2)(1-2\beta)$, then we must also have $\mathrm{er}_{P'_{k,\zeta,t}}(\hat{h}') - \inf_{h\in\mathbb{C}_k} \mathrm{er}_{P'_{k,\zeta,t}}(h) \leq (\zeta/2)(1-2\beta)$. Since $n < \frac{\beta k \ln\left(\frac{1}{4\delta}\right)}{3(1-2\beta)^2}$, Lemma 25 implies that (for this choice of $P'_{k,\zeta,t}$) $\mathcal{A}'(n)$ achieves the latter guarantee with probability strictly less than $1-\delta$, and therefore the corresponding $P_{k,\zeta,t} \in \mathrm{RR}(k,\zeta,\beta)$ is such that $\mathcal{A}(n)$ has probability strictly less than $1-\delta$ of achieving $\mathrm{er}_{P_{k,\zeta,t}}(\hat{h}) - \inf_{h\in\mathbb{C}} \mathrm{er}_{P_{k,\zeta,t}}(h) \leq (\zeta/2)(1-2\beta)$. Since this argument applies to any active learning algorithm $\mathcal{A}$, the result follows. ■

A.3 Finite Approximation of VC Classes

For a given probability measure $\mathcal{P}$ over $\mathcal{X}$, Adams and Nobel (2012) have proven that for any $\tau > 0$, if $d < \infty$, there exist disjoint measurable sets $A_1,\ldots,A_k$ (for some $k \in \mathbb{N}$) with $\bigcup_i A_i = \mathcal{X}$ such that, $\forall h \in \mathbb{C}$, $\mathcal{P}\left(\bigcup\{A_i : \exists x,y \in A_i \text{ s.t. } h(x) \neq h(y)\}\right) < \tau$: that is, every $h \in \mathbb{C}$ is constant on all of the sets A_i, except a few of them whose total probability is at most τ. This property has implications for bracketing behavior in VC classes, and was proven in the context of establishing uniform laws of large numbers for VC classes under stationary ergodic processes (see also Adams and Nobel, 2010; van Handel, 2013).

For our purposes, this result has the appealing feature that it allows one to effectively *discretize* the space $\mathcal{X}$ by partitioning it into subsets, with the guarantee that with high probability over the random choice of a point x, any other point y in the same cell in the partition as x will have $f^{\star}_{\mathcal{P}_{XY}}(x) = f^{\star}_{\mathcal{P}_{XY}}(y)$, for any $\mathcal{P}_{XY} \in \bigcup_{\nu\in[0,1/2)} \mathrm{BE}(\nu)$. However, before we can make use of this property, we must first address the fact that the construction of these sets A_i by Adams and Nobel (2012) requires a strong dependence on $\mathcal{P}$, to the extent that it is not obvious that this dependence can be supplanted by a data-dependent construction. However, it turns out that if we relax the requirement that the classifiers be *constant* in these cells, instead settling for being *nearly-constant*, then it is straightforward to construct a partition $A_1, \ldots, A_k$ satisfying the requirement. Specifically, we have the following result.

Lemma 27 *Fix any $\tau, \delta \in (0,1)$, and let $m_{\tau,\delta} = \left\lceil \frac{c}{\tau}\left(d\mathrm{Log}\left(\frac{1}{\tau}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right\rceil$ (for c as in Lemma 21). For any probability measure $\mathcal{P}$ over $\mathcal{X}$, if $X'_1, \ldots, X'_{m_{\tau,\delta}}$ are independent $\mathcal{P}$-distributed random variables, then with probability at least $1-\delta$, letting $\mathbb{C}_{\tau,\delta} = \mathbb{C}[(X'_1, \ldots, X'_{m_{\tau,\delta}})]$ (as defined in Section 7.3), the collection of disjoint sets*

$$J_{\tau,\delta} = \left\{ \bigcap_{g \in \mathbb{C}[(X'_1,\ldots,X'_{m_{\tau,\delta}})]} \mathcal{X}_g : \forall g \in \mathbb{C}_{\tau,\delta}, \mathcal{X}_g \in \{\{x : g(x) = +1\}, \{x : g(x) = -1\}\} \right\}$$

is a partition of $\mathcal{X}$ with the property that, $\forall h \in \mathbb{C}$,

$$\sum_{A \in J_{\tau,\delta}} \min_{y \in \mathcal{Y}} \mathcal{P}(x \in A : h(x) = y) \le \tau,$$

and $\forall \varepsilon > 0$, $\forall h \in \mathbb{C}$,

$$\mathcal{P}\left(\bigcup \left\{ A \in J_{\tau,\delta} : \min_{y \in \mathcal{Y}} \mathcal{P}(x \in A : h(x) = y) > \varepsilon \mathcal{P}(A) \right\} \right) \le \frac{\tau}{\varepsilon}.$$

Proof By Lemma 21, with probability at least $1-\delta$, $\mathbb{C}_{\tau,\delta}$ is a τ-cover of $\mathbb{C}$. Furthermore, note that for every $g \in \mathbb{C}_{\tau,\delta}$ and every $A \in J_{\tau,\delta}$, either every $x \in A$ has $g(x) = +1$ or every $x \in A$ has $g(x) = -1$ (i.e., g is constant on A). Therefore, $\forall h \in \mathbb{C}$,

$$\begin{aligned}
\sum_{A \in J_{\tau,\delta}} \min_{y \in \mathcal{Y}} \mathcal{P}(x \in A : h(x) = y) &\le \sum_{A \in J_{\tau,\delta}} \min_{g \in \mathbb{C}_{\tau,\delta}} \mathcal{P}(x \in A : h(x) \neq g(x)) \\
&\le \min_{g \in \mathbb{C}_{\tau,\delta}} \sum_{A \in J_{\tau,\delta}} \mathcal{P}(x \in A : h(x) \neq g(x)) = \min_{g \in \mathbb{C}_{\tau,\delta}} \mathcal{P}(x : h(x) \neq g(x)) \le \tau.
\end{aligned}$$

The final claim follows by Markov's inequality, since on the above event, $\forall \varepsilon > 0$, $\forall h \in \mathbb{C}$,

$$\begin{aligned}
&\mathcal{P}\left(\bigcup \left\{ A \in J_{\tau,\delta} : \min_{y \in \mathcal{Y}} \mathcal{P}(x \in A : h(x) = y) > \varepsilon \mathcal{P}(A) \right\} \right) \\
&= \mathcal{P}\left(\bigcup \left\{ A \in J_{\tau,\delta} : \mathcal{P}(A) > 0, \min_{y \in \mathcal{Y}} \mathcal{P}(x \in A : h(x) = y) > \varepsilon \mathcal{P}(A) \right\} \right) \\
&= \mathcal{P}\left(\bigcup \left\{ A \in J_{\tau,\delta} : \mathcal{P}(A) > 0, \min_{y \in \mathcal{Y}} \frac{\mathcal{P}(x \in A : h(x) = y)}{\mathcal{P}(A)} > \varepsilon \right\} \right) \\
&\le \frac{1}{\varepsilon} \sum_{A \in J_{\tau,\delta}} \mathcal{P}(A) \min_{y \in \mathcal{Y}} \frac{\mathcal{P}(x \in A : h(x) = y)}{\mathcal{P}(A)} = \frac{1}{\varepsilon} \sum_{A \in J_{\tau,\delta}} \min_{y \in \mathcal{Y}} \mathcal{P}(x \in A : h(x) = y) \le \frac{\tau}{\varepsilon}.
\end{aligned}$$

■

Appendix B. Proofs for Results in Section 5

This section provides proofs of the main results of this article.

B.1 The Realizable Case

We begin with the particularly-simple case of Theorem 3.

Proof of Theorem 3 The lower bounds proportional to d and $\mathrm{Log}\left(\min\left\{\frac{1}{\varepsilon}, |\mathbb{C}|\right\}\right)$ are due to Kulkarni, Mitter, and Tsitsiklis (1993) (lower bound in terms of the covering numbers) in conjunction with Kulkarni (1989); Kulkarni, Mitter, and Tsitsiklis (1993) (lower bounds on the worst-case covering numbers). Specifically, Kulkarni, Mitter, and Tsitsiklis (1993) study the problem of learning from arbitrary binary-valued queries. Since active learning receives binary responses in the binary classification setting, it is a special case of this type of algorithm. In particular, for any probability measure $\mathcal{P}$ over $\mathcal{X}$, and $\varepsilon \in (0,1)$, let $\mathcal{N}(\varepsilon, \mathbb{C}, \mathcal{P})$ denote the minimum cardinality $|\mathcal{H}|$ over all ε-covers $\mathcal{H}$ of $\mathbb{C}$ (under the $\mathcal{P}(\mathrm{DIS}(\{\cdot,\cdot\}))$ pseudometric), or else $\mathcal{N}(\varepsilon, \mathbb{C}, \mathcal{P}) = \infty$ if no finite ε-cover of $\mathbb{C}$ exists. Then the lower bound of Kulkarni, Mitter, and Tsitsiklis (1993, Theorem 3) implies that, $\forall \varepsilon, \delta \in (0, 1/2)$,

$$\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \geq \sup_{\mathcal{P}} \left\lceil \log_2 \left((1-\delta) \mathcal{N}(2\varepsilon, \mathbb{C}, \mathcal{P}) \right) \right\rceil. \tag{9}$$

Furthermore, the construction in the proof of Kulkarni, Mitter, and Tsitsiklis (1993, Lemma 2) implies that $\sup_{\mathcal{P}} \mathcal{N}(2\varepsilon, \mathbb{C}, \mathcal{P}) \geq \min\left\{\left\lfloor \frac{1}{4\varepsilon} \right\rfloor, |\mathbb{C}|\right\}$, so that combined with (9), we have

$$\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \geq \left\lceil \log_2 \left((1-\delta) \min\left\{ \left\lfloor \frac{1}{4\varepsilon} \right\rfloor, |\mathbb{C}| \right\} \right) \right\rceil.$$

For $\delta \in (0, 1/3)$ and $\varepsilon \in (0, 1/8)$, and since $|\mathbb{C}| \geq 3$ (by assumption, intended to focus on nontrivial cases to simplify the expressions), the right hand side is at least $\frac{1}{4}\mathrm{Log}\left(\min\left\{\frac{1}{\varepsilon}, |\mathbb{C}|\right\}\right)$. Furthermore, if $d < 162$, this already implies that for any $\varepsilon \in (0, 1/3)$ and $\delta \in (0, 1/3)$, $\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \geq \frac{1}{4}\ln(3) \geq \frac{d}{648}$. Otherwise, in the case that $d \geq 162$, Kulkarni (1989, Proposition 3) proves that, if $\varepsilon \in (0, 1/9)$, $\sup_{\mathcal{P}} \mathcal{N}(2\varepsilon, \mathbb{C}, \mathcal{P}) \geq \exp\left\{2\left(\frac{1}{2} - 4\varepsilon\right)^2 d\right\} \geq \exp\{d/162\}$. Combined with (9), this implies that for $\varepsilon \in (0, 1/9)$ and $\delta \in (0, 1/3)$, if $d \geq 162$, then

$$\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \geq \left\lceil \log_2 \left(\frac{2}{3} e^{d/162} \right) \right\rceil \geq \frac{d}{162} \log_2(e) - \log_2\left(\frac{3}{2}\right) \geq \frac{d}{162} \log_2\left(\frac{2e}{3}\right) \geq \frac{d}{189}.$$

Thus, regardless of the value of d, we have $\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \geq \frac{d}{648}$.

For the final part of the proof of the lower bound, a lower bound proportional to $\mathfrak{s} \wedge \frac{1}{\varepsilon}$ may be credited to Dasgupta (2005, 2004). It can be proven as follows. Let $x_1, \ldots, x_{\mathfrak{s}}$ and $h_0, h_1, \ldots, h_{\mathfrak{s}}$ be as in Definition 2, let $t = \mathfrak{s} \wedge \left\lceil \frac{1-\varepsilon}{\varepsilon} \right\rceil$, and let us restrict the discussion to those $t+1$ distributions $\mathcal{P}_{XY} \in \mathrm{RE}$ such that the marginal distribution $\mathcal{P}$ of $\mathcal{P}_{XY}$ over $\mathcal{X}$ is uniform on $\{x_1, \ldots, x_t\}$, and $f^{\star}_{\mathcal{P}_{XY}} \in \{h_0, h_1, \ldots, h_t\}$. Then for any active learning algorithm $\mathcal{A}$, for any $n \leq t/2$, let Q_i denote the (possibly random) set of (at most n) points X_i that $\mathcal{A}(n)$ requests the labels of, given that $f^{\star}_{\mathcal{P}_{XY}} = h_i$ (for $i \in \{0, \ldots, t\}$), and let $\hat{h}_i$ denote the classifier returned by $\mathcal{A}(n)$ in this case. Since the marginal distribution of $\mathcal{P}_{XY}$ over $\mathcal{X}$ is fixed to $\mathcal{P}$ for all $t+1$ of these $\mathcal{P}_{XY}$ distributions, we may consider the sequence $X_1, X_2, \ldots$ of i.i.d. $\mathcal{P}$-distributed random variables to be identical over these $t+1$ possible choices of $\mathcal{P}_{XY}$, without affecting the distributions of Q_i and $\hat{h}_i$ (see Kallenberg, 2002). Thus, we may note that $\hat{h}_i = \hat{h}_0$ whenever $x_i \notin Q_0$, since $x_i \notin Q_0$ implies that all of the labels observed by the algorithm are identical to those that would be observed if $f^\star_{\mathcal{P}_{XY}} = h_0$ instead of $f^\star_{\mathcal{P}_{XY}} = h_i$. Now, if it holds that $\mathbb{P}\left(\mathcal{P}\left(x : \hat{h}_0(x) \neq h_0(x)\right) > \varepsilon\right) \leq \delta$, then since every x_i with $i \leq t$ has $\mathcal{P}(\{x_i\}) > \varepsilon$, we have that $\mathbb{P}\left(\forall i \in \{1,\ldots,t\}, \hat{h}_0(x_i) = h_0(x_i)\right) \geq 1-\delta$. But if this holds, then it must also be true that

$$\begin{aligned}
&\max_{i\in\{1,\ldots,t\}} \mathbb{P}\left(\mathcal{P}(x : \hat{h}_i(x) \neq h_i(x)) > \varepsilon\right) \geq \frac{1}{t}\sum_{i=1}^{t} \mathbb{P}\left(\mathcal{P}(x : \hat{h}_i(x) \neq h_i(x)) > \varepsilon\right) \\
&\geq \frac{1}{t}\sum_{i=1}^{t} \mathbb{P}\left(\hat{h}_i(x_i) = h_0(x_i)\right) = \frac{1}{t}\mathbb{E}\left[\sum_{i=1}^{t} \mathbb{1}\left[\hat{h}_i(x_i) = h_0(x_i)\right]\right] \\
&\geq \frac{1}{t}\mathbb{E}\left[\sum_{i=1}^{t} \mathbb{1}\left[x_i \notin Q_0\right] \mathbb{1}\left[\hat{h}_i(x_i) = h_0(x_i)\right]\right] = \frac{1}{t}\mathbb{E}\left[\sum_{i=1}^{t} \mathbb{1}\left[x_i \notin Q_0\right] \mathbb{1}\left[\hat{h}_0(x_i) = h_0(x_i)\right]\right] \\
&\geq \frac{1}{t}\mathbb{E}\left[\mathbb{1}\left[\forall i \in \{1,\ldots,t\}, \hat{h}_0(x_i) = h_0(x_i)\right]\sum_{i=1}^{t} \mathbb{1}\left[x_i \notin Q_0\right]\right] \\
&\geq \frac{1}{t}\mathbb{E}\left[\mathbb{1}\left[\forall i \in \{1,\ldots,t\}, \hat{h}_0(x_i) = h_0(x_i)\right](t-n)\right] \\
&= \frac{t-n}{t}\mathbb{P}\left(\forall i \in \{1,\ldots,t\}, \hat{h}_0(x_i) = h_0(x_i)\right) \geq \frac{t-n}{t}(1-\delta) \geq \frac{1-\delta}{2} \geq \frac{1}{3} > \delta.
\end{aligned}$$

Thus, when $n \leq t/2$, at least one of these $t+1$ distributions $\mathcal{P}_{XY}$ (all of which are in RE) has $\mathbb{P}\left(\mathrm{er}_{\mathcal{P}_{XY}}(\mathcal{A}(n)) > \varepsilon\right) > \delta$. Since this argument holds for any $\mathcal{A}$, we have that $\Lambda_{\mathrm{RE}}(\varepsilon,\delta) > t/2 = \frac{1}{2}\min\left\{\mathfrak{s}, \left\lceil\frac{1-\varepsilon}{\varepsilon}\right\rceil\right\} \geq \frac{4}{9}\min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}$. Combined with the lower bounds proportional d and $\mathrm{Log}\left(\min\left\{\frac{1}{\varepsilon}, |\mathbb{C}|\right\}\right)$ established above, this completes the proof of the lower bound in Theorem 3.

The proof of the upper bound is in three parts. The first part, establishing the $\frac{d}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ upper bound, is a straightforward application of Lemma 22. The second part, establishing the $\frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ upper bound, is directly based on techniques of Hanneke (2007a); Hegedüs (1995). Finally, and most involved, is the third part, establishing the $\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ upper bound. This part is partly based on a recent technique of Wiener, Hanneke, and El-Yaniv (2014) for analyzing disagreement-based active learning (which refines an earlier technique of El-Yaniv and Wiener, 2010, 2012). Here, we modify this technique by using an ε-net in place of random samples, thereby refining logarithmic factors, and entirely eliminating the dependence on δ in the label complexity.

Fix any $\varepsilon, \delta \in (0,1)$. We begin with the $\frac{d}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ upper bound. Let $m = \left\lceil\frac{c'd}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right\rceil$ and $\ell = \left\lceil\frac{c'}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right\rceil$, for c' as in Lemma 22. Define

$$\hat{i} = \operatorname*{argmin}_{i\in\{1,\ldots,\lceil\log_2(2/\delta)\rceil\}} \max_{\substack{h,g\in\mathbb{C}: \\ \sum_{j=m(i-1)+1}^{mi} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X_j)=0}} \sum_{j=m\lceil\log_2(2/\delta)\rceil+1}^{m\lceil\log_2(2/\delta)\rceil+\ell} \mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X_j).$$

Consider an active learning algorithm which, given a budget $n \in \mathbb{N}$, requests the labels Y_t for $t \in \left\{m\left(\hat{i}-1\right)+1,\ldots,m\left(\hat{i}-1\right)+\min\{m,n\}\right\}$, and returns any classifier $\hat{h}_n \in \mathbb{C}$ with

$\sum_{t=m(\hat{i}-1)+1}^{m(\hat{i}-1)+\min\{m,n\}} \mathbb{1}\left[\hat{h}_n(X_t) \neq Y_t\right] = 0$ if such a classifier exists (and otherwise returns an arbitrary classifier). Note that, for $\mathcal{P}_{XY} \in \mathrm{RE}$, $\sum_{t=m(\hat{i}-1)+1}^{m(\hat{i}-1)+\min\{m,n\}} \mathbb{1}\left[f^{\star}_{\mathcal{P}_{XY}}(X_t) \neq Y_t\right] = 0$ with probability one, and since $f^{\star}_{\mathcal{P}_{XY}} \in \mathbb{C}$, the classifier $\hat{h}_n$ will have $\sum_{t=m(\hat{i}-1)+1}^{m(\hat{i}-1)+\min\{m,n\}} \mathbb{1}\left[\hat{h}_n(X_t) \neq Y_t\right] = 0$ with probability one. Furthermore, this implies $\sum_{t=m(\hat{i}-1)+1}^{m(\hat{i}-1)+\min\{m,n\}} \mathbb{1}\left[\hat{h}_n(X_t) \neq f^{\star}_{\mathcal{P}_{XY}}(X_t)\right] = 0$ with probability one. Additionally, Lemma 22 implies that, with probability at least $1-\delta$, the set $\left\{X_t : t \in \left\{m\left(\hat{i}-1\right)+1, \ldots, m\hat{i}\right\}\right\}$ is an ε-net of $\mathcal{P}$ for $\{\mathrm{DIS}(\{h,g\}) : h,g \in \mathbb{C}\}$. Since both $\hat{h}_n, f^{\star}_{\mathcal{P}_{XY}} \in \mathbb{C}$, this implies that if $n \geq m$, then with probability at least $1-\delta$, $\mathcal{P}\left(\mathrm{DIS}\left(\left\{\hat{h}_n, f^{\star}_{\mathcal{P}_{XY}}\right\}\right)\right) \leq \varepsilon$. Since $\mathcal{P}_{XY} \in \mathrm{RE}$, $\mathrm{er}_{\mathcal{P}_{XY}}\left(\hat{h}_n\right) = \mathcal{P}\left(\mathrm{DIS}\left(\left\{\hat{h}_n, f^{\star}_{\mathcal{P}_{XY}}\right\}\right)\right)$. Thus, if $n \geq m$, then with probability at least $1-\delta$, $\mathrm{er}_{\mathcal{P}_{XY}}\left(\hat{h}_n\right) \leq \varepsilon$. Since this holds for any $\mathcal{P}_{XY} \in \mathrm{RE}$, we have established that $\Lambda_{\mathrm{RE}}(\varepsilon,\delta) \leq m \leq \frac{2c'd}{\varepsilon}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$. This also completes the proof of the entire upper bound in Theorem 3 in the case $\mathfrak{s} = \infty$; for this reason, for the remainder of the proof below, we restrict our attention to the case $\mathfrak{s} < \infty$.

Next, we turn to proving the $\frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ upper bound, based on a technique of Hanneke (2007a); Hegedüs (1995) (see also Hellerstein, Pillaipakkamnatt, Raghavan, and Wilkins, 1996 for related ideas), except using an ε-net in place of the random samples used by Hanneke (2007a). Let m and $\hat{i}$ be defined as above, and denote $\mathcal{U} = \left\{X_t : t \in \left\{m\left(\hat{i}-1\right)+1, \ldots, m\hat{i}\right\}\right\}$. The technique is based on using a general algorithm for *Exact* learning with membership queries, treating $\mathcal{U}$ as the instance space, and $\mathbb{C}[\mathcal{U}]$ as the concept space (where $\mathbb{C}[\mathcal{U}]$ is as defined in Section 7.3). Specifically, for any finite set $V \subseteq \mathbb{C}$ and any $x \in \mathcal{X}$, let $h_{\mathrm{maj}(V)}(x) = \mathrm{argmax}_{y\in\mathcal{Y}} |\{h \in V : h(x) = y\}|$ (breaking ties arbitrarily); $h_{\mathrm{maj}(V)}$ is called the *majority vote classifier*. In this context, the following algorithm is due to Hegedüs (1995) (see Section 7.3 for the definition of "specifying set").

MEMB-HALVING-2
Input: label budget n
Output: classifier $\hat{h}_n$

0. $V \leftarrow \mathbb{C}[\mathcal{U}]$, $t \leftarrow 0$
1. While $|V| \geq 2$ and $t < n$
2. $\quad \hat{h} \leftarrow h_{\mathrm{maj}(V)}$
3. $\quad$ Let $k = \mathrm{TD}(\hat{h}, \mathbb{C}[\mathcal{U}], \mathcal{U})$
4. $\quad$ Let $\{X_{j_1}, \ldots, X_{j_k}\} \in \mathcal{U}^k$ be a minimal specifying set for $\hat{h}$ on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$
5. $\quad$ Repeat
6. $\quad\quad$ Let $\hat{j} = \underset{j\in\{j_1,\ldots,j_k\}}{\mathrm{argmin}} |\{g \in V : g(X_j) = \hat{h}(X_j)\}|$
7. $\quad\quad$ Request the label $Y_{\hat{j}}$, let $t \leftarrow t+1$
8. $\quad\quad$ $V \leftarrow \{h \in V : h(X_{\hat{j}}) = Y_{\hat{j}}\}$
9. $\quad$ Until $\hat{h}(X_{\hat{j}}) \neq Y_{\hat{j}}$ or $|V| \leq 1$ or $t = n$
10. Return any $\hat{h}_n$ in V (or $\hat{h}_n$ arbitrary if $V = \emptyset$)

Fix any $\mathcal{P}_{XY} \in \mathrm{RE}$, and note that we have $f^{\star}_{\mathcal{P}_{XY}} \in \mathbb{C}$, so that $\exists h^* \in \mathbb{C}[\mathcal{U}]$ with $h^*(x) = f^{\star}_{\mathcal{P}_{XY}}(x)$, $\forall x \in \mathcal{U}$. Since $Y_j = f^{\star}_{\mathcal{P}_{XY}}(X_j)$ for every j with probability one in this case, we have that with probability one the set V will be nonempty in Step 10, so that $\hat{h}_n$ is chosen from V; in particular, we have $h^*(X_j) = Y_j$ for every $X_j \in \mathcal{U}$, and hence $h^* \in V$ in Step 10. Furthermore, when this is the case, Hegedüs (1995) proves that, letting $\mathrm{XTD}(\mathbb{C}[\mathcal{U}],\mathcal{U}) = \max_{h:\mathcal{X}\to\mathcal{Y}} \mathrm{TD}(h,\mathbb{C}[\mathcal{U}],\mathcal{U})$ (see Section 7.3), if

$$n \geq 2\frac{\mathrm{XTD}(\mathbb{C}[\mathcal{U}],\mathcal{U})}{1 \vee \log_2(\mathrm{XTD}(\mathbb{C}[\mathcal{U}],\mathcal{U}))} \log_2(|\mathbb{C}[\mathcal{U}]|),$$

then the classifier $\hat{h}_n$ returned by MEMB-HALVING-2 satisfies $\hat{h}_n = h^*$, so that $\hat{h}_n(x) = f^{\star}_{\mathcal{P}_{XY}}(x)$ for every $x \in \mathcal{U}$.[14] Since $\mathrm{XTD}(\mathbb{C}[\mathcal{U}],\mathcal{U}) \leq \mathrm{XTD}(\mathbb{C}, m)$, and Theorem 13 implies $\mathrm{XTD}(\mathbb{C}, m) = \mathfrak{s} \wedge m \leq \mathfrak{s}$, and since $\mathrm{Log}(\mathrm{XTD}(\mathbb{C}[\mathcal{U}],\mathcal{U})) \leq 1 \vee \log_2(\mathrm{XTD}(\mathbb{C}[\mathcal{U}],\mathcal{U}))$ and $x \mapsto \frac{x}{\mathrm{Log}(x)}$ is nondecreasing on $\mathbb{N} \cup \{0\}$, and the VC-Sauer Lemma (Vapnik and Chervonenkis, 1971; Sauer, 1972) implies $|\mathbb{C}[\mathcal{U}]| \leq \left(\frac{em}{d}\right)^d$, we have that for any $n \geq 2\frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})} \log_2\left(\frac{em}{d}\right)$, if $\forall j, f^{\star}_{\mathcal{P}_{XY}}(X_j) = Y_j$, then $\hat{h}_n(x) = f^{\star}_{\mathcal{P}_{XY}}(x)$ for every $x \in \mathcal{U}$. Thus, for $n \geq 2\frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})} \log_2\left(\frac{em}{d}\right)$, with probability one the classifier $\hat{h}_n$ returned by MEMB-HALVING-2 has $\hat{h}_n(x) = f^{\star}_{\mathcal{P}_{XY}}(x)$ for every $x \in \mathcal{U}$. Furthermore, as proven above, with probability at least $1-\delta$, $\mathcal{U}$ is an ε-net of $\mathcal{P}$ for $\{\mathrm{DIS}(\{h,g\}) : h, g \in \mathbb{C}\}$. Thus, since $f^{\star}_{\mathcal{P}_{XY}}, \hat{h}_n \in \mathbb{C}$, by a union bound we have that for any $n \geq 2\frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})} \log_2\left(\frac{em}{d}\right)$, with probability at least $1-\delta$, $\mathcal{P}(\mathrm{DIS}(\{f^{\star}_{\mathcal{P}_{XY}}, \hat{h}_n\})) \leq \varepsilon$. Since $\mathcal{P}_{XY} \in \mathrm{RE}$, this implies $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) = \mathcal{P}(\mathrm{DIS}(\{f^{\star}_{\mathcal{P}_{XY}}, \hat{h}_n\})) \leq \varepsilon$ as well. Thus, since this reasoning holds for any $\mathcal{P}_{XY} \in \mathrm{RE}$, we have established that

$$\Lambda_{\mathrm{RE}}(\varepsilon,\delta) \leq 2\frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})} \log_2\left(\frac{em}{d}\right) \leq 16\mathrm{Log}\left(2ec'\right)\frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})}\mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

Finally, we establish the $\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ upper bound, as follows. Note that, since $|\mathbb{C}| \geq 2$, we must have $\mathfrak{s} \geq 1$. Fix any $\mathcal{P}_{XY} \in \mathrm{RE}$. Let $\mathcal{T} = \{\mathrm{DIS}(V_{S,h}) : S \in \bigcup_{m\in\mathbb{N}} \mathcal{X}^m, h \text{ a classifier}\}$, and for each $x_1, \ldots, x_{\mathfrak{s}} \in \mathcal{X}$ and $y_1, \ldots, y_{\mathfrak{s}} \in \mathcal{Y}$, define

$$\phi_{\mathfrak{s}}(x_1, \ldots, x_{\mathfrak{s}}, y_1, \ldots, y_{\mathfrak{s}}) = \mathrm{DIS}(\{g \in \mathbb{C} : \forall i \leq \mathfrak{s}, g(x_i) = y_i\}) \in \mathcal{T}.$$

Let $\tilde{c}'$ be as in Lemma 24, and define $\delta' = \delta/\left(2\lceil\log_2(1/\varepsilon)\rceil\right)$, $\ell = \lceil 2\tilde{c}'\left(\mathfrak{s}\mathrm{Log}(3\tilde{c}') + \mathrm{Log}(1/\delta')\right)\rceil$, $m = \lceil 2\tilde{c}'\mathfrak{s}\rceil$, and $\tilde{j} = \lceil(2m\lceil\log_2(2/\delta')\rceil + 2\ell)/\varepsilon\rceil$. Consider the following algorithm.

14. The two cases not covered by the theorem of Hegedüs (1995) are the case $|\mathbb{C}[\mathcal{U}]| = 1$, for which the algorithm returns the sole element of $\mathbb{C}[\mathcal{U}]$ (which must agree with $f^{\star}_{\mathcal{P}_{XY}}$ on $\mathcal{U}$) without requesting any labels, and the case $|\mathbb{C}[\mathcal{U}]| = 2$, for which one can easily verify that $\mathrm{XTD}(\mathbb{C}[\mathcal{U}],\mathcal{U}) = 1$ and that the algorithm returns a classifier with the claimed property after requesting exactly one label.

Algorithm 0
Input: label budget n
Output: classifier $\hat{h}_n$

0. $V_0 \leftarrow \mathbb{C}$, $\bar{j}_0 = 0$
1. For $k = 1, 2, \ldots, \lfloor n/m \rfloor$
2. If $|\{j \in \{\bar{j}_{k-1}+1, \ldots, \bar{j}_{k-1}+\tilde{j}\} : X_j \in \mathrm{DIS}(V_{k-1})\}| < m\lceil \log_2(2/\delta') \rceil + \ell$
3. Return any $\hat{h}_n \in V_{k-1}$ (or an arbitrary classifier $\hat{h}_n$ if $V_{k-1} = \emptyset$)
4. Let $j_{k,1}, \ldots, j_{k,m\lceil \log_2(2/\delta')\rceil+\ell}$ denote the $m\lceil \log_2(2/\delta') \rceil + \ell$ smallest indices in the set $\{j \in \{\bar{j}_{k-1}+1, \ldots, \bar{j}_{k-1}+\tilde{j}\} : X_j \in \mathrm{DIS}(V_{k-1})\}$ (in increasing order)
5. Let $\bar{j}_k = j_{k,m\lceil \log_2(2/\delta')\rceil+\ell}$
6. For each $i \in \mathbb{N}$, let

$$I_i = \Bigg\{ (i_1, \ldots, i_{\mathfrak{s}}, y_1, \ldots, y_{\mathfrak{s}}) \in \mathbb{N}^{\mathfrak{s}} \times \mathcal{Y}^{\mathfrak{s}} : m(i-1) < i_1 \leq \cdots \leq i_{\mathfrak{s}} \leq mi, \\ \sum_{t=m(i-1)+1}^{mi} \mathbb{1}_{\phi_{\mathfrak{s}}(X_{j_{k,i_1}}, \ldots, X_{j_{k,i_{\mathfrak{s}}}}, y_1, \ldots, y_{\mathfrak{s}})}(X_{j_{k,t}}) = 0 \Bigg\}$$

7. Let

$$\hat{i}_k = \operatorname*{argmin}_{i \in \{1, \ldots, \lceil \log_2(2/\delta') \rceil\}} \max_{(i_1, \ldots, i_{\mathfrak{s}}, y_1, \ldots, y_{\mathfrak{s}}) \in I_i} \sum_{t=m\lceil \log_2(2/\delta')\rceil+1}^{m\lceil \log_2(2/\delta')\rceil+\ell} \mathbb{1}_{\phi_{\mathfrak{s}}(X_{j_{k,i_1}}, \ldots, X_{j_{k,i_{\mathfrak{s}}}}, y_1, \ldots, y_{\mathfrak{s}})}(X_{j_{k,t}})$$

8. Request the label $Y_{j_{k,t}}$ for each $t \in \left\{ m\left(\hat{i}_k - 1\right) + 1, \ldots, m\hat{i}_k \right\}$
9. Let $V_k \leftarrow \left\{ g \in V_{k-1} : \forall t \in \left\{ m\left(\hat{i}_k - 1\right) + 1, \ldots, m\hat{i}_k \right\}, g(X_{j_{k,t}}) = Y_{j_{k,t}} \right\}$
10. Return any $\hat{h}_n \in V_{\lfloor n/m \rfloor}$

Fix any $k \in \{1, \ldots, \lfloor n/m \rfloor\}$. In the event that V_{k-1} is defined, let

$$M_k = \left| \left\{ j \in \left\{ \bar{j}_{k-1}+1, \ldots, \bar{j}_{k-1}+\tilde{j} \right\} : X_j \in \mathrm{DIS}(V_{k-1}) \right\} \right|.$$

By a Chernoff bound (applied under the conditional distribution given V_{k-1} and $\bar{j}_{k-1}$) and the law of total probability (integrating out V_{k-1} and $\bar{j}_{k-1}$), there is an event E'_k of probability at least $1-\delta'$, on which, if V_{k-1} is defined and satisfies

$$\mathcal{P}(\mathrm{DIS}(V_{k-1})) \geq 2\tilde{j}^{-1}\left(m\lceil \log_2(2/\delta') \rceil + \ell \right), \tag{10}$$

then $M_k \geq (1/2)\tilde{j}\mathcal{P}(\mathrm{DIS}(V_{k-1})) \geq m\lceil \log_2(2/\delta') \rceil + \ell$, in which case the algorithm will execute Steps 4-9 for this particular value of k, and in particular, the set V_k is defined. In this case, denote $\mathcal{U}_k = \left\{ X_{j_{k,t}} : t \in \left\{ m\left(\hat{i}_k - 1\right) + 1, \ldots, m\hat{i}_k \right\} \right\}$, which is well-defined in this case.

Next note that, on the event that V_{k-1} is defined, the M_k samples

$$\left\{ X_j : j \in \left\{ \bar{j}_{k-1}+1, \ldots, \bar{j}_{k-1}+\tilde{j} \right\}, X_j \in \mathrm{DIS}(V_{k-1}) \right\}$$

are conditionally independent given V_{k-1}, $\bar{j}_{k-1}$, and M_k, each having conditional distribution $\mathcal{P}(\cdot|\mathrm{DIS}(V_{k-1}))$. Thus, applying Lemma 24 under the conditional distribution given V_{k-1}, $\bar{j}_{k-1}$, and M_k, combined with the law of total probability (integrating out V_{k-1}, $\bar{j}_{k-1}$, and M_k), we have that there exists an event E_k of probability at least $1-\delta'$, on which, if V_{k-1} is defined, and $M_k \geq m\lceil \log_2(2/\delta')\rceil + \ell$, then $\mathcal{U}_k$ is a $\frac{1}{2}$-net of $\mathcal{P}(\cdot|\mathrm{DIS}(V_{k-1}))$ for

$$\left\{\phi_{\mathfrak{s}}(X_{j_{k,i_1}},\ldots,X_{j_{k,i_{\mathfrak{s}}}},y_1,\ldots,y_{\mathfrak{s}}) : m\left(\hat{i}_k - 1\right)+1 < i_1 \leq \cdots \leq i_{\mathfrak{s}} \leq m\hat{i}_k, y_1,\ldots,y_{\mathfrak{s}} \in \mathcal{Y}\right\}. \tag{11}$$

Together, we have that on $E_k \cap E'_k$, if V_{k-1} is defined and satisfies (10), then $\mathcal{U}_k$ is a $\frac{1}{2}$-net of $\mathcal{P}(\cdot|\mathrm{DIS}(V_{k-1}))$ for the collection (11).

In particular, Theorem 13 implies that, for any $x_1,\ldots,x_m \in \mathcal{X}^m$ and classifier $f \in \mathbb{C}$, $\exists i_1,\ldots,i_{\mathfrak{s}} \in \{1,\ldots,m\}$ such that $\{g \in \mathbb{C} : \forall j \leq \mathfrak{s}, g(x_{i_j}) = f(x_{i_j})\} = \{g \in \mathbb{C} : \forall i \leq m, g(x_i) = f(x_i)\}$ (see the discussion in Section 7.3.1), and since the left hand side is invariant to permutations of the i_j values, without loss of generality we may take $i_1 \leq \cdots \leq i_{\mathfrak{s}}$. This implies that on $E_k \cap E'_k$, if V_{k-1} is defined and satisfies (10), then $\exists i'_1,\ldots,i'_{\mathfrak{s}} \in \left\{m\left(\hat{i}_k-1\right)+1,\ldots,m\hat{i}_k\right\}$ with $i'_1 \leq \cdots \leq i'_{\mathfrak{s}}$ such that

$$\begin{aligned}
&\phi_{\mathfrak{s}}(X_{j_{k,i'_1}},\ldots,X_{j_{k,i'_{\mathfrak{s}}}},f(X_{j_{k,i'_1}}),\ldots,f(X_{j_{k,i'_{\mathfrak{s}}}}))\\
&= \mathrm{DIS}\left(\left\{g \in \mathbb{C} : \forall t \in \left\{m\left(\hat{i}_k-1\right)+1,\ldots,m\hat{i}_k\right\}, g(X_{j_{k,t}}) = f(X_{j_{k,t}})\right\}\right) = \mathrm{DIS}(V_{\mathcal{U}_k,f}),
\end{aligned}$$

so that

$$\begin{aligned}
&\mathrm{DIS}(V_{\mathcal{U}_k,f}) \in\\
&\quad\left\{\phi_{\mathfrak{s}}(X_{j_{k,i_1}},\ldots,X_{j_{k,i_{\mathfrak{s}}}},y_1,\ldots,y_{\mathfrak{s}}) : m\left(\hat{i}_k-1\right) < i_1 \leq \cdots \leq i_{\mathfrak{s}} \leq m\hat{i}_k, y_1,\ldots,y_{\mathfrak{s}} \in \mathcal{Y}\right\}.
\end{aligned}$$

But we certainly have $\mathrm{DIS}(V_{\mathcal{U}_k,f}) \cap \mathcal{U}_k = \emptyset$. Thus, by the $\frac{1}{2}$-net property, on the event $E_k \cap E'_k$, if V_{k-1} is defined and satisfies (10), then every $f \in \mathbb{C}$ has

$$\mathcal{P}\left(\mathrm{DIS}(V_{\mathcal{U}_k,f})\middle|\mathrm{DIS}(V_{k-1})\right) \leq \frac{1}{2}. \tag{12}$$

Also note that, since $\mathcal{P}_{XY} \in \mathrm{RE}$, we have $f^\star_{\mathcal{P}_{XY}} \in \mathbb{C}$, and furthermore that there is an event E of probability one, on which $\forall j, Y_j = f^\star_{\mathcal{P}_{XY}}(X_j)$. In particular, on E, if V_{k-1} and V_k are defined, then $V_k = V_{\mathcal{U}_k,f^\star_{\mathcal{P}_{XY}}} \cap V_{k-1}$, which implies $\mathrm{DIS}(V_k) = \mathrm{DIS}\left(V_{\mathcal{U}_k,f^\star_{\mathcal{P}_{XY}}} \cap V_{k-1}\right) \subseteq \mathrm{DIS}(V_{k-1})$. Thus, applying (12) with $f = f^\star_{\mathcal{P}_{XY}}$, we have that on the event $E \cap E_k \cap E'_k$, if V_{k-1} is defined and satisfies (10), then V_k is defined and satisfies

$$\begin{aligned}
\mathcal{P}(\mathrm{DIS}(V_k)) &= \mathcal{P}(\mathrm{DIS}(V_k)|\mathrm{DIS}(V_{k-1}))\mathcal{P}(\mathrm{DIS}(V_{k-1}))\\
&\leq \mathcal{P}\left(\mathrm{DIS}\left(V_{\mathcal{U}_k,f^\star_{\mathcal{P}_{XY}}}\right)\middle|\mathrm{DIS}(V_{k-1})\right)\mathcal{P}(\mathrm{DIS}(V_{k-1})) \leq \frac{1}{2}\mathcal{P}(\mathrm{DIS}(V_{k-1})).
\end{aligned}$$

Now suppose $\lfloor n/m \rfloor \geq \lceil \log_2(1/\varepsilon)\rceil$. Applying the above to every $k \leq \lceil \log_2(1/\varepsilon)\rceil$, we have that there exist events E'_k and E_k for each $k \in \{1,\ldots,\lceil \log_2(1/\varepsilon)\rceil\}$, each of probability at least $1-\delta'$, such that on the event $E \cap \bigcap_{k=1}^{\lceil \log_2(1/\varepsilon)\rceil} E'_k \cap E_k$, every $k \in \{1,\ldots,\lceil \log_2(1/\varepsilon)\rceil\}$ with

V_{k-1} defined either has $\mathcal{P}(\mathrm{DIS}(V_{k-1})) < 2\tilde{j}^{-1}(m\lceil\log_2(2/\delta')\rceil + \ell)$ or else V_k is defined and satisfies $\mathcal{P}(\mathrm{DIS}(V_k)) \leq \frac{1}{2}\mathcal{P}(\mathrm{DIS}(V_{k-1}))$. Since $V_0 = \mathbb{C}$ is defined, by induction we have that on the event $E \cap \bigcap_{k=1}^{\lceil\log_2(1/\varepsilon)\rceil} E'_k \cap E_k$, either some $k \in \{1, \ldots, \lceil\log_2(1/\varepsilon)\rceil\}$ has V_{k-1} defined and satisfies $\mathcal{P}(\mathrm{DIS}(V_{k-1})) < 2\tilde{j}^{-1}(m\lceil\log_2(2/\delta')\rceil + \ell)$, or else every $k \in \{1, \ldots, \lceil\log_2(1/\varepsilon)\rceil\}$ has V_k defined and satisfying $\mathcal{P}(\mathrm{DIS}(V_k)) \leq \frac{1}{2}\mathcal{P}(\mathrm{DIS}(V_{k-1}))$. In particular, in this latter case, since $\mathcal{P}(\mathrm{DIS}(V_0)) \leq 1$, by induction we have $\mathcal{P}(\mathrm{DIS}(V_{\lceil\log_2(1/\varepsilon)\rceil})) \leq 2^{-\lceil\log_2(1/\varepsilon)\rceil} \leq \varepsilon$.

Also note that $2\tilde{j}^{-1}(m\lceil\log_2(2/\delta')\rceil + \ell) \leq \varepsilon$. Thus, denoting by $\hat{k}$ the largest $k \leq \lfloor n/m\rfloor$ for which V_k is defined (which also implies V_k is defined for every $k \in \{0, \ldots, \hat{k}\}$), on the event $E \cap \bigcap_{k=1}^{\lceil\log_2(1/\varepsilon)\rceil} E'_k \cap E_k$, either some $k \leq (\hat{k}+1) \wedge \lceil\log_2(1/\varepsilon)\rceil$ has $\mathcal{P}(\mathrm{DIS}(V_{k-1})) < \varepsilon$, so that (since $k \mapsto V_k$ is nonincreasing for $k \leq \hat{k}$) $\mathcal{P}(\mathrm{DIS}(V_{\hat{k}})) \leq \mathcal{P}(\mathrm{DIS}(V_{k-1})) < \varepsilon$, or else $\hat{k} \geq \lceil\log_2(1/\varepsilon)\rceil$, so that $\mathcal{P}(\mathrm{DIS}(V_{\hat{k}})) \leq \mathcal{P}(\mathrm{DIS}(V_{\lceil\log_2(1/\varepsilon)\rceil})) \leq \varepsilon$. Thus, on the event $E \cap \bigcap_{k=1}^{\lceil\log_2(1/\varepsilon)\rceil} E'_k \cap E_k$, in any case we have $\mathcal{P}(\mathrm{DIS}(V_{\hat{k}})) \leq \varepsilon$. Furthermore, by the realizable case assumption, we have $f^{\star}_{\mathcal{P}_{XY}} \in V_0$, and if $f^{\star}_{\mathcal{P}_{XY}} \in V_{k-1}$ in Step 9, then (on the event E) $f^{\star}_{\mathcal{P}_{XY}} \in V_k$ as well. Thus, by induction, on the event E, $f^{\star}_{\mathcal{P}_{XY}} \in V_{\hat{k}}$. In particular, this also implies $V_{\hat{k}} \neq \emptyset$ on E, so that there exist valid choices of $\hat{h}_n$ in $V_{\hat{k}}$ upon reaching the "Return" step (Step 3, if $\hat{k} < \lfloor n/m\rfloor$, or Step 10, if $\hat{k} = \lfloor n/m\rfloor$). Thus, $\hat{h}_n \in V_{\hat{k}}$ as well on E, so that on the event E we have $\left\{x : \hat{h}_n(x) \neq f^{\star}_{\mathcal{P}_{XY}}(x)\right\} \subseteq \mathrm{DIS}(V_{\hat{k}})$. Therefore, on the event $E \cap \bigcap_{k=1}^{\lceil\log_2(1/\varepsilon)\rceil} E'_k \cap E_k$, we have

$$\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) = \mathcal{P}\left(x : \hat{h}_n(x) \neq f^{\star}_{\mathcal{P}_{XY}}(x)\right) \leq \mathcal{P}\left(\mathrm{DIS}\left(V_{\hat{k}}\right)\right) \leq \varepsilon.$$

Finally, by a union bound, the event $E \cap \bigcap_{k=1}^{\lceil\log_2(1/\varepsilon)\rceil} E'_k \cap E_k$ has probability at least $1 - \lceil\log_2(1/\varepsilon)\rceil 2\delta' = 1 - \delta$. Noting that the above argument holds for any $\mathcal{P}_{XY} \in \mathrm{RE}$, and that the condition $\lfloor n/m\rfloor \geq \lceil\log_2(1/\varepsilon)\rceil$ is satisfied for any $n \geq 9\tilde{c}'\mathfrak{s}\mathrm{Log}(1/\varepsilon)$, this completes the proof that $\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \leq 9\tilde{c}'\mathfrak{s}\mathrm{Log}(1/\varepsilon) \lesssim \mathfrak{s}\mathrm{Log}(1/\varepsilon)$. ■

B.2 The Noisy Cases

To extend the above ideas to noisy settings, we make use of a novel modification of a technique of Kääriäinen (2006). We first partition the data sequence into three parts. For $m \in \mathbb{N}$, let $X^1_m = X_{3(m-1)+1}$, $X^2_m = X_{3(m-1)+2}$, and let $X^3_m = X_{3m}$ and $Y^3_m = Y_{3m}$; also denote $\mathbb{X}_1 = \{X^1_m\}_{m=1}^{\infty}$, $\mathbb{X}_2 = \{X^2_m\}_{m=1}^{\infty}$, $\mathbb{X}_3 = \{X^3_m\}_{m=1}^{\infty}$, $\mathbb{Y}_3 = \{Y^3_m\}_{m=1}^{\infty}$, and $\mathcal{Z} = \{(X_m, Y_m)\}_{m=1}^{\infty}$. Additionally, it will simplify some of the proofs to further partition $\mathbb{X}_3$ and $\mathbb{Y}_3$, as follows. Fix any bijection $\phi : \mathbb{N}^2 \to \mathbb{N}$, and for each $m, \ell \in \mathbb{N}$, let $X^3_{m,\ell} = X^3_{\phi(m,\ell)}$ and $Y^3_{m,\ell} = Y^3_{\phi(m,\ell)}$.

Fix values $\varepsilon, \delta \in (0,1)$, and let $\hat{\gamma}_\varepsilon$ be a value in $[\varepsilon/2, 1]$. Let $k_\varepsilon = \lceil\log_2(8/\hat{\gamma}_\varepsilon)\rceil$, and for each $k \in \{2, \ldots, k_\varepsilon\}$, define

$$\tilde{m}_k = \left\lceil \frac{16\max\{c, 8\}k_\varepsilon}{2^k\varepsilon}\left(d\mathrm{Log}\left(\frac{2k_\varepsilon}{\varepsilon}\right) + \mathrm{Log}\left(\frac{64k_\varepsilon}{\delta}\right)\right)\right\rceil,$$

for c as in Lemma 21. Also define $\tilde{m}_{k_\varepsilon+1} = 0$, $\tilde{m} = \tilde{m}_2$. and $q_{\varepsilon,\delta} = 2 + \left\lceil 2^{2k_\varepsilon+4}\ln\left(\frac{32\tilde{m}2^{2k_\varepsilon+3}}{\delta}\right)\right\rceil$. Also, for each $m \in \{1, \ldots, \tilde{m}\}$, define $\tilde{k}_m = \max\{k \in \{2, \ldots, k_\varepsilon\} : m \leq \tilde{m}_k\}$ and let $\tilde{q}_m = 2^{3+2\tilde{k}_m}\ln(32\tilde{m}q_{\varepsilon,\delta}/\delta)$. Fix a value $\tau = \frac{\delta\varepsilon}{512\tilde{m}}$. Let $J_{\tau,\delta/2}$ be as in Lemma 27, applied to the sequence $X'_m = X^1_m$; to simplify notation, in this section we abbreviate $J = J_{\tau,\delta/2}$. Also, for each $x \in \mathcal{X}$, denote by $J(x)$ the (unique) set $A \in J$ with $x \in A$, and for each $m \in \{1, \ldots, \tilde{m}\}$, we abbreviate $J_m = J(X^2_m)$. Now consider the following algorithm.

Algorithm 1
Input: label budget n
Output: classifier $\hat{h}_n$

---

0. $V_0 \leftarrow \mathbb{C}$, $t \leftarrow 0$, $m \leftarrow 0$
1. While $t < n$ and $m < \tilde{m}$
2. $\quad m \leftarrow m + 1$
3. $\quad$ If $X^2_m \in \mathrm{DIS}(V_{m-1})$
4. $\quad\quad$ Run Subroutine 1 with arguments $(n - t, m)$;
$\quad\quad$ let (q, y) be the returned values; let $t \leftarrow t + q$
5. $\quad\quad$ If $y \neq 0$ and $\exists h \in V_{m-1}$ with $h(X^2_m) = y$
6. $\quad\quad\quad$ Let $V_m \leftarrow \{h \in V_{m-1} : h(X^2_m) = y\}$
7. $\quad\quad$ Else let $V_m \leftarrow V_{m-1}$
8. $\quad$ Else let $V_m \leftarrow V_{m-1}$
9. Return any $\hat{h}_n \in V_m$

Subroutine 1
Input: label budget n, data point index m
Output: query counter q, value y

---

0. $\sigma_{m,0} \leftarrow 0$, $q \leftarrow 0$, $\ell_{m,0} \leftarrow 0$
1. Repeat
2. $\quad$ Let $\ell_{m,q+1} \leftarrow \min\{\ell > \ell_{m,q} : X^3_{m,\ell} \in J_m\}$ (or $\ell_{m,q+1} \leftarrow 1$ if this set is empty)
3. $\quad$ Request the label $Y^3_{m,\ell_{m,q+1}}$; let $\sigma_{m,q+1} \leftarrow \sigma_{m,q} + Y^3_{m,\ell_{m,q+1}}$; let $q \leftarrow q + 1$
4. $\quad$ If $|\sigma_{m,q}| \geq 3\sqrt{2q \ln(32\tilde{m} q_{\varepsilon,\delta}/\delta)}$
5. $\quad\quad$ Return $(q, \mathrm{sign}(\sigma_{m,q}))$
6. $\quad$ Else if $q \geq \min\{n, \tilde{q}_m\}$
7. $\quad\quad$ Return $(q, 0)$

In this algorithm, the first part of the data (namely, $\mathbb{X}_1$) is used to partition the space via Lemma 27, so that each cell of the partition has $f^\star_{\mathcal{P}_{XY}}$ nearly-constant within it (assuming $f^\star_{\mathcal{P}_{XY}} \in \mathbb{C}$). The second part, $\mathbb{X}_2$, is used to simulate a commonly-studied active learning algorithm for the realizable case (namely, the algorithm of Cohn, Atlas, and Ladner, 1994), with two significant modifications. First, instead of directly requesting the label of a point, we use samples from the third part of the data (i.e., $\mathbb{X}_3$) that co-occur in the same cell of the partition as the would-be query point, repeatedly requesting for labels from that cell and using the majority vote of these returned labels in place of the label of the original point. Second, we discard a point X^2_m if we cannot identify a clear majority label within a certain number of queries, which decreases as the algorithm runs. Since this second modification often ends up rejecting more samples in cells with higher noise rates than those with lower noise rates, this effectively alters the marginal distribution over $\mathcal{X}$, shifting the distribution to favor less-noisy regions.

For the remainder of Appendix B.2, we fix an arbitrary probability measure $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$ with $f^{\star}_{\mathcal{P}_{XY}} \in \mathbb{C}$, and as usual, we denote by $\mathcal{P}(\cdot) = \mathcal{P}_{XY}(\cdot \times \mathcal{Y})$ the marginal of $\mathcal{P}_{XY}$ over $\mathcal{X}$. For any $x \in \mathcal{X}$, define $\gamma_x = \left|\eta(x; \mathcal{P}_{XY}) - \frac{1}{2}\right|$, and define

$$\gamma_{\varepsilon} = \sup\left\{\gamma \in (0, 1/2] : \gamma \mathcal{P}(x : \gamma_x \leq \gamma) \leq \varepsilon/2\right\}.$$

Also, for the remainder of Appendix B.2, we suppose $\hat{\gamma}_{\varepsilon}$ is chosen to be in the range $[\varepsilon/2, \gamma_{\varepsilon}]$. For each $A \in J$, define

$$y_A = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathcal{P}\left(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) = y\right) = \operatorname{sign}\left(\int_A f^{\star}_{\mathcal{P}_{XY}} \mathrm{d}\mathcal{P}\right),$$

and if $\mathcal{P}(A) > 0$, define $\eta(A; \mathcal{P}_{XY}) = \mathcal{P}_{XY}(A \times \{1\} | A \times \mathcal{Y})$ (i.e., the average value of $\eta(x; \mathcal{P}_{XY})$ over $x \in A$), and let $\gamma_A = \left|\eta(A; \mathcal{P}_{XY}) - \frac{1}{2}\right|$. For completeness, for any $A \in J$ with $\mathcal{P}(A) = 0$, define $\eta(A; \mathcal{P}_{XY}) = 1/2$ and $\gamma_A = 0$. Additionally, for each $n \in \mathbb{N} \cup \{\infty\}$ and $m \in \mathbb{N}$, let $(\hat{q}_{n,m}, \hat{y}_{n,m})$ denote the return values of Subroutine 1 when run with arguments (n, m).

Denote by E_1 the $\mathbb{X}_1$-measurable event of probability at least $1 - \delta/2$ implied by Lemma 27, on which every $h \in \mathbb{C}$ has

$$\sum_{A \in J} \min_{y \in \mathcal{Y}} \mathcal{P}\left(x \in A : h(x) = y\right) \leq \tau \tag{13}$$

and $\forall \gamma > 0$,

$$\mathcal{P}\left(\bigcup\left\{A \in J : \min_{y \in \mathcal{Y}} \mathcal{P}\left(x \in A : h(x) = y\right) > \gamma \mathcal{P}(A)\right\}\right) \leq \frac{\tau}{\gamma}. \tag{14}$$

We now proceed to characterize the behaviors of Subroutine 1 and Algorithm 1 via the following sequence of lemmas.

Lemma 28 *There exists a $(\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3)$-measurable event E_0 of probability 1, on which $\forall m \in \{1, \ldots, \tilde{m}\}$, $\mathcal{P}(J_m) > 0$ and $|\{\ell \in \mathbb{N} : X^3_{m,\ell} \in J_m\}| = \infty$.*

Proof For each m, since each $A \in J$ with $\mathcal{P}(A) = 0$ has $\mathbb{P}(X^2_m \in A) = 0$, and J has finite size, a union bound implies $\mathbb{P}(\mathcal{P}(J_m) = 0) = 0$. The strong law of large numbers (applied under the conditional distribution given J_m) and the law of total probability implies that $\frac{1}{\ell}\sum_{j=1}^{\ell} \mathbb{1}_{J_m}(X^3_{m,j}) \to \mathcal{P}(J_m)$ with probability 1, so that when $\mathcal{P}(J_m) > 0$, $\sum_{j=1}^{\ell} \mathbb{1}_{J_m}(X^3_{m,j}) \to \infty$. Finally, a union bound implies $\mathbb{P}(\exists m \leq \tilde{m} : \mathcal{P}(J_m) = 0$ or $|\{\ell \in \mathbb{N} : X^3_{m,\ell} \in J_m\}| < \infty) \leq \sum_{m=1}^{\tilde{m}} \mathbb{P}(\mathcal{P}(J_m) = 0) + \mathbb{P}(\mathcal{P}(J_m) > 0$ and $|\{\ell \in \mathbb{N} : X^3_{m,\ell} \in J_m\}| < \infty) = 0$. ■

Lemma 29 *There exists a $(\mathbb{X}_1, \mathbb{X}_2)$-measurable event E_2 of probability at least $1 - \tau\tilde{m} \geq 1 - \delta/512$ such that, on $E_1 \cap E_2$, every $m \in \{1, \ldots, \tilde{m}\}$ has $f^{\star}_{\mathcal{P}_{XY}}(X^2_m) = y_{J_m}$.*

Proof Noting that, on E_1, (13) implies that

$$\begin{aligned}
\mathcal{P}\left(x : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_{J(x)}\right) &= \sum_{A \in J} \mathcal{P}\left(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A\right) \\
&= \sum_{A \in J} \min_{y \in \mathcal{Y}} \mathcal{P}\left(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) = y\right) \leq \tau,
\end{aligned}$$

the result follows by a union bound. ∎

Lemma 30 *There exists a $(\mathbb{X}_1, \mathbb{X}_2)$-measurable event E_3 of probability at least $1 - \frac{128\tau}{\varepsilon}\tilde{m} \geq 1 - \delta/4$ such that, on $E_1 \cap E_3$, every $m \in \{1, \ldots, \tilde{m}\}$ has $\mathcal{P}\left(x \in J_m : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_{J_m}\right) \leq \frac{\varepsilon}{128}\mathcal{P}(J_m)$.*

Proof Noting that, on E_1, (14) implies that

$$\begin{aligned}
&\mathcal{P}\left(x : \mathcal{P}\left(x' \in J(x) : f^{\star}_{\mathcal{P}_{XY}}(x') \neq y_{J(x)}\right) > \frac{\varepsilon}{128}\mathcal{P}(J(x))\right) \\
&= \mathcal{P}\left(\bigcup\left\{A \in J : \mathcal{P}\left(x' \in A : f^{\star}_{\mathcal{P}_{XY}}(x') \neq y_A\right) > \frac{\varepsilon}{128}\mathcal{P}(A)\right\}\right) \\
&= \mathcal{P}\left(\bigcup\left\{A \in J : \min_{y \in \mathcal{Y}} \mathcal{P}\left(x' \in A : f^{\star}_{\mathcal{P}_{XY}}(x') = y\right) > \frac{\varepsilon}{128}\mathcal{P}(A)\right\}\right) \leq \frac{128\tau}{\varepsilon},
\end{aligned}$$

the result follows by a union bound. ∎

Lemma 31 $\forall A \in J$,

$$\mathcal{P}_{XY}\left(A \times \{y_A\}\right) \geq \frac{1}{2}\mathcal{P}(A) + \int_A \gamma_x \mathcal{P}(\mathrm{d}x) - \mathcal{P}\left(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A\right).$$

Proof Any $A \in J$ has

$$\begin{aligned}
\mathcal{P}_{XY}\left(A \times \{y_A\}\right) &\geq \int_A \mathbb{1}[f^{\star}_{\mathcal{P}_{XY}}(x) = y_A]\left(\frac{1}{2} + \gamma_x\right)\mathcal{P}(\mathrm{d}x) \\
&\geq \int_A \left(\frac{1}{2} + \gamma_x\right)\mathcal{P}(\mathrm{d}x) - \mathcal{P}\left(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A\right) \\
&= \frac{1}{2}\mathcal{P}(A) + \int_A \gamma_x \mathcal{P}(\mathrm{d}x) - \mathcal{P}\left(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A\right).
\end{aligned}$$

∎

Lemma 32 *On the event $E_0 \cap E_1 \cap E_3$, every $m \in \{1, \ldots, \tilde{m}\}$ with $\gamma_{J_m} > \varepsilon/128$ has $\mathcal{P}_{XY}(J_m \times \{y_{J_m}\}) > \mathcal{P}_{XY}(J_m \times \{-y_{J_m}\})$, and every $m \in \{1, \ldots, \tilde{m}\}$ with $\int_{J_m} \gamma_x \mathcal{P}(\mathrm{d}x) > (\varepsilon/2)\mathcal{P}(J_m)$ has*

$$\int_{J_m} \gamma_x \mathcal{P}(\mathrm{d}x) \geq \gamma_{J_m}\mathcal{P}(J_m) \geq \frac{63}{64}\int_{J_m} \gamma_x \mathcal{P}(\mathrm{d}x) > \frac{63}{128}\varepsilon\mathcal{P}(J_m). \tag{15}$$

Proof Jensen's inequality implies we always have $\gamma_A \mathcal{P}(A) \leq \int_A \gamma_x \mathcal{P}(\mathrm{d}x)$. In particular, this implies that any $A \in J$ with $\mathcal{P}(A) > 0$ and $\mathcal{P}(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A) \leq \frac{\varepsilon}{128}\mathcal{P}(A)$ and $\gamma_A > \varepsilon/128$ has $\int_A \gamma_x \mathcal{P}(\mathrm{d}x) - \mathcal{P}(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A) \geq \gamma_A \mathcal{P}(A) - \mathcal{P}(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A) > (\varepsilon/128)\mathcal{P}(A) - (\varepsilon/128)\mathcal{P}(A) = 0$, so that Lemma 31 implies $\mathcal{P}_{XY}(A \times \{y_A\}) > \frac{1}{2}\mathcal{P}(A)$, and therefore $\mathcal{P}_{XY}(A \times \{y_A\}) > \mathcal{P}_{XY}(A \times \{-y_A\})$. Since Lemmas 28 and 30 imply that, on

$E_0 \cap E_1 \cap E_3$, for every $m \in \{1,\ldots,\tilde{m}\}$, $\mathcal{P}(J_m) > 0$ and $\mathcal{P}(x \in J_m : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_{J_m}) \leq \frac{\varepsilon}{128}\mathcal{P}(J_m)$, we have established the first claim in the lemma statement.

For the second claim, the first inequality follows by Jensen's inequality. For the second inequality, note that any $A \in J$ has $\gamma_A \mathcal{P}(A) \geq \mathcal{P}_{XY}(A \times \{y_A\}) - \frac{1}{2}\mathcal{P}(A)$, so that Lemma 31 implies $\gamma_A\mathcal{P}(A) \geq \int_A \gamma_x \mathcal{P}(\mathrm{d}x) - \mathcal{P}(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A)$. Therefore, since Lemma 30 implies that, on $E_1 \cap E_3$, every $m \in \{1,\ldots,\tilde{m}\}$ has $\mathcal{P}(x \in J_m : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_{J_m}) \leq \frac{\varepsilon}{128}\mathcal{P}(J_m)$, we have that on $E_1 \cap E_3$, any $m \in \{1,\ldots,\tilde{m}\}$ with $\int_{J_m} \gamma_x \mathcal{P}(\mathrm{d}x) > (\varepsilon/2)\mathcal{P}(J_m)$ has $\mathcal{P}(x \in J_m : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_{J_m}) \leq \frac{1}{64}\int_{J_m} \gamma_x \mathcal{P}(\mathrm{d}x)$, so that $\gamma_{J_m}\mathcal{P}(J_m) \geq \int_{J_m} \gamma_x\mathcal{P}(\mathrm{d}x) - \mathcal{P}(x \in J_m : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_{J_m}) \geq \frac{63}{64}\int_{J_m}\gamma_x\mathcal{P}(\mathrm{d}x)$. The final inequality then follows by the assumption that $\int_{J_m}\gamma_x\mathcal{P}(\mathrm{d}x) > (\varepsilon/2)\mathcal{P}(J_m)$. ∎

Lemma 33 *On* E_1, $\forall \gamma > (1/4)\gamma_\varepsilon$,

$$\mathcal{P}\left(\bigcup\{A \in J : \gamma_A \leq \gamma\}\right) \leq 3\mathcal{P}(x : \gamma_x < 4\gamma),$$

and $\forall \gamma \in (0, (1/4)\gamma_\varepsilon]$,

$$\mathcal{P}\left(\bigcup\{A \in J : \gamma_A \leq \gamma\}\right) \leq \frac{3\varepsilon}{2\gamma_\varepsilon}.$$

Proof By Markov's inequality, for any $\gamma > 0$, any $A \in J$ with $\int_A \gamma_x\mathcal{P}(\mathrm{d}x) \leq \gamma\mathcal{P}(A)$ must have $\mathcal{P}(x \in A : \gamma_x \geq 2\gamma) \leq \frac{1}{2}\mathcal{P}(A)$, which implies $\mathcal{P}(x \in A : \gamma_x < 2\gamma) \geq \frac{1}{2}\mathcal{P}(A)$. Therefore,

$$\mathcal{P}\left(\bigcup\left\{A \in J : \int_A \gamma_x\mathcal{P}(\mathrm{d}x) \leq \gamma\mathcal{P}(A)\right\}\right)$$
$$\leq \mathcal{P}\left(\bigcup\left\{A \in J : \mathcal{P}(x \in A : \gamma_x < 2\gamma) \geq \frac{1}{2}\mathcal{P}(A)\right\}\right) \leq 2\mathcal{P}(x : \gamma_x < 2\gamma), \quad (16)$$

where the last inequality is due to Markov's inequality.

Also, for every $\gamma > 0$, since $\gamma_A\mathcal{P}(A) \geq \mathcal{P}_{XY}(A \times \{y_A\}) - \frac{1}{2}\mathcal{P}(A)$,

$$\mathcal{P}\left(\bigcup\{A \in J : \gamma_A \leq \gamma\}\right) = \mathcal{P}\left(\bigcup\{A \in J : \gamma_A\mathcal{P}(A) \leq \gamma\mathcal{P}(A)\}\right)$$
$$\leq \mathcal{P}\left(\bigcup\left\{A \in J : \mathcal{P}_{XY}(A \times \{y_A\}) - \frac{1}{2}\mathcal{P}(A) \leq \gamma\mathcal{P}(A)\right\}\right).$$

Lemma 31 implies $\mathcal{P}_{XY}(A \times \{y_A\}) - \frac{1}{2}\mathcal{P}(A) \geq \int_A \gamma_x\mathcal{P}(\mathrm{d}x) - \mathcal{P}(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A)$, so that the above is at most

$$\mathcal{P}\left(\bigcup\left\{A \in J : \int_A \gamma_x\mathcal{P}(\mathrm{d}x) \leq \gamma\mathcal{P}(A) + \mathcal{P}(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A)\right\}\right).$$

By a union bound, this is at most

$$\mathcal{P}\left(\bigcup\left\{A \in J : \int_A \gamma_x\mathcal{P}(\mathrm{d}x) \leq 2\gamma\mathcal{P}(A)\right\}\right)$$
$$+\mathcal{P}\left(\bigcup\{A \in J : \mathcal{P}(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A) > \gamma\mathcal{P}(A)\}\right). \quad (17)$$

On E_1, (14) implies that

$$\mathcal{P}\left(\bigcup\left\{A \in J : \mathcal{P}(x \in A : f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_A) > \gamma\mathcal{P}(A)\right\}\right) \leq \frac{\tau}{\gamma} < \frac{\varepsilon}{8\gamma}.$$

Furthermore, by (16),

$$\mathcal{P}\left(\bigcup\left\{A \in J : \int_A \gamma_x \mathcal{P}(\mathrm{d}x) \leq 2\gamma\mathcal{P}(A)\right\}\right) \leq 2\mathcal{P}(x : \gamma_x < 4\gamma).$$

Using these two inequalities to bound the two terms in (17), we have that

$$\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \leq \gamma\right\}\right) \leq 2\mathcal{P}(x : \gamma_x < 4\gamma) + \frac{\varepsilon}{8\gamma}.$$

By definition of γ_ε, if $\gamma > (1/4)\gamma_\varepsilon$, we must have $4\gamma\mathcal{P}(x : \gamma_x < 4\gamma) \geq \gamma_\varepsilon\mathcal{P}(x : \gamma_x \leq \gamma_\varepsilon) \geq \varepsilon/2$, so that $\frac{\varepsilon}{8\gamma} \leq \mathcal{P}(x : \gamma_x < 4\gamma)$, which implies

$$2\mathcal{P}(x : \gamma_x < 4\gamma) + \frac{\varepsilon}{8\gamma} \leq 3\mathcal{P}(x : \gamma_x < 4\gamma),$$

which establishes the first claim. On the other hand, if $0 < \gamma \leq (1/4)\gamma_\varepsilon$, we have $4\gamma\mathcal{P}(x : \gamma_x < 4\gamma) \leq \varepsilon/2$, so that $2\mathcal{P}(x : \gamma_x < 4\gamma) \leq \frac{\varepsilon}{4\gamma}$, which implies

$$2\mathcal{P}(x : \gamma_x < 4\gamma) + \frac{\varepsilon}{8\gamma} \leq \frac{3\varepsilon}{8\gamma}.$$

This establishes the second claim, since (combined with monotonicity of probabilities) it implies

$$\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \leq \gamma\right\}\right) \leq \mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \leq (1/4)\gamma_\varepsilon\right\}\right) \leq \frac{3\varepsilon}{2\gamma_\varepsilon}.$$

■

Lemma 34 *On E_1, $\forall h \in \mathbb{C}$,*

$$\mathrm{er}_{\mathcal{P}_{XY}}(h) - \mathrm{er}_{\mathcal{P}_{XY}}(f^{\star}_{\mathcal{P}_{XY}}) \leq 5\tau + \int \mathbb{1}[h(x) \neq f^{\star}_{\mathcal{P}_{XY}}(x)] 2\gamma_{J(x)}\mathcal{P}(\mathrm{d}x).$$

Proof For any $h \in \mathbb{C}$, we generally have

$$\mathrm{er}_{\mathcal{P}_{XY}}(h) - \mathrm{er}_{\mathcal{P}_{XY}}(f^{\star}_{\mathcal{P}_{XY}}) = \int \mathbb{1}[h(x) \neq f^{\star}_{\mathcal{P}_{XY}}(x)] 2\gamma_x\mathcal{P}(\mathrm{d}x).$$

For each $A \in J$, let $y^h_A = \mathrm{argmax}_{y \in \mathcal{Y}}\,\mathcal{P}(x : h(x) = y)$. $\forall x \in \mathcal{X}$, $\mathbb{1}[h(x) \neq f^{\star}_{\mathcal{P}_{XY}}(x)]2\gamma_x \leq 1$. Therefore,

$$\int \mathbb{1}[h(x) \neq f^{\star}_{\mathcal{P}_{XY}}(x)] 2\gamma_x\mathcal{P}(\mathrm{d}x) \leq \mathcal{P}\left(x : h(x) \neq y^h_{J(x)} \text{ or } f^{\star}_{\mathcal{P}_{XY}}(x) \neq y_{J(x)}\right)$$
$$+ \int_{\left\{x : h(x) = y^h_{J(x)}, f^{\star}_{\mathcal{P}_{XY}}(x) = y_{J(x)}\right\}} \mathbb{1}[y^h_{J(x)} \neq y_{J(x)}] 2\gamma_x\mathcal{P}(\mathrm{d}x). \quad (18)$$

By a union bound,

$$\mathcal{P}\left(x : h(x) \neq y^h_{J(x)} \text{ or } f^\star_{\mathcal{P}_{XY}}(x) \neq y_{J(x)}\right) \leq \mathcal{P}\left(x : h(x) \neq y^h_{J(x)}\right) + \mathcal{P}\left(x : f^\star_{\mathcal{P}_{XY}}(x) \neq y_{J(x)}\right).$$

Furthermore, on E_1, (13) implies the right hand side is at most 2τ. Combining this with (18) implies

$$\mathrm{er}_{\mathcal{P}_{XY}}(h) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq 2\tau + \int_{\left\{x : h(x) = y^h_{J(x)}, f^\star_{\mathcal{P}_{XY}}(x) = y_{J(x)}\right\}} \mathbb{1}[y^h_{J(x)} \neq y_{J(x)}] 2\gamma_x \mathcal{P}(\mathrm{d}x). \quad (19)$$

Also,

$$\begin{aligned}
&\int_{\left\{x : h(x) = y^h_{J(x)}, f^\star_{\mathcal{P}_{XY}}(x) = y_{J(x)}\right\}} \mathbb{1}[y^h_{J(x)} \neq y_{J(x)}] 2\gamma_x \mathcal{P}(\mathrm{d}x) \\
&= \sum_{A \in J : y^h_A \neq y_A} \int_{\left\{x \in A : h(x) = y^h_A, f^\star_{\mathcal{P}_{XY}}(x) = y_A\right\}} 2\gamma_x \mathcal{P}(\mathrm{d}x) \leq \sum_{A \in J : y^h_A \neq y_A} \int_{\left\{x \in A : f^\star_{\mathcal{P}_{XY}}(x) = y_A\right\}} 2\gamma_x \mathcal{P}(\mathrm{d}x).
\end{aligned}$$

Since $f^\star_{\mathcal{P}_{XY}}(x) = \mathrm{sign}(2\eta(x; \mathcal{P}_{XY}) - 1)$ for every $x \in \mathcal{X}$, any measurable $C \subseteq \mathcal{X}$ has

$$\mathcal{P}_{XY}\left((x, y) : x \in C, y = f^\star_{\mathcal{P}_{XY}}(x)\right) = \int_C \left(\frac{1}{2} + \gamma_x\right) \mathcal{P}(\mathrm{d}x).$$

Therefore, for each $A \in J$,

$$\begin{aligned}
\gamma_A \mathcal{P}(A) &\geq \mathcal{P}_{XY}(A \times \{y_A\}) - \frac{1}{2}\mathcal{P}(A) \geq \mathcal{P}_{XY}\left(\left\{x \in A : f^\star_{\mathcal{P}_{XY}}(x) = y_A\right\} \times \{y_A\}\right) - \frac{1}{2}\mathcal{P}(A) \\
&= \int_{\left\{x \in A : f^\star_{\mathcal{P}_{XY}}(x) = y_A\right\}} \left(\frac{1}{2} + \gamma_x\right) \mathcal{P}(\mathrm{d}x) - \frac{1}{2}\mathcal{P}(A) \\
&= \int_{\left\{x \in A : f^\star_{\mathcal{P}_{XY}}(x) = y_A\right\}} \gamma_x \mathcal{P}(\mathrm{d}x) - \frac{1}{2}\mathcal{P}\left(x \in A : f^\star_{\mathcal{P}_{XY}}(x) \neq y_A\right).
\end{aligned}$$

Therefore,

$$\sum_{A \in J : y^h_A \neq y_A} \int_{\left\{x \in A : f^\star_{\mathcal{P}_{XY}}(x) = y_A\right\}} 2\gamma_x \mathcal{P}(\mathrm{d}x) \leq \sum_{A \in J : y^h_A \neq y_A} \mathcal{P}\left(x \in A : f^\star_{\mathcal{P}_{XY}}(x) \neq y_A\right) + 2\gamma_A \mathcal{P}(A).$$

On E_1, (13) implies that the right hand side is at most

$$\tau + \sum_{A \in J : y^h_A \neq y_A} 2\gamma_A \mathcal{P}(A).$$

Combining this with (19), we have that on E_1,

$$\mathrm{er}_{\mathcal{P}_{XY}}(h) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq 3\tau + \sum_{A \in J : y^h_A \neq y_A} 2\gamma_A \mathcal{P}(A). \quad (20)$$

For each $A \in J$ and $x \in A$, if $y_A^h \neq y_A$, then either $h(x) \neq f^\star_{\mathcal{P}_{XY}}(x)$ holds, or else one of $h(x) \neq y_A^h$ or $f^\star_{\mathcal{P}_{XY}}(x) \neq y_A$ holds. Thus, any $A \in J$ with $y_A^h \neq y_A$ has

$$\begin{aligned}\mathcal{P}(A) &\leq \int_A \left(\mathbb{1}\left[h(x) \neq f^\star_{\mathcal{P}_{XY}}(x)\right] + \mathbb{1}\left[h(x) \neq y_A^h\right] + \mathbb{1}\left[f^\star_{\mathcal{P}_{XY}}(x) \neq y_A\right] \right) \mathcal{P}(\mathrm{d}x) \\ &= \mathcal{P}\left(x \in A : h(x) \neq y_A^h\right) + \mathcal{P}\left(x \in A : f^\star_{\mathcal{P}_{XY}}(x) \neq y_A\right) + \int_A \mathbb{1}\left[h(x) \neq f^\star_{\mathcal{P}_{XY}}(x)\right] \mathcal{P}(\mathrm{d}x).\end{aligned}$$

Combined with (20), this implies that on E_1,

$$\begin{aligned}&\mathrm{er}_{\mathcal{P}_{XY}}(h) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \\ &\quad\leq 3\tau + \sum_{A \in J : y_A^h \neq y_A} 2\gamma_A \Bigg(\mathcal{P}\left(x \in A : h(x) \neq y_A^h\right) + \mathcal{P}\left(x \in A : f^\star_{\mathcal{P}_{XY}}(x) \neq y_A\right) \\ &\qquad\qquad + \int_A \mathbb{1}\left[h(x) \neq f^\star_{\mathcal{P}_{XY}}(x)\right] \mathcal{P}(\mathrm{d}x) \Bigg).\end{aligned}$$

Since $2\gamma_A \leq 1$, the right hand side is at most

$$\begin{aligned}&3\tau + \sum_{A \in J} \mathcal{P}\left(x \in A : h(x) \neq y_A^h\right) + \sum_{A \in J} \mathcal{P}\left(x \in A : f^\star_{\mathcal{P}_{XY}}(x) \neq y_A\right) \\ &\qquad + \sum_{A \in J : y_A^h \neq y_A} 2\gamma_A \int_A \mathbb{1}\left[h(x) \neq f^\star_{\mathcal{P}_{XY}}(x)\right] \mathcal{P}(\mathrm{d}x),\end{aligned}$$

and on E_1, (13) implies this is at most

$$\begin{aligned}&5\tau + \sum_{A \in J : y_A^h \neq y_A} 2\gamma_A \int_A \mathbb{1}\left[h(x) \neq f^\star_{\mathcal{P}_{XY}}(x)\right] \mathcal{P}(\mathrm{d}x) \\ &\leq 5\tau + \sum_{A \in J} \int_A \mathbb{1}\left[h(x) \neq f^\star_{\mathcal{P}_{XY}}(x)\right] 2\gamma_A \mathcal{P}(\mathrm{d}x) = 5\tau + \int \mathbb{1}\left[h(x) \neq f^\star_{\mathcal{P}_{XY}}(x)\right] 2\gamma_{J(x)} \mathcal{P}(\mathrm{d}x).\end{aligned}$$

■

Lemma 35 *There is a $\mathcal{Z}$-measurable event E_4 of probability at least $1 - \delta/32$ such that, on $\bigcap_{j=0}^{4} E_j$, $\forall k \in \{2, \ldots, k_\varepsilon\}$, $\forall m \in \{\tilde{m}_{k+1} + 1, \ldots, \tilde{m}_k\}$, $\forall n \in \mathbb{N} \cup \{\infty\}$, $\hat{y}_{n,m} \in \{0, f^\star_{\mathcal{P}_{XY}}(X_m^2)\}$, $\hat{q}_{n,m} \leq \left\lceil \frac{8}{\max\{\gamma_{J_m}^2, 2^{-2k}\}} \ln\left(\frac{32\tilde{m} q_{\varepsilon,\delta}}{\delta}\right) \right\rceil$, and if $\gamma_{J_m} \geq 2^{-k}$ then $\hat{y}_{\infty,m} = f^\star_{\mathcal{P}_{XY}}(X_m^2)$.*

Proof Since $\hat{q}_{n,m} \leq \hat{q}_{\infty,m}$, and $\hat{y}_{n,m} = 0$ whenever $\hat{q}_{n,m} < \hat{q}_{\infty,m}$, it suffices to show the claims hold for $\hat{q}_{\infty,m}$ and $\hat{y}_{\infty,m}$ for each $m \in \{1, \ldots, \tilde{m}\}$.

For each $m \in \{1, \ldots, \tilde{m}\}$, let $\ell_{m,1}, \ell_{m,2}, \ldots$ denote the increasing infinite subsequence of values $\ell \in \mathbb{N}$ with $X_{m,\ell}^3 \in J_m$, guaranteed to exist by Lemma 28 on E_0; also, for each $q \in \mathbb{N}$, define

$\sigma_{m,q} = \sum_{j=1}^{q} Y^3_{m,\ell_{m,j}}$. Note that these definitions of $\ell_{m,q}$ and $\sigma_{m,q}$ agree with those defined in Subroutine 1 for each $q \leq \hat{q}_{\infty,m}$. Let E_4 denote the event that E_0 occurs and that $\forall m \in \{1, \ldots, \tilde{m}\}$, $\forall q \in \{1, \ldots, q_{\varepsilon,\delta}\}$,

$$|\sigma_{m,q} - q(2\eta(J_m; \mathcal{P}_{XY}) - 1)| \leq \sqrt{2q \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}. \tag{21}$$

For each $m \in \{1, \ldots, \tilde{m}\}$ and $q \in \{1, \ldots, q_{\varepsilon,\delta}\}$, Lemma 28 and Hoeffding's inequality imply that (21) holds with conditional probability (given J_m) at least $1 - \delta/(32\tilde{m}q_{\varepsilon,\delta})$. The law of total probability and a union bound over values of m and q then imply that E_4 has probability at least $1 - \delta/32$.

Now fix any $k \in \{2, \ldots, k_\varepsilon\}$ and $m \in \{\tilde{m}_{k+1} + 1, \ldots, \tilde{m}_k\}$. Since $\tilde{k}_m = k$, the condition in Step 6 guarantees $\hat{q}_{\infty,m} \leq \left\lceil 2^{2k+3} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\right\rceil$. Furthermore, if $\gamma_{J_m} \geq 2^{-k}$, then for

$$q = \left\lceil \frac{8}{\gamma^2_{J_m}} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\right\rceil,$$

we have

$$2q\gamma_m \geq 4\sqrt{2q \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}.$$

In particular, recalling that $2q\gamma_{J_m} = |q(2\eta(J_m; \mathcal{P}_{XY}) - 1)|$, we have

$$|q(2\eta(J_m; \mathcal{P}_{XY}) - 1)| \geq 4\sqrt{2q \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}. \tag{22}$$

Since $q_{\varepsilon,\delta} \geq \left\lceil 2^{2k+3} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\right\rceil \geq q$, the event E_4 implies that (21) holds, so that

$$\sigma_{m,q} \geq q(2\eta(J_m; \mathcal{P}_{XY}) - 1) - \sqrt{2q \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}.$$

Thus, if $q(2\eta(J_m; \mathcal{P}_{XY}) - 1) \geq 4\sqrt{2q \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}$, the condition in Step 4 will imply $\hat{q}_{\infty,m} \leq q$, and since $q \leq \tilde{q}_m$, that $\hat{y}_{\infty,m} \in \mathcal{Y}$. Likewise, (21) implies

$$\sigma_{m,q} \leq q(2\eta(J_m; \mathcal{P}_{XY}) - 1) + \sqrt{2q \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)},$$

so that $q(2\eta(J_m; \mathcal{P}_{XY}) - 1) \leq -4\sqrt{2q \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}$ would also suffice to imply $\hat{q}_{\infty,m} \leq q$ and $\hat{y}_{\infty,m} \in \mathcal{Y}$ via the condition in Step 4. Thus, since (22) implies one of these two conditions holds, we have that on E_4, if $\gamma_{J_m} \geq 2^{-k}$ then $\hat{q}_{\infty,m} \leq \left\lceil \frac{8}{\gamma^2_{J_m}} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\right\rceil$ and $\hat{y}_{\infty,m} \in \mathcal{Y}$.

It remains only to show that $\hat{y}_{\infty,m} \in \{0, f^\star_{\mathcal{P}_{XY}}(X^2_m)\}$. This clearly holds if the return value originates in Step 7, so we need only consider the case where Subroutine 1 reaches Step 5. Due to the condition in Step 6, this cannot occur for a value of $q > q_{\varepsilon,\delta}$ (since $\tilde{q}_m \leq \tilde{q}_1 \leq q_{\varepsilon,\delta}$), so let us consider any value of $q \in \{1,\ldots,q_{\varepsilon,\delta}\}$, and suppose $|\sigma_{m,q}| \geq 3\sqrt{2q\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}$. On the event E_4, (21) implies that if $\sigma_{m,q} \geq 3\sqrt{2q\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}$, then $q(2\eta(J_m;\mathcal{P}_{XY})-1) \geq \sigma_{m,q} - \sqrt{2q\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)} \geq 2\sqrt{2q\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)} > 0$, and likewise if $\sigma_{m,q} \leq -3\sqrt{2q\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}$, then $q(2\eta(J_m;\mathcal{P}_{XY})-1) \leq \sigma_{m,q} + \sqrt{2q\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)} \leq -2\sqrt{2q\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)} < 0$; thus, since $|2\eta(J_m;\mathcal{P}_{XY})-1| = 2\gamma_{J_m}$, if $|\sigma_{m,q}| \geq 3\sqrt{2q\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}$, then

$$\gamma_{J_m} \geq \sqrt{\frac{2}{q}\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)} \tag{23}$$

and $\mathrm{sign}(2\eta(J_m;\mathcal{P}_{XY})-1) = \mathrm{sign}(\sigma_{m,q})$. In particular, since $q \leq q_{\varepsilon,\delta} \leq 2^{2k_\varepsilon+5}\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)$, this implies

$$\gamma_{J_m} \geq \sqrt{\frac{2}{q_{\varepsilon,\delta}}\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)} \geq 2^{-k_\varepsilon-2} > \varepsilon/128.$$

Therefore, Lemma 32 implies that on $\bigcap_{j=0}^4 E_j$, $\mathrm{sign}(2\eta(J_m;\mathcal{P}_{XY})-1) = y_{J_m}$; combined with the above, this implies $\mathrm{sign}(\sigma_{m,q}) = y_{J_m}$. Furthermore, Lemma 29 implies that on $\bigcap_{j=0}^4 E_j$, $y_{J_m} = f^\star_{\mathcal{P}_{XY}}(X_m^2)$, so that $\mathrm{sign}(\sigma_{m,q}) = f^\star_{\mathcal{P}_{XY}}(X_m^2)$. In particular, recall that if $\hat{y}_{\infty,m} \in \mathcal{Y}$, then $|\sigma_{m,\hat{q}_{\infty,m}}| \geq 3\sqrt{2\hat{q}_{\infty,m}\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)}$. Thus, since the condition in Step 6 implies $\hat{q}_{\infty,m} \leq \tilde{q}_m \leq q_{\varepsilon,\delta}$, we have that on $\bigcap_{j=0}^4 E_j$, if $\hat{y}_{\infty,m} \in \mathcal{Y}$, then $\hat{y}_{\infty,m} = f^\star_{\mathcal{P}_{XY}}(X_m^2)$. This completes the proof that $\hat{y}_{\infty,m} \in \{0, f^\star_{\mathcal{P}_{XY}}(X_m^2)\}$ on $\bigcap_{j=0}^4 E_j$. Since we established above that $\hat{y}_{\infty,m} \in \mathcal{Y}$ if $\gamma_{J_m} \geq 2^{-k}$ on E_4, this also completes the proof that $\hat{y}_{\infty,m} = f^\star_{\mathcal{P}_{XY}}(X_m^2)$ when $\gamma_{J_m} \geq 2^{-k}$ on $\bigcap_{j=0}^4 E_j$. ■

Lemma 36 *There exists an $(\mathbb{X}_1,\mathbb{X}_2)$-measurable event E_5 of probability at least $1-\delta/64$ such that, on E_5, for every $k \in \{2,\ldots,k_\varepsilon\}$ with $\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k},2^{1-k}\right]\right\}\right) \geq 2^{k-3}\varepsilon/k_\varepsilon$,*

$$\left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : \gamma_{J_m} \in \left[2^{-k},2^{1-k}\right]\right\}\right| \geq (1/2)\tilde{m}_k\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k},2^{1-k}\right]\right\}\right).$$

Proof Fix any $k \in \{2,\ldots,k_\varepsilon\}$. First, note that a Chernoff bound (under the conditional distribution given J) implies that, with conditional probability (given J) at least

$$1-\exp\left\{-\frac{\tilde{m}_k}{8}\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k},2^{1-k}\right]\right\}\right)\right\},$$

we have

$$\left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : \gamma_{J_m} \in \left[2^{-k},2^{1-k}\right]\right\}\right| \geq \frac{\tilde{m}_k}{2}\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k},2^{1-k}\right]\right\}\right). \tag{24}$$

If $\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right) \geq 2^{k-3}\varepsilon/k_\varepsilon$, then

$$\begin{aligned}
&\exp\left\{-\frac{\tilde{m}_k}{8}\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right)\right\} \\
&\leq \exp\left\{-\frac{8k_\varepsilon}{2^k\varepsilon}\mathrm{Log}\left(\frac{64k_\varepsilon}{\delta}\right)2^{k-3}\varepsilon/k_\varepsilon\right\} = \exp\left\{-\mathrm{Log}\left(\frac{64k_\varepsilon}{\delta}\right)\right\} = \frac{\delta}{64k_\varepsilon}.
\end{aligned}$$

Thus, by the law of total probability, there is an event $G_5(k)$ of probability at least $1-\delta/(64k_\varepsilon)$ such that, on $G_5(k)$, if $\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right) \geq 2^{k-3}\varepsilon/k_\varepsilon$, then (24) holds. This holds for all $k \in \{2, \ldots, k_\varepsilon\}$ on the event $E_5 = \bigcap_{k=2}^{k_\varepsilon} G_5(k)$, which has probability at least $1 - \delta/64$ by a union bound. ■

We are now ready to apply the above results to characterize the behavior of Algorithm 1. For simplicity, we begin with the case of an infinite budget n, so that the algorithm proceeds until $m = \tilde{m}$; later, we discuss sufficient finite sizes of n to retain this behavior.

Lemma 37 *Consider running Algorithm 1 with budget ∞. On the event $\bigcap_{j=0}^{4} E_j$, $\forall k \in \{2, \ldots, k_\varepsilon\}$, $\forall m \in \{1, \ldots, \tilde{m}_k\}$, $f^\star_{\mathcal{P}_{XY}} \in V_m$ and*

$$V_m \subseteq \left\{h \in \mathbb{C} : \forall m' \leq m \text{ with } \gamma_{J_{m'}} \geq 2^{-k}, h(X^2_{m'}) = f^\star_{\mathcal{P}_{XY}}(X^2_{m'})\right\}.$$

Proof Fix any $k \in \{2, \ldots, k_\varepsilon\}$. We proceed by induction. The claim is clearly satisfied for $V_0 = \mathbb{C}$. Now take as the inductive hypothesis that, for some $m \in \{1, \ldots, \tilde{m}_k\}$, $f^\star_{\mathcal{P}_{XY}} \in V_{m-1} \subseteq \left\{h \in \mathbb{C} : \forall m' \leq m-1 \text{ with } \gamma_{J_{m'}} \geq 2^{-k}, h(X^2_{m'}) = f^\star_{\mathcal{P}_{XY}}(X^2_{m'})\right\}$.

If $X^2_m \notin \mathrm{DIS}(V_{m-1})$, then we have $V_m = V_{m-1}$, so that $f^\star_{\mathcal{P}_{XY}} \in V_m$ as well. Furthermore, since $f^\star_{\mathcal{P}_{XY}} \in V_{m-1}$, the fact that $X^2_m \notin \mathrm{DIS}(V_{m-1})$ implies that every $h \in V_m$ has $h(X^2_m) = f^\star_{\mathcal{P}_{XY}}(X^2_m)$. Therefore,

$$\begin{aligned}
V_m &= V_{m-1} \cap \left\{h \in \mathbb{C} : h(X^2_m) = f^\star_{\mathcal{P}_{XY}}(X^2_m)\right\} \\
&\subseteq \left\{h \in \mathbb{C} : \forall m' \leq m-1 \text{ with } \gamma_{J_{m'}} \geq 2^{-k}, h(X^2_{m'}) = f^\star_{\mathcal{P}_{XY}}(X^2_{m'})\right\} \\
&\qquad\qquad\qquad\qquad \cap \left\{h \in \mathbb{C} : h(X^2_m) = f^\star_{\mathcal{P}_{XY}}(X^2_m)\right\} \\
&\subseteq \left\{h \in \mathbb{C} : \forall m' \leq m \text{ with } \gamma_{J_{m'}} \geq 2^{-k}, h(X^2_{m'}) = f^\star_{\mathcal{P}_{XY}}(X^2_{m'})\right\}.
\end{aligned}$$

Next, consider the case that $X^2_m \in \mathrm{DIS}(V_{m-1})$. Lemma 35 implies that on $\bigcap_{j=0}^{4} E_j$, $\hat{y}_{\infty,m} \in \{0, f^\star_{\mathcal{P}_{XY}}(X^2_m)\}$. If $\hat{y}_{\infty,m} = 0$, then $V_m = V_{m-1}$, so that $f^\star_{\mathcal{P}_{XY}} \in V_m$ by the inductive hypothesis. Furthermore, since $k \leq \tilde{k}_m$, Lemma 35 implies that on $\bigcap_{j=0}^{4} E_j$, if $\gamma_{J_m} \geq 2^{-k}$ then $\hat{y}_{\infty,m} \neq 0$; thus, if $\hat{y}_{\infty,m} = 0$, we have $\gamma_{J_m} < 2^{-k}$, so that

$$\begin{aligned}
V_m = V_{m-1} &\subseteq \left\{h \in \mathbb{C} : \forall m' \leq m-1 \text{ with } \gamma_{J_{m'}} \geq 2^{-k}, h(X^2_{m'}) = f^\star_{\mathcal{P}_{XY}}(X^2_{m'})\right\} \\
&= \left\{h \in \mathbb{C} : \forall m' \leq m \text{ with } \gamma_{J_{m'}} \geq 2^{-k}, h(X^2_{m'}) = f^\star_{\mathcal{P}_{XY}}(X^2_{m'})\right\}.
\end{aligned}$$

On the other hand, if $\hat{y}_{\infty,m} = f^{\star}_{\mathcal{P}_{XY}}(X^2_m)$, then since $f^{\star}_{\mathcal{P}_{XY}} \in V_{m-1}$ by the inductive hypothesis, the condition in Step 5 will be satisfied, so that we have $V_m = \left\{h \in V_{m-1} : h(X^2_m) = f^{\star}_{\mathcal{P}_{XY}}(X^2_m)\right\}$. In particular, this implies $f^{\star}_{\mathcal{P}_{XY}} \in V_m$ as well, and combined with the inductive hypothesis, we have

$$\begin{aligned} V_m &= V_{m-1} \cap \left\{h \in \mathbb{C} : h(X^2_m) = f^{\star}_{\mathcal{P}_{XY}}(X^2_m)\right\} \\ &\subseteq \left\{h \in \mathbb{C} : \forall m' \leq m \text{ with } \gamma_{J_{m'}} \geq 2^{-k}, h(X^2_{m'}) = f^{\star}_{\mathcal{P}_{XY}}(X^2_{m'})\right\}. \end{aligned}$$

The result follows by the principle of induction. ■

In particular, this implies the following result.

Lemma 38 *There exists an event E_6 of probability at least $1 - \delta/64$ such that, on $\bigcap_{j=0}^{6} E_j$, the classifier $\hat{h}_\infty$ produced by running Algorithm 1 with budget ∞ has* $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_\infty) - \mathrm{er}_{\mathcal{P}_{XY}}(f^{\star}_{\mathcal{P}_{XY}}) \leq \varepsilon$.

Proof Fix any $k \in \{2, \ldots, k_\varepsilon\}$ and let $\hat{\ell}_k = \left\lceil (1/2)\tilde{m}_k \mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right)\right\rceil$. Note that

$$\hat{\ell}_k \geq \frac{8ck_\varepsilon \mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right)}{2^k \varepsilon}\left(d\mathrm{Log}\left(\frac{8k_\varepsilon \mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right)}{2^k\varepsilon}\right) + \mathrm{Log}\left(\frac{64k_\varepsilon}{\delta}\right)\right),$$

for c as in Lemma 21. Let $\hat{m}_k = \min\left\{m \in \mathbb{N} : \sum_{m'=1}^{m} \mathbb{1}_{\left[2^{-k}, 2^{1-k}\right]}(\gamma_{J_{m'}}) = \hat{\ell}_k\right\} \cup \{\infty\}$. Note that, if $\hat{m}_k < \infty$, then the sequence $\left\{X^2_m : 1 \leq m \leq \hat{m}_k, \gamma_{J_m} \in \left[2^{-k}, 2^{1-k}\right]\right\}$ is conditionally i.i.d. (given J and $\hat{m}_k$), with conditional distributions $\mathcal{P}\left(\cdot \middle| \bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right)$. Applying Lemma 21 to these samples implies that there exists an event of conditional probability (given J and $\hat{m}_k$) at least $1 - \delta/(64k_\varepsilon)$ on which, if $\hat{m}_k < \infty$ and $\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right) > \frac{2^k\varepsilon}{8k_\varepsilon}$, letting

$$\mathcal{H}_k = \left\{h \in \mathbb{C} : \forall m \leq \hat{m}_k \text{ with } \gamma_{J_m} \in \left[2^{-k}, 2^{1-k}\right], h(X^2_m) = f^{\star}_{\mathcal{P}_{XY}}(X^2_m)\right\},$$

every $h \in \mathcal{H}_k$ has

$$\mathcal{P}\left(x : h(x) \neq f^{\star}_{\mathcal{P}_{XY}}(x) \middle| \gamma_{J(x)} \in \left[2^{-k}, 2^{1-k}\right]\right) \leq \frac{2^k\varepsilon}{8k_\varepsilon \mathcal{P}\left(\bigcup\{A \in J : \gamma_A \in [2^{-k}, 2^{1-k}]\}\right)},$$

which implies

$$\mathcal{P}\left(x : h(x) \neq f^{\star}_{\mathcal{P}_{XY}}(x) \text{ and } \gamma_{J(x)} \in \left[2^{-k}, 2^{1-k}\right]\right) \leq \frac{2^k\varepsilon}{8k_\varepsilon}.$$

By the law of total probability and a union bound, there exists an event E_6 of probability at least $1 - \delta/64$ on which this holds for every $k \in \{2, \ldots, k_\varepsilon\}$.

Lemma 37 implies that, on $\bigcap_{j=0}^{4} E_j$, $\forall k \in \{2, \ldots, k_\varepsilon\}$,

$$V_{\tilde{m}} \subseteq V_{\tilde{m}_k} \subseteq \left\{h \in \mathbb{C} : \forall m \leq \tilde{m}_k \text{ with } \gamma_{J_m} \geq 2^{-k}, h(X^2_m) = f^{\star}_{\mathcal{P}_{XY}}(X^2_m)\right\}.$$

Lemma 36 implies that, on E_5, $\forall k \in \{2,\ldots,k_\varepsilon\}$, if $\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right) > \frac{2^k\varepsilon}{8k_\varepsilon}$, then $\left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : \gamma_{J_m} \in \left[2^{-k}, 2^{1-k}\right]\right\}\right| \geq (1/2)\tilde{m}_k \mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right)$, so that $\hat{m}_k \leq \tilde{m}_k$. In particular, this implies $\hat{m}_k < \infty$ and

$$\left\{h \in \mathbb{C} : \forall m \leq \tilde{m}_k \text{ with } \gamma_{J_m} \geq 2^{-k}, h(X_m^2) = f^\star_{\mathcal{P}_{XY}}(X_m^2)\right\} \subseteq \mathcal{H}_k.$$

Combining the above three results, we have that on $\bigcap_{j=0}^{6} E_j$, for every $k \in \{2,\ldots,k_\varepsilon\}$ with $\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right) > \frac{2^k\varepsilon}{8k_\varepsilon}$, $V_{\tilde{m}} \subseteq \mathcal{H}_k$, and therefore every $h \in V_{\tilde{m}}$ has

$$\mathcal{P}\left(x : h(x) \neq f^\star_{\mathcal{P}_{XY}}(x) \text{ and } \gamma_{J(x)} \in \left[2^{-k}, 2^{1-k}\right]\right) \leq \frac{2^k\varepsilon}{8k_\varepsilon}.$$

Furthermore, for every $k \in \{2,\ldots,k_\varepsilon\}$ with $\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right) \leq \frac{2^k\varepsilon}{8k_\varepsilon}$, we also have that every $h \in V_{\tilde{m}}$ satisfies

$$\begin{aligned}\mathcal{P}&\left(x : h(x) \neq f^\star_{\mathcal{P}_{XY}}(x) \text{ and } \gamma_{J(x)} \in \left[2^{-k}, 2^{1-k}\right]\right) \\ &\qquad\qquad \leq \mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \in \left[2^{-k}, 2^{1-k}\right]\right\}\right) \leq \frac{2^k\varepsilon}{8k_\varepsilon}.\end{aligned}$$

Combined with Lemma 34, we have that on $\bigcap_{j=0}^{6} E_j$, every $h \in V_{\tilde{m}}$ has

$$\begin{aligned}\mathrm{er}_{\mathcal{P}_{XY}}(h) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) &\leq 5\tau + \int \mathbb{1}\left[h(x) \neq f^\star_{\mathcal{P}_{XY}}(x)\right] 2\gamma_{J(x)} \mathcal{P}(\mathrm{d}x) \\ &\leq 5\tau + 2^{1-k_\varepsilon}\mathcal{P}\left(x : h(x) \neq f^\star_{\mathcal{P}_{XY}}(x) \text{ and } \gamma_{J(x)} \leq 2^{-k_\varepsilon}\right) \\ &\quad + \sum_{k=2}^{k_\varepsilon} 2^{2-k}\mathcal{P}\left(x : h(x) \neq f^\star_{\mathcal{P}_{XY}}(x) \text{ and } \gamma_{J(x)} \in \left[2^{-k}, 2^{1-k}\right]\right) \\ &\leq 5\tau + 2^{1-k_\varepsilon}\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \leq 2^{-k_\varepsilon}\right\}\right) + \sum_{k=2}^{k_\varepsilon} 2^{2-k}\frac{2^k\varepsilon}{8k_\varepsilon}. \qquad (25)\end{aligned}$$

Next, note that $\sum_{k=2}^{k_\varepsilon} 2^{2-k}\frac{2^k\varepsilon}{8k_\varepsilon} = (k_\varepsilon - 1)\frac{\varepsilon}{2k_\varepsilon} \leq \frac{\varepsilon}{2}$. Furthermore, since $2^{-k_\varepsilon} \leq \hat{\gamma}_\varepsilon/8 < \gamma_\varepsilon/4$, Lemma 33 implies that, on E_1,

$$\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \leq 2^{-k_\varepsilon}\right\}\right) \leq \frac{3\varepsilon}{2\gamma_\varepsilon}.$$

Plugging these facts into (25) reveals that, on $\bigcap_{j=0}^{6} E_j$, $\forall h \in V_{\tilde{m}}$,

$$\mathrm{er}_{\mathcal{P}_{XY}}(h) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq 5\tau + 2^{1-k_\varepsilon}\frac{3\varepsilon}{2\gamma_\varepsilon} + \frac{\varepsilon}{2} \leq 5\tau + \frac{3}{8}\varepsilon + \frac{\varepsilon}{2} \leq \frac{453}{512}\varepsilon < \varepsilon.$$

The result follows by noting that, when the budget is set to ∞, Algorithm 1 definitely reaches $m = \tilde{m}$ before halting, so that $\hat{h}_\infty \in V_{\tilde{m}}$. ■

The only remaining question is how many label requests the algorithm makes in the process of producing this $\hat{h}_\infty$, so that taking a budget n of at least this size is equivalent to having an infinite budget. This question is addressed by the following sequence of lemmas.

Lemma 39 *Consider running Algorithm 1 with budget ∞. There exists an event E_7 of probability at least $1-\delta/64$ such that, on $E_1 \cap E_7$, $\forall k \in \{2,\ldots,k_\varepsilon\}$,*

$$\left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : \gamma_{J_m} \leq 2^{1-k}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right|$$
$$\leq 17 \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{2\hat{\gamma}_\varepsilon}\right\}\tilde{m}_k.$$

Proof Fix any $k \in \{2,\ldots,k_\varepsilon\}$. By a Chernoff bound (applied under the conditional given J) and the law of total probability, there is an event $G_7(k)$ of probability at least $1-\frac{\delta}{64k_\varepsilon}$, on which

$$\left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : \gamma_{J_m} \leq 2^{1-k}\right\}\right| \leq \log_2\left(\frac{64k_\varepsilon}{\delta}\right) + 2e\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \leq 2^{1-k}\right\}\right)\tilde{m}_k.$$

Lemma 33 implies that, on E_1,

$$\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \leq 2^{1-k}\right\}\right) \leq \max\left\{3\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{3\varepsilon}{2\gamma_\varepsilon}\right\}.$$

Therefore, on $E_1 \cap G_7(k)$,

$$\left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : \gamma_{J_m} \leq 2^{1-k}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \leq \left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : \gamma_{J_m} \leq 2^{1-k}\right\}\right|$$
$$\leq \log_2\left(\frac{64k_\varepsilon}{\delta}\right) + 6e\max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{2\gamma_\varepsilon}\right\}\tilde{m}_k. \tag{26}$$

Furthermore, since $\hat{\gamma}_\varepsilon \leq \gamma_\varepsilon$, and

$$\frac{\varepsilon}{2\hat{\gamma}_\varepsilon}\tilde{m}_k \geq \frac{64}{2^{k_\varepsilon}\hat{\gamma}_\varepsilon}\mathrm{Log}\left(\frac{64k_\varepsilon}{\delta}\right) \geq 4\mathrm{Log}\left(\frac{64k_\varepsilon}{\delta}\right) \geq 2\log_2\left(\frac{64k_\varepsilon}{\delta}\right),$$

(26) is at most

$$\left(6e+\frac{1}{2}\right)\max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{2\hat{\gamma}_\varepsilon}\right\}\tilde{m}_k \leq 17\max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{2\hat{\gamma}_\varepsilon}\right\}\tilde{m}_k.$$

Defining $E_7 = \bigcap_{k=2}^{k_\varepsilon} G_7(k)$, a union bound implies E_7 has probability at least $1-\delta/64$, and the result follows. ■

Lemma 40 *Consider running Algorithm 1 with budget ∞. There exists an event E_8 of probability at least $1-\delta/64$ such that, on $E_8 \cap \bigcap_{j=0}^{4} E_j$, $\forall \bar{k} \in \{3,\ldots,k_\varepsilon\}$, $\forall k \in \{2,\ldots,\bar{k}-1\}$,*

$$\left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right|$$
$$\leq 6e\max\left\{\mathcal{P}\left(x : \gamma_x < 2^{2-\bar{k}}\right), \frac{\varepsilon}{2\gamma_\varepsilon}\right\}\tilde{m}_k$$
$$+ 91\tilde{c}\left(2^{1+\bar{k}-k} + \mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\right)\left(6\mathfrak{s}\mathrm{Log}\left(\frac{128c}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right),$$

for c as in Lemma 21 and $\tilde{c}$ as in Lemma 23.

Proof The claim trivially holds if $\mathfrak{s} = \infty$, so for the remainder of the proof we suppose $\mathfrak{s} < \infty$. Fix any $\bar{k} \in \{3, \ldots, k_\varepsilon\}$ and $k \in \{2, \ldots, \bar{k} - 1\}$, and note that

$$\begin{aligned} &\left|\{m \in \{1, \ldots, \tilde{m}_k\} : X_m^2 \in \mathrm{DIS}(V_{m-1})\}\right| \\ &\quad \leq \left|\left\{m \in \{1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \leq 2^{-\bar{k}}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\ &\quad + \left|\left\{m \in \{1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \geq 2^{-\bar{k}}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right|. \end{aligned} \tag{27}$$

We proceed to bound each term on the right hand side. A Chernoff bound (applied under the conditional distribution given J) and the law of total probability imply that, on an event $G_8^{(i)}(\bar{k}, k)$ of probability at least $1 - \frac{\delta}{256k_\varepsilon^2}$,

$$\begin{aligned} &\left|\left\{m \in \{1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \leq 2^{-\bar{k}}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\ &\qquad \leq \log_2\left(\frac{256k_\varepsilon^2}{\delta}\right) + 2e\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \leq 2^{-\bar{k}}\right\}\right)\tilde{m}_k, \end{aligned}$$

and Lemma 33 implies that, on E_1, this is at most

$$\log_2\left(\frac{256k_\varepsilon^2}{\delta}\right) + 6e\max\left\{\mathcal{P}\left(x : \gamma_x < 2^{2-\bar{k}}\right), \frac{\varepsilon}{2\gamma_\varepsilon}\right\}\tilde{m}_k.$$

Now we turn to bounding the second term on the right hand side of (27). We proceed in two steps, noting that monotonicity of $m \mapsto \mathrm{DIS}(V_m)$ implies

$$\begin{aligned} &\left|\left\{m \in \{1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \geq 2^{-\bar{k}}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\ &\quad \leq \left|\left\{m \in \{1, \ldots, \tilde{m}_{\bar{k}}\} : \gamma_{J_m} \geq 2^{-\bar{k}}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\ &\quad + \left|\left\{m \in \{\tilde{m}_{\bar{k}} + 1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \geq 2^{-\bar{k}}, X_m^2 \in \mathrm{DIS}(V_{\tilde{m}_{\bar{k}}})\right\}\right|. \end{aligned} \tag{28}$$

We start with the first term on the right side of (28). Let $L = \left|\left\{m \in \{1, \ldots, \tilde{m}_{\bar{k}}\} : \gamma_{J_m} \geq 2^{-\bar{k}}\right\}\right|$, and let $\ell_1, \ldots, \ell_L$ denote the increasing subsequence of values $\ell \in \{1, \ldots, \tilde{m}_{\bar{k}}\}$ with $\gamma_{J_\ell} \geq 2^{-\bar{k}}$. Also, let $\tilde{j}_{\bar{k}} = \max\{1, \lceil\log_2(\tilde{m}_{\bar{k}}/(\mathfrak{s} + \mathrm{Log}(1/\delta)))\rceil\}$, let $M_0 = 0$, and for each $j \in \mathbb{N}$, let

$$M_j = \left\lceil \tilde{c}2^j\left(\mathfrak{s}\mathrm{Log}\left(2^j\right) + \mathrm{Log}\left(\frac{256k_\varepsilon^2\tilde{j}_{\bar{k}}}{\delta}\right)\right)\right\rceil,$$

for $\tilde{c}$ as in Lemma 23. Let $V_0^\star = \mathbb{C}$, and for each $i \leq L$, let

$$V_i^\star = \left\{h \in \mathbb{C} : \forall j \in \{1, \ldots, i\}, h(X_{\ell_j}^2) = f_{\mathcal{P}_{XY}}^\star(X_{\ell_j}^2)\right\}.$$

Let $\phi_\mathfrak{s}$ be the function mapping any $\mathcal{U} \in \mathcal{X}^\mathfrak{s}$ to the set $\mathrm{DIS}(\{h \in \mathbb{C} : \forall x \in \mathcal{U}, h(x) = f_{\mathcal{P}_{XY}}^\star(x)\})$. Fix any $j \in \mathbb{N}$. By Theorem 13, if $M_j \leq L$, then there exist $i_1, \ldots, i_\mathfrak{s} \in \{1, \ldots, M_j\}$ such that $\{h \in \mathbb{C} : \forall r \in \{1, \ldots, \mathfrak{s}\}, h(X_{\ell_{i_r}}^2) = f_{\mathcal{P}_{XY}}^\star(X_{\ell_{i_r}}^2)\} = V_{M_j}^\star$ (see the discussion in Section 7.3.1). In particular, for this choice of $i_1, \ldots, i_\mathfrak{s}$, we have $\phi_\mathfrak{s}(X_{\ell_{i_1}}^2, \ldots, X_{\ell_{i_\mathfrak{s}}}^2) = \mathrm{DIS}(V_{M_j}^\star)$; furthermore, since

$\phi_{\mathfrak{s}}$ is permutation-invariant, we can take $i_1 \leq \cdots \leq i_{\mathfrak{s}}$ without loss of generality. Also note that $X^2_{\ell_1}, \ldots, X^2_{\ell_{M_j \wedge L}}$ are conditionally independent (given L and J), each with conditional distribution $\mathcal{P}\left(\cdot \middle| \bigcup \left\{A \in J : \gamma_A \geq 2^{-\bar{k}}\right\}\right)$. Since (when $M_j \leq L$) $\{X^2_{\ell_1}, \ldots, X^2_{\ell_{M_j}}\} \cap \phi_{\mathfrak{s}}(X^2_{\ell_{i_1}}, \ldots, X^2_{\ell_{i_{\mathfrak{s}}}}) = \{X^2_{\ell_1}, \ldots, X^2_{\ell_{M_j}}\} \cap \mathrm{DIS}(V^\star_{M_j}) = \emptyset$, Lemma 23 (applied under the conditional distribution given L and J) and the law of total probability imply that, on an event $G_8^{(ii)}(\bar{k}, k, j)$ of probability at least $1 - \frac{\delta}{256 k_\varepsilon^2 \tilde{j}_{\bar{k}}}$, if $M_j \leq L$, then

$$\mathcal{P}\left(\mathrm{DIS}(V^\star_{M_j}) \middle| \bigcup \left\{A \in J : \gamma_A \geq 2^{-\bar{k}}\right\}\right) \leq 2^{-j}. \tag{29}$$

Furthermore, this clearly holds for $j = 0$ as well. Since $\mathcal{P}\left(\mathrm{DIS}\left(V^\star_{i-1}\right) \middle| \bigcup \left\{A \in J : \gamma_A \geq 2^{-\bar{k}}\right\}\right)$ is nonincreasing in i, for every $j \geq 0$ with $M_j < L$, and every $i \in \{M_j + 1, \ldots, M_{j+1} \wedge L\}$, on $G_8^{(ii)}(\bar{k}, k, j)$, $\mathcal{P}\left(\mathrm{DIS}\left(V^\star_{i-1}\right) \middle| \bigcup \left\{A \in J : \gamma_A \geq 2^{-\bar{k}}\right\}\right) \leq 2^{-j}$. Since every $j \geq \tilde{j}_{\bar{k}}$ has $M_j \geq \tilde{m}_{\bar{k}} \geq L$, this holds simultaneously for every j with $M_j < L$ on $\bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1} G_8^{(ii)}(\bar{k}, k, j)$.

Now note that, conditioned on J and L,

$$\left\{\mathbb{1}_{\mathrm{DIS}(V^\star_{i-1})}\left(X^2_{\ell_i}\right) - \mathcal{P}\left(\mathrm{DIS}\left(V^\star_{i-1}\right) \middle| \bigcup \left\{A \in J : \gamma_A \geq 2^{-\bar{k}}\right\}\right)\right\}_{i=1}^{L}$$

is a martingale difference sequence with respect to $X^2_{\ell_1}, \ldots, X^2_{\ell_L}$. Therefore, Bernstein's inequality for martingales (e.g., McDiarmid, 1998, Theorem 3.12), applied under the conditional distribution given J and L, along with the law of total probability, imply that there exists an event $G_8^{(iii)}(\bar{k}, k)$ of probability at least $1 - \frac{\delta}{256 k_\varepsilon^2}$ such that, on $G_8^{(iii)}(\bar{k}, k) \cap \bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1} G_8^{(ii)}(\bar{k}, k, j)$,

$$\begin{aligned}
\sum_{i=1}^{L} \mathbb{1}_{\mathrm{DIS}(V^\star_{i-1})}\left(X^2_{\ell_i}\right) &\leq \log_2\left(\frac{256 k_\varepsilon^2}{\delta}\right) + 2e \sum_{j=0}^{\tilde{j}_{\bar{k}}-1} 2^{-j}(M_{j+1} - M_j) \\
&\leq \log_2\left(\frac{256 k_\varepsilon^2}{\delta}\right) + 4e + 4e\tilde{c}\left(\mathfrak{s}\mathrm{Log}\left(2^{\tilde{j}_{\bar{k}}}\right) + \mathrm{Log}\left(\frac{256 k_\varepsilon^2 \tilde{j}_{\bar{k}}}{\delta}\right)\right)\tilde{j}_{\bar{k}} \\
&\leq 8e\tilde{c}\left(\mathfrak{s}\tilde{j}_{\bar{k}} + \mathrm{Log}\left(\frac{256 k_\varepsilon^2}{\delta}\right)\right)\tilde{j}_{\bar{k}}.
\end{aligned}$$

By Lemma 37, on $\bigcap_{j=0}^{4} E_j$, $\forall m \in \{1, \ldots, \tilde{m}_{\bar{k}}\}$,

$$V_m \subseteq \left\{h \in \mathbb{C} : \forall m' \leq m \text{ with } \gamma_{J_{m'}} \geq 2^{-\bar{k}}, h(X^2_{m'}) = f^\star_{\mathcal{P}_{XY}}(X^2_{m'})\right\}.$$

In particular, this implies $V_{\ell_i - 1} \subseteq V^\star_{i-1}$ for all $i \leq L$. Therefore, on $\bigcap_{j=0}^{4} E_j \cap G_8^{(iii)}(\bar{k}, k) \cap \bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1} G_8^{(ii)}(\bar{k}, k, j)$,

$$\begin{aligned}
\left|\left\{m \in \{1, \ldots, \tilde{m}_{\bar{k}}\} : \gamma_{J_m} \geq 2^{-\bar{k}}, X^2_m \in \mathrm{DIS}(V_{m-1})\right\}\right| &= \sum_{i=1}^{L} \mathbb{1}_{\mathrm{DIS}(V_{\ell_i - 1})}\left(X^2_{\ell_i}\right) \\
&\leq \sum_{i=1}^{L} \mathbb{1}_{\mathrm{DIS}(V^\star_{i-1})}\left(X^2_{\ell_i}\right) \leq 8e\tilde{c}\left(\mathfrak{s}\tilde{j}_{\bar{k}} + \mathrm{Log}\left(\frac{256 k_\varepsilon^2}{\delta}\right)\right)\tilde{j}_{\bar{k}}.
\end{aligned} \tag{30}$$

Next, we turn to bounding the second term on the right hand side of (28). A Chernoff bound (applied under the conditional distribution given $V_{\tilde{m}_{\bar{k}}}$ and J) and the law of total probability imply that there is an event $G_8^{(iv)}(\bar{k},k)$ of probability at least $1-\frac{\delta}{256k_\varepsilon^2}$, on which

$$\left|\left\{m\in\{\tilde{m}_{\bar{k}}+1,\ldots,\tilde{m}_k\}:\gamma_{J_m}\geq 2^{-\bar{k}}, X_m^2\in \mathrm{DIS}(V_{\tilde{m}_{\bar{k}}})\right\}\right| \\ \leq \log_2\left(\frac{256k_\varepsilon^2}{\delta}\right)+2e\mathcal{P}\left(\mathrm{DIS}\left(V_{\tilde{m}_{\bar{k}}}\right)\cap\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)\tilde{m}_k. \quad (31)$$

Also, by a Chernoff bound (applied under the conditional distribution given J), with probability at least

$$1-\exp\left\{-(1/8)\mathcal{P}\left(\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)\tilde{m}_{\bar{k}}\right\},$$

we have

$$L\geq (1/2)\tilde{m}_{\bar{k}}\mathcal{P}\left(\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right). \quad (32)$$

If $\mathcal{P}\left(\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)\geq \frac{8}{\tilde{m}_{\bar{k}}}\mathrm{Log}\left(\frac{256k_\varepsilon}{\delta}\right)$, then

$$\exp\left\{-(1/8)\mathcal{P}\left(\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)\tilde{m}_{\bar{k}}\right\}\leq\frac{\delta}{256k_\varepsilon}.$$

Thus, by the law of total probability, there is an event $G_8^{(v)}(\bar{k})$ of probability at least $1-\frac{\delta}{256k_\varepsilon}$, on which, if $\mathcal{P}\left(\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)\geq\frac{8}{\tilde{m}_{\bar{k}}}\mathrm{Log}\left(\frac{256k_\varepsilon}{\delta}\right)$, then (32) holds. Let

$$\hat{j}=\max\left\{j\in\{0,1,\ldots,\tilde{j}_{\bar{k}}-1\}:M_j\leq(1/2)\tilde{m}_{\bar{k}}\mathcal{P}\left(\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)\right\},$$

and note that

$$\hat{j}\geq\left\lfloor\log_2\left(\frac{\tilde{m}_{\bar{k}}\mathcal{P}\left(\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)}{4\tilde{c}\left(2\mathfrak{s}\mathrm{Log}\left(2^{\tilde{j}_{\bar{k}}}\right)+\mathrm{Log}\left(\frac{256k_\varepsilon^2}{\delta}\right)\right)}\right)\right\rfloor. \quad (33)$$

(29) implies that on $\bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1}G_8^{(ii)}(\bar{k},k,j)$, if (32) holds, we have

$$\mathcal{P}\left(\mathrm{DIS}\left(V_L^\star\right)\middle|\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)\leq 2^{-\hat{j}}.$$

Furthermore, Lemma 37 implies that, on $\bigcap_{j=0}^4 E_j$, $V_{\tilde{m}_{\bar{k}}}\subseteq V_L^\star$. Altogether, on $\bigcap_{j=0}^4 E_j\cap G_8^{(v)}(\bar{k})\cap\bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1}G_8^{(ii)}(\bar{k},k,j)$, if $\mathcal{P}\left(\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)\geq\frac{8}{\tilde{m}_{\bar{k}}}\mathrm{Log}\left(\frac{256k_\varepsilon}{\delta}\right)$, then

$$\mathcal{P}\left(\mathrm{DIS}\left(V_{\tilde{m}_{\bar{k}}}\right)\cap\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right)\leq 2^{-\hat{j}}\mathcal{P}\left(\bigcup\left\{A\in J:\gamma_A\geq 2^{-\bar{k}}\right\}\right) \\ \leq\frac{8\tilde{c}}{\tilde{m}_{\bar{k}}}\left(2\mathfrak{s}\mathrm{Log}\left(2^{\tilde{j}_{\bar{k}}}\right)+\mathrm{Log}\left(\frac{256k_\varepsilon^2}{\delta}\right)\right),$$

where the last inequality is by (33). Otherwise, if $\mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \geq 2^{-\bar{k}}\right\}\right) < \frac{8}{\tilde{m}_{\bar{k}}}\mathrm{Log}\left(\frac{256k_{\varepsilon}}{\delta}\right)$, then in any case we have

$$\mathcal{P}\left(\mathrm{DIS}\left(V_{\tilde{m}_{\bar{k}}}\right) \cap \bigcup\left\{A \in J : \gamma_A \geq 2^{-\bar{k}}\right\}\right) \leq \mathcal{P}\left(\bigcup\left\{A \in J : \gamma_A \geq 2^{-\bar{k}}\right\}\right)$$
$$< \frac{8}{\tilde{m}_{\bar{k}}}\mathrm{Log}\left(\frac{256k_{\varepsilon}}{\delta}\right) \leq \frac{8\tilde{c}}{\tilde{m}_{\bar{k}}}\left(2\mathfrak{s}\mathrm{Log}\left(2^{\tilde{j}_{\bar{k}}}\right) + \mathrm{Log}\left(\frac{256k_{\varepsilon}^2}{\delta}\right)\right).$$

Combined with (31), this implies that on $\bigcap_{j=0}^{4} E_j \cap G_8^{(iv)}(\bar{k},k) \cap G_8^{(v)}(\bar{k}) \cap \bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1} G_8^{(ii)}(\bar{k},k,j)$,

$$\begin{aligned}
&\left|\left\{m \in \{\tilde{m}_{\bar{k}}+1,\ldots,\tilde{m}_k\} : \gamma_{J_m} \geq 2^{-\bar{k}}, X_m^2 \in \mathrm{DIS}(V_{\tilde{m}_{\bar{k}}})\right\}\right| \\
&\leq \log_2\left(\frac{256k_{\varepsilon}^2}{\delta}\right) + 16e\tilde{c}\frac{\tilde{m}_k}{\tilde{m}_{\bar{k}}}\left(2\mathfrak{s}\mathrm{Log}\left(2^{\tilde{j}_{\bar{k}}}\right) + \mathrm{Log}\left(\frac{256k_{\varepsilon}^2}{\delta}\right)\right) \\
&\leq 32e\tilde{c}\frac{\tilde{m}_k}{\tilde{m}_{\bar{k}}}\left(\mathfrak{s}\tilde{j}_{\bar{k}} + \mathrm{Log}\left(\frac{256k_{\varepsilon}^2}{\delta}\right)\right) \leq 64e\tilde{c}2^{\bar{k}-k}\left(\mathfrak{s}\tilde{j}_{\bar{k}} + \mathrm{Log}\left(\frac{256k_{\varepsilon}^2}{\delta}\right)\right).
\end{aligned}$$

Plugging this and (30) into (28), we have that on $\bigcap_{j=0}^{4} E_j \cap G_8^{(iii)}(\bar{k},k) \cap G_8^{(iv)}(\bar{k},k) \cap G_8^{(v)}(\bar{k}) \cap \bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1} G_8^{(ii)}(\bar{k},k,j)$,

$$\begin{aligned}
&\left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : \gamma_{J_m} \geq 2^{-\bar{k}}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\
&\leq 8e\tilde{c}\left(\mathfrak{s}\tilde{j}_{\bar{k}} + \mathrm{Log}\left(\frac{256k_{\varepsilon}^2}{\delta}\right)\right)\tilde{j}_{\bar{k}} + 64e\tilde{c}2^{\bar{k}-k}\left(\mathfrak{s}\tilde{j}_{\bar{k}} + \mathrm{Log}\left(\frac{256k_{\varepsilon}^2}{\delta}\right)\right) \\
&= 8e\tilde{c}\left(2^{3+\bar{k}-k} + \tilde{j}_{\bar{k}}\right)\left(\mathfrak{s}\tilde{j}_{\bar{k}} + \mathrm{Log}\left(\frac{256k_{\varepsilon}^2}{\delta}\right)\right).
\end{aligned}$$

Combined with the above result bounding the first term in (27), we have that on $\bigcap_{j=0}^{4} E_j \cap G_8^{(i)}(\bar{k},k) \cap G_8^{(iii)}(\bar{k},k) \cap G_8^{(iv)}(\bar{k},k) \cap G_8^{(v)}(\bar{k}) \cap \bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1} G_8^{(ii)}(\bar{k},k,j)$,

$$\begin{aligned}
&\left|\left\{m \in \{1,\ldots,\tilde{m}_k\} : X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\
&\leq \log_2\left(\frac{256k_{\varepsilon}^2}{\delta}\right) + 6e\max\left\{\mathcal{P}\left(x : \gamma_x < 2^{2-\bar{k}}\right), \frac{\varepsilon}{2\gamma_{\varepsilon}}\right\}\tilde{m}_k \\
&\qquad + 8e\tilde{c}\left(2^{3+\bar{k}-k} + \tilde{j}_{\bar{k}}\right)\left(\mathfrak{s}\tilde{j}_{\bar{k}} + \mathrm{Log}\left(\frac{256k_{\varepsilon}^2}{\delta}\right)\right) \\
&\leq 6e\max\left\{\mathcal{P}\left(x : \gamma_x < 2^{2-\bar{k}}\right), \frac{\varepsilon}{2\gamma_{\varepsilon}}\right\}\tilde{m}_k + (1+8e\tilde{c})\left(2^{3+\bar{k}-k} + \tilde{j}_{\bar{k}}\right)\left(\mathfrak{s}\tilde{j}_{\bar{k}} + \mathrm{Log}\left(\frac{256k_{\varepsilon}^2}{\delta}\right)\right).
\end{aligned} \tag{34}$$

Noting that $\mathfrak{s} \geq d$, a bit of algebra reveals that

$$\frac{\tilde{m}_{\bar{k}}}{\mathfrak{s} + \mathrm{Log}(1/\delta)} \leq \frac{32ck_{\varepsilon}}{\varepsilon}\mathrm{Log}\left(\frac{128k_{\varepsilon}^2}{\varepsilon}\right) \leq \frac{2^9ck_{\varepsilon}^2}{\varepsilon^{3/2}},$$

so that

$$\tilde{j}_{\bar{k}} \leq \log_2\left(\frac{2^{10}ck_{\varepsilon}^2}{\varepsilon^{3/2}}\right) \leq \frac{3}{2}\mathrm{Log}\left(\frac{2^{10}ck_{\varepsilon}^2}{\varepsilon^{3/2}}\right),$$

and therefore

$$
\begin{aligned}
&(1+8e\tilde{c})\left(2^{3+\bar{k}-k}+\tilde{j}_{\bar{k}}\right)\left(\mathfrak{s}\tilde{j}_{\bar{k}}+\mathrm{Log}\left(\frac{256k_\varepsilon^2}{\delta}\right)\right)\\
&\leq (1+8e\tilde{c})\left(2^{3+\bar{k}-k}+\frac{3}{2}\mathrm{Log}\left(\frac{2^{10}ck_\varepsilon^2}{\varepsilon^{3/2}}\right)\right)\left(\frac{3}{2}\mathfrak{s}\mathrm{Log}\left(\frac{2^{10}ck_\varepsilon^2}{\varepsilon^{3/2}}\right)+\mathrm{Log}\left(\frac{256k_\varepsilon^2}{\delta}\right)\right)\\
&\leq (1+8e\tilde{c})\left(2^{3+\bar{k}-k}+\frac{3}{2}\mathrm{Log}\left(\frac{2^{10}ck_\varepsilon^2}{\varepsilon^{3/2}}\right)\right)\left(\frac{3}{2}\mathfrak{s}\mathrm{Log}\left(\frac{2^{16}ck_\varepsilon^4}{\varepsilon^{3/2}}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right).
\end{aligned}
$$

Furthermore, since $k_\varepsilon \leq \sqrt{32/\varepsilon}$, this is at most

$$
\begin{aligned}
&(1+8e\tilde{c})\left(2^{3+\bar{k}-k}+\frac{3}{2}\mathrm{Log}\left(\frac{2^{15}c}{\varepsilon^{5/2}}\right)\right)\left(\frac{3}{2}\mathfrak{s}\mathrm{Log}\left(\frac{2^{26}c}{\varepsilon^{7/2}}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\\
&\leq 91\tilde{c}\left(2^{1+\bar{k}-k}+\mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\right)\left(6\mathfrak{s}\mathrm{Log}\left(\frac{128c}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right).
\end{aligned}
$$

Plugging this into (34), we have that on $\bigcap_{j=0}^{4} E_j \cap G_8^{(i)}(\bar{k},k) \cap G_8^{(iii)}(\bar{k},k) \cap G_8^{(iv)}(\bar{k},k) \cap G_8^{(v)}(\bar{k}) \cap \bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1} G_8^{(ii)}(\bar{k},k,j)$,

$$
\begin{aligned}
&\left|\left\{m\in\{1,\ldots,\tilde{m}_k\} : X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right|\\
&\qquad\qquad \leq 6e\max\left\{\mathcal{P}\left(x:\gamma_x<2^{2-\bar{k}}\right),\frac{\varepsilon}{2\gamma_\varepsilon}\right\}\tilde{m}_k\\
&\qquad\qquad + 91\tilde{c}\left(2^{1+\bar{k}-k}+\mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\right)\left(6\mathfrak{s}\mathrm{Log}\left(\frac{128c}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right). \qquad (35)
\end{aligned}
$$

Letting

$$
E_8=\bigcap_{\bar{k}=3}^{k_\varepsilon}\left(G_8^{(v)}(\bar{k})\cap\bigcap_{k=2}^{\bar{k}-1}G_8^{(i)}(\bar{k},k)\cap G_8^{(iii)}(\bar{k},k)\cap G_8^{(iv)}(\bar{k},k)\cap\bigcap_{j=1}^{\tilde{j}_{\bar{k}}-1}G_8^{(ii)}(\bar{k},k,j)\right),
$$

we have that (35) holds for all $\bar{k}\in\{3,\ldots,k_\varepsilon\}$ and $k\in\{2,\ldots,\bar{k}-1\}$ on the event $E_8\cap\bigcap_{j=0}^{4}E_j$. A union bound implies that E_8 has probability at least

$$
\begin{aligned}
&1-\sum_{\bar{k}=3}^{k_\varepsilon}\left(\frac{\delta}{256k_\varepsilon}+\sum_{k=2}^{\bar{k}-1}\left(3\frac{\delta}{256k_\varepsilon^2}+\sum_{j=1}^{\tilde{j}_{\bar{k}}-1}\frac{\delta}{256k_\varepsilon^2\tilde{j}_{\bar{k}}}\right)\right)\\
&\qquad\qquad \geq 1-\frac{\delta}{256}-\sum_{\bar{k}=3}^{k_\varepsilon}(\bar{k}-2)\frac{\delta}{64k_\varepsilon^2}\geq 1-\frac{\delta}{256}-\frac{\delta}{128}>1-\frac{\delta}{64}.
\end{aligned}
$$

■

We can now state a sufficient size on the budget n so that, with high probability, Algorithm 1 reaches $m=\tilde{m}$, so that the returned $\hat{h}_n$ is equivalent to the $\hat{h}_\infty$ classifier from Lemma 38, which therefore satisfies the same guarantee on its error rate.

Lemma 41 *There exists a finite universal constant $\bar{c} \geq 1$ such that, on the event $\bigcap_{j=0}^{8} E_j$, for any $\bar{k} \in \{2, \ldots, k_\varepsilon\}$, for any n of size at least*

$$\begin{aligned} &\bar{c}\mathbb{1}[\bar{k} > 2]2^{2\bar{k}} \left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right) \mathrm{Log}\left(\frac{1}{\varepsilon}\right) \\ &+ \bar{c}\sum_{k=\bar{k}}^{k_\varepsilon} \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{\hat{\gamma}_\varepsilon}\right\} \frac{2^k}{\varepsilon} \left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right) \mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right), \end{aligned} \tag{36}$$

running Algorithm 1 with budget n results in at most n label requests, and the returned classifier $\hat{h}_n$ satisfies $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq \varepsilon$. Furthermore, the event $\bigcap_{j=0}^{8} E_j$ has probability at least $1 - \delta$.

Proof The value of t keeps the running total of the number of label requests made by the algorithm after each call to Subroutine 1. Furthermore, within each execution of Subroutine 1, the value $t + q$ represents the running total of the number of label requests made by the algorithm so far. Since the $n - t$ budget argument to Subroutine 1 ensures that it halts (in Step 6) if ever $t + q = n$, and since the first condition in Step 1 of Algorithm 1 ensures that Algorithm 1 halts if ever $t = n$, we are guaranteed that the algorithm never requests a number of labels larger than the budget n.

We will show that taking n of the stated size suffices for the result by showing that this size suffices to reproduce the behavior of the infinite budget execution of Algorithm 1. Due to the condition $m < \tilde{m}$ in Step 1 of Algorithm 1, the final value of t obtained when running Algorithm 1 with budget ∞ may be expressed as

$$\sum_{m=1}^{\tilde{m}} \hat{q}_{\infty,m} \mathbb{1}_{\mathrm{DIS}(V_{m-1})}\left(X_m^2\right).$$

Lemma 35 implies that, on $\bigcap_{j=0}^{8} E_j$, this is at most

$$\begin{aligned} &\sum_{m=1}^{\tilde{m}} \left\lceil \frac{8}{\max\{\gamma_{J_m}^2, 2^{-2\tilde{k}_m}\}} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\right\rceil \mathbb{1}_{\mathrm{DIS}(V_{m-1})}\left(X_m^2\right) \\ &\leq \sum_{m=1}^{\tilde{m}} \sum_{k=2}^{\tilde{k}_m} \mathbb{1}\left[\gamma_{J_m} \leq 2^{1-k}\right] 2^{2k+4} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right) \mathbb{1}_{\mathrm{DIS}(V_{m-1})}\left(X_m^2\right). \end{aligned}$$

The summation in this last expression is over all $m \in \{1, \ldots, \tilde{m}\}$ and $k \in \{2, \ldots, k_\varepsilon\}$ such that $k \leq \tilde{k}_m$, which is equivalent to those $m \in \{1, \ldots, \tilde{m}\}$ and $k \in \{2, \ldots, k_\varepsilon\}$ such that $m \leq \tilde{m}_k$. Therefore, exchanging the order of summation, this expression is equal to

$$\begin{aligned} &\sum_{k=2}^{k_\varepsilon} \sum_{m=1}^{\tilde{m}_k} \mathbb{1}\left[\gamma_{J_m} \leq 2^{1-k}\right] 2^{2k+4} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right) \mathbb{1}_{\mathrm{DIS}(V_{m-1})}\left(X_m^2\right) \\ &= \sum_{k=2}^{k_\varepsilon} 2^{2k+4} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right) \left|\left\{m \in \{1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \leq 2^{1-k}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right|. \end{aligned} \tag{37}$$

Fix any value $\bar{k} \in \{2, \ldots, k_\varepsilon\}$. For any $k \in \left\{\bar{k}, \ldots, k_\varepsilon\right\}$, Lemma 39 implies that, on $\bigcap_{j=0}^{8} E_j$,

$$\left|\left\{m \in \{1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \leq 2^{1-k}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\ \leq 17 \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{2\hat{\gamma}_\varepsilon}\right\} \tilde{m}_k.$$

This implies

$$\begin{aligned}
&\sum_{k=\bar{k}}^{k_\varepsilon} 2^{2k+4} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right) \left|\left\{m \in \{1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \leq 2^{1-k}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\
&\leq \sum_{k=\bar{k}}^{k_\varepsilon} 2^{2k+9} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right) \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{2\hat{\gamma}_\varepsilon}\right\} \tilde{m}_k \\
&\leq \sum_{k=\bar{k}}^{k_\varepsilon} \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{2\hat{\gamma}_\varepsilon}\right\} \frac{2^{k+17} c k_\varepsilon}{\varepsilon} \left(d\mathrm{Log}\left(\frac{2k_\varepsilon}{\varepsilon}\right) + \mathrm{Log}\left(\frac{64k_\varepsilon}{\delta}\right)\right) \mathrm{Log}\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right) \\
&\leq \sum_{k=\bar{k}}^{k_\varepsilon} \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{2\hat{\gamma}_\varepsilon}\right\} \frac{2^{k+25} c \mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right)}{\varepsilon} \left(d\mathrm{Log}\left(\frac{64}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{32cd}{\varepsilon\delta}\right),
\end{aligned} \tag{38}$$

where this last inequality is based on the fact that $k_\varepsilon \leq \sqrt{32/\varepsilon}$, combined with some simple algebra. If $\bar{k} > 2$, for any $k \in \left\{2, \ldots, \bar{k} - 1\right\}$, Lemma 40 implies that, on $\bigcap_{j=0}^{8} E_j$,

$$\begin{aligned}
&\left|\left\{m \in \{1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \leq 2^{1-k}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\
&\qquad \leq 6e \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{2-\bar{k}}\right), \frac{\varepsilon}{2\gamma_\varepsilon}\right\} \tilde{m}_k \\
&\qquad + 91\tilde{c}\left(2^{1+\bar{k}-k} + \mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\right)\left(6\mathfrak{s}\mathrm{Log}\left(\frac{128c}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).
\end{aligned}$$

This implies

$$\begin{aligned}
&\sum_{k=2}^{\bar{k}-1} 2^{2k+4} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right) \left|\left\{m \in \{1, \ldots, \tilde{m}_k\} : \gamma_{J_m} \leq 2^{1-k}, X_m^2 \in \mathrm{DIS}(V_{m-1})\right\}\right| \\
&\leq \sum_{k=2}^{\bar{k}-1} 2^{2k+9} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right) \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{2-\bar{k}}\right), \frac{\varepsilon}{2\gamma_\varepsilon}\right\} \tilde{m}_k \\
&+ \sum_{k=2}^{\bar{k}-1} 2^{2k+11} \tilde{c} \ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right) \left(2^{1+\bar{k}-k} + \mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\right)\left(6\mathfrak{s}\mathrm{Log}\left(\frac{128c}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).
\end{aligned}$$

Since

$$\begin{aligned}\sum_{k=2}^{\bar{k}-1} 2^{2k}\tilde{m}_k &\leq \sum_{k=2}^{\bar{k}-1} \frac{2^{k+8}ck_\varepsilon}{\varepsilon}\left(d\mathrm{Log}\left(\frac{2k_\varepsilon}{\varepsilon}\right)+\mathrm{Log}\left(\frac{64k_\varepsilon}{\delta}\right)\right)\\ &\leq \frac{2^{\bar{k}+8}ck_\varepsilon}{\varepsilon}\left(d\mathrm{Log}\left(\frac{2k_\varepsilon}{\varepsilon}\right)+\mathrm{Log}\left(\frac{64k_\varepsilon}{\delta}\right)\right)\\ &\leq \frac{2^{\bar{k}+12}c\mathrm{Log}(1/\hat{\gamma}_\varepsilon)}{\varepsilon}\left(d\mathrm{Log}\left(\frac{64}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\end{aligned}$$

and

$$\sum_{k=2}^{\bar{k}-1} 2^{2k}\left(2^{1+\bar{k}-k}+\mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\right) \leq 2^{2\bar{k}}\left(2+\mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\right) \leq 2^{2\bar{k}+1}\mathrm{Log}\left(\frac{64c}{\varepsilon}\right),$$

we have that

$$\begin{aligned}&\sum_{k=2}^{\bar{k}-1} 2^{2k+4}\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\left|\left\{m\in\{1,\ldots,\tilde{m}_k\}:\gamma_{J_m}\leq 2^{1-k}, X_m^2\in\mathrm{DIS}(V_{m-1})\right\}\right|\\ &\leq 2^9\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\max\left\{\mathcal{P}\left(x:\gamma_x<2^{2-\bar{k}}\right),\frac{\varepsilon}{2\gamma_\varepsilon}\right\}\sum_{k=2}^{\bar{k}-1}2^{2k}\tilde{m}_k\\ &\quad+2^{11}\tilde{c}\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\left(6\mathfrak{s}\mathrm{Log}\left(\frac{128c}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\sum_{k=2}^{\bar{k}-1}2^{2k}\left(2^{1+\bar{k}-k}+\mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\right)\\ &\leq 2^9\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\max\left\{\mathcal{P}\left(x:\gamma_x<2^{2-\bar{k}}\right),\frac{\varepsilon}{2\gamma_\varepsilon}\right\}\frac{2^{\bar{k}+12}c\mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right)}{\varepsilon}\left(d\mathrm{Log}\left(\frac{64}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\\ &\quad+2^{11}\tilde{c}\ln\left(\frac{32\tilde{m}q_{\varepsilon,\delta}}{\delta}\right)\left(6\mathfrak{s}\mathrm{Log}\left(\frac{128c}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)2^{2\bar{k}+1}\mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\\ &\leq \max\left\{\mathcal{P}\left(x:\gamma_x<2^{2-\bar{k}}\right),\frac{\varepsilon}{2\gamma_\varepsilon}\right\}\frac{2^{\bar{k}+25}c\mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right)}{\varepsilon}\left(d\mathrm{Log}\left(\frac{64}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{32cd}{\varepsilon\delta}\right)\\ &\quad+2^{2\bar{k}+16}\tilde{c}\left(6\mathfrak{s}\mathrm{Log}\left(\frac{128c}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{64c}{\varepsilon}\right)\mathrm{Log}\left(\frac{32cd}{\varepsilon\delta}\right).\end{aligned}$$

Plugging this and (38) into (37) reveals that, on $\bigcap_{j=0}^{8} E_j$, if $\bar{k}>2$,

$$
\begin{aligned}
&\sum_{m=1}^{\tilde{m}} \hat{q}_{\infty,m} \mathbb{1}_{\mathrm{DIS}(V_{m-1})}\left(X_m^2\right) \\
&\leq \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{2-\bar{k}}\right), \frac{\varepsilon}{2\gamma_\varepsilon}\right\} \frac{2^{\bar{k}+25} c \mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right)}{\varepsilon} \left(d\mathrm{Log}\left(\frac{64}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{32cd}{\varepsilon\delta}\right) \\
&+ 2^{2\bar{k}+16} \tilde{c} \left(6\mathfrak{s}\mathrm{Log}\left(\frac{128c}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{64c}{\varepsilon}\right) \mathrm{Log}\left(\frac{32cd}{\varepsilon\delta}\right) \\
&+ \sum_{k=\bar{k}}^{k_\varepsilon} \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{2\hat{\gamma}_\varepsilon}\right\} \frac{2^{k+25} c \mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right)}{\varepsilon} \left(d\mathrm{Log}\left(\frac{64}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{32cd}{\varepsilon\delta}\right). \\
&\leq \bar{c} 2^{2\bar{k}} \left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right) \mathrm{Log}\left(\frac{1}{\varepsilon}\right) \\
&+ \bar{c} \sum_{k=\bar{k}}^{k_\varepsilon} \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{\hat{\gamma}_\varepsilon}\right\} \frac{2^k}{\varepsilon} \left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right) \mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right),
\end{aligned}
$$

for an appropriate finite universal constant $\bar{c} \geq 1$. Furthermore, if $\bar{k} = 2$, (38) and (37) already imply that, on $\bigcap_{j=0}^{8} E_j$,

$$
\begin{aligned}
&\sum_{m=1}^{\tilde{m}} \hat{q}_{\infty,m} \mathbb{1}_{\mathrm{DIS}(V_{m-1})}\left(X_m^2\right) \\
&\leq \bar{c} \sum_{k=\bar{k}}^{k_\varepsilon} \max\left\{\mathcal{P}\left(x : \gamma_x < 2^{3-k}\right), \frac{\varepsilon}{\hat{\gamma}_\varepsilon}\right\} \frac{2^k}{\varepsilon} \left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right) \mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right),
\end{aligned}
$$

again for $\bar{c} \geq 1$ chosen appropriately large.

Therefore, for a choice of $\bar{c}$ as above, on $\bigcap_{j=0}^{8} E_j$, for any $\bar{k} \in \{2, \ldots, k_\varepsilon\}$, the final value of t obtained when running Algorithm 1 with budget ∞ is at most (36). Since running Algorithm 1 with a finite budget n only returns a different $\hat{h}_n$ from the $\hat{h}_\infty$ returned by the infinite-budget execution if t would exceed n in the infinite-budget execution, this implies that taking any n of size at least (36) suffices to produce identical output to the infinite-budget execution, on the event $\bigcap_{j=0}^{8} E_j$: that is, $\hat{h}_n = \hat{h}_\infty$. Therefore, since Lemma 38 implies that, on $\bigcap_{j=0}^{8} E_j$, $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_\infty) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq \varepsilon$, we conclude that for n of size at least (36), on $\bigcap_{j=0}^{8} E_j$, $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq \varepsilon$.

Finally, by a union bound, the event $\bigcap_{j=0}^{8} E_j$ has probability at least

$$
1 - 0 - \frac{\delta}{2} - \frac{\delta}{512} - \frac{\delta}{4} - \frac{\delta}{32} - 4\frac{\delta}{64} > 1 - \delta.
$$

■

We can obtain the upper bounds for Theorems 4, 5, and 7 from Section 5 by straightforward applications of Lemma 41. Note that, due to the choice of $\hat{\gamma}_\varepsilon$ in each of these proofs, Algorithm 1 is not adaptive to the noise parameters. It is conceivable that this dependence can be removed by a model selection procedure (see Balcan and Hanneke, 2012; Hanneke, 2011, for discussions related to this). However, we do not discuss this further here, leaving this important issue for future work. The upper bounds for Theorems 6 and 8 are based on known results for other algorithms in the literature, though the lower bound for Theorem 6 is new here. The remainder of this section provides the details of these proofs.

Proof of Theorem 4 Fix any $\beta \in [0, 1/2)$, $\varepsilon, \delta \in (0, 1)$, and $\mathcal{P}_{XY} \in \mathrm{BN}(\beta)$. Any $\gamma < 1/2 - \beta$ has $\mathcal{P}(x : \gamma_x \leq \gamma) = 0$, and since we always have $\gamma_\varepsilon \geq \varepsilon/2$, we must have $\gamma_\varepsilon \geq \max\{1/2 - \beta, \varepsilon/2\}$. We may therefore take $\hat{\gamma}_\varepsilon = \max\{1/2 - \beta, \varepsilon/2\}$. Therefore, taking $\bar{k} = k_\varepsilon$ in Lemma 41, the first term in (36) is at most

$$\frac{2^{10}\bar{c}}{(1-2\beta)^2}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right),$$

while the second term in (36) is at most

$$\bar{c}\max\left\{\mathcal{P}\left(x : \gamma_x < \hat{\gamma}_\varepsilon\right), \frac{\varepsilon}{\hat{\gamma}_\varepsilon}\right\}\frac{16}{\hat{\gamma}_\varepsilon\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right).$$

Since $\mathcal{P}\left(x : \gamma_x < 1/2 - \beta\right) = 0 < \frac{\varepsilon}{1/2-\beta}$ and $\mathcal{P}\left(x : \gamma_x < \varepsilon/2\right) \leq 1 < 2 = \frac{\varepsilon}{\varepsilon/2}$, we have that $\mathcal{P}\left(x : \gamma_x < \hat{\gamma}_\varepsilon\right) < \frac{\varepsilon}{\hat{\gamma}_\varepsilon}$, so that the above is at most

$$\frac{64\bar{c}}{(1-2\beta)^2}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{2}{(1-2\beta)\vee\varepsilon}\right).$$

Therefore, recalling that $\mathfrak{s} \geq d$, since Lemma 41 implies that, with any budget n at least the size of the sum of these two terms, Algorithm 1 produces a classifier $\hat{h}_n$ with $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq \varepsilon$ with probability at least $1-\delta$, and requests a number of labels at most n, we have that

$$\begin{aligned}\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) &\leq \frac{2^{10}\bar{c}}{(1-2\beta)^2}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\\ &\quad+ \frac{64\bar{c}}{(1-2\beta)^2}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{2}{(1-2\beta)\vee\varepsilon}\right)\\ &\lesssim \frac{1}{(1-2\beta)^2}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right).\end{aligned}$$

On the other hand, Giné and Koltchinskii (2006) have shown that for the passive learning method of *empirical risk minimization*, producing the classifier $\check{h}_n = \mathrm{argmin}_{h\in\mathbb{C}}\sum_{m=1}^{n}\mathbb{1}[h(X_m) \neq Y_m]$, if n is of size at least

$$\frac{\check{c}}{(1-2\beta)\varepsilon}\left(d\mathrm{Log}\left(\theta_{\mathcal{P}_{XY}}\left(\frac{\varepsilon}{1-2\beta}\right)\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right),$$

for an appropriate finite universal constant $\check{c}$, then with probability at least $1-\delta$, $\mathrm{er}_{\mathcal{P}_{XY}}(\check{h}_n) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq \varepsilon$. Therefore, since Theorem 10 implies $\theta_{\mathcal{P}_{XY}}(\varepsilon/(1-2\beta)) \leq \theta_{\mathcal{P}_{XY}}((\varepsilon/(1-2\beta)) \wedge 1) \leq \min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon} \vee 1\right\}$, we have

$$\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \lesssim \frac{1}{(1-2\beta)\varepsilon}\left(d\mathrm{Log}\left(\min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon}\right\}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

Together, these two bounds on $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta)$ imply the following upper bound, simply by choosing whichever of these two methods has the smaller corresponding bound for the given values of ε, δ, β, d, and $\mathfrak{s}$.

$$\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \lesssim \min \begin{cases} \frac{1}{(1-2\beta)^2}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right) \\ \frac{1}{(1-2\beta)\varepsilon}\left(d\mathrm{Log}\left(\min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon}\right\}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \end{cases}.$$

The statement of the upper bound in Theorem 4 represents a relaxation of this, in that it is slightly larger (in the logarithmic factors), the intention being that it is a simpler expression to state. To arrive at this relaxation, we note that $\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right) \leq \mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon\delta}\right)$, and $d\mathrm{Log}\left(\min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon}\right\}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right) \leq d\mathrm{Log}\left(\frac{1}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$, so that the above is at most

$$\frac{1}{(1-2\beta)^2}\min\left\{\mathfrak{s}, \frac{(1-2\beta)d}{\varepsilon}\right\}\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right).$$

Next, we turn to establishing the lower bound. Fix $\varepsilon \in (0,(1-2\beta)/24)$ and $\delta \in (0,1/24]$. First note that taking $\zeta = \frac{2\varepsilon}{1-2\beta}$ and $k = \min\{\mathfrak{s}-1, \lfloor 1/\zeta \rfloor\}$ in Lemma 26, we have $\mathrm{RR}(k,\zeta,\beta) \subseteq \mathrm{BN}(\beta)$, so that Lemma 26 implies

$$\begin{aligned} \Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) &\geq \Lambda_{\mathrm{RR}(k,\zeta,\beta)}(\varepsilon,\delta) = \Lambda_{\mathrm{RR}(k,\zeta,\beta)}((\zeta/2)(1-2\beta),\delta) \geq \frac{\beta(k-1)\ln\left(\frac{1}{4\delta}\right)}{3(1-2\beta)^2} \\ &\geq \min\left\{\mathfrak{s}-2, \frac{1-2\zeta}{\zeta}\right\}\frac{\beta\ln\left(\frac{1}{4\delta}\right)}{3(1-2\beta)^2} = \frac{\beta}{(1-2\beta)^2}\min\left\{\mathfrak{s}-2, \frac{1-2\beta-4\varepsilon}{2\varepsilon}\right\}\ln\left(\frac{1}{4\delta}\right). \\ &\geq \frac{\beta}{8(1-2\beta)^2}\min\left\{\mathfrak{s}-2, \frac{1-2\beta}{\varepsilon}\right\}\mathrm{Log}\left(\frac{1}{\delta}\right). \end{aligned} \tag{39}$$

Additionally, based on techniques of Kääriäinen (2006); Beygelzimer, Dasgupta, and Langford (2009); Hanneke (2011), the recent article of Hanneke (2014) contains the following lower bound (in the proof of Theorem 4.3 there), for $\varepsilon \in (0,(1-2\beta)/24)$ and $\delta \in (0,1/24]$.

$$\begin{aligned} \Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) &\geq \max\left\{2\left\lfloor \frac{1-(1-2\beta)^2}{2(1-2\beta)^2}\ln\left(\frac{1}{8\delta(1-2\delta)}\right)\right\rfloor, \frac{d-1}{6}\left\lfloor \frac{1-(1-2\beta)^2}{2(1-2\beta)^2}\ln\left(\frac{9}{8}\right)\right\rfloor\right\} \\ &\geq \max\left\{2\left\lfloor \frac{\beta}{(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{8\delta}\right)\right\rfloor, \frac{d-1}{6}\left\lfloor \frac{\beta}{10(1-2\beta)^2}\right\rfloor\right\}. \end{aligned}$$

If $\frac{\beta}{(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{8\delta}\right) \geq 1$, then $2\left\lfloor \frac{\beta}{(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{8\delta}\right)\right\rfloor \geq \frac{\beta}{(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{8\delta}\right) \geq \frac{\beta}{3(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{\delta}\right)$, so that $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \gtrsim \frac{\beta}{(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{\delta}\right)$. Otherwise, if $\frac{\beta}{(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{8\delta}\right) < 1$, then since $\mathrm{RE} \subseteq \mathrm{BN}(\beta)$, and $|\mathbb{C}| \geq 2$ implies $d \geq 1 > \frac{\beta}{(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{8\delta}\right)$, Theorem 3 (proven above) implies we still have $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \geq \Lambda_{\mathrm{RE}}(\varepsilon,\delta) \gtrsim \frac{\beta}{(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{\delta}\right)$ in this case. When $d=1$, these observations further imply $\Lambda_{\mathrm{BN}(\beta)} \gtrsim \frac{d\beta}{(1-2\beta)^2}$. On the other hand, if $d > 1$, and if $\frac{\beta}{10(1-2\beta)^2} \geq 1$, then $\frac{d-1}{6}\left\lfloor \frac{\beta}{10(1-2\beta)^2}\right\rfloor \geq \frac{d}{240}\frac{\beta}{(1-2\beta)^2}$, so that $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \gtrsim \frac{d\beta}{(1-2\beta)^2}$. Otherwise, if $\frac{\beta}{10(1-2\beta)^2} < 1$, then since $\mathrm{RE} \subseteq \mathrm{BN}(\beta)$, Theorem 3 implies we still have $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \geq \Lambda_{\mathrm{RE}}(\varepsilon,\delta) \gtrsim d \gtrsim \frac{d\beta}{(1-2\beta)^2}$ in this case as well. If $\beta > 1/4$, then $\frac{d\beta}{(1-2\beta)^2} \geq \frac{d}{4(1-2\beta)^2} \gtrsim \frac{d}{(1-2\beta)^2}$, so that $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \delta) \gtrsim \frac{d}{(1-2\beta)^2}$. Otherwise, if $\beta \leq 1/4$, then $\frac{1}{(1-2\beta)^2} \leq 4$, so that Theorem 3 implies $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \delta) \geq \Lambda_{\mathrm{RE}}(\varepsilon, \delta) \gtrsim d \gtrsim \frac{d}{(1-2\beta)^2}$. Altogether, we have that

$$\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \delta) \gtrsim \frac{1}{(1-2\beta)^2} \max\left\{\beta \mathrm{Log}\left(\frac{1}{\delta}\right), d\right\}. \tag{40}$$

When $\mathfrak{s} \leq 2$, $\min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon}\right\} \leq 2$, so that (40) trivially implies

$$\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \delta) \gtrsim \frac{1}{(1-2\beta)^2} \max\left\{\min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon}\right\} \beta \mathrm{Log}\left(\frac{1}{\delta}\right), d\right\}. \tag{41}$$

Otherwise, when $\mathfrak{s} \geq 3$, we have $\mathfrak{s} - 2 \geq \mathfrak{s}/3$, so that $\min\left\{\mathfrak{s}-2, \frac{1-2\beta}{\varepsilon}\right\} \geq \frac{1}{3}\min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon}\right\}$. Combined with (39) and (40), this implies (41) holds in this case as well. ■

Proof of Theorem 5 We begin with the upper bounds. Fix any $a \in [1, \infty)$, $\alpha \in (0,1)$, $\varepsilon, \delta \in (0,1)$, and $\mathcal{P}_{XY} \in \mathrm{TN}(a, \alpha)$. For any $\gamma \leq \left(\frac{\varepsilon}{2a'}\right)^{1-\alpha}$, by definition of $\mathrm{TN}(a, \alpha)$, we have $\gamma \mathcal{P}(x : \gamma_x \leq \gamma) \leq a' \gamma^{1/(1-\alpha)} \leq \varepsilon/2$. Therefore, since we always have $\gamma_\varepsilon \geq \varepsilon/2$, we have $\gamma_\varepsilon \geq \max\left\{\left(\frac{\varepsilon}{2a'}\right)^{1-\alpha}, \frac{\varepsilon}{2}\right\}$, so that we can take $\hat{\gamma}_\varepsilon = \max\left\{\left(\frac{\varepsilon}{2a'}\right)^{1-\alpha}, \frac{\varepsilon}{2}\right\}$.

Therefore, taking $\bar{k} = 2$ in Lemma 41 implies that, with any budget n of size at least

$$\bar{c} \sum_{k=2}^{k_\varepsilon} \max\left\{\min\left\{a' 2^{(3-k)\alpha/(1-\alpha)}, 1\right\}, \frac{\varepsilon}{\hat{\gamma}_\varepsilon}\right\} \frac{2^k}{\varepsilon} \left(d \mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right) \mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right), \tag{42}$$

Algorithm 1 produces a classifier $\hat{h}_n$ with $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq \varepsilon$ with probability at least $1-\delta$, and requests a number of labels at most n. This implies $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon, \delta)$ is at most (42).

First note that

$$\begin{aligned}\sum_{k=2}^{k_\varepsilon} \frac{\varepsilon}{\hat{\gamma}_\varepsilon} \frac{2^k}{\varepsilon} &\leq \frac{2^{1+k_\varepsilon}}{\hat{\gamma}_\varepsilon} = \frac{2^{\lceil \log_2(16/\hat{\gamma}_\varepsilon)\rceil}}{\hat{\gamma}_\varepsilon} \leq \frac{32}{\hat{\gamma}_\varepsilon^2} \leq 32 \min\left\{\left(2a'\right)^{2-2\alpha} \varepsilon^{2\alpha-2}, 4\varepsilon^{-2}\right\} \\ &= 32 \min\left\{(2-2\alpha)^{2-2\alpha}(2\alpha)^{2\alpha} a^2 \varepsilon^{2\alpha-2}, 4\varepsilon^{-2}\right\} \leq 128 \min\left\{a^2 \varepsilon^{2\alpha-2}, \varepsilon^{-2}\right\}.\end{aligned} \tag{43}$$

Furthermore, since $\varepsilon^{-2} < a^2\varepsilon^{2\alpha-2}$ only if $\varepsilon > a^{-1/\alpha}$, this is at most $128 \min\left\{a^2\varepsilon^{2\alpha-2}, a^{1/\alpha}\varepsilon^{-1}\right\}$. Also, for $\alpha \geq 1/2$, letting $k_{(a,\alpha)} = \left\lceil \log_2\left(8\left(a'\right)^{(1-\alpha)/\alpha}\right)\right\rceil$, we have $k_{(a,\alpha)} \geq 2$. Additionally, for $\alpha \geq 1/2$, $2^{k\frac{1-2\alpha}{1-\alpha}}$ is nonincreasing in k. In particular, if $k_{(a,\alpha)} = 2$, then

$$\sum_{k=2}^{k_\varepsilon} \min\left\{a' 2^{(3-k)\alpha/(1-\alpha)}, 1\right\} \frac{2^k}{\varepsilon} \leq \sum_{k=k_{(a,\alpha)}}^{k_\varepsilon} \frac{8a'}{\varepsilon} 2^{(k-3)\frac{1-2\alpha}{1-\alpha}} \leq \frac{8k_\varepsilon}{\varepsilon}(a')^{\frac{1-\alpha}{\alpha}}.$$

Otherwise, if $k_{(a,\alpha)} \geq 3$, then

$$\sum_{k=2}^{k_\varepsilon} \min\left\{a' 2^{(3-k)\alpha/(1-\alpha)}, 1\right\} \frac{2^k}{\varepsilon} \leq \sum_{k=2}^{k_{(a,\alpha)}-1} \frac{2^k}{\varepsilon} + \sum_{k=k_{(a,\alpha)}}^{k_\varepsilon} \frac{8a'}{\varepsilon} 2^{(k-3)\frac{1-2\alpha}{1-\alpha}}.$$
$$\leq \frac{16}{\varepsilon}(a')^{\frac{1-\alpha}{\alpha}} + \frac{8(k_\varepsilon - 2)}{\varepsilon}(a')^{\frac{1-\alpha}{\alpha}} = \frac{8k_\varepsilon}{\varepsilon}(a')^{\frac{1-\alpha}{\alpha}}.$$

Furthermore, since $(1-\alpha)^{\frac{1-\alpha}{\alpha}} \leq 1$, we have

$$\frac{8k_\varepsilon}{\varepsilon}(a')^{\frac{1-\alpha}{\alpha}} = \frac{8k_\varepsilon}{\varepsilon}(1-\alpha)^{\frac{1-\alpha}{\alpha}}(2\alpha)a^{1/\alpha} \leq \frac{16k_\varepsilon}{\varepsilon}a^{1/\alpha}.$$

Therefore, in either case, when $\alpha \geq 1/2$, (42) is at most

$$\bar{c}\left(16k_\varepsilon a^{1/\alpha}\varepsilon^{-1} + 128a^{1/\alpha}\varepsilon^{-1}\right)\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right)$$
$$\leq 767\bar{c}\frac{a^{1/\alpha}}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}^2\left(\frac{1}{\varepsilon}\right),$$

which is therefore an upper bound on $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta)$ in this case.

Otherwise, if $\alpha \leq 1/2$, then $2^{k\frac{1-2\alpha}{1-\alpha}}$ is nondecreasing in k, so that

$$\sum_{k=2}^{k_\varepsilon} \min\left\{a' 2^{(3-k)\alpha/(1-\alpha)}, 1\right\} \frac{2^k}{\varepsilon} \leq \sum_{k=2}^{k_\varepsilon} 8a' 2^{(k-3)\frac{1-2\alpha}{1-\alpha}} \frac{1}{\varepsilon} \leq (k_\varepsilon - 1)8a' 2^{(k_\varepsilon-3)\frac{1-2\alpha}{1-\alpha}} \frac{1}{\varepsilon}$$
$$\leq (k_\varepsilon - 1)8a'\left(\frac{2}{\hat{\gamma}_\varepsilon}\right)^{\frac{1-2\alpha}{1-\alpha}} \frac{1}{\varepsilon} \leq (k_\varepsilon - 1)8a' 2^{\frac{1-2\alpha}{1-\alpha}}\left(2a'\right)^{1-2\alpha}\left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \leq (k_\varepsilon - 1)32\left(\frac{a'}{\varepsilon}\right)^{2-2\alpha}$$
$$= (k_\varepsilon - 1)32(1-\alpha)^{2-2\alpha}(2\alpha)^{2\alpha}a^2\varepsilon^{2\alpha-2} \leq (k_\varepsilon - 1)32a^2\varepsilon^{2\alpha-2}.$$

Therefore, (42) is at most

$$\bar{c}\left((k_\varepsilon - 1)32a^2\varepsilon^{2\alpha-2} + 128a^2\varepsilon^{2\alpha-2}\right)\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right)$$
$$\leq 832\bar{c}a^2\varepsilon^{2\alpha-2}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}^2\left(\frac{1}{\varepsilon}\right).$$

In particular, this implies $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta)$ is at most this large when $\alpha \leq 1/2$. Furthermore, this completes the proof of the upper bound for the cases where either $\alpha \leq 1/2$, or $\alpha \geq 1/2$ and $\frac{\mathfrak{s}}{d} \geq \frac{1}{a^{1/\alpha}\varepsilon}$.

Next, consider the remaining case that $\alpha \geq 1/2$ and $\frac{\mathfrak{s}}{d} < \frac{1}{a^{1/\alpha}\varepsilon}$. In particular, this requires that $\mathfrak{s} < \infty$, and since $\mathfrak{s} \geq d$, that $\varepsilon < a^{-1/\alpha}$. In this case, let us take

$$\bar{k} = 3 + \left\lceil (1-\alpha)\log_2\left(\frac{k_\varepsilon a'}{8\varepsilon}\frac{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)}{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)}\right)\right\rceil.$$

Since $\mathfrak{s} \geq d$, we have $\frac{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)} \leq \frac{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)}{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)} = \frac{\mathfrak{s}}{d}$, so that, since $\frac{\mathfrak{s}}{d} < \frac{1}{a^{1/\alpha}\varepsilon}$, we have $\frac{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)} < \frac{1}{a^{1/\alpha}\varepsilon}$. A bit of algebra reveals that, in this case, $\bar{k} \geq 2$. Therefore, in this case, Lemma 41 implies that, with any budget n of size at least

$$\begin{aligned}
&\bar{c}2^{2\bar{k}}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\\
&\bar{c}\sum_{k=\bar{k}}^{k_\varepsilon}\max\left\{\min\left\{a'2^{(3-k)\alpha/(1-\alpha)},1\right\},\frac{\varepsilon}{\hat{\gamma}_\varepsilon}\right\}\frac{2^k}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right),
\end{aligned} \tag{44}$$

Algorithm 1 produces a classifier $\hat{h}_n$ with $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq \varepsilon$ with probability at least $1-\delta$, and requests a number of labels at most n. This implies $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta)$ is at most (44).

Now note that

$$\begin{aligned}
&2^{2\bar{k}}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\\
&\leq 256\left(\frac{k_\varepsilon a'}{8\varepsilon}\frac{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}\right)^{2-2\alpha}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\\
&\leq 1024a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(\frac{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}\right)^{2\alpha-1}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}^{2-2\alpha}\left(\frac{1}{\varepsilon}\right).
\end{aligned}$$

Also, since $\alpha \geq 1/2$, $2^{k\frac{1-2\alpha}{1-\alpha}}$ is nonincreasing in k, so that

$$\begin{aligned}
&\sum_{k=\bar{k}}^{k_\varepsilon}a'2^{(3-k)\alpha/(1-\alpha)}\frac{2^k}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\\
&\leq \frac{8a'k_\varepsilon}{\varepsilon}2^{(\bar{k}-3)\frac{1-2\alpha}{1-\alpha}}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\\
&\leq \frac{8a'k_\varepsilon}{\varepsilon}\left(\frac{k_\varepsilon a'}{8\varepsilon}\frac{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}\right)^{1-2\alpha}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\\
&\leq 256a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(\frac{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}\right)^{2\alpha-1}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}^{2-2\alpha}\left(\frac{1}{\varepsilon}\right).
\end{aligned}$$

Furthermore, by (43),

$$\begin{aligned}
&\bar{c}\sum_{k=\bar{k}}^{k_\varepsilon}\frac{\varepsilon}{\hat{\gamma}_\varepsilon}\frac{2^k}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right) \leq 128a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\\
&\leq 128a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(\frac{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}\right)^{2\alpha-1}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}^{2-2\alpha}\left(\frac{1}{\varepsilon}\right).
\end{aligned}$$

Therefore, since $\mathrm{Log}\left(\frac{1}{\gamma_\varepsilon}\right) \le \mathrm{Log}\left(\frac{2}{\varepsilon}\right) \le 2\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$, (44) is at most

$$2^{11}\bar{c}a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(\frac{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}\right)^{2\alpha-1}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}^{3-2\alpha}\left(\frac{1}{\varepsilon}\right). \tag{45}$$

The upper bound for the case $\alpha \ge 1/2$ and $\frac{\mathfrak{s}}{d} < \frac{1}{a^{1/\alpha}\varepsilon}$ then follows by further relaxing this (purely to simplify the theorem statement), noting that $\mathrm{Log}^{3-2\alpha}\left(\frac{1}{\varepsilon}\right) \le \mathrm{Log}^2\left(\frac{1}{\varepsilon}\right)$, and $\frac{\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)}{d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)} \le \frac{\mathfrak{s}}{d}$.

Next, we turn to establishing the lower bound. Fix any $a \in [4,\infty)$, $\alpha \in (0,1)$, $\delta \in (0,1/24]$, and $\varepsilon \in \left(0, 1/(24a^{1/\alpha})\right)$. For this range of values, the recent article of Hanneke (2014) proves a lower bound of

$$\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta) \gtrsim a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\left(d+\mathrm{Log}\left(\frac{1}{\delta}\right)\right),$$

based on techniques of Kääriäinen (2006); Beygelzimer, Dasgupta, and Langford (2009); Hanneke (2011). It remains only to establish the remaining term in the lower bound for the case when $\alpha > 1/2$, via Lemma 26. In the cases that $\mathfrak{s} \le 2$, this term is implied by the above $a^2\varepsilon^{2\alpha-2}\mathrm{Log}\left(\frac{1}{\delta}\right)$ lower bound. For the remainder of the proof, suppose $\mathfrak{s} \ge 3$ and $\alpha > 1/2$. Let

$$k = \min\left\{\mathfrak{s}-1, \left\lfloor\frac{(a')^{\frac{\alpha-1}{\alpha}}}{\varepsilon}\right\rfloor, \left\lfloor\frac{a'}{\varepsilon}4^{-\frac{1}{1-\alpha}}\right\rfloor\right\},$$

$\beta = \frac{1}{2} - \left(\frac{k\varepsilon}{a'}\right)^{1-\alpha}$, and $\zeta = \frac{2\varepsilon}{1-2\beta}$; note that $\zeta \in (0,1]$, $\beta \in [0,1/2)$, and $2 \le k \le \min\{\mathfrak{s}-1, \lfloor 1/\zeta\rfloor\}$; in particular, the fact that $k \le \lfloor 1/\zeta \rfloor$ is established by concavity of the $x \mapsto \frac{(a')^{\alpha-1}}{\varepsilon^\alpha}x^{1-\alpha}$ function, which equals x at both $x = 0$ and $x = x_0 = \frac{(a')^{\frac{\alpha-1}{\alpha}}}{\varepsilon}$; since this function is $1/\zeta$ at $x = k$, and $0 < k \le x_0$, concavity of the function implies $1/\zeta \ge k$, and integrality of k implies $\lfloor 1/\zeta\rfloor \ge k$ as well. Also note that any $\mathcal{P}_{XY} \in \mathrm{RR}(k,\zeta,\beta)$ has a marginal distribution $\mathcal{P}$ such that

$$\begin{aligned}\mathcal{P}\left(x : |\eta(x;\mathcal{P}_{XY}) - 1/2| \le 1/2-\beta\right) &= k\zeta = k\varepsilon\frac{2}{1-2\beta} \\ &= a'\left(1/2-\beta\right)^{\frac{1}{1-\alpha}}\frac{2}{1-2\beta} = a'\left(1/2-\beta\right)^{\frac{\alpha}{1-\alpha}}.\end{aligned}$$

Since every point x in the support of $\mathcal{P}_{k,\zeta}$ has either $|\eta(x;\mathcal{P}_{XY})-1/2| = 1/2-\beta$ or $|\eta(x;\mathcal{P}_{XY})-1/2| = 1/2$, this implies that any $\gamma \in [1/2-\beta, 1/2)$ has $\mathcal{P}\left(x : |\eta(x;\mathcal{P}_{XY})-1/2| \le \gamma\right) = \mathcal{P}\left(x : |\eta(x;\mathcal{P}_{XY})-1/2| \le 1/2-\beta\right) = a'\left(1/2-\beta\right)^{\alpha/(1-\alpha)} \le a'\gamma^{\alpha/(1-\alpha)}$, while any $\gamma \ge 1/2$ always has $\mathcal{P}\left(x : |\eta(x;\mathcal{P}_{XY})-1/2| \le \gamma\right) = 1 \le a'\gamma^{\alpha/(1-\alpha)}$. Furthermore, any $\gamma \in (0, 1/2-\beta)$ has $\mathcal{P}(x : |\eta(x;\mathcal{P}_{XY})-1/2| \le \gamma) = 0 \le a'\gamma^{\alpha/(1-\alpha)}$. Thus, $\mathcal{P}_{XY} \in \mathrm{TN}(a,\alpha)$ as well. Since this holds for every $\mathcal{P}_{XY} \in \mathrm{RR}(k,\zeta,\beta)$, this implies $\mathrm{RR}(k,\zeta,\beta) \subseteq \mathrm{TN}(a,\alpha)$. Therefore, Lemma 26 implies

$$\begin{aligned}\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta) \ge \Lambda_{\mathrm{RR}(k,\zeta,\beta)}(\varepsilon,\delta) &= \Lambda_{\mathrm{RR}(k,\zeta,\beta)}((\zeta/2)(1-2\beta),\delta) \\ &\ge \frac{\beta(k-1)\ln\left(\frac{1}{4\delta}\right)}{3(1-2\beta)^2} \gtrsim \frac{\beta(k-1)}{(1-2\beta)^2}\mathrm{Log}\left(\frac{1}{\delta}\right).\end{aligned} \tag{46}$$

Finally, note that

$$\frac{\beta(k-1)}{(1-2\beta)^2} = \left(\frac{1}{2} - \left(\frac{k\varepsilon}{a'}\right)^{1-\alpha}\right)\frac{1}{4}\left(\frac{a'}{k\varepsilon}\right)^{2-2\alpha}(k-1) \geq \frac{1}{16}\left(\frac{a'}{\varepsilon}\right)^{2-2\alpha} k^{2\alpha-2}(k-1)$$
$$\geq \frac{1}{32}\left(\frac{a'}{\varepsilon}\right)^{2-2\alpha}(k-1)^{2\alpha-1} \geq \frac{a^2}{64}\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}(k-1)^{2\alpha-1}. \tag{47}$$

Since $a \geq 4$,

$$\left(a'\right)^{\frac{\alpha-1}{\alpha}} = a'\left(a'\right)^{-1/\alpha} = a'(1-\alpha)^{-1/\alpha}(2\alpha)^{-1/(1-\alpha)}a^{-\frac{1}{\alpha(1-\alpha)}}$$
$$\leq a'(1-\alpha)^{-1/\alpha}(2\alpha)^{-1/(1-\alpha)}4^{-\frac{1}{\alpha(1-\alpha)}} = a'\left(4^{1/\alpha}(1-\alpha)^{(1-\alpha)/\alpha}(2\alpha)\right)^{-1/(1-\alpha)}.$$

One can easily verify that $4^{1/\alpha}(1-\alpha)^{(1-\alpha)/\alpha}(2\alpha) \geq 6$ for $\alpha \in (1/2, 1)$ (with minimum achieved at $\alpha = 3/4$), so that $a'\left(4^{1/\alpha}(1-\alpha)^{(1-\alpha)/\alpha}(2\alpha)\right)^{-1/(1-\alpha)} \leq a'6^{-1/(1-\alpha)} \leq a'4^{-1/(1-\alpha)}$. Thus, $\frac{(a')^{\frac{\alpha-1}{\alpha}}}{\varepsilon} \leq \frac{a'}{\varepsilon}4^{-\frac{1}{1-\alpha}}$, so that the third term in the definition of k is redundant. Therefore, (47) is at least

$$\frac{a^2}{64}\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\min\left\{\mathfrak{s}-2, \frac{(a')^{\frac{\alpha-1}{\alpha}}}{\varepsilon}-2\right\}^{2\alpha-1} \geq \frac{a^2}{64}\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\min\left\{\mathfrak{s}-2, \frac{1}{2a^{1/\alpha}\varepsilon}-2\right\}^{2\alpha-1}$$
$$\geq \frac{a^2}{64}\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\min\left\{\frac{\mathfrak{s}}{3}, \frac{1}{3a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1} \geq \frac{a^2}{192}\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\min\left\{\mathfrak{s}, \frac{1}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1}.$$

Plugging this into (46) completes the proof. ■

As an aside, we note that it is possible to improve the logarithmic factors in the upper bound in Theorem 5. One clear refinement comes from using (45) directly (rather than relaxing the factor depending on $\mathfrak{s}$). We can further reduce the bound by another logarithmic factor when α is bounded away from $1/2$ by noting that the summations of terms $2^{(k-3)\frac{1-2\alpha}{1-\alpha}}$ in the above proof are geometric in that case. We also note that, for very large values of a, the bounds (proven below) for $\Lambda_{\mathrm{BE}(1/2)}(\varepsilon, \delta)$ may be more informative than those derived above.

Proof of Theorem 6 The technique leading to Lemma 41 does not apply to $\mathrm{BC}(a,\alpha)$, since we are not guaranteed $f^{\star}_{\mathcal{P}_{XY}} \in \mathbb{C}$ for $\mathcal{P}_{XY} \in \mathrm{BC}(a,\alpha)$. We therefore base the upper bounds in Theorem 6 directly on existing results in the literature, in combination with Theorem 10. Thus, the proof of this upper bound does not provide any new insights on improving the design of active learning algorithms for distributions in $\mathrm{BC}(a,\alpha)$. Rather, it merely re-expresses the known results, in terms of the star number instead of a distribution-dependent complexity measure. The lower bounds are directly inherited from Theorem 5.

Fix any $a \in [1,\infty)$, $\alpha \in [0,1]$, and $\varepsilon, \delta \in (0,1)$. Following the work of Hanneke (2009a, 2011) and Koltchinskii (2010), the recent work of Hanneke and Yang (2012) studies an algorithm proposed by Hanneke (2012) (a modified variant of the A^2 algorithm of Balcan, Beygelzimer, and Langford, 2006, 2009), and shows that there exists a finite universal constant $\mathring{c} \geq 1$ such that, for any $\mathcal{P}_{XY} \in \mathrm{BC}(a,\alpha)$, for any budget n of size at least

$$\mathring{c}a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\theta_{\mathcal{P}_{XY}}(a\varepsilon^{\alpha})\left(d\mathrm{Log}\left(\theta_{\mathcal{P}_{XY}}(a\varepsilon^{\alpha})\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right), \tag{48}$$

the algorithm produces a classifier $\hat{h}_n$ with $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) - \inf_{h\in\mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) \leq \varepsilon$ with probability at least $1-\delta/4$, and requests a number of labels at most n (see also Hanneke, 2009b,a, 2011, 2012, 2014; Koltchinskii, 2010, for similar results for related methods). By Theorem 10, when $a\varepsilon^{\alpha} \leq 1$, (48) is at most

$$\mathring{c}a^2 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^{\alpha}}\right\} \left(d\mathrm{Log}\left(\min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^{\alpha}}\right\}\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right) \mathrm{Log}\left(\frac{1}{\varepsilon}\right), \quad (49)$$

which is therefore an upper bound on $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta)$. We can also extend this to the case $a\varepsilon^{\alpha} > 1$ as follows. Vapnik and Chervonenkis (1971); Vapnik (1982, 1998) have proven that the sample complexity of passive learning satisfies

$$\mathcal{M}_{\mathrm{AG}(1)}(\varepsilon,\delta) \lesssim \frac{1}{\varepsilon^2}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

In the case $a\varepsilon^{\alpha} > 1$, this is at most

$$\begin{aligned} & a\left(\frac{1}{\varepsilon}\right)^{2-\alpha}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \\ & = a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^{\alpha}}\right\}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \\ & \leq a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^{\alpha}}\right\}\left(d\mathrm{Log}\left(\min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^{\alpha}}\right\}\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right). \end{aligned}$$

Therefore, since $\Lambda_{\mathrm{AG}(1)}(\varepsilon,\delta) \leq \mathcal{M}_{\mathrm{AG}(1)}(\varepsilon,\delta)$ and $\mathrm{BC}(a,\alpha) \subseteq \mathrm{AG}(1)$, we may conclude that, regardless of whether $a\varepsilon^{\alpha}$ is greater than or less than 1, we have that $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta)$ is bounded by a value proportional to (49). To match the form of the upper bound stated in Theorem 6, we can simply relax this, noting that $d\mathrm{Log}\left(\min\left\{\mathfrak{s}, \frac{1}{a\varepsilon^{\alpha}}\right\}\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right) \leq 2d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right) \leq 2d\mathrm{Log}\left(\frac{1}{\varepsilon\delta}\right)$.

Next, turning to the lower bound, recall that $\mathrm{TN}(a,\alpha) \subseteq \mathrm{BC}(a,\alpha)$, so that $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta) \leq \Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta)$ (Mammen and Tsybakov, 1999; Tsybakov, 2004). Thus, the lower bound in Theorem 5 (proven above) for $\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta)$ also applies to $\Lambda_{\mathrm{BC}(a,\alpha)}(\varepsilon,\delta)$. ■

Proof of Theorem 7 Again, we begin with the upper bound. Fix any $\nu \in [0,1/2]$, $\varepsilon,\delta \in (0,1)$, and $\mathcal{P}_{XY} \in \mathrm{BE}(\nu)$. The case $\nu = 0$ is already addressed by the upper bound in Theorem 3; we therefore focus the remainder of the proof on the case of $\nu > 0$. For $(X,Y) \sim \mathcal{P}_{XY}$, any $x \in \mathcal{X}$ has $1 - 2\mathbb{P}(Y \neq f^{\star}_{\mathcal{P}_{XY}}(X)|X=x) = 2\gamma_x$. Therefore, for any $\gamma \in [0,1/2)$, any $x \in \mathcal{X}$ with $\gamma_x \leq \gamma$ has $\mathbb{P}(Y \neq f^{\star}_{\mathcal{P}_{XY}}(X)|X=x) \geq 1/2-\gamma$. Thus, Markov's inequality implies

$$\mathcal{P}(x : \gamma_x \leq \gamma) \leq \mathcal{P}(x : \mathbb{P}(Y \neq f^{\star}_{\mathcal{P}_{XY}}(X)|X=x) \geq 1/2-\gamma) \leq \frac{2}{1-2\gamma}\mathrm{er}_{\mathcal{P}_{XY}}(f^{\star}_{\mathcal{P}_{XY}}) \leq \frac{2\nu}{1-2\gamma}. \quad (50)$$

In particular, this implies that for $\gamma \leq \frac{\varepsilon}{4\nu+2\varepsilon}$, $\gamma\mathcal{P}(x : \gamma_x \leq \gamma) \leq \frac{2\nu\gamma}{1-2\gamma} \leq \frac{2\nu/(2\nu+\varepsilon)}{1-\varepsilon/(2\nu+\varepsilon)}\frac{\varepsilon}{2} = \frac{\varepsilon}{2}$. Thus, $\gamma_{\varepsilon} \geq \frac{\varepsilon}{4\nu+2\varepsilon}$. We can therefore take $\hat{\gamma}_{\varepsilon} = \max\left\{\frac{\varepsilon}{4\nu+2\varepsilon}, \frac{\varepsilon}{2}\right\}$.

Also note that any $\gamma \geq 0$ has $\mathcal{P}(x : \gamma_x \leq \gamma) \leq 1$, so that together with (50), we have $\mathcal{P}(x : \gamma_x \leq \gamma) \leq \frac{2\nu}{1-\min\{2\gamma, 1-2\nu\}}$. Now taking $\bar{k} = 2$, Lemma 41 implies that, with any budget n of size at least

$$\bar{c} \sum_{k=2}^{k_\varepsilon} \max\left\{ \frac{2\nu}{1 - \min\{2^{4-k}, 1-2\nu\}}, \frac{\varepsilon}{\hat{\gamma}_\varepsilon} \right\} \frac{2^k}{\varepsilon} \left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right) \right) \mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right) \mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right), \tag{51}$$

Algorithm 1 produces a classifier $\hat{h}_n$ with $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq \varepsilon$ with probability at least $1-\delta$, and requests a number of labels at most n. This implies $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta)$ is at most (51). Now note that

$$\sum_{k=2}^{k_\varepsilon} \frac{\varepsilon}{\hat{\gamma}_\varepsilon} \frac{2^k}{\varepsilon} \leq \frac{1}{\hat{\gamma}_\varepsilon} 2^{1+k_\varepsilon} \leq 512 \left(\frac{\nu+\varepsilon}{\varepsilon}\right)^2. \tag{52}$$

Next, we have

$$\begin{aligned} &\sum_{k=2}^{k_\varepsilon} \frac{2\nu}{1-\min\{2^{4-k}, 1-2\nu\}} \frac{2^k}{\varepsilon} \leq \frac{28}{\varepsilon} + \sum_{k=5}^{k_\varepsilon} \frac{2\nu}{1-2^{4-k}} \frac{2^k}{\varepsilon} \leq \frac{28}{\varepsilon} + \sum_{k=5}^{k_\varepsilon} \frac{4\nu}{\varepsilon} 2^k \\ &\leq \frac{28}{\varepsilon} + \frac{4\nu}{\varepsilon} 2^{1+k_\varepsilon} \leq \frac{28}{\varepsilon} + \frac{128\nu}{\varepsilon\hat{\gamma}_\varepsilon} \leq \frac{28}{\varepsilon} + 512\left(\frac{\nu+\varepsilon}{\varepsilon}\right)^2. \end{aligned}$$

Therefore, (51) is at most

$$\begin{aligned} &2^{10}\bar{c}\left(\left(\frac{\nu+\varepsilon}{\varepsilon}\right)^2 + \frac{1}{\varepsilon}\right)\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right) \\ &\quad \leq 2^{10} 3\bar{c}\left(\left(\frac{\nu+\varepsilon}{\varepsilon}\right)^2 + \frac{1}{\varepsilon}\right)\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{\nu+\varepsilon}{\varepsilon}\right). \end{aligned} \tag{53}$$

Next, consider taking $\bar{k} = 5$. Lemma 41 implies that, with any budget n of size at least

$$\begin{aligned} &2^{10}\bar{c}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right) \\ &\quad + \bar{c}\sum_{k=5}^{k_\varepsilon} \max\left\{\frac{2\nu}{1-2^{4-k}}, \frac{\varepsilon}{\hat{\gamma}_\varepsilon}\right\} \frac{2^k}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right), \end{aligned} \tag{54}$$

Algorithm 1 produces a classifier $\hat{h}_n$ with $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) - \mathrm{er}_{\mathcal{P}_{XY}}(f^\star_{\mathcal{P}_{XY}}) \leq \varepsilon$ with probability at least $1-\delta$, and requests a number of labels at most n. This implies $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta)$ is at most (54). As above, we have

$$\sum_{k=5}^{k_\varepsilon} \frac{2\nu}{1-2^{4-k}} \frac{2^k}{\varepsilon} \leq 512\left(\frac{\nu+\varepsilon}{\varepsilon}\right)^2.$$

Combined with (52), this implies (54) is at most

$$
\begin{aligned}
&2^{10}\bar{c}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\\
&+2^{10}\bar{c}\left(\frac{\nu+\varepsilon}{\varepsilon}\right)^2\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\hat{\gamma}_\varepsilon}\right)\\
&\leq 2^{10}\bar{c}\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\\
&+2^{10}3\bar{c}\left(\frac{\nu+\varepsilon}{\varepsilon}\right)^2\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{\nu+\varepsilon}{\varepsilon}\right). \qquad (55)
\end{aligned}
$$

In particular, when $\left(\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right) < \frac{3}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\mathrm{Log}\left(\frac{\nu+\varepsilon}{\varepsilon}\right)$, this is smaller than (53). Thus, the minimum of these two expressions upper bounds $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta)$.

To simplify the expression of this bound into the form given in the statement of Theorem 7, we note that $d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right) \leq d\mathrm{Log}\left(\frac{1}{\varepsilon\delta}\right)$, $\mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right) \leq \mathfrak{s}\mathrm{Log}\left(\frac{1}{\varepsilon\delta}\right)$, $\mathrm{Log}\left(\frac{\nu+\varepsilon}{\varepsilon}\right) \leq \mathrm{Log}\left(\frac{1}{\varepsilon}\right)$,$\left(\frac{\nu+\varepsilon}{\varepsilon}\right)^2 \leq 4\frac{\max\{\nu,\varepsilon\}^2}{\varepsilon^2} \leq 4\left(\frac{\nu^2}{\varepsilon^2}+1\right)$, and $d \leq \min\left\{\mathfrak{s},\frac{d}{\varepsilon}\right\}$, so that the minimum of (53) and (55) is at most

$$
\begin{aligned}
2^{12}3\bar{c}&\left(\left(\frac{\nu^2}{\varepsilon^2}+1\right)d+\min\left\{\mathfrak{s},\frac{d}{\varepsilon}\right\}\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\\
&\leq 2^{13}3\bar{c}\left(\frac{\nu^2}{\varepsilon^2}d+\min\left\{\mathfrak{s},\frac{d}{\varepsilon}\right\}\right)\mathrm{Log}\left(\frac{d}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon\delta}\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right).
\end{aligned}
$$

This completes the proof of the upper bound.

Next, we turn to establishing the lower bound. Fix $\nu \in [0,1/2)$, $\varepsilon \in (0,(1-2\nu)/24)$, and $\delta \in (0,1/24]$. Based on the works of Kääriäinen (2006); Hanneke (2007a); Beygelzimer, Dasgupta, and Langford (2009), the recent article of Hanneke (2014) contains the following lower bound (in the proof of Theorem 4.3 there), letting $\gamma = \frac{12\varepsilon}{\nu+12\varepsilon}$.

$$
\begin{aligned}
\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) &\geq \max\left\{2\left\lfloor\frac{1-\gamma^2}{2\gamma^2}\ln\left(\frac{1}{8\delta(1-2\delta)}\right)\right\rfloor,\frac{d-1}{6}\left\lfloor\frac{1-\gamma^2}{2\gamma^2}\ln\left(\frac{9}{8}\right)\right\rfloor\right\}\\
&\geq \max\left\{2\left\lfloor\frac{1-\gamma^2}{2\gamma^2}\ln\left(\frac{1}{8\delta}\right)\right\rfloor,\frac{d-1}{6}\left\lfloor\frac{1-\gamma^2}{17\gamma^2}\right\rfloor\right\} \qquad (56)
\end{aligned}
$$

If $\frac{1-\gamma^2}{2\gamma^2}\ln\left(\frac{1}{8\delta}\right) \geq 1$, then $2\left\lfloor\frac{1-\gamma^2}{2\gamma^2}\ln\left(\frac{1}{8\delta}\right)\right\rfloor \geq \frac{1-\gamma^2}{2\gamma^2}\ln\left(\frac{1}{8\delta}\right)$, so that (56) implies $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \gtrsim \frac{1-\gamma^2}{\gamma^2}\mathrm{Log}\left(\frac{1}{\delta}\right)$. Otherwise, if $\frac{1-\gamma^2}{2\gamma^2}\ln\left(\frac{1}{8\delta}\right) < 1$, then since $\mathrm{RE} \subseteq \mathrm{BE}(\nu)$, and $|\mathbb{C}| \geq 2$ implies $d \geq 1 > \frac{1-\gamma^2}{2\gamma^2}\ln\left(\frac{1}{8\delta}\right)$, Theorem 3 (proven above) implies $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \geq \Lambda_{\mathrm{RE}}(\varepsilon,\delta) \gtrsim d \gtrsim \frac{1-\gamma^2}{\gamma^2}\mathrm{Log}\left(\frac{1}{\delta}\right)$ in this case as well. If $d=1$, these observations further imply $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \gtrsim d\frac{1-\gamma^2}{\gamma^2}$. On the other hand, if $d \geq 2$, and if $\frac{1-\gamma^2}{17\gamma^2} \geq 1$, then $\frac{d-1}{6}\left\lfloor\frac{1-\gamma^2}{17\gamma^2}\right\rfloor \geq \frac{d}{408}\frac{1-\gamma^2}{\gamma^2}$, so that (56) implies $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \gtrsim d\frac{1-\gamma^2}{\gamma^2}$. Otherwise, if $\frac{1-\gamma^2}{17\gamma^2} < 1$, then since $\mathrm{RE} \subseteq \mathrm{BE}(\nu)$, Theorem 3 implies we still have $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \geq \Lambda_{\mathrm{RE}}(\varepsilon,\delta) \gtrsim d \gtrsim d\frac{1-\gamma^2}{\gamma^2}$ in this case as well. Altogether, we have that

$$
\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \gtrsim \frac{1-\gamma^2}{\gamma^2}\max\left\{d,\mathrm{Log}\left(\frac{1}{\delta}\right)\right\} \gtrsim \frac{1-\gamma^2}{\gamma^2}\left(d+\mathrm{Log}\left(\frac{1}{\delta}\right)\right). \qquad (57)
$$

When $\nu \geq 12\varepsilon$, $\gamma \leq 1/2$, so that (57) implies

$$\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \gtrsim \frac{1}{\gamma^2}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) = \left(\frac{\nu+12\varepsilon}{12\varepsilon}\right)^2\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) \gtrsim \frac{\nu^2}{\varepsilon^2}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right).$$

Otherwise, if $\nu < 12\varepsilon$, then

$$\frac{1-\gamma^2}{\gamma^2} = \frac{(1-\gamma)(1+\gamma)}{\gamma^2} = \left(\frac{\nu+12\varepsilon}{12\varepsilon}\right)^2\left(\frac{\nu}{\nu+12\varepsilon}\right)\left(\frac{\nu+24\varepsilon}{\nu+12\varepsilon}\right) \geq \frac{\nu}{\nu+12\varepsilon} \geq \frac{\nu}{12\varepsilon} \geq \frac{\nu^2}{144\varepsilon^2}. \tag{58}$$

Therefore, if $\nu < 12\varepsilon$, (57) implies that $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \gtrsim \frac{\nu^2}{\varepsilon^2}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)$ in this case as well. It remains only to establish the final term in the lower bound. For this, we simply note that $\mathrm{RE} \subseteq \mathrm{BE}(\nu)$, so that Theorem 3 implies $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \geq \Lambda_{\mathrm{RE}}(\varepsilon,\delta) \gtrsim \min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}$. Combining these results implies

$$\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \gtrsim \max\left\{\frac{\nu^2}{\varepsilon^2}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right), \min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}\right\} \gtrsim \frac{\nu^2}{\varepsilon^2}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right) + \min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}.$$

■

Examining the proof of the lower bound for $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta)$, we note that this argument also establishes a slightly stronger lower bound in the case $\varepsilon > \nu$. Specifically, if we use the expression just left of the right-most inequality in (58), rather than the right-most expression, we find that we can add a term $\frac{\nu}{\varepsilon}\mathrm{Log}\left(\frac{1}{\delta}\right)$ to the stated lower bound. This term can be larger than the stated term $\frac{\nu^2}{\varepsilon^2}\mathrm{Log}\left(\frac{1}{\delta}\right)$ when $\varepsilon > \nu$. Additionally, since $\mathrm{RE} \subseteq \mathrm{BE}(\nu)$, we can of course also add a term d to the stated lower bound, which again would increase the bound when $\varepsilon > \nu$.

Proof of Theorem 8 Again, we begin with the upper bounds. As with the proof of Theorem 6, we cannot use the technique leading to Lemma 41; we turn instead to a simple combination of an upper bound from the literature, combined with Theorem 10.

Fix any $\nu \in [0,1]$ and $\varepsilon, \delta \in (0,1)$. Following the work of Hanneke (2007b); Dasgupta, Hsu, and Monteleoni (2007); Koltchinskii (2010), the recent work of Hanneke (2014) studies a modified variant of the A^2 algorithm of Balcan, Beygelzimer, and Langford (2006, 2009), showing that there exists a finite universal constant $\ddot{c} \geq 1$ such that, for any $\mathcal{P}_{XY} \in \mathrm{AG}(\nu)$, for any budget n of size at least

$$\ddot{c}\theta_{\mathcal{P}_{XY}}(\nu+\varepsilon)\left(\frac{\nu^2}{\varepsilon^2} + \mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right)\left(d\mathrm{Log}\left(\theta_{\mathcal{P}_{XY}}(\nu+\varepsilon)\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right), \tag{59}$$

the algorithm produces a classifier $\hat{h}_n$ with $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}_n) - \inf_{h\in\mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) \leq \varepsilon$ with probability at least $1-\delta$, and requests a number of labels at most n (see also Dasgupta, Hsu, and Monteleoni, 2007; Beygelzimer, Dasgupta, and Langford, 2009, for similar results for related methods). By Theorem 10,

$$\theta_{\mathcal{P}_{XY}}(\nu+\varepsilon) = \theta_{\mathcal{P}_{XY}}((\nu+\varepsilon)\wedge 1) \leq \min\left\{\mathfrak{s}, \frac{1}{(\nu+\varepsilon)\wedge 1}\right\} \leq \min\left\{\mathfrak{s}, \frac{2}{\nu+\varepsilon}\right\} \leq 2\min\left\{\mathfrak{s}, \frac{1}{\nu+\varepsilon}\right\},$$

while $\mathrm{Log}\left(\theta_{\mathcal{P}_{XY}}(\nu+\varepsilon)\right) \leq \mathrm{Log}\left(\min\left\{\mathfrak{s}, \frac{1}{\nu+\varepsilon}\right\} \vee 1\right) = \mathrm{Log}\left(\min\left\{\mathfrak{s}, \frac{1}{\nu+\varepsilon}\right\}\right)$. Therefore, (59) is at most

$$2\ddot{c}\min\left\{\mathfrak{s}, \frac{1}{\nu+\varepsilon}\right\}\left(\frac{\nu^2}{\varepsilon^2} + \mathrm{Log}\left(\frac{1}{\varepsilon}\right)\right)\left(d\mathrm{Log}\left(\min\left\{\mathfrak{s}, \frac{1}{\nu+\varepsilon}\right\}\right) + \mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\right),$$

which is therefore an upper bound on $\Lambda_{\mathrm{AG}(\nu)}(\varepsilon,\delta)$. To match the form of the upper bound stated in Theorem 8, we can relax this by noting that $d\mathrm{Log}\left(\min\left\{\mathfrak{s},\frac{1}{\nu+\varepsilon}\right\}\right)+\mathrm{Log}\left(\frac{\mathrm{Log}(1/\varepsilon)}{\delta}\right)\leq 2d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ $+\mathrm{Log}\left(\frac{1}{\delta}\right)\leq 2d\mathrm{Log}\left(\frac{1}{\varepsilon\delta}\right)$, while $\frac{\nu^2}{\varepsilon^2}+\mathrm{Log}\left(\frac{1}{\varepsilon}\right)\leq\left(\frac{\nu^2}{\varepsilon^2}+1\right)\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$.

To prove the lower bound in Theorem 8, we note that $\mathrm{BE}(\nu)\subseteq\mathrm{AG}(\nu)$ for $\nu\in[0,1/2)$, so that $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta)\leq\Lambda_{\mathrm{AG}(\nu)}(\varepsilon,\delta)$. Thus, the lower bound on $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta)$ in Theorem 7 (proven above) also applies to $\Lambda_{\mathrm{AG}(\nu)}(\varepsilon,\delta)$. ■

Appendix C. Proofs for Results in Section 7

This section provides proofs of the equivalences between complexity measures stated in Section 7.

C.1 The Disagreement Coefficient

Here we present the proof of Theorem 10. First, we have a helpful lemma, which allows us to restrict focus to *finitely discrete* probability measures. Let Π denote the set of probability measures $\mathcal{P}$ on $\mathcal{X}$ such that $\exists m\in\mathbb{N}$ and a sequence $\{z_i\}_{i=1}^m$ in $\mathcal{X}$ for which $\mathcal{P}(\{z_i:i\in\{1,\ldots,m\}\})=1$.

Lemma 42 *If* $\mathfrak{s}<\infty$*, then* $\forall\varepsilon\in(0,1]$, $\hat{\hat{\theta}}(\varepsilon)=\sup_{\mathcal{P}\in\Pi}\sup_{h\in\mathbb{C}}\theta_{h,\mathcal{P}}(\varepsilon)$.

Proof Suppose $\mathfrak{s}<\infty$, and fix any $\varepsilon\in(0,1]$. Since $\mathcal{P}_{XY}$ ranges over all probability measures over $\mathcal{X}\times\mathcal{Y}$ in the definition of $\hat{\hat{\theta}}(\varepsilon)$, including all those in RE with marginal $\mathcal{P}$ over $\mathcal{X}$ contained in Π (in which case, $\theta_{\mathcal{P}_{XY}}(\varepsilon)=\theta_{f^\star_{\mathcal{P}_{XY}},\mathcal{P}}(\varepsilon)$), we always have $\sup_{P\in\Pi}\sup_{h\in\mathbb{C}}\theta_{h,P}(\varepsilon)\leq\hat{\hat{\theta}}(\varepsilon)$. Thus, it suffices to show that we also have $\sup_{\mathcal{P}\in\Pi}\sup_{h\in\mathbb{C}}\theta_{h,\mathcal{P}}(\varepsilon)\geq\hat{\hat{\theta}}(\varepsilon)$.

The result trivially holds if $\hat{\hat{\theta}}(\varepsilon)=1$, since *every* $\mathcal{P}$ and h have $\theta_{h,\mathcal{P}}(\varepsilon)\geq 1$. To address the nontrivial case, suppose $\hat{\hat{\theta}}(\varepsilon)>1$. Fix any $\gamma_1,\gamma_2,\gamma_3\in(0,1)$. Fix any $\mathcal{P}_{XY}$ with $\theta_{\mathcal{P}_{XY}}(\varepsilon)>1$, and as usual denote $\mathcal{P}(\cdot)=\mathcal{P}_{XY}(\cdot\times\mathcal{Y})$. Also let $h^*_{\mathcal{P}_{XY}}$ be as in Definition 9, so that $\theta_{\mathcal{P}_{XY}}(\varepsilon)=\theta_{h^*_{\mathcal{P}_{XY}},\mathcal{P}}(\varepsilon)$. Let $r_\varepsilon\in(\varepsilon,1]$ be such that $\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h^*_{\mathcal{P}_{XY}},r_\varepsilon)))}{r_\varepsilon}\geq(1-\gamma_1)\theta_{\mathcal{P}_{XY}}(\varepsilon)$ (which exists, by the definition of the supremum, combined with the fact that $1<\theta_{\mathcal{P}_{XY}}(\varepsilon)\leq 1/\varepsilon<\infty$). Also let $h\in\mathbb{C}$ have $\mathcal{P}(x:h(x)\neq h^*_{\mathcal{P}_{XY}}(x))\leq\gamma_3 r_\varepsilon$, which exists by the definition of $h^*_{\mathcal{P}_{XY}}$.

Let $m=\left\lceil\frac{8}{\gamma_2^2r_\varepsilon^2}\left(10d\mathrm{Log}\left(\frac{8e}{\gamma_2^2r_\varepsilon^2}\right)+\mathrm{Log}(24)\right)\right\rceil$, which is a finite natural number, since $d\leq\mathfrak{s}<\infty$. It follows from Lemma 20 and Lemma 18 that, for $X'_1,\ldots,X'_m$ independent $\mathcal{P}$-distributed random variables, with probability at least $2/3$, every $g\in\mathbb{C}$ has $\frac{1}{m}\sum_{i=1}^m\mathbb{1}_{\mathrm{DIS}(\{h,g\})}(X'_i)\leq\mathcal{P}(x:h(x)\neq g(x))+\gamma_2 r_\varepsilon\leq\mathcal{P}(x:h^*_{\mathcal{P}_{XY}}(x)\neq g(x))+(\gamma_3+\gamma_2)r_\varepsilon$. Furthermore, by Hoeffding's inequality, we also have that with probability at least $2/3$, $\frac{1}{m}\sum_{i=1}^m\mathbb{1}_{\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h^*_{\mathcal{P}_{XY}},r_\varepsilon))}(X'_i)\geq\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h^*_{\mathcal{P}_{XY}},r_\varepsilon)))-\gamma_2 r_\varepsilon$. By a union bound, both of these events happen with probability at least $1/3$. In particular, this implies $\exists z_1,\ldots,z_m\in\mathcal{X}$ such that, letting $\hat{\mathcal{P}}$ be the probability measure with $\hat{\mathcal{P}}(A)=\frac{1}{m}\sum_{i=1}^m\mathbb{1}_A(z_m)$ for all measurable $A\subseteq\mathcal{X}$, we have, $\forall g\in\mathbb{C}$, $\hat{\mathcal{P}}(\mathrm{DIS}(\{h,g\}))\leq\mathcal{P}(\mathrm{DIS}(\{h^*_{\mathcal{P}_{XY}},g\}))+(\gamma_3+\gamma_2)r_\varepsilon$, and furthermore $\hat{\mathcal{P}}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h^*_{\mathcal{P}_{XY}},r_\varepsilon)))\geq\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h^*_{\mathcal{P}_{XY}},r_\varepsilon)))-\gamma_2 r_\varepsilon$. This further implies that $\mathrm{B}_{\mathcal{P}}(h^*_{\mathcal{P}_{XY}},r_\varepsilon)\subseteq\mathrm{B}_{\hat{\mathcal{P}}}(h,(1+\gamma_3+\gamma_2)r_\varepsilon)$, and thus

$$\begin{aligned}\hat{\mathcal{P}}(\mathrm{DIS}(\mathrm{B}_{\hat{\mathcal{P}}}(h,(1+\gamma_3+\gamma_2)r_\varepsilon))) &\geq \hat{\mathcal{P}}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h^*_{\mathcal{P}_{XY}},r_\varepsilon))) \geq \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h^*_{\mathcal{P}_{XY}},r_\varepsilon))) - \gamma_2 r_\varepsilon \\ &\geq (1-\gamma_1)\theta_{\mathcal{P}_{XY}}(\varepsilon)r_\varepsilon - \gamma_2 r_\varepsilon \geq (1-\gamma_1-\gamma_2)\theta_{\mathcal{P}_{XY}}(\varepsilon)r_\varepsilon.\end{aligned}$$

Therefore,

$$\theta_{h,\hat{\mathcal{P}}}(\varepsilon) \geq \frac{\hat{\mathcal{P}}(\mathrm{DIS}(\mathrm{B}_{\hat{\mathcal{P}}}(h,(1+\gamma_3+\gamma_2)r_\varepsilon)))}{(1+\gamma_3+\gamma_2)r_\varepsilon} \geq \frac{1-\gamma_1-\gamma_2}{1+\gamma_3+\gamma_2}\theta_{\mathcal{P}_{XY}}(\varepsilon).$$

Noting that $\hat{\mathcal{P}}(\{z_1,\ldots,z_m\}) = 1$, so that $\hat{\mathcal{P}} \in \Pi$, since $\mathcal{P}_{XY}$ was arbitrary, we have established that $\forall \mathcal{P}_{XY}$, $\exists P \in \Pi$ and $h \in \mathbb{C}$ such that $\theta_{h,P}(\varepsilon) \geq \frac{1-\gamma_1-\gamma_2}{1+\gamma_3+\gamma_2}\theta_{\mathcal{P}_{XY}}(\varepsilon)$. Since this holds for any choices of $\gamma_1,\gamma_2,\gamma_3 \in (0,1)$, taking the limits as $\gamma_1 \to 0$, $\gamma_3 \to 0$, and $\gamma_2 \to 0$, we have $\sup_{P\in\Pi}\sup_{h\in\mathbb{C}}\theta_{h,P}(\varepsilon) \geq \hat{\hat{\theta}}(\varepsilon)$. ■

In fact, it is easy to show (based on the first part of the proof below) that the "$\mathfrak{s} < \infty$" constraint is unnecessary in Lemma 42, though this is not important for our purposes. We are now ready for the proof of Theorem 10.

Proof of Theorem 10 First, we prove $\hat{\hat{\theta}}(\varepsilon) \geq \mathfrak{s} \wedge \frac{1}{\varepsilon}$. Toward this end, let $\{x_i\}_{i=1}^{\mathfrak{s}}$ and $\{h_i\}_{i=0}^{\mathfrak{s}}$ be as in Definition 2, and let $m = \mathfrak{s} \wedge \left\lceil \frac{1}{\varepsilon} \right\rceil$. Let $\mathcal{P}$ be a probability measure on $\mathcal{X}$ with $\mathcal{P}(\{x_i\}) = 1/m$ for each $i \in \{1,\ldots,m\}$. In particular, this implies that every $i \in \{1,\ldots,m\}$ has $\mathcal{P}(x : h_i(x) \neq h_0(x)) = 1/m$, so that $h_i \in \mathrm{B}_{\mathcal{P}}(h_0, 1/m)$. Since clearly $h_0 \in \mathrm{B}_{\mathcal{P}}(h_0,1/m)$ as well, and every $i \in \{1,\ldots,m\}$ has $x_i \in \mathrm{DIS}(\{h_i,h_0\})$, every $r > 1/m$ has $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h_0,r))) = \mathcal{P}(\{x_i : i \in \{1,\ldots,m\}\}) = 1$. Therefore, letting $\mathcal{P}_{XY}$ be the distribution in RE with $f^\star_{\mathcal{P}_{XY}} = h_0$ and marginal $\mathcal{P}$ over $\mathcal{X}$,

$$\begin{aligned}\hat{\hat{\theta}}(\varepsilon) \geq \theta_{\mathcal{P}_{XY}}(\varepsilon) = \theta_{h_0,\mathcal{P}}(\varepsilon) &\geq \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h_0,\max\{1/m,\varepsilon\})))}{\max\{1/m,\varepsilon\}} \\ &= \frac{1}{\max\{1/m,\varepsilon\}} = m \wedge \frac{1}{\varepsilon} = \mathfrak{s} \wedge \frac{1}{\varepsilon}.\end{aligned}$$

Next, we prove that $\hat{\hat{\theta}}(\varepsilon) \leq \mathfrak{s} \wedge \frac{1}{\varepsilon}$. That $\hat{\hat{\theta}}(\varepsilon) \leq \frac{1}{\varepsilon}$ follows directly from the definition, and the fact that probabilities are at most 1: that is, any $\mathcal{P}$ and h have $\sup_{r>\varepsilon}\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r)))}{r} \leq \sup_{r>\varepsilon}\frac{1}{r} = \frac{1}{\varepsilon}$. Therefore, it remains only to show that $\hat{\hat{\theta}}(\varepsilon) \leq \mathfrak{s}$ when $\mathfrak{s} < \frac{1}{\varepsilon}$. Furthermore, Lemma 42 implies that it suffices to show that $\sup_{\mathcal{P}\in\Pi}\sup_{h\in\mathbb{C}}\theta_{h,\mathcal{P}}(\varepsilon) \leq \mathfrak{s}$ in this case. Toward this end, suppose $\mathfrak{s} < \frac{1}{\varepsilon}$. We first stratify the set Π based on the size of the support, defining, for each $m \in \mathbb{N}$, $\Pi_m = \{\mathcal{P} \in \Pi : \exists z_1,\ldots,z_m \in \mathcal{X} \text{ s.t. } \mathcal{P}(\{z_1,\ldots,z_m\}) = 1\}$. Thus, Π_m is the set of probability measures on $\mathcal{X}$ for which the support of the probability mass function has cardinality at most m.

We now proceed by induction on m. As a base case, fix any $m \leq \mathfrak{s}$, any classifier h, and any $\mathcal{P} \in \Pi_m$, and let $z_1,\ldots,z_m \in \mathcal{X}$ be such that $\mathcal{P}(\{z_1,\ldots,z_m\}) = 1$. For any $r \in [1/\mathfrak{s}, 1]$, $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r)))/r \leq 1/r \leq \mathfrak{s}$. Furthermore (following an argument of Hanneke, 2014), for any $r \in (\varepsilon, 1/\mathfrak{s})$, for any $g \in \mathbb{C}$ with $\mathcal{P}(x : g(x) \neq h(x)) \leq r$, every $z \in \mathcal{X}$ with $\mathcal{P}(\{z\}) > r$ has $\mathcal{P}(x : g(x) \neq h(x)) < \mathcal{P}(\{z\})$, so that $g(z) = h(z)$; thus, $z \notin \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r))$. We therefore have that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r))) \leq \mathcal{P}(x : \mathcal{P}(\{x\}) \leq r) = \sum_{i=1}^m \mathbb{1}[\mathcal{P}(\{z_i\}) \leq r]\,\mathcal{P}(\{z_i\}) \leq r|\{i \in \{1,\ldots,m\} : \mathcal{P}(\{z_i\}) \leq r\}|$. Therefore, $\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r)))}{r} \leq |\{i \in \{1,\ldots,m\} : \mathcal{P}(\{z_i\}) \leq r\}| \leq m \leq \mathfrak{s}$, so that (since $\mathfrak{s} \geq 1$, due to the assumption that $|\mathbb{C}| \geq 2$), we have $\theta_{h,\mathcal{P}}(\varepsilon) \leq \mathfrak{s}$.

Now take as an inductive hypothesis that, for some $m \in \mathbb{N}$ with $m > \mathfrak{s}$, we have

$$\sup_{\mathcal{P} \in \Pi_{m-1}} \sup_{h \in \mathbb{C}} \theta_{h,\mathcal{P}}(\varepsilon) \leq \mathfrak{s}.$$

Fix any $h \in \mathbb{C}$, $r > \varepsilon$, and $\mathcal{P} \in \Pi_m$, and let $z_1, \ldots, z_m \in \mathcal{X}$ be such that $\mathcal{P}(\{z_1, \ldots, z_m\}) = 1$. If $\exists i, j \in \{1, \ldots, m\}$ with $i \neq j$ and $z_i = z_j$, or if some $j \in \{1, \ldots, m\}$ has $\mathcal{P}(\{z_j\}) = 0$, then since either of these has $\mathcal{P}(\{z_k : k \in \{1, \ldots, m\} \setminus \{j\}\}) = 1$, we would also have $\mathcal{P} \in \Pi_{m-1}$, so that $\theta_{h,\mathcal{P}}(\varepsilon) \leq \mathfrak{s}$ by the inductive hypothesis. To handle the remaining nontrivial cases, suppose the $z_1, \ldots, z_m$ are all distinct, and $\min_{i \in \{1,\ldots,m\}} \mathcal{P}(\{z_i\}) > 0$. Furthermore, note that, since $m > \mathfrak{s}$, $\{z_1, \ldots, z_m\}$ cannot be a star set for $\mathbb{C}$.

We now consider three cases. First, consider the case that $\exists k \in \{1, \ldots, m\}$ with $z_k \notin \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))$. In this case, define a probability measure $\mathcal{P}'$ over $\mathcal{X}$ such that, for any measurable $A \subseteq \mathcal{X} \setminus \{z_k\}$, $\mathcal{P}'(A) = \mathcal{P}'(A \cup \{z_k\}) = \mathcal{P}(A)/(1 - \mathcal{P}(\{z_k\}))$. Note that this is a well-defined probability measure, since $m \geq 2$ and $\min_{i \in \{1,\ldots,m\}} \mathcal{P}(\{z_i\}) > 0$, so that $\mathcal{P}(\mathcal{X} \setminus \{z_k\}) = 1 - \mathcal{P}(\{z_k\}) > 0$. Also note that (since $h \in \mathrm{B}_{\mathcal{P}}(h, r)$) any $g \in \mathrm{B}_{\mathcal{P}}(h, r)$ has $g(z_k) = h(z_k)$, so that $\mathcal{P}'(x : g(x) \neq h(x)) = \mathcal{P}(x : g(x) \neq h(x))/(1 - \mathcal{P}(\{z_k\})) \leq r/(1 - \mathcal{P}(\{z_k\}))$. Therefore, $\mathrm{B}_{\mathcal{P}'}(h, r/(1 - \mathcal{P}(\{z_k\}))) \supseteq \mathrm{B}_{\mathcal{P}}(h, r)$, and since $z_k \notin \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))$, $\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r/(1 - \mathcal{P}(\{z_k\}))))) \geq \mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))) = \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r)))/(1 - \mathcal{P}(\{z_k\}))$. Thus,

$$\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))) \leq (1 - \mathcal{P}(\{z_k\}))\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r/(1 - \mathcal{P}(\{z_k\}))))). \tag{60}$$

Noting that $\mathcal{P}'(\{z_i : i \in \{1, \ldots, m\} \setminus \{k\}\}) = \mathcal{P}(\{z_1, \ldots, z_m\} \setminus \{z_k\})/(1 - \mathcal{P}(\{z_k\})) = 1$, we have that $\mathcal{P}' \in \Pi_{m-1}$. Therefore, by the inductive hypothesis and the fact that $r/(1 - \mathcal{P}(\{z_k\})) > r > \varepsilon$,

$$\begin{aligned}
\mathcal{P}'\left(\mathrm{DIS}\left(\mathrm{B}_{\mathcal{P}'}\left(h, \frac{r}{1 - \mathcal{P}(\{z_k\})}\right)\right)\right) &\leq \theta_{h,\mathcal{P}'}(\varepsilon)\frac{r}{1 - \mathcal{P}(\{z_k\})} \\
&\leq \sup_{P \in \Pi_{m-1}} \sup_{h' \in \mathbb{C}} \theta_{h',P}(\varepsilon)\frac{r}{1 - \mathcal{P}(\{z_k\})} \leq \frac{\mathfrak{s}r}{1 - \mathcal{P}(\{z_k\})}.
\end{aligned}$$

Combined with (60), this further implies that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))) \leq (1 - \mathcal{P}(\{z_k\}))\mathfrak{s}r/(1 - \mathcal{P}(\{z_k\})) = \mathfrak{s}r$.

Next, consider a second case, where $\{z_1, \ldots, z_m\} \subseteq \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))$, and $\exists j, k \in \{1, \ldots, m\}$ with $j \neq k$ such that, $\forall g \in \mathrm{B}_{\mathcal{P}}(h, r)$, $g(z_k) \neq h(z_k) \Rightarrow g(z_j) \neq h(z_j)$. In this case, define a probability measure $\mathcal{P}'$ over $\mathcal{X}$ such that, for any measurable $A \subseteq \mathcal{X} \setminus \{z_j, z_k\}$, $\mathcal{P}'(A) = \mathcal{P}(A)$, $\mathcal{P}'(A \cup \{z_j\}) = \mathcal{P}(A)$, and $\mathcal{P}'(A \cup \{z_k\}) = \mathcal{P}'(A \cup \{z_j, z_k\}) = \mathcal{P}(A \cup \{z_j, z_k\})$: in other words, $\mathcal{P}'$ has a probability mass function $x \mapsto \mathcal{P}'(\{x\})$ equal to $x \mapsto \mathcal{P}(\{x\})$ everywhere, except that $\mathcal{P}'(\{z_j\}) = 0$ and $\mathcal{P}'(\{z_k\}) = \mathcal{P}(\{z_j\}) + \mathcal{P}(\{z_k\})$. Note that, for any $g \in \mathrm{B}_{\mathcal{P}}(h, r)$ with $g(z_k) = h(z_k)$, $\mathcal{P}'(x : g(x) \neq h(x)) = \mathcal{P}(x : g(x) \neq h(x)) - \mathbb{1}[g(z_j) \neq h(z_j)]\mathcal{P}(\{z_j\}) \leq \mathcal{P}(x : g(x) \neq h(x)) \leq r$. Furthermore, any $g \in \mathrm{B}_{\mathcal{P}}(h, r)$ with $g(z_k) \neq h(z_k)$ also has $g(z_j) \neq h(z_j)$, so that $\mathcal{P}'(x : g(x) \neq h(x)) = \mathcal{P}(x : g(x) \neq h(x)) \leq r$. Therefore, $\mathrm{B}_{\mathcal{P}'}(h, r) \supseteq \mathrm{B}_{\mathcal{P}}(h, r)$. Since $z_j, z_k \in \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))$, this further implies that $z_j, z_k \in \mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r))$. Therefore, by definition of $\mathcal{P}'$ and monotonicity of measures, $\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r))) = \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r))) \geq \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r)))$. Noting that $\mathcal{P}'(\{z_i : i \in \{1, \ldots, m\} \setminus \{j\}\}) = \mathcal{P}(\{z_1, \ldots, z_m\}) = 1$, we have $\mathcal{P}' \in \Pi_{m-1}$, and therefore (by the inductive hypothesis), $\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r))) \leq \theta_{h,\mathcal{P}'}(\varepsilon)r \leq \sup_{P \in \Pi_{m-1}} \sup_{h' \in \mathbb{C}} \theta_{h',P}(\varepsilon)r \leq \mathfrak{s}r$. Thus, since we established above that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))) \leq \mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h, r)))$, we have that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))) \leq \mathfrak{s}r$.

Finally, consider a third case (the complement of the first and second), in which $\{z_1,\ldots,z_m\} \subseteq \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r))$, but $\nexists j,k \in \{1,\ldots,m\}$ with $j \neq k$ such that, $\forall g \in \mathrm{B}_{\mathcal{P}}(h,r)$, $g(z_k) \neq h(z_k) \Rightarrow g(z_j) \neq h(z_j)$. In particular, note that the first condition (which is, in fact, redundant, but included for clarity) implies $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r))) = 1$. In this case, since (as above) $\{z_1,\ldots,z_m\}$ is not a star set for $\mathbb{C}$, $\exists i \in \{1,\ldots,m\}$ such that $\forall g \in \mathbb{C}$ with $g(z_i) \neq h(z_i)$, $\exists j \in \{1,\ldots,m\} \setminus \{i\}$ with $g(z_j) \neq h(z_j)$ as well; fix any such $i \in \{1,\ldots,m\}$. Since $\{z_1,\ldots,z_m\} \subseteq \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r))$, we have $z_i \in \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r))$. Thus, we may let $g_i \in \mathrm{B}_{\mathcal{P}}(h,r)$ be such that $g_i(z_i) \neq h(z_i)$, and let $j \in \{1,\ldots,m\} \setminus \{i\}$ be such that $g_i(z_j) \neq h(z_j)$ (which exists, by our choice of i). Let $\mathcal{P}'$ be a probability measure over $\mathcal{X}$ such that, for all measurable $A \subseteq \mathcal{X} \setminus \{z_i,z_j\}$, $\mathcal{P}'(A) = \mathcal{P}(A)$, $\mathcal{P}'(A \cup \{z_i\}) = \mathcal{P}(A)$, and $\mathcal{P}'(A \cup \{z_j\}) = \mathcal{P}'(A \cup \{z_i,z_j\}) = \mathcal{P}(A \cup \{z_i,z_j\})$: in other words, $\mathcal{P}'$ has a probability mass function $x \mapsto \mathcal{P}'(\{x\})$ equal to $x \mapsto \mathcal{P}(\{x\})$ everywhere, except that $\mathcal{P}'(\{z_i\}) = 0$ and $\mathcal{P}'(\{z_j\}) = \mathcal{P}(\{z_i\}) + \mathcal{P}(\{z_j\})$. Note that, for any measurable set $A \subseteq \mathcal{X}$ with $\{z_i,z_j\} \subseteq A$, $\mathcal{P}'(A) = \mathcal{P}(A)$. In particular, since $\{z_i,z_j\} \subseteq \mathrm{DIS}(\{g_i,h\})$, $\mathcal{P}'(\mathrm{DIS}(\{g_i,h\})) = \mathcal{P}(\mathrm{DIS}(\{g_i,h\})) \leq r$, so that $g_i \in \mathrm{B}_{\mathcal{P}'}(h,r)$, and therefore (since $h \in \mathrm{B}_{\mathcal{P}'}(h,r)$ as well) $\{z_i,z_j\} \subseteq \mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h,r))$. Furthermore, for any $k \in \{1,\ldots,m\} \setminus \{i,j\}$, by the property characterizing this third case, and since $z_k \in \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r))$, $\exists g \in \mathrm{B}_{\mathcal{P}}(h,r)$ with $g(z_k) \neq h(z_k)$ and $g(z_j) = h(z_j)$, so that $\mathcal{P}'(\mathrm{DIS}(\{g,h\})) = \mathcal{P}(\mathrm{DIS}(\{g,h\}) \setminus \{z_i\}) \leq \mathcal{P}(\mathrm{DIS}(\{g,h\})) \leq r$ (i.e., $g \in \mathrm{B}_{\mathcal{P}'}(h,r)$), and therefore (since $h \in \mathrm{B}_{\mathcal{P}'}(h,r)$ as well) $z_k \in \mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h,r))$ as well. Altogether, we have that $\{z_1,\ldots,z_m\} \subseteq \mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h,r))$. Therefore, since $\{z_i,z_j\} \subseteq \mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h,r))$, the definition of $\mathcal{P}'$ implies $\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h,r))) = \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h,r))) \geq \mathcal{P}(\{z_1,\ldots,z_m\}) = 1 = \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r)))$. Noting that $\mathcal{P}'(\{z_k : k \in \{1,\ldots,m\} \setminus \{i\}\}) = \mathcal{P}(\{z_1,\ldots,z_m\}) = 1$, we have that $\mathcal{P}' \in \Pi_{m-1}$, and therefore (by the inductive hypothesis), $\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h,r))) \leq \theta_{h,\mathcal{P}'}(\varepsilon) r \leq \sup_{P \in \Pi_{m-1}} \sup_{h' \in \mathbb{C}} \theta_{h',P}(\varepsilon) r \leq \mathfrak{s} r$. Since $\mathcal{P}'(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}'}(h,r))) = 1 = \mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r)))$, we have that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r))) \leq \mathfrak{s} r$ as well.

Thus, in all three cases, we have that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h,r))) \leq \mathfrak{s} r$. Since this holds for every $r > \varepsilon$, and $|\mathbb{C}| \geq 2$ implies $\mathfrak{s} \geq 1$, we have that $\theta_{h,\mathcal{P}}(\varepsilon) \leq \mathfrak{s}$. Since this holds for every $h \in \mathbb{C}$ and $\mathcal{P} \in \Pi_m$, we have established that $\sup_{\mathcal{P} \in \Pi_m} \sup_{h \in \mathbb{C}} \theta_{h,\mathcal{P}}(\varepsilon) \leq \mathfrak{s}$, which completes the inductive step. It follows by the principle of induction that $\sup_{\mathcal{P} \in \Pi_m} \sup_{h \in \mathbb{C}} \theta_{h,\mathcal{P}}(\varepsilon) \leq \mathfrak{s}$ for every $m \in \mathbb{N}$, and therefore, since $\Pi = \bigcup_m \Pi_m$, $\sup_{\mathcal{P} \in \Pi} \sup_{h \in \mathbb{C}} \theta_{h,\mathcal{P}}(\varepsilon) \leq \mathfrak{s}$.

The claim that $\hat{\hat{\theta}}(0) = \mathfrak{s}$ follows as a limiting case, due to continuity of the supremum from below. Specifically, fix any sequence $\{A_n\}_{n=1}^{\infty}$ of nonempty subsets of $\mathbb{R}$. For each $m \in \mathbb{N}$, $\bigcup_n A_n \supseteq A_m$, so $\sup \bigcup_n A_n \geq \sup A_m$ (allowing the supremum to take the value ∞ where appropriate), and since this holds for every such m, we have $\sup \bigcup_n A_n \geq \sup_n \sup A_n$ Furthermore, $\forall a \in \bigcup_n A_n$, $\exists m \in \mathbb{N}$ s.t. $a \in A_m$, so that $\sup_n \sup A_n \geq \sup A_m \geq a$, and therefore (since this holds for every such a) $\sup_n \sup A_n \geq \sup \bigcup_n A_n$. Thus, $\sup \bigcup_n A_n = \sup_n \sup A_n$. In particular, taking (for each $n \in \mathbb{N}$)

$$A_n = \left\{ \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h^*_{\mathcal{P}_{XY}}, r)))}{r} \vee 1 : r > 1/n, \mathcal{P}_{XY} \in \mathrm{AG}(1) \right\},$$

(where, as usual, $\mathcal{P}(\cdot) = \mathcal{P}_{XY}(\cdot \times \mathcal{Y})$ denotes the marginal of $\mathcal{P}_{XY}$ over $\mathcal{X}$), and noting that $\sup \bigcup_n A_n = \hat{\hat{\theta}}(0)$ and $\forall n \in \mathbb{N}$, $\sup A_n = \hat{\hat{\theta}}(1/n)$, we have that $\hat{\hat{\theta}}(0) = \sup_n \hat{\hat{\theta}}(1/n) = \sup_n \mathfrak{s} \wedge n = \mathfrak{s}$. ∎

C.2 The Splitting Index

Here we present the proof of Theorem 12. First, we introduce a quantity related to $\hat{\dot{\rho}}(\varepsilon)$, but slightly simpler. For $\varepsilon, \tau \in (0,1]$ and any probability measure $\mathcal{P}$ over $\mathcal{X}$, define

$$\bar{\rho}_{\mathcal{P}}(\varepsilon; \tau) = \sup \left\{ \rho \in [0,1] : \mathbb{C} \text{ is } (\rho, \varepsilon, \tau)\text{-splittable under } \mathcal{P} \right\},$$

and let

$$\bar{\rho}(\varepsilon) = \inf_{P} \lim_{\tau \to 0} \bar{\rho}_P(\varepsilon; \tau).$$

In the arguments below, we will see that $\lfloor 1/\bar{\rho}(\varepsilon) \rfloor = \lfloor 1/\hat{\dot{\rho}}(\varepsilon) \rfloor$, so that it suffices to work with this simpler quantity. We begin with a lemma which allows us to restrict our focus (in part of the proof) to finitely discrete probability measures. Recall the definition of Π from Appendix C.1 above.

Lemma 43 *If $d < \infty$, then $\forall \varepsilon \in (0,1]$, $\bar{\rho}(\varepsilon) \geq \lim_{\gamma \to 0} \inf_{P \in \Pi} \lim_{\tau \to 0} \bar{\rho}_P((1-\gamma)\varepsilon; \tau)$.*

Proof Suppose $d < \infty$, and fix any $\varepsilon \in (0,1]$. Fix arbitrary values $\gamma_1, \gamma_2 \in (0,1)$, and let

$$m = \left\lceil \frac{8}{\gamma_2^2 \varepsilon^2} \left(10 d \mathrm{Log}\left(\frac{8e}{\gamma_2^2 \varepsilon^2} \right) + \mathrm{Log}(24) \right) \right\rceil,$$

which is a finite natural number. Fix any probability measure $\mathcal{P}$ over $\mathcal{X}$, and any $\tau \in (0, 1/(3m))$, and note that $\tau' \mapsto \bar{\rho}_{\mathcal{P}}(\varepsilon; \tau')$ is nonincreasing, so that $\bar{\rho}_{\mathcal{P}}(\varepsilon; \tau) \leq \lim_{\tau' \to 0} \bar{\rho}_{\mathcal{P}}(\varepsilon; \tau')$. For brevity, denote $\bar{\rho} = \bar{\rho}_{\mathcal{P}}(\varepsilon; \tau)$. Since $\mathbb{C}$ is not $(\gamma_1 + \bar{\rho}, \varepsilon, \tau)$-splittable under $\mathcal{P}$, let $Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \mathcal{P}(x : f(x) \neq g(x)) \geq \varepsilon\}$ be a finite set such that $\mathcal{P}(x : \mathrm{Split}(Q, x) \geq (\gamma_1 + \bar{\rho})\,|Q|) < \tau$.

Let $X_1', \ldots, X_m'$ be independent $\mathcal{P}$-distributed random variables. Lemma 20 and Lemma 18 imply that, with probability at least $2/3$, $\forall f, g \in \mathbb{C}$,

$$\left| \mathcal{P}(x : f(x) \neq g(x)) - \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left[f(X_i') \neq g(X_i') \right] \right| \leq \gamma_2 \varepsilon.$$

Furthermore, by a union bound, with probability at least $1 - m\mathcal{P}(x : \mathrm{Split}(Q, x) \geq (\gamma_1 + \bar{\rho})\,|Q|) > 1 - m\tau > 1 - m(1/(3m)) = 2/3$, every $i \in \{1, \ldots, m\}$ has $\mathrm{Split}(Q, X_i') < (\gamma_1 + \bar{\rho})|Q|$. By a union bound, both of the above events occur with probability at least $1/3$. In particular, this implies $\exists z_1, \ldots, z_m \in \mathcal{X}$ such that, letting $\hat{\mathcal{P}}$ be the probability measure with $\hat{\mathcal{P}}(A) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_A(z_m)$ for all measurable $A \subseteq \mathcal{X}$, we have, $\forall f, g \in \mathbb{C}$, $\left| \mathcal{P}(x : f(x) \neq g(x)) - \hat{\mathcal{P}}(x : f(x) \neq g(x)) \right| \leq \gamma_2 \varepsilon$, and $\hat{\mathcal{P}}(x : \mathrm{Split}(Q, x) \geq (\gamma_1 + \bar{\rho})|Q|) = 0$.

For any $\{f, g\} \in Q$, we have $\hat{\mathcal{P}}(x : f(x) \neq g(x)) \geq \mathcal{P}(x : f(x) \neq g(x)) - \gamma_2 \varepsilon \geq (1 - \gamma_2)\varepsilon$. Therefore, $\mathbb{C}$ is not $(\gamma_1 + \bar{\rho}, (1 - \gamma_2)\varepsilon, \tau')$-splittable under $\hat{\mathcal{P}}$ for any $\tau' > 0$, which implies $\lim_{\tau' \to 0} \bar{\rho}_{\hat{\mathcal{P}}}((1 - \gamma_2)\varepsilon; \tau') \leq \gamma_1 + \bar{\rho}_{\mathcal{P}}(\varepsilon; \tau)$. Since $\hat{\mathcal{P}} \in \Pi$, we have

$$\inf_{P \in \Pi} \lim_{\tau' \to 0} \bar{\rho}_P((1 - \gamma_2)\varepsilon; \tau') \leq \gamma_1 + \bar{\rho}_{\mathcal{P}}(\varepsilon; \tau) \leq \gamma_1 + \lim_{\tau' \to 0} \bar{\rho}_{\mathcal{P}}(\varepsilon; \tau').$$

Since this holds for any $\gamma_1 \in (0,1)$, taking the limit as $\gamma_1 \to 0$ implies

$$\inf_{P \in \Pi} \lim_{\tau' \to 0} \bar{\rho}_P((1 - \gamma_2)\varepsilon; \tau') \leq \lim_{\tau' \to 0} \bar{\rho}_{\mathcal{P}}(\varepsilon; \tau').$$

Furthermore, since this holds for any $\gamma_2 \in (0,1)$ and any $\mathcal{P}$, we have

$$\lim_{\gamma_2 \to 0} \inf_{P \in \Pi} \lim_{\tau' \to 0} \bar{\rho}_P((1-\gamma_2)\varepsilon; \tau') \leq \inf_{P} \lim_{\tau' \to 0} \bar{\rho}_P(\varepsilon; \tau') = \bar{\rho}(\varepsilon).$$

■

We are now ready for the proof of Theorem 12.

Proof of Theorem 12 We first establish that $\mathfrak{s} \wedge \left\lfloor \frac{1}{\varepsilon} \right\rfloor \leq \left\lfloor \frac{1}{\hat{\hat{\rho}}(\varepsilon)} \right\rfloor$ for any $\varepsilon \in (0,1]$. The proof of this fact was implicitly established in the original work of Dasgupta (2005, Corollary 3), but we include the argument here for completeness. Let $\{x_i\}_{i=1}^{\mathfrak{s}}$ and $\{h_i\}_{i=0}^{\mathfrak{s}}$ be as in Definition 2, and let $m = \mathfrak{s} \wedge \left\lfloor \frac{1}{\varepsilon} \right\rfloor$. Let $\Delta = 1/m$, and note that $\Delta \geq 1/\left\lfloor \frac{1}{\varepsilon} \right\rfloor \geq \varepsilon$. As in the proof of Theorem 10, let $\mathcal{P}$ be a probability measure on $\mathcal{X}$ with $\mathcal{P}(\{x_i\}) = 1/m$ for each $i \in \{1, \ldots, m\}$. Thus, every $i \in \{1, \ldots, m\}$ has $\mathcal{P}(x : h_i(x) \neq h_0(x)) = \Delta$, so that $h_i \in \mathrm{B}_{\mathcal{P}}(h_0, \Delta) \subseteq \mathrm{B}_{\mathcal{P}}(h_0, 4\Delta)$, and the finite set $Q = \{\{h_0, h_i\} : i \in \{1, \ldots, m\}\}$ satisfies $Q \subseteq \{\{f,g\} \subseteq \mathrm{B}_{\mathcal{P}}(h_0, 4\Delta) : \mathcal{P}(x : f(x) \neq g(x)) \geq \Delta\}$. In particular, since $\mathcal{P}(\mathcal{X} \setminus \{x_1, \ldots, x_m\}) = 0$, and every $i \in \{1, \ldots, m\}$ has $\mathrm{Split}(Q, x_i) = 1 = \frac{1}{m}|Q|$, we have $\mathcal{P}\left(x : \mathrm{Split}(Q,x) > \frac{1}{m}|Q|\right) = 0$. Thus, for any $\rho > \frac{1}{m}$, and any $\tau > 0$, $\mathrm{B}_{\mathcal{P}}(h_0, 4\Delta)$ is not (ρ, Δ, τ)-splittable. Therefore, $\hat{\hat{\rho}}(\varepsilon) \leq \lim_{\tau \to 0} \rho_{h_0, \mathcal{P}}(\varepsilon; \tau) \leq \frac{1}{m}$, which implies $\frac{1}{\hat{\hat{\rho}}(\varepsilon)} \geq m$; since $m \in \mathbb{N}$, it follows that $\left\lfloor \frac{1}{\hat{\hat{\rho}}(\varepsilon)} \right\rfloor \geq m$.

Next, we prove that $\left\lfloor \frac{1}{\hat{\hat{\rho}}(\varepsilon)} \right\rfloor \leq \mathfrak{s} \wedge \left\lfloor \frac{1}{\varepsilon} \right\rfloor$ for any $\varepsilon \in (0,1]$. Since, for every $h \in \mathbb{C}$, every probability measure $\mathcal{P}$ over $\mathcal{X}$, and every $\Delta \geq \varepsilon$, every finite $Q \subseteq \{\{f,g\} \subseteq \mathrm{B}_{\mathcal{P}}(h, 4\Delta) : \mathcal{P}(x : f(x) \neq g(x)) \geq \Delta\}$ also has $Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \mathcal{P}(x : f(x) \neq g(x)) \geq \varepsilon\}$, we have $\bar{\rho}(\varepsilon) \leq \hat{\hat{\rho}}(\varepsilon)$. Thus, it suffices to show $\left\lfloor \frac{1}{\bar{\rho}(\varepsilon)} \right\rfloor \leq \mathfrak{s} \wedge \left\lfloor \frac{1}{\varepsilon} \right\rfloor$.

That $\bar{\rho}(\varepsilon) \geq \varepsilon$ was established by Dasgupta (2005, Lemma 1); we repeat the argument here for completeness. Fix any probability measure $\mathcal{P}$ over $\mathcal{X}$ and any $\varepsilon, \tau \in (0,1]$ with $\tau < \varepsilon$. Fix any finite set $Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \mathcal{P}(x : f(x) \neq g(x)) \geq \varepsilon\}$. If $Q = \emptyset$, then trivially $\mathcal{P}(x : \mathrm{Split}(Q,x) \geq \varepsilon|Q|) = 1 \geq \tau$. Otherwise, if $Q \neq \emptyset$, letting $X \sim \mathcal{P}$,

$$\mathbb{E}[\mathrm{Split}(Q,X)] \geq \mathbb{E}\left[\sum_{\{f,g\} \in Q} \mathbb{1}[f(Z) \neq g(Z)]\right] = \sum_{\{f,g\} \in Q} \mathcal{P}(x : f(x) \neq g(x)) \geq |Q|\varepsilon.$$

Furthermore, since $\mathrm{Split}(Q,x) \leq |Q|$,

$$\begin{aligned}
&\mathbb{E}[\mathrm{Split}(Q,X)] \\
&= \mathbb{E}\left[\mathbb{1}[\mathrm{Split}(Q,X) \geq (\varepsilon-\tau)|Q|]\mathrm{Split}(Q,X)\right] + \mathbb{E}\left[\mathbb{1}[\mathrm{Split}(Q,X) < (\varepsilon-\tau)|Q|]\mathrm{Split}(Q,X)\right] \\
&< \mathcal{P}\left(x : \mathrm{Split}(Q,x) \geq (\varepsilon-\tau)|Q|\right)|Q| + (\varepsilon-\tau)|Q|.
\end{aligned}$$

Together, these inequalities imply

$$|Q|\varepsilon < \mathcal{P}\left(x : \mathrm{Split}(Q,x) \geq (\varepsilon-\tau)|Q|\right)|Q| + (\varepsilon-\tau)|Q|.$$

Subtracting $(\varepsilon-\tau)|Q|$ from both sides and dividing by $|Q|$, we have

$$\tau < \mathcal{P}\left(x : \mathrm{Split}(Q,x) \geq (\varepsilon-\tau)|Q|\right).$$

Since this holds for any such Q, we have that $\mathbb{C}$ is $((\varepsilon-\tau),\varepsilon,\tau)$-splittable under $\mathcal{P}$, so that $\bar{\rho}_{\mathcal{P}}(\varepsilon;\tau) \geq \varepsilon - \tau$. Since this holds for every choice of $\mathcal{P}$, we have that

$$\bar{\rho}(\varepsilon) = \inf_{\mathcal{P}} \lim_{\tau\to 0} \bar{\rho}_{\mathcal{P}}(\varepsilon;\tau) \geq \lim_{\tau\to 0} \varepsilon - \tau = \varepsilon,$$

from which it immediately follows that $\left\lfloor \frac{1}{\bar{\rho}(\varepsilon)} \right\rfloor \leq \left\lfloor \frac{1}{\varepsilon} \right\rfloor$.

It remains only to show that $\left\lfloor \frac{1}{\bar{\rho}(\varepsilon)} \right\rfloor \leq \mathfrak{s}$. In particular, since this trivially holds when $\mathfrak{s} = \infty$, for the remainder of the proof we suppose $\mathfrak{s} < \infty$. As argued in Section 4, we have $d \leq \mathfrak{s}$, so that this also implies $d < \infty$. Thus, Lemma 43 implies that $\bar{\rho}(\varepsilon) \geq \lim_{\gamma\to 0} \inf_{\mathcal{P}\in\Pi} \lim_{\tau\to 0} \bar{\rho}_{\mathcal{P}}((1-\gamma)\varepsilon;\tau)$. Therefore, if we can establish that, for every $\varepsilon \in (0,1]$ and $\mathcal{P} \in \Pi$, $\lim_{\tau\to 0} \bar{\rho}_{\mathcal{P}}(\varepsilon;\tau) \geq 1/\mathfrak{s}$, then we would have that for every $\varepsilon \in (0,1]$,

$$\left\lfloor \frac{1}{\bar{\rho}(\varepsilon)} \right\rfloor \leq \frac{1}{\bar{\rho}(\varepsilon)} \leq \lim_{\gamma\to 0} \sup_{\mathcal{P}\in\Pi} \frac{1}{\lim_{\tau\to 0} \bar{\rho}_{\mathcal{P}}((1-\gamma)\varepsilon;\tau)} \leq \mathfrak{s},$$

which would thereby complete the proof.

Toward this end, fix any $\varepsilon \in (0,1]$, and for each $\mathcal{P} \in \Pi$, denote $\tau_{\mathcal{P}} = \min\{\mathcal{P}(\{x\}) : x \in \mathcal{X}, \mathcal{P}(\{x\}) > 0\}$; in particular, note that (since $\mathcal{P} \in \Pi$) $0 < \tau_{\mathcal{P}} \leq 1$, and therefore also that, $\forall \varepsilon \in (0,1]$, $\lim_{\tau\to 0} \bar{\rho}_{\mathcal{P}}(\varepsilon;\tau) \geq \bar{\rho}_{\mathcal{P}}(\varepsilon;\tau_{\mathcal{P}})$ (in fact, they are equal). Furthermore, denoting $\operatorname{supp}(\mathcal{P}) = \{x \in \mathcal{X} : \mathcal{P}(\{x\}) > 0\}$, every $x \in \operatorname{supp}(\mathcal{P})$ has $\mathcal{P}(\{x\}) \geq \tau_{\mathcal{P}}$, while $\mathcal{P}(\mathcal{X} \setminus \operatorname{supp}(\mathcal{P})) = 0$. Thus, for any finite $Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \mathcal{P}(x : f(x) \neq g(x)) \geq \varepsilon\}$, and any $\rho \in [0,1]$, $\mathcal{P}(x : \operatorname{Split}(Q,x) \geq \rho|Q|) \geq \tau_{\mathcal{P}}$ if and only if $\max_{x\in\operatorname{supp}(\mathcal{P})} \operatorname{Split}(Q,x) \geq \rho|Q|$. Furthermore, since $\mathcal{P}(\mathcal{X} \setminus \operatorname{supp}(\mathcal{P})) = 0$, for any $\varepsilon \in (0,1]$, every $\{f,g\} \subseteq \mathbb{C}$ with $\mathcal{P}(x : f(x) \neq g(x)) \geq \varepsilon$ must have $\operatorname{DIS}(\{f,g\}) \cap \operatorname{supp}(\mathcal{P}) \neq \emptyset$. Thus, defining

$$\mathring{\rho}_{\mathcal{P}} = \sup\Bigg\{\rho \in [0,1] : \forall \text{ finite } Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \operatorname{DIS}(\{f,g\}) \cap \operatorname{supp}(\mathcal{P}) \neq \emptyset\},$$
$$\max_{x\in\operatorname{supp}(\mathcal{P})} \operatorname{Split}(Q,x) \geq \rho|Q|\Bigg\},$$

we have $\mathring{\rho}_{\mathcal{P}} \leq \bar{\rho}_{\mathcal{P}}(\varepsilon;\tau_{\mathcal{P}})$ for all $\varepsilon \in (0,1]$ (in fact, they are equal for $\varepsilon \leq \tau_{\mathcal{P}}$). Thus, it suffices to show that $\inf_{\mathcal{P}\in\Pi} \mathring{\rho}_{\mathcal{P}} \geq 1/\mathfrak{s}$. Now partition the set Π by the sizes of the supports, defining, for each $m \in \mathbb{N}$, $\Pi_m = \{\mathcal{P} \in \Pi : |\operatorname{supp}(\mathcal{P})| = m\}$ (this is slightly different from the definition used in the proof of Theorem 10). Note that, for any $\mathcal{P} \in \Pi$, the value of $\mathring{\rho}_{\mathcal{P}}$ is entirely determined by $\operatorname{supp}(\mathcal{P})$. Thus, defining, $\forall m \in \mathbb{N}$ with $m \leq |\mathcal{X}|$,

$$\mathring{\rho}_m = \inf_{\mathcal{X}_m \subseteq \mathcal{X} : |\mathcal{X}_m| = m} \sup\Bigg\{\rho \in [0,1] : \forall \text{ finite } Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \operatorname{DIS}(\{f,g\}) \cap \mathcal{X}_m \neq \emptyset\},$$
$$\max_{x\in\mathcal{X}_m} \operatorname{Split}(Q,x) \geq \rho|Q|\Bigg\},$$

we have $\inf_{\mathcal{P}\in\Pi_m} \mathring{\rho}_{\mathcal{P}} \geq \mathring{\rho}_m$ (in fact, they are equal). Thus, since $\Pi = \bigcup_{m\in\mathbb{N}} \Pi_m$, we have $\inf_{\mathcal{P}\in\Pi} \mathring{\rho}_{\mathcal{P}} = \inf_{m\in\mathbb{N}:m\leq|\mathcal{X}|} \inf_{\mathcal{P}\in\Pi_m} \mathring{\rho}_{\mathcal{P}} \geq \inf_{m\in\mathbb{N}:m\leq|\mathcal{X}|} \mathring{\rho}_m$. Therefore, it suffices to show that $\mathring{\rho}_m \geq 1/\mathfrak{s}$ for all $m \in \mathbb{N}$ with $m \leq |\mathcal{X}|$.

We proceed by induction on $m \in \mathbb{N}$ with $m \leq |\mathcal{X}|$, combined with a nested inductive argument on Q. As base cases (for induction on m), consider any $m \leq \mathfrak{s}$. Fix any $\mathcal{X}_m \subseteq \mathcal{X}$ with $|\mathcal{X}_m| = m$

(noting that $m \leq \mathfrak{s}$ implies $m \leq |\mathcal{X}|$, since $\mathfrak{s} \leq |\mathcal{X}|$ immediately follows from Definition 2). Also fix any finite set $Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \mathrm{DIS}(\{f,g\}) \cap \mathcal{X}_m \neq \emptyset\}$. Since $\forall \{f,g\} \in Q$, $\exists x \in \mathcal{X}_m$ such that $f(x) \neq g(x)$, the pigeonhole principle implies $\exists x \in \mathcal{X}_m$ with $|\{\{f,g\} \in Q : f(x) \neq g(x)\}| \geq |Q|/|\mathcal{X}_m| = |Q|/m$. For this x, we have $\mathrm{Split}(Q,x) \geq |\{\{f,g\} \in Q : f(x) \neq g(x)\}| \geq (1/m)|Q| \geq (1/\mathfrak{s})|Q|$. Since this holds for any such choice of Q and $\mathcal{X}_m$, we have that $\mathring{\rho}_m \geq 1/\mathfrak{s}$.

If $|\mathcal{X}| = \mathfrak{s}$, this completes the proof. Otherwise, take as an inductive hypothesis that, for some $m \in \mathbb{N}$ with $\mathfrak{s} < m \leq |\mathcal{X}|$, $\mathring{\rho}_{m-1} \geq 1/\mathfrak{s}$. Fix any $\mathcal{X}_m \subseteq \mathcal{X}$ with $|\mathcal{X}_m| = m$. We now introduce a nested inductive argument on Q (based on the partial ordering induced by the subset relation). As a base case, if $Q = \emptyset$, then trivially $\max_{x \in \mathcal{X}_m} \mathrm{Split}(Q,x) = 0 = (1/\mathfrak{s})|Q|$. Now take as a nested inductive hypothesis that, for some nonempty finite set $Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \mathrm{DIS}(\{f,g\}) \cap \mathcal{X}_m \neq \emptyset\}$, for every strict subset $R \subset Q$, $\max_{x \in \mathcal{X}_m} \mathrm{Split}(R,x) \geq (1/\mathfrak{s})|R|$.

First, consider the case in which $\exists x \in \mathcal{X}_m$ such that $x \notin \bigcup_{\{f,g\} \in Q} \mathrm{DIS}(\{f,g\})$. In this case, every $\{f,g\} \in Q$ has $\mathrm{DIS}(\{f,g\}) \cap (\mathcal{X}_m \setminus \{x\}) = \mathrm{DIS}(\{f,g\}) \cap \mathcal{X}_m \neq \emptyset$, so that $Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \mathrm{DIS}(\{f,g\}) \cap (\mathcal{X}_m \setminus \{x\}) \neq \emptyset\}$. Therefore, since $|\mathcal{X}_m \setminus \{x\}| = m-1$, by definition of $\mathring{\rho}_{m-1}$ we have $\max_{x' \in \mathcal{X}_m} \mathrm{Split}(Q,x') \geq \max_{x' \in \mathcal{X}_m \setminus \{x\}} \mathrm{Split}(Q,x') \geq \mathring{\rho}_{m-1}|Q|$. Combined with the inductive hypothesis (for m), this implies $\max_{x' \in \mathcal{X}_m} \mathrm{Split}(Q,x') \geq (1/\mathfrak{s})|Q|$.

Now consider the remaining case, in which $\forall x \in \mathcal{X}_m$, $\exists \{f_x, g_x\} \in Q$ with $x \in \mathrm{DIS}(\{f_x, g_x\})$. Since $\{f_x, g_x\} \notin Q_x^y$ for every $y \in \mathcal{Y}$ and $x \in \mathcal{X}_m$, we have $\max_{x \in \mathcal{X}_m} \mathrm{Split}(Q,x) \geq 1$. We proceed by a kind of set-covering argument, as follows. For each $x \in \mathcal{X}_m$, denote $y_x = \mathrm{argmax}_{y \in \mathcal{Y}} |Q_x^y|$ (breaking ties arbitrarily), and denote $S_x = \{x' \in \mathcal{X}_m : \{f_x, g_x\} \notin Q_{x'}^{y_{x'}}\}$. Let z_1 be any element of $\mathcal{X}_m$. Then, for integers $i \geq 2$, inductively define z_i as any element of $\mathcal{X}_m \setminus \bigcup_{j=1}^{i-1} S_{z_j}$, up until the smallest index $i \in \mathbb{N}$ for which $\mathcal{X}_m \setminus \bigcup_{j=1}^{i} S_{z_i} = \emptyset$; denote by I this smallest i with $\mathcal{X}_m \setminus \bigcup_{j=1}^{i} S_{z_i} = \emptyset$. Note that, since $\{f_x, g_x\} \notin Q_x^{y_x}$ (and hence $x \in S_x$) for each $x \in \mathcal{X}_m$, every z_i is distinct, which further implies that $I \leq m$ (and in particular, that I exists). Furthermore, since any $i \in \{1, \ldots, I\}$ and $x \in \mathcal{X}_m$ with $\{f_x, g_x\} = \{f_{z_i}, g_{z_i}\}$ have $S_x = S_{z_i}$, and therefore $x \in S_{z_i}$, $\nexists j > i$ with $z_j = x$. Thus, we also have that $\{f_{z_i}, g_{z_i}\} \neq \{f_{z_j}, g_{z_j}\}$ for every $i, j \in \{1, \ldots, I\}$ with $i \neq j$.

Now let $i_1 = I$, and for integers $k \geq 2$, inductively define

$$i_k = \max\left\{ i \in \{1, \ldots, i_{k-1} - 1\} : \left(S_{z_i} \setminus \bigcup_{j=1}^{i-1} S_{z_j} \right) \setminus \bigcup_{j=1}^{k-1} S_{z_{i_j}} \neq \emptyset \right\},$$

up to the smallest index $k \in \mathbb{N}$ with $\left\{ i \in \{1, \ldots, i_k - 1\} : \left(S_{z_i} \setminus \bigcup_{j=1}^{i-1} S_{z_j} \right) \setminus \bigcup_{j=1}^{k} S_{z_{i_j}} \neq \emptyset \right\} = \emptyset$; denote by K this final value of k (which must exist, since $i_{k+1} \in \mathbb{N}$ is defined and strictly smaller than i_k for any k for which this set is nonempty; in particular, $1 \leq K \leq I$). Finally, let $x_1 = z_{i_1}$, and for each $k \in \{1, \ldots, K\}$, let x_k denote any element of $\left(S_{z_{i_k}} \setminus \bigcup_{j=1}^{i_k - 1} S_{z_j} \right) \setminus \bigcup_{j=1}^{k-1} S_{z_{i_j}}$, which is nonempty by definition of i_k.

We first establish, by induction, that $\bigcup_{k=1}^{K} S_{z_{i_k}} = \mathcal{X}_m$. By construction, we have $\bigcup_{i=1}^{I} S_{z_i} = \mathcal{X}_m$. Furthermore, for any $i \in \{1, \ldots, I\}$, if $\bigcup_{j \leq i} S_{z_j} \cup \bigcup_{1 \leq k \leq K : i_k \geq i+1} S_{z_{i_k}} = \mathcal{X}_m$, then either $i \in \{i_1, \ldots, i_K\}$, in which case $\bigcup_{j<i} S_{z_j} \cup \bigcup_{1 \leq k \leq K : i_k \geq i} S_{z_{i_k}} = \bigcup_{j \leq i} S_{z_j} \cup \bigcup_{1 \leq k \leq K : i_k \geq i+1} S_{z_{i_k}} = \mathcal{X}_m$, or else $i \notin \{i_1, \ldots, i_K\}$, which (by definition of the i_k sequence) implies $S_{z_i} \subseteq \bigcup_{j=1}^{i-1} S_{z_j} \cup \bigcup_{1 \leq k \leq K : i_k \geq i+1} S_{z_{i_k}}$, so that $\bigcup_{j<i} S_{z_j} \cup \bigcup_{1 \leq k \leq K : i_k \geq i} S_{z_{i_k}} = \bigcup_{j<i} S_{z_j} \cup \bigcup_{1 \leq k \leq K : i_k \geq i+1} S_{z_{i_k}} = \bigcup_{j \leq i} S_{z_j} \cup \bigcup_{1 \leq k \leq K : i_k \geq i+1} S_{z_{i_k}} = \mathcal{X}_m$. By induction, we have that $\bigcup_{k=1}^{K} S_{z_{i_k}} = \bigcup_{j<1} S_{z_j} \cup$

$\bigcup_{1\leq k\leq K:i_k\geq 1} S_{z_{i_k}} = \mathcal{X}_m$. In other words, $\forall x \in \mathcal{X}_m$, $\exists k(x) \in \{1,\ldots,K\}$ with $\{f_{z_{i_{k(x)}}}, g_{z_{i_{k(x)}}}\} \notin Q_x^{y_x}$.

In particular, letting $R = Q \setminus \{\{f_{z_{i_k}}, g_{z_{i_k}}\} : k \in \{1,\ldots,K\}\}$, we have that $\forall x \in \mathcal{X}_m$, $\{f_{z_{i_{k(x)}}}, g_{z_{i_{k(x)}}}\} \in (Q \setminus R) \setminus (Q_x^{y_x} \setminus R)$ while $Q_x^{y_x} \setminus R \subseteq Q \setminus R$, so that $|Q \setminus R| - |Q_x^{y_x} \setminus R| \geq 1$. Therefore, $\forall x \in \mathcal{X}_m$,

$$\begin{aligned}\operatorname{Split}(R,x) &= |R| - \max_{y\in\mathcal{Y}} |R_x^y| \leq |R| - |R_x^{y_x}| = |R| - |R \cap Q_x^{y_x}| \\ &= (|Q| - |Q \setminus R|) - (|Q_x^{y_x}| - |Q_x^{y_x} \setminus R|) = (|Q| - |Q_x^{y_x}|) - (|Q \setminus R| - |Q_x^{y_x} \setminus R|) \\ &\leq |Q| - |Q_x^{y_x}| - 1 = |Q| - \max_{y\in\mathcal{Y}} |Q_x^y| - 1 = \operatorname{Split}(Q,x) - 1. \end{aligned} \tag{61}$$

Since $K \geq 1$, we may note that R is a strict subset of Q, so that the (nested) inductive hypothesis implies that $\max_{x\in\mathcal{X}_m} \operatorname{Split}(R,x) \geq (1/\mathfrak{s})|R|$. Combined with (61), this implies

$$\max_{x\in\mathcal{X}_m} \operatorname{Split}(Q,x) \geq \max_{x\in\mathcal{X}_m} \operatorname{Split}(R,x) + 1 \geq (1/\mathfrak{s})|R| + 1. \tag{62}$$

Next, we argue that $K \leq \mathfrak{s}$, by proving that $\{x_1,\ldots,x_K\}$ is a star set for $\mathbb{C}$. By definition of z_I, we have $z_I \in \mathcal{X}_m \setminus \bigcup_{j=1}^{I-1} S_{z_j} \subseteq \mathcal{X}_m \setminus \bigcup_{k=2}^{K} S_{z_{i_k}}$. Furthermore, $z_I \in S_{z_I}$, so that $z_I \in S_{z_I} \setminus \bigcup_{k=2}^{K} S_{z_{i_k}}$. Since $x_1 = z_{i_1} = z_I$, we have $x_1 \in S_{z_{i_1}} \setminus \bigcup_{k=2}^{K} S_{z_{i_k}}$. Also, for each $k \in \{2,\ldots,K\}$, by definition, $x_k \in \left(S_{z_{i_k}} \setminus \bigcup_{j=1}^{i_k-1} S_{z_j}\right) \setminus \bigcup_{j=1}^{k-1} S_{z_{i_j}} \subseteq \left(S_{z_{i_k}} \setminus \bigcup_{j=k+1}^{K} S_{z_{i_j}}\right) \setminus \bigcup_{j=1}^{k-1} S_{z_{i_j}} = S_{z_{i_k}} \setminus \bigcup_{1\leq j\leq K:j\neq k} S_{z_{i_j}}$. Therefore, every $k \in \{1,\ldots,K\}$ has $x_k \in S_{z_{i_k}} \setminus \bigcup_{1\leq j\leq K:j\neq k} S_{z_{i_j}}$. In particular, for every $k \in \{1,\ldots,K\}$, since $x_k \in S_{z_{i_k}}$, we have $\{f_{z_{i_k}}, g_{z_{i_k}}\} \notin Q_{x_k}^{y_{x_k}}$, so that $\exists h_k \in \{f_{z_{i_k}}, g_{z_{i_k}}\}$ with $h_k(x_k) \neq y_{x_k}$. Furthermore, for every $j \in \{1,\ldots,K\} \setminus \{k\}$, since $x_j \notin S_{z_{i_k}}$, we have $\{f_{z_{i_k}}, g_{z_{i_k}}\} \in Q_{x_j}^{y_{x_j}}$, so that $f_{z_{i_k}}(x_j) = g_{z_{i_k}}(x_j) = y_{x_j}$, and in particular, $h_k(x_j) = y_{x_j}$. Also, since we have chosen $x_1 = z_{i_1}$, so that $x_1 \in \operatorname{DIS}(\{f_{z_{i_1}}, g_{z_{i_1}}\})$, $\exists h_0 \in \{f_{z_{i_1}}, g_{z_{i_1}}\}$ with $h_0(x_1) \neq h_1(x_1)$: that is, $h_0(x_1) = y_{x_1}$. Thus, since $f_{z_{i_1}}(x_j) = g_{z_{i_1}}(x_j) = y_{x_j}$ for every $j \in \{2,\ldots,K\}$, we have that $h_0(x_k) = y_{x_k}$ for every $k \in \{1,\ldots,K\}$. Altogether, we have that every $k \in \{1,\ldots,K\}$ has $h_k(x_k) \neq h_0(x_k)$, while every $j \in \{1,\ldots,K\} \setminus \{k\}$ has $h_k(x_j) = h_0(x_j)$. In other words, $\forall k \in \{1,\ldots,K\}$, $\operatorname{DIS}(\{h_0,h_k\}) \cap \{x_1,\ldots,x_K\} = \{x_k\}$: that is, $\{x_1,\ldots,x_K\}$ is a star set for $\mathbb{C}$, witnessed by $\{h_0,h_1,\ldots,h_K\}$. In particular, this implies $K \leq \mathfrak{s}$.

Therefore, since $|Q \setminus R| = K$ (by distinctness of the pairs $\{f_{z_i}, g_{z_i}\}$ argued above), (62) implies

$$\max_{x\in\mathcal{X}_m} \operatorname{Split}(Q,x) \geq (1/\mathfrak{s})|R| + \frac{K}{\mathfrak{s}} = (1/\mathfrak{s})(|R| + |Q \setminus R|) = (1/\mathfrak{s})|Q|.$$

By the principle of induction (on Q), we have $\max_{x\in\mathcal{X}_m} \operatorname{Split}(Q,x) \geq (1/\mathfrak{s})|Q|$ for every finite set $Q \subseteq \{\{f,g\} \subseteq \mathbb{C} : \operatorname{DIS}(\{f,g\}) \cap \mathcal{X}_m \neq \emptyset\}$. Since this holds for any choice of $\mathcal{X}_m$ with $|\mathcal{X}_m| = m$, we have $\mathring{\rho}_m \geq 1/\mathfrak{s}$. By the principle of induction (on m), we have established that $\mathring{\rho}_m \geq 1/\mathfrak{s}$ for every $m \in \mathbb{N}$ with $m \leq |\mathcal{X}|$, which completes the proof of the theorem. ■

C.3 The Teaching Dimension

Here we give the proofs of results from Section 7.3. We first prove that every minimal specifying set is a star set (Lemma 14). In fact, we establish a slightly stronger claim here (which also applies to local minima), stated formally as follows.

Lemma 44 *Fix any $h : \mathcal{X} \to \mathcal{Y}$, $m \in \mathbb{N}$, $\mathcal{U} \in \mathcal{X}^m$, and any specifying set S for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$. If $\forall x \in S$, $S \setminus \{x\}$ is not a specifying set for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$, then S is a star set for $\mathbb{C} \cup \{h\}$ centered at h.*

Proof Fix an arbitrary sequence $\mathcal{U} = \{x_1, \ldots, x_m\} \in \mathcal{X}^m$ and any $h : \mathcal{X} \to \mathcal{Y}$. Let $t \geq \mathrm{TD}(h, \mathbb{C}[\mathcal{U}], \mathcal{U})$, and let $i_1, \ldots, i_t \in \{1, \ldots, m\}$ be such that $S = \{x_{i_1}, \ldots, x_{i_t}\}$ is a specifying set for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$. First note that, if $\exists j \in \{1, \ldots, t\}$ such that every $g \in V_{S \setminus \{x_{i_j}\}, h}$ has $g(x_{i_j}) = h(x_{i_j})$ (which includes the case $V_{S \setminus \{x_{i_j}\}, h} = \emptyset$), then $V_{S \setminus \{x_{i_j}\}, h} = V_{S,h}$, so that $|V_{S \setminus \{x_{i_j}\}, h} \cap \mathbb{C}[\mathcal{U}]| = |V_{S,h} \cap \mathbb{C}[\mathcal{U}]| \leq 1$; thus, $S \setminus \{x_{i_j}\}$ is also a specifying set for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$.

Therefore, if S is such that $\forall j \leq t$, $S \setminus \{x_{i_j}\}$ is *not* a specifying set for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$, then $\forall j \in \{1, \ldots, t\}$, $\exists h_j \in V_{S \setminus \{x_{i_j}\}, h}$ with $h_j(x_{i_j}) \neq h(x_{i_j})$; noting that "$h_j \in V_{S \setminus \{x_{i_j}\}, h}$" is equivalent to saying "$h_j(x_{i_k}) = h(x_{i_k})$ for every $k \in \{1, \ldots, t\} \setminus \{j\}$," this precisely matches the definition of a star set in Section 4: that is, we have proven that $\{x_{i_1}, \ldots, x_{i_t}\}$ is a star set for $\mathbb{C} \cup \{h\}$, witnessed by $\{h, h_1, \ldots, h_t\}$, and hence centered at h. ■

Proof of Lemma 14 Lemma 14 follows immediately from Lemma 44 by noting that, for any *minimal* specifying set S for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$, $\forall x \in S$, $|S \setminus \{x\}| < \mathrm{TD}(h, \mathbb{C}[\mathcal{U}], \mathcal{U})$, so that $S \setminus \{x\}$ cannot possibly be a specifying set for h on $\mathcal{U}$ with respect to $\mathbb{C}[\mathcal{U}]$. ■

We are now ready for the proof of Theorem 13.

Proof of Theorem 13 Fix any $m \in \mathbb{N}$. First, note that for $\{x_i\}_{i=1}^{\mathfrak{s}}$ and $\{h_i\}_{i=0}^{\mathfrak{s}}$ as in Definition 2, letting $\mathcal{U} = \{x_1, \ldots, x_{\min\{\mathfrak{s}, m\}}\}$, for any positive integer $i \leq \min\{\mathfrak{s}, m\}$, any subsequence $S \subseteq \mathcal{U}$ with $x_i \notin S$ has $\{h_0, h_i\} \subseteq V_{S, h_0}$. Thus, since $x_i \in \mathcal{U}$, and $h_0(x_i) \neq h_i(x_i)$, we have $|V_{S,h_0} \cap \mathbb{C}[\mathcal{U}]| \geq 2$. Since this is true for every such $i \leq \min\{\mathfrak{s}, m\}$, every $S \subseteq \mathcal{U}$ without $\{x_1, \ldots, x_{\min\{\mathfrak{s}, m\}}\} \subseteq S$ has $|V_{S,h_0} \cap \mathbb{C}[\mathcal{U}]| \geq 2$. Therefore, $\mathrm{TD}(h_0, \mathbb{C}[\mathcal{U}], \mathcal{U}) \geq \min\{\mathfrak{s}, m\}$. Thus, by the definitions of XTD and TD, monotonicity of maximization in the set maximized over, and monotonicity of $t \mapsto \mathrm{TD}(\mathbb{C}, t)$,[15] we have

$$\mathrm{XTD}(\mathbb{C}, m) \geq \mathrm{TD}(\mathbb{C}, m) \geq \mathrm{TD}(\mathbb{C}, \min\{\mathfrak{s}, m\}) \geq \mathrm{TD}(h_0, \mathbb{C}[\mathcal{U}], \mathcal{U}) \geq \min\{\mathfrak{s}, m\}.$$

Furthermore, it follows immediately from the definition that $\mathrm{XTD}(\mathbb{C}, m) \leq m$. Note that this completes the proof in the case that $\mathfrak{s} \geq m$. To address the remaining case, for the remainder of the proof, we suppose $\mathfrak{s} \leq m$, and focus on establishing $\mathrm{XTD}(\mathbb{C}, m) \leq \mathfrak{s}$.

For this, we proceed by induction on m, taking as a base case the fact that $\mathrm{XTD}(\mathbb{C}, \mathfrak{s}) \leq \mathfrak{s}$, which trivially follows from the definition of XTD. Now take as an inductive hypothesis that for some $m > \mathfrak{s}$, we have $\mathrm{XTD}(\mathbb{C}, m-1) \leq \mathfrak{s}$. Fix any sequence $\mathcal{U}_m = \{x_1, \ldots, x_m\} \in \mathcal{X}^m$, and $h : \mathcal{X} \to \mathcal{Y}$, and denote $\mathcal{U}_{m-1} = \{x_1, \ldots, x_{m-1}\}$. Let $t \in \mathbb{N} \cup \{0\}$ and $S \in \mathcal{U}_{m-1}^t$ be such that S is a minimal specifying set for h on $\mathcal{U}_{m-1}$ with respect to $\mathbb{C}[\mathcal{U}_{m-1}]$. If $|S| \geq \mathrm{TD}(h, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m)$, then since S is a *minimal* specifying set for h on $\mathcal{U}_{m-1}$ with respect to $\mathbb{C}[\mathcal{U}_{m-1}]$, we have $|S| = \mathrm{TD}(h, \mathbb{C}[\mathcal{U}_{m-1}], \mathcal{U}_{m-1}) \leq \mathrm{XTD}(\mathbb{C}, m-1) \leq \mathfrak{s}$ by the inductive hypothesis; thus, in this case we have $\mathrm{TD}(h, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m) \leq |S| \leq \mathfrak{s}$. On the other hand, suppose

15. $\forall S \in \mathcal{X}^t$, $\forall x \in S$, $\forall h$, $\mathrm{TD}(h, \mathbb{C}[S \cup \{x\}], S \cup \{x\}) = \mathrm{TD}(h, \mathbb{C}[S], S)$. Thus, $\mathrm{TD}(\mathbb{C}, t+1) = \max_{h \in \mathbb{C}} \max_{S \in \mathcal{X}^t} \max_{x \in \mathcal{X}} \mathrm{TD}(h, \mathbb{C}[S \cup \{x\}], S \cup \{x\}) \geq \max_{h \in \mathbb{C}} \max_{S \in \mathcal{X}^t} \max_{x \in S} \mathrm{TD}(h, \mathbb{C}[S \cup \{x\}], S \cup \{x\}) = \max_{h \in \mathbb{C}} \max_{S \in \mathcal{X}^t} \mathrm{TD}(h, \mathbb{C}[S], S) = \mathrm{TD}(\mathbb{C}, t)$.

$|S| < \mathrm{TD}(h, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m)$. In this case, since S is a specifying set for h on $\mathcal{U}_{m-1}$ with respect to $\mathbb{C}[\mathcal{U}_{m-1}]$, we have $\mathrm{DIS}(V_{S,h}) \cap \mathcal{U}_m \subseteq (\mathrm{DIS}(V_{S,h}) \cap \mathcal{U}_{m-1}) \cup \{x_m\} = \{x_m\}$. But since $|S| < \mathrm{TD}(h, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m)$, S cannot be a specifying set for h on $\mathcal{U}_m$ with respect to $\mathbb{C}[\mathcal{U}_m]$, so that $\mathrm{DIS}(V_{S,h}) \cap \mathcal{U}_m \neq \emptyset$. Therefore, $\mathrm{DIS}(V_{S,h}) \cap \mathcal{U}_m = \{x_m\}$. In particular, this implies that $S \cup \{x_m\}$ is a specifying set for h on $\mathcal{U}_m$ with respect to $\mathbb{C}[\mathcal{U}_m]$, and in particular, must be a *minimal* such specifying set, since $|S \cup \{x_m\}| = |S| + 1 \leq \mathrm{TD}(h, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m)$. Therefore, Lemma 14 implies that $S \cup \{x_m\}$ is a star set for $\mathbb{C} \cup \{h\}$ centered at h. If $h \in \mathbb{C}$, this already implies that $|S \cup \{x_m\}| \leq \mathfrak{s}$; furthermore, we can argue that this remains the case even if $h \notin \mathbb{C}$, as follows. Since $x_m \in \mathrm{DIS}(V_{S,h})$, we have $V_{S \cup \{x_m\}, h} \neq \emptyset$, so that $\exists g_0 \in \mathbb{C}$ such that $\forall x \in S \cup \{x_m\}$, $g_0(x) = h(x)$. Therefore, $S \cup \{x_m\}$ is also a star set for $\mathbb{C}$ centered at g_0, so that $|S \cup \{x_m\}| \leq \mathfrak{s}$. In particular, since $S \cup \{x_m\}$ is a minimal specifying set for h on $\mathcal{U}_m$ with respect to $\mathbb{C}[\mathcal{U}_m]$, we have $|S \cup \{x_m\}| = \mathrm{TD}(h, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m)$, so that $\mathrm{TD}(h, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m) \leq \mathfrak{s}$ in this case as well. Thus, in either case, we have $\mathrm{TD}(h, \mathbb{C}[\mathcal{U}_m], \mathcal{U}_m) \leq \mathfrak{s}$. Maximizing over the choice of h and $\{x_1, \ldots, x_m\}$, we have $\mathrm{XTD}(\mathbb{C}, m) \leq \mathfrak{s}$, which completes the inductive step. The result now follows by the principle of induction. ■

Next, we prove Theorem 15.

Proof of Theorem 15 Fix any $m \in \mathbb{N}$ and $\delta \in [0, 1]$. Let $\{x_i\}_{i=1}^{\mathfrak{s}}$ and $\{h_i\}_{i=0}^{\mathfrak{s}}$ be as in Definition 2, and let $\mathcal{U} = \{x_1, \ldots, x_{\min\{\mathfrak{s},m\}}\}$ and $\mathcal{G} = \{h_i : i \in \{0, \ldots, \min\{\mathfrak{s}, m\}\}$. As in the proof of Theorem 13, for any positive integer $i \leq \min\{\mathfrak{s}, m\}$, any subsequence $S \subseteq \mathcal{U}$ with $x_i \notin S$ has $\{h_0, h_i\} \subseteq V_{S,h_0}$. Thus, since $x_i \in \mathcal{U}$ for every $i \leq \min\{\mathfrak{s}, m\}$, and every h_i realizes a distinct classification of $\mathcal{U}$ ($i \leq \min\{\mathfrak{s}, m\}$), we have $|V_{S,h_0} \cap \mathcal{G}[\mathcal{U}]| \geq |\{i \in \{1, \ldots, \min\{\mathfrak{s}, m\}\} : x_i \notin S\}| + 1 \geq \min\{\mathfrak{s}, m\} - |S| + 1$. In particular, to have $|V_{S,h_0} \cap \mathcal{G}[\mathcal{U}]| \leq \delta |\mathcal{G}[\mathcal{U}]| + 1 = \delta(\min\{\mathfrak{s}, m\} + 1) + 1$, we must have $|S| \geq (1 - \delta) \min\{\mathfrak{s}, m\} - \delta$. Therefore, $\mathrm{XPTD}(h_0, \mathcal{G}[\mathcal{U}], \mathcal{U}, \delta) \geq (1 - \delta) \min\{\mathfrak{s}, m\} - \delta$. By definition of $\mathrm{XPTD}(\mathcal{H}, m, \delta)$ and the fact that $\mathcal{G} \subseteq \mathbb{C}$, and since $t \mapsto \mathrm{XPTD}(\mathcal{H}, t, \delta)$ is nondecreasing (since $\forall S \in \mathcal{X}^t, \forall x \in S, \forall h, \mathrm{XPTD}(h, \mathcal{H}[S \cup \{x\}], S \cup \{x\}, \delta) = \mathrm{XPTD}(h, \mathcal{H}[S], S, \delta)$), this further implies

$$\begin{aligned}\max_{\mathcal{H} \subseteq \mathbb{C}} \mathrm{XPTD}(\mathcal{H}, m, \delta) &\geq \mathrm{XPTD}(\mathcal{G}, m, \delta) \geq \mathrm{XPTD}(\mathcal{G}, \min\{\mathfrak{s}, m\}, \delta) \\ &\geq \mathrm{XPTD}(h_0, \mathcal{G}[\mathcal{U}], \mathcal{U}, \delta) \geq (1 - \delta) \min\{\mathfrak{s}, m\} - \delta \geq (1-2\delta) \min\{\mathfrak{s}, m\},\end{aligned}$$

where this last inequality is due to the assumption that $|\mathbb{C}| \geq 3$ (Section 2), which implies $\mathfrak{s} \geq 1$. Since $\mathrm{XPTD}(\cdot, m, \delta) \in \mathbb{N} \cup \{0\}$, this further implies $\max_{\mathcal{H} \subseteq \mathbb{C}} \mathrm{XPTD}(\mathcal{H}, m, \delta) \geq \lceil (1 - 2\delta) \min\{\mathfrak{s}, m\} \rceil$ when $\delta \leq 1/2$.

To establish the right inequality, fix any $\mathcal{H} \subseteq \mathbb{C}$, let $\mathcal{U} \in \mathcal{X}^m$ and $h : \mathcal{X} \to \mathcal{Y}$ be such that $\mathrm{XPTD}(h, \mathcal{H}[\mathcal{U}], \mathcal{U}, \delta) = \mathrm{XPTD}(\mathcal{H}, m, \delta)$, and let $S \subseteq \mathcal{U}$ be a minimal specifying set for h on $\mathcal{U}$ with respect to $\mathcal{H}[\mathcal{U}]$. If $\delta = 0$ or $|S| < \frac{1+\delta}{\delta}$, then $|S| - 1 < \left(1 - \frac{\delta}{1+\delta}\right) |S| \leq |S|$, so that $\mathrm{XPTD}(h, \mathcal{H}[\mathcal{U}], \mathcal{U}, \delta) \leq |S| = \left\lceil \left(1 - \frac{\delta}{1+\delta}\right) |S| \right\rceil$. Otherwise, suppose $\delta > 0$ and $|S| \geq \frac{1+\delta}{\delta}$, and let $k = \left\lfloor |S| / \left\lfloor \frac{\delta}{1+\delta} |S| \right\rfloor \right\rfloor$, and note that $k \geq 1$. Let $R_1, \ldots, R_k$ denote disjoint subsequences of S with each $|R_i| = \left\lfloor \frac{\delta}{1+\delta} |S| \right\rfloor$, which must exist since minimality of S guarantees that its elements are distinct. Note that, for each $i \in \{1, \ldots, k\}$, $(V_{S \setminus R_i, h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}]$ is the set of classifiers g in $\mathcal{H}[\mathcal{U}]$ with $\mathrm{DIS}(\{g, h\}) \cap (S \setminus R_i) = \emptyset$ but $\mathrm{DIS}(\{g, h\}) \cap R_i \neq \emptyset$; in particular, for any $i, j \in \{1, \ldots, k\}$ with $i \neq j$, since $R_j \subseteq S \setminus R_i$ and $R_i \subseteq S \setminus R_j$, $(V_{S \setminus R_i, h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}]$ and $(V_{S \setminus R_j, h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}]$

are disjoint. Thus, since $\mathcal{H}[\mathcal{U}] \supseteq (V_{S,h} \cap \mathcal{H}[\mathcal{U}]) \cup \bigcup_{i=1}^{k} (V_{S \setminus R_i,h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}]$, we have

$$\begin{aligned}
|\mathcal{H}[\mathcal{U}]| &\geq \left| (V_{S,h} \cap \mathcal{H}[\mathcal{U}]) \cup \bigcup_{i=1}^{k} (V_{S \setminus R_i,h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}] \right| \\
&= |V_{S,h} \cap \mathcal{H}[\mathcal{U}]| + \sum_{i=1}^{k} \left| (V_{S \setminus R_i,h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}] \right| \geq \sum_{i=1}^{k} \left| (V_{S \setminus R_i,h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}] \right| \\
&\qquad \geq k \min_{i \in \{1,\ldots,k\}} \left| (V_{S \setminus R_i,h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}] \right|.
\end{aligned}$$

Thus, letting $i^* = \operatorname{argmin}_{i \in \{1,\ldots,k\}} \left| (V_{S \setminus R_i,h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}] \right|$, we have $\left| (V_{S \setminus R_{i^*},h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}] \right| \leq \frac{1}{k} |\mathcal{H}[\mathcal{U}]|$. Furthermore, since S is a specifying set for h on $\mathcal{U}$ with respect to $\mathcal{H}[\mathcal{U}]$, $|V_{S,h} \cap \mathcal{H}[\mathcal{U}]| \leq 1$, so that (since $V_{S,h} \subseteq V_{S \setminus R_{i^*},h}$)

$$\begin{aligned}
\left| V_{S \setminus R_{i^*},h} \cap \mathcal{H}[\mathcal{U}] \right| &= \left| \left((V_{S \setminus R_{i^*},h} \setminus V_{S,h}) \cap \mathcal{H}[\mathcal{U}] \right) \cup (V_{S,h} \cap \mathcal{H}[\mathcal{U}]) \right| \\
&= \left| \left(V_{S \setminus R_{i^*},h} \setminus V_{S,h} \right) \cap \mathcal{H}[\mathcal{U}] \right| + |V_{S,h} \cap \mathcal{H}[\mathcal{U}]| \leq \frac{1}{k} |\mathcal{H}[\mathcal{U}]| + 1.
\end{aligned}$$

Also, since

$$\frac{1}{k} \leq \frac{1}{\left\lfloor \frac{1+\delta}{\delta} \right\rfloor} \leq \frac{1}{\frac{1+\delta}{\delta} - 1} = \delta,$$

this implies $|V_{S \setminus R_{i^*},h} \cap \mathcal{H}[\mathcal{U}]| \leq \delta |\mathcal{H}[\mathcal{U}]| + 1$, so that $\mathrm{XPTD}(h, \mathcal{H}[\mathcal{U}], \mathcal{U}, \delta) \leq |S \setminus R_{i^*}|$. Furthermore, since $R_{i^*} \subseteq S$, $|S \setminus R_{i^*}| = |S| - |R_{i^*}| = |S| - \left\lfloor \frac{\delta}{1+\delta} |S| \right\rfloor = \left\lceil \left(1 - \frac{\delta}{1+\delta} \right) |S| \right\rceil$.

Thus, for any $\delta \in [0,1]$ and regardless of the size of $|S|$, we have $\mathrm{XPTD}(h, \mathcal{H}[\mathcal{U}], \mathcal{U}, \delta) \leq \left\lceil \left(1 - \frac{\delta}{1+\delta} \right) |S| \right\rceil$. Furthermore, since S is a minimal specifying set for h on $\mathcal{U}$ with respect to $\mathcal{H}[\mathcal{U}]$, we have $|S| \leq \mathrm{XTD}(\mathcal{H}, m) \leq \mathrm{XTD}(\mathbb{C}, m)$, and Theorem 13 implies $\mathrm{XTD}(\mathbb{C}, m) = \min\{\mathfrak{s}, m\}$. Therefore, $\mathrm{XPTD}(h, \mathcal{H}[\mathcal{U}], \mathcal{U}, \delta) \leq \left\lceil \left(1 - \frac{\delta}{1+\delta} \right) \min\{\mathfrak{s}, m\} \right\rceil$. Maximizing the left hand side over the choice of h, $\mathcal{H}$, and $\mathcal{U}$ completes the proof. ■

C.4 The Doubling Dimension

We now present the proof of Theorem 17

Proof of Theorem 17 For the lower bound, fix any $\varepsilon \in (0,1]$, and take $\{x_i\}_{i=1}^{\mathfrak{s}}$ and $\{h_i\}_{i=0}^{\mathfrak{s}}$ as in Definition 2, and let $m = \mathfrak{s} \wedge \left\lfloor \frac{1}{\varepsilon} \right\rfloor$. Let $\mathcal{P}$ be a probability measure on $\mathcal{X}$ with $\mathcal{P}(\{x_i\}) = 1/m$ for each $i \in \{1, \ldots, m\}$. Thus, $\{h_0, h_1, \ldots, h_m\} \subseteq \mathrm{B}_{\mathcal{P}}(h_0, 1/m)$. Furthermore, for any $i \in \{0, \ldots, m\}$ and any classifier g with $\mathcal{P}(x : g(x) \neq h_i(x)) \leq 1/(2m)$, we must have $g(x_j) = h_i(x_j)$ for every $j \in \{1, \ldots, m\}$. Therefore, any $\frac{1}{2m}$-cover of $\mathrm{B}_{\mathcal{P}}(h_0, 1/m)$ must contain classifiers $g_0, \ldots, g_m$ with $\forall i \in \{0, \ldots, m\}, \forall j \in \{1, \ldots, m\}$, $g_i(x_j) = h_i(x_j)$. Thus, since each h_i (with $i \leq m$) realizes a distinct classification of $\{x_1, \ldots, x_m\}$, it follows that $\mathcal{N}(1/(2m), \mathrm{B}_{\mathcal{P}}(h_0, 1/m), \mathcal{P}) \geq m + 1$. Noting that $1/m \geq \varepsilon$, we have that

$$\sup_{P} \sup_{h \in \mathbb{C}} D_{h,P}(\varepsilon) \geq D_{h_0,\mathcal{P}}(\varepsilon) \geq \log_2 \left(\mathcal{N} \left(\frac{1}{2m}, \mathrm{B}_{\mathcal{P}} \left(h_0, \frac{1}{m} \right), \mathcal{P} \right) \right) \geq \log_2 (m+1) \geq \log_2 \left(\mathfrak{s} \wedge \frac{1}{\varepsilon} \right).$$

For the remaining term in the lower bound (i.e., d), we modify an argument of Kulkarni (1989, Proposition 3). If $d < 5$, then $d \lesssim \mathrm{Log}\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right)$, so that the lower bound follows from the above. Otherwise, suppose $d \geq 5$. We first let $\{x_1', \ldots, x_d'\}$ denote a set of d points in $\mathcal{X}$ shattered by $\mathbb{C}$, and we let G denote the set of classifiers $g \in \mathbb{C}[\{x_1', \ldots, x_d'\}]$ with $g(x_d') = -1$ and $\sum_{i=1}^{d-1} \mathbb{1}[g(x_i') = +1] = \left\lfloor \frac{d-1}{4} \right\rfloor$. For any $g \in G$, note that, if H is a classifier sampled uniformly at random from G, a Chernoff bound (for sampling without replacement) implies

$$\mathbb{P}\left(\sum_{i=1}^{d-1} \mathbb{1}[H(x_i') = g(x_i')] \geq \frac{d-1}{8}\right) \leq \exp\left\{-\frac{d-1}{48}\right\}.$$

Thus, there are at most $|G| \exp\left\{-\frac{d-1}{48}\right\}$ elements $h \in G$ with $\sum_{i=1}^{d-1} \mathbb{1}[h(x_i') = g(x_i')] \geq \frac{d-1}{8}$. Now take $\mathcal{H}_0 = \{\}$, and take as an inductive hypothesis that, for some positive integer $k < 1 + \exp\left\{\frac{d-1}{48}\right\}$, there is a set $\mathcal{H}_{k-1} \subseteq G$ with $|\mathcal{H}_{k-1}| = k-1$ such that $\forall h, g \in \mathcal{H}_{k-1}$ with $h \neq g$, $\sum_{i=1}^{d-1} \mathbb{1}[h(x_i') = g(x_i')] < \frac{d-1}{8}$. Since $|\mathcal{H}_{k-1}| \cdot |G| \exp\left\{-\frac{d-1}{48}\right\} < |G|$, $\exists g_k \in G$ such that $\forall h \in \mathcal{H}_{k-1}$, $\sum_{i=1}^{d-1} \mathbb{1}[h(x_i') = g_k(x_i')] < \frac{d-1}{8}$. Thus, defining $\mathcal{H}_k = \mathcal{H}_{k-1} \cup \{g_k\}$ extends the inductive hypothesis. By induction, this establishes the existence of a set $\mathcal{H} \subseteq G$ with $|\mathcal{H}| \geq \exp\left\{\frac{d-1}{48}\right\}$ such that $\forall h, g \in \mathcal{H}$ with $h \neq g$, $\sum_{i=1}^{d-1} \mathbb{1}[h(x_i') = g(x_i')] < \frac{d-1}{8}$. Fix any $\varepsilon \in (0, 1/4]$ and let $\mathcal{P}$ denote a probability measure over $\mathcal{X}$ with $\mathcal{P}(\{x_i'\}) = \frac{4\varepsilon}{d-1}$ for each $i \in \{1, \ldots, d-1\}$, and $\mathcal{P}(\{x_d'\}) = 1 - 4\varepsilon$. Note that any $h, g \in G$ with $\sum_{i=1}^{d-1} \mathbb{1}[h(x_i') = g(x_i')] < \frac{d-1}{8}$ have $\mathcal{P}(x : h(x) \neq g(x)) > \frac{d-1}{4} \frac{4\varepsilon}{d-1} = \varepsilon$. Thus, $\mathcal{H}$ is an *ε-packing* under the $L_1(\mathcal{P})$ pseudometric. Recall that this implies $|\mathcal{H}| \leq \mathcal{N}(\varepsilon/2, G, \mathcal{P})$ (Kolmogorov and Tikhomirov, 1959, 1961). Furthermore, note that any $g \in G$ has $\mathcal{P}(x : g(x) = +1) = \left\lfloor \frac{d-1}{4} \right\rfloor \frac{4\varepsilon}{d-1} \leq \varepsilon$. Thus, letting $h_- \in \mathbb{C}$ be such that $\forall i \in \{1, \ldots, d\}$, $h_-(x_i') = -1$ (which exists, by shatterability of $x_1', \ldots, x_d'$), we have $G \subseteq \mathrm{B}_{\mathcal{P}}(h_-, \varepsilon)$. Therefore, $\mathcal{N}(\varepsilon/2, G, \mathcal{P}) \leq \mathcal{N}(\varepsilon/2, \mathrm{B}_{\mathcal{P}}(h_-, \varepsilon), \mathcal{P})$. Altogether, we have that

$$d \lesssim \frac{d-1}{48} \log_2(e) \leq \log_2(|\mathcal{H}|) \leq \log_2\left(\mathcal{N}(\varepsilon/2, \mathrm{B}_{\mathcal{P}}(h_-, \varepsilon), \mathcal{P})\right) \leq D_{h_-, \mathcal{P}}(\varepsilon) \leq \sup_P \sup_{h \in \mathbb{C}} D_{h,P}(\varepsilon).$$

For the upper bound, fix any $h \in \mathbb{C}$, any probability measure $\mathcal{P}$ over $\mathcal{X}$, and any $\varepsilon \in (0, 1]$, and fix any value $r \in [\varepsilon, 1]$. Recall that any maximal subset $G_r \subseteq \mathrm{B}_{\mathcal{P}}(h, r)$ of classifiers in $\mathrm{B}_{\mathcal{P}}(h, r)$ with $\min_{f,g \in G_r : f \neq g} \mathcal{P}(x : f(x) \neq g(x)) > r/2$ (called a maximal $(r/2)$-packing of $\mathrm{B}_{\mathcal{P}}(h, r)$) is also an $(r/2)$-cover of $\mathrm{B}_{\mathcal{P}}(h, r)$ (see e.g., Kolmogorov and Tikhomirov, 1959, 1961). Thus, we have that $\mathcal{N}\left(\frac{r}{2}, \mathrm{B}_{\mathcal{P}}(h, r), \mathcal{P}\right) \leq |G_r|$, for any such set G_r. Let $m = \left\lceil \frac{4}{r} \ln(|G_r|) \right\rceil$, and let $X_1, X_2, \ldots, X_m$ be independent $\mathcal{P}$-distributed random variables. Let E_1 denote the event that $\forall f, g \in G_r$ with $f \neq g$, $\exists i \in \{1, \ldots, m\}$ with $f(X_i) \neq g(X_i)$. For any $f, g \in G_r$ with $f \neq g$, $\mathbb{P}(\exists i \in \{1, \ldots, m\} : f(X_i) \neq g(X_i)) = 1 - (1 - \mathcal{P}(x : f(x) \neq g(x)))^m > 1 - (1 - r/2)^m > 1 - e^{-mr/2} \geq 1 - 1/|G_r|^2$. Therefore, by a union bound, $\mathbb{P}(E_1) > 1 - \binom{|G_r|}{2} \frac{1}{|G_r|^2} \geq \frac{1}{2}$. In particular, note that on the event E_1, the elements of G_r realize distinct classifications of the sequence $(X_1, \ldots, X_m)$, so that (since $G_r \subseteq \mathrm{B}_{\mathcal{P}}(h, r)$) $|G_r|$ is upper bounded by the number of distinct classifications of $(X_1, \ldots, X_m)$ realized by classifiers in $\mathrm{B}_{\mathcal{P}}(h, r)$. Furthermore, since all classifiers in $\mathrm{B}_{\mathcal{P}}(h, r)$ agree on the classification of any points $X_i \notin \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))$, and $\mathrm{B}_{\mathcal{P}}(h, r) \subseteq \mathbb{C}$, we have that $|G_r|$ is upper bounded by the number of distinct classifications of $\{X_1, \ldots, X_m\} \cap \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))$ realized by classifiers in $\mathbb{C}$.

By a Chernoff bound, on an event E_2 of probability at least $1/2$,

$$|\{X_1, \ldots, X_m\} \cap \mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r))| \leq 1 + 2e\mathcal{P}(\mathrm{DIS}(\mathrm{B}_{\mathcal{P}}(h, r)))m.$$

By the definition of the disagreement coefficient, this is at most $1 + 2e\theta_{h,\mathcal{P}}(r)rm \leq 1 + 2e + 8e\theta_{h,\mathcal{P}}(r)\ln(|G_r|)$, which, if $|G_r| \geq 3$, is at most $11e\theta_{h,\mathcal{P}}(r)\ln(|G_r|)$. By a union bound, the event $E_1 \cap E_2$ has probability strictly greater than 0. Thus, letting $m' = \lceil 11e\theta_{h,\mathcal{P}}(r)\ln(|G_r|)\rceil$, there exists a sequence $x_1, \ldots, x_{m'} \in \mathcal{X}$ such that $|G_r|$ is at most the max of 2 and the number of distinct classifications of $\{x_1, \ldots, x_{m'}\}$ realized by classifiers in $\mathbb{C}$. In the case $|G_r| \geq 3$, this latter value is at most $\left(\frac{em'}{d}\right)^d \leq \left(\frac{22e^2\theta_{h,\mathcal{P}}(r)\ln(|G_r|)}{d}\right)^d$ by the VC-Sauer lemma (Vapnik and Chervonenkis, 1971; Sauer, 1972).

Taking the logarithm, we have that

$$\ln(|G_r|) \leq \max\left\{\ln(2), d\ln\left(22e^2\theta_{h,\mathcal{P}}(r)\right) + d\ln\left(\frac{\ln(|G_r|)}{d}\right)\right\},$$

which implies (see e.g., Vidyasagar, 2003, Corollary 4.1)

$$\ln(|G_r|) < \max\left\{1, 2d\ln\left(22e^2\theta_{h,\mathcal{P}}(r)\right)\right\} = 2d\ln\left(22e^2\theta_{h,\mathcal{P}}(r)\right).$$

Dividing both sides by $\ln(2)$, altogether we have that

$$\begin{aligned} D_{h,\mathcal{P}}(\varepsilon) &= \sup_{r\in[\varepsilon,1]} \log_2\left(\mathcal{N}\left(\frac{r}{2}, \mathrm{B}_{\mathcal{P}}(h,r), \mathcal{P}\right)\right) \leq \sup_{r\in[\varepsilon,1]} \log_2\left(|G_r|\right) \\ &\leq \sup_{r\in[\varepsilon,1]} 2d\log_2\left(22e^2\theta_{h,\mathcal{P}}(r)\right) = 2d\log_2\left(22e^2\theta_{h,\mathcal{P}}(\varepsilon)\right). \end{aligned}$$

In particular, by Theorem 10, this is at most $2d\log_2\left(22e^2\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right)\right)$, so that maximizing the left hand side over the choice of $h \in \mathbb{C}$ and $\mathcal{P}$ completes the proof. ■

Appendix D. Examples Spanning the Gaps

In this section, taking d and $\mathfrak{s}$ as fixed values in $\mathbb{N}$ (with $d \geq 3$ and $\mathfrak{s} \geq 4d$), and taking $\mathcal{X} = \mathbb{N}$, we establish that the upper bounds in Theorems 3, 4, 5, and 7 are all tight (up to universal constant and logarithmic factors) when we take $\mathbb{C} = \{x \mapsto 2\mathbb{1}_S(x) - 1 : S \subseteq \{1, \ldots, \mathfrak{s}\}, |S| \leq d\}$, and that the lower bounds in these theorems are all tight (up to logarithmic factors) when we take $\mathbb{C} = \{x \mapsto 2\mathbb{1}_S(x) - 1 : S \in 2^{\{1,\ldots,d\}} \cup \{\{i\} : d+1 \leq i \leq \mathfrak{s}\}\}$. One can easily verify that, in both cases, the VC dimension is indeed d, and the star number is indeed $\mathfrak{s}$.

D.1 The Upper Bounds are Sometimes Tight

We begin with the upper bounds. In this case, take

$$\mathbb{C} = \{x \mapsto 2\mathbb{1}_S(x) - 1 : S \subseteq \{1, \ldots, \mathfrak{s}\}, |S| \leq d\}. \tag{63}$$

For this hypothesis class, we argue that the lower bounds can be increased to match the upper bounds (up to logarithmic factors). We begin with a general lemma.

For each $i \in \{1, \ldots, d\}$, let $\mathcal{X}_i = \{\lfloor \mathfrak{s}/d\rfloor(i-1)+1, \ldots, \lfloor \mathfrak{s}/d\rfloor i\}$, $\mathbb{C}_i = \{x \mapsto 2\mathbb{1}_{\{t\}}(x) - 1 : t \in \mathcal{X}_i\} \cup \{x \mapsto -1\}$, and let $\mathbb{D}_i$ be a finite nonempty set of probability measures P_i on $\mathcal{X} \times \mathcal{Y}$ such that $P_i(\mathcal{X}_i \times \mathcal{Y}) = 1$ (i.e., with marginal over $\mathcal{X}$ supported only on $\mathcal{X}_i$). Let

$\mathbb{D} = \left\{ \frac{1}{d} \sum_{i=1}^{d} P_i : \forall i \in \{1, \ldots, d\}, P_i \in \mathbb{D}_i \right\}$. Note that for any choices of $P_i \in \mathbb{D}_i$ for each $i \in \{1, \ldots, d\}$, letting $P = \frac{1}{d} \sum_{i=1}^{d} P_i$, we have that $\forall i \in \{1, \ldots, d\}$, $\forall x \in \mathcal{X}_i$ with $P_i(\{x\} \times \mathcal{Y}) > 0$,

$$\begin{aligned} P(\{(x,+1)\} | \{x\} \times \mathcal{Y}) = \frac{P(\{(x,+1)\})}{P(\{x\} \times \mathcal{Y})} &= \frac{\frac{1}{d} \sum_{j=1}^{d} P_j(\{(x,+1)\})}{\frac{1}{d} \sum_{j=1}^{d} P_j(\{x\} \times \mathcal{Y})} \\ &= \frac{P_i(\{(x,+1)\})}{P_i(\{x\} \times \mathcal{Y})} = P_i(\{(x,+1)\} | \{x\} \times \mathcal{Y}), \end{aligned}$$

so that the conditional distribution of Y given $X = x$ (for $(X,Y) \sim P$) is specified by the conditional of Y' given $X' = x$ for $(X', Y') \sim P_i$, for the value i with $x \in \mathcal{X}_i$. Furthermore, since any $x \in \mathcal{X}_i$ has $P(\{x\} \times \mathcal{Y}) = 0$ if and only if $P_i(\{x\} \times \mathcal{Y}) = 0$, without loss we may define $P(\{(x,+1)\} | \{x\} \times \mathcal{Y}) = P_i(\{(x,+1)\} | \{x\} \times \mathcal{Y})$ for any such x. For each $i \in \{1, \ldots, d\}$ and $\varepsilon, \delta \in (0,1)$, let $\Lambda_i(\varepsilon, \delta)$ denote the minimax label complexity under $\mathbb{D}_i$ with respect to $\mathbb{C}_i$ (i.e., the value of $\Lambda_{\mathbb{D}_i}(\varepsilon, \delta)$ when $\mathbb{C} = \mathbb{C}_i$). The value $\Lambda_{\mathbb{D}}(\varepsilon, \delta)$ remains defined as usual (i.e., with respect to the set $\mathbb{C}$ specified in (63)).

Lemma 45 *Fix any* $\gamma \in (2/d, 1)$, $\varepsilon \in (0, \gamma/4)$, *and* $\delta \in \left(0, \frac{\gamma}{4-\gamma}\right)$. *If* $\min_{i \in \{1,\ldots,d\}} \Lambda_i((4/\gamma)\varepsilon, \gamma) \geq 2$, *then*

$$\Lambda_{\mathbb{D}}(\varepsilon, \delta) \geq (\gamma/4) d \min_{i \in \{1,\ldots,d\}} \Lambda_i((4/\gamma)\varepsilon, \gamma).$$

Proof Fix any $n \in \mathbb{N}$ with $n < (\gamma/4) d \min_{i \in \{1,\ldots,d\}} \Lambda_i((4/\gamma)\varepsilon, \gamma)$. Denote $n' = \left\lceil \frac{n}{(\gamma/2)d} \right\rceil$, and note that $n' \leq n$ and $n' < \min_{i \in \{1,\ldots,d\}} \Lambda_i((4/\gamma)\varepsilon, \gamma)$. For each $i \in \{1, \ldots, d\}$, let $P_i \in \mathbb{D}_i$, and denote $g_i^* = \operatorname{argmin}_{g \in \mathbb{C}_i} \operatorname{er}_{P_i}(g)$ (breaking ties arbitrarily). We will later optimize over the choice of these P_i. Also let $g^* = \sum_{i=1}^{d} g_i^* \mathbb{1}_{\mathcal{X}_i}$, the classifier that predicts with g_i^* on each respective $\mathcal{X}_i$ set; note that, since each g_i^* classifies at most one point as $+1$, we have $g^* \in \mathbb{C}$. Denote $P = \frac{1}{d} \sum_{i=1}^{d} P_i$. Let $\hat{h}_P$ denote the (random) classifier produced by $\mathcal{A}(n)$ when $\mathcal{P}_{XY} = P$. Note that if $\sum_{i=1}^{d} \mathbb{1}\left[\operatorname{er}_{P_i}\left(\hat{h}_P\right) - \operatorname{er}_{P_i}(g_i^*) > (4/\gamma)\varepsilon\right] > (\gamma/4)d$, then

$$\begin{aligned} \operatorname{er}_P\left(\hat{h}_P\right) - \inf_{h \in \mathbb{C}} \operatorname{er}_P(h) &= \frac{1}{d} \sum_{i=1}^{d} \operatorname{er}_{P_i}\left(\hat{h}_P\right) - \inf_{h \in \mathbb{C}} \frac{1}{d} \sum_{i=1}^{d} \operatorname{er}_{P_i}(h) \\ &\geq \frac{1}{d} \sum_{i=1}^{d} \operatorname{er}_{P_i}\left(\hat{h}_P\right) - \frac{1}{d} \sum_{i=1}^{d} \operatorname{er}_{P_i}(g^*) = \frac{1}{d} \sum_{i=1}^{d} \left(\operatorname{er}_{P_i}\left(\hat{h}_P\right) - \operatorname{er}_{P_i}(g_i^*)\right) \\ &\geq \frac{1}{d} \sum_{i=1}^{d} \mathbb{1}\left[\operatorname{er}_{P_i}\left(\hat{h}_P\right) - \operatorname{er}_{P_i}(g_i^*) > (4/\gamma)\varepsilon\right] (4/\gamma)\varepsilon > \varepsilon. \end{aligned}$$

Therefore,

$$\mathbb{P}\left(\mathrm{er}_P\left(\hat{h}_P\right) - \inf_{h\in\mathbb{C}} \mathrm{er}_P(h) > \varepsilon\right) \geq \mathbb{P}\left(\sum_{i=1}^{d} \mathbb{1}\left[\mathrm{er}_{P_i}\left(\hat{h}_P\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right] > (\gamma/4)d\right)$$
$$= 1 - \mathbb{P}\left(\sum_{i=1}^{d} \mathbb{1}\left[\mathrm{er}_{P_i}\left(\hat{h}_P\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right] \leq (\gamma/4)d\right)$$
$$= 1 - \mathbb{P}\left(\sum_{i=1}^{d} \left(1 - \mathbb{1}\left[\mathrm{er}_{P_i}\left(\hat{h}_P\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right]\right) \geq (1-\gamma/4)d\right)$$
$$\geq 1 - \frac{1}{(1-\gamma/4)d} \sum_{i=1}^{d} \left(1 - \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}_P\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right)\right)$$
$$= -\frac{\gamma}{4-\gamma} + \frac{4}{4-\gamma}\frac{1}{d}\sum_{i=1}^{d} \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}_P\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right), \tag{64}$$

where the second inequality is due to Markov's inequality and linearity of expectations.

Now note that there is a simple reduction from the problem of learning with $\mathbb{C}_i$ under P_i to the problem of learning with $\mathbb{C}$ under P. Specifically, for a given i.i.d. P_i-distributed sequence $(X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), \ldots$, we can construct an i.i.d. P-distributed sequence $(X_1', Y_1'), (X_2', Y_2'), \ldots$ as follows. For each $j \in \{1, \ldots, d\} \setminus \{i\}$, let $(X_{j1}, Y_{j1}), (X_{j2}, Y_{j2}), \ldots \sim P_j$ be i.i.d., and independent over j, and all independent from the (X_{it}, Y_{it}) sequence. Let $j_1, j_2, \ldots$ be independent $\mathrm{Uniform}(\{1, \ldots, d\})$ random variables (also independent from the above sequences). Then for each $t \in \mathbb{N}$, let $r_t = \sum_{s=1}^{t} \mathbb{1}[j_s = j_t]$, and define $(X_t', Y_t') = (X_{j_t r_t}, Y_{j_t r_t})$. One can easily verify that this these (X_t', Y_t') are independent and P-distributed. Now we can construct an active learning algorithm for the problem of learning with $\mathbb{C}_i$ under P_i, given the budget $n' \leq n$, as follows. We execute the algorithm $\mathcal{A}(n)$. If at any time it requests the label Y_t' of some X_t' in the sequence such that $j_t \neq i$, then we simply use the value $Y_t' = Y_{j_t r_t}$ (which, for the purpose of this reduction, is considered an accessible quantity). Otherwise, if $\mathcal{A}(n)$ requests the label Y_t' of some X_t' in the sequence such that $j_t = i$, then our algorithm will request the label $Y_{i r_t}$ and provide that as the value of Y_t' to be used in the execution of $\mathcal{A}(n)$. If at any time $\mathcal{A}(n)$ has already requested n' labels Y_t' such that $j_t = i$, and attempts to request another label Y_t' with $j_t = i$, our algorithm simply returns an arbitrary classifier, and this is considered a "failure" event. Otherwise, upon termination of $\mathcal{A}(n)$, our algorithm halts and returns the classifier $\mathcal{A}(n)$ produces. Note that this is a valid active learning algorithm for the problem of learning $\mathbb{C}_i$ under P_i with budget n', since the algorithm requests at most n' labels from the P_i-distributed sequence. In particular, in this reduction, we are thinking of the samples (X_t', Y_t') with $j_t \neq i$ as simply part of the internal randomness of the learning algorithm.

Let $\hat{h}_{P,i}'$ denote the classifier returned by the algorithm constructed via this reduction. Furthermore, if we consider also the classifier $\hat{h}_{P,i}$ returned by $\mathcal{A}(n)$ when run (unmodified) on the P-distributed sequence $(X_1', Y_1'), (X_2', Y_2'), \ldots$, and denote by $n_{P,i}'$ the number of labels Y_t' with $j_t = i$ that this unmodified $\mathcal{A}(n)$ requests, then on the event that $n_{P,i}' \leq n'$, we have $\hat{h}_{P,i}' = \hat{h}_{P,i}$. Additionally, let $n_{P,i}$ denote the number of labels Y_t requested by $\mathcal{A}(n)$ with $X_t \in \mathcal{X}_i$ (when $\mathcal{A}(n)$ is run with the sequence $\{(X_t, Y_t)\}_{t=1}^{\infty}$), and note that the sequences $\{(X_t', Y_t')\}_{t=1}^{\infty}$ and $\{(X_t, Y_t)\}_{t=1}^{\infty}$ are distributionally equivalent, so that $(\hat{h}_{P,i}, n_{P,i}')$ and $(\hat{h}_P, n_{P,i})$ are distributionally equivalent as well. Therefore,

$$\mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}_P\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right) \geq \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}_P\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon \text{ and } n_{P,i} \leq n'\right)$$
$$= \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}_{P,i}\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon \text{ and } n'_{P,i} \leq n'\right)$$
$$= \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon \text{ and } n'_{P,i} \leq n'\right)$$
$$= \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right) - \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon \text{ and } n'_{P,i} > n'\right)$$
$$\geq \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right) - \mathbb{P}\left(n'_{P,i} > n'\right)$$
$$= \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right) - \mathbb{P}\left(n_{P,i} > n'\right)$$
$$\geq \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right) - \frac{\mathbb{E}[n_{P,i}]}{n'},$$

where this last inequality is due to Markov's inequality.

Applying this to every $i \in \{1, \ldots, d\}$, this implies

$$\frac{1}{d}\sum_{i=1}^{d} \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}_P\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right)$$
$$\geq -\frac{1}{dn'}\sum_{i=1}^{d} \mathbb{E}[n_{P,i}] + \frac{1}{d}\sum_{i=1}^{d} \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right).$$

By linearity of the expectation, $\frac{1}{dn'}\sum_{i=1}^{d} \mathbb{E}[n_{P,i}] = \frac{1}{dn'}\mathbb{E}\left[\sum_{i=1}^{d} n_{P,i}\right] \leq \frac{n}{dn'} \leq \frac{\gamma}{2}$, so that the above is at least

$$-\frac{\gamma}{2} + \frac{1}{d}\sum_{i=1}^{d} \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \mathrm{er}_{P_i}\left(g_i^*\right) > (4/\gamma)\varepsilon\right).$$

Plugging this into (64), we have that

$$\mathbb{P}\left(\mathrm{er}_P\left(\hat{h}_P\right) - \inf_{h\in\mathbb{C}} \mathrm{er}_P(h) > \varepsilon\right) \geq -\frac{3\gamma}{4-\gamma} + \frac{4}{4-\gamma}\frac{1}{d}\sum_{i=1}^{d} \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \mathrm{er}_{P_i}(g_i^*) > (4/\gamma)\varepsilon\right).$$

The above strategy, producing $\hat{h}'_{P,i}$, is a valid active learning algorithm (with budget n') for any choices of the probability measures P_j, $j \in \{1, \ldots, d\} \setminus \{i\}$. We may therefore consider its behavior if we choose these at random. Specifically, for any probability measure $\Pi^{\setminus i}$ over $\times_{j\neq i}\mathbb{D}_j$, let $\{\tilde{P}_{j,\Pi^{\setminus i}}\}_{j\neq i} \sim \Pi^{\setminus i}$, and for any $P_i \in \mathbb{D}_i$, let $\tilde{P}_{\Pi^{\setminus i},P_i} = \frac{1}{d}P_i + \frac{1}{d}\sum_{j\neq i}\tilde{P}_{j,\Pi^{\setminus i}}$. Then $\hat{h}'_{\tilde{P}_{\Pi^{\setminus i},P_i},i}$ is the output of a valid active learning algorithm (with budget n'); in particular, here we are considering the $\tilde{P}_{j,\Pi^{\setminus i}}$ as internal random variables to the algorithm (along with their corresponding (X_{jt}, Y_{jt}) samples used in the algorithm, which are now considered conditionally independent given $\{\tilde{P}_{j,\Pi^{\setminus i}}\}_{j\neq i}$, where each (X_{jt}, Y_{jt}) has conditional distribution $\tilde{P}_{j,\Pi^{\setminus i}}$): that is, random variables that are independent from the data sequence $(X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), \ldots$. Now note that, since $n' < \Lambda_i((4/\gamma)\varepsilon, \gamma)$,

$$\max_{P_i\in\mathbb{D}_i} \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{\tilde{P}_{\Pi^{\setminus i},P_i},i}\right) - \inf_{g\in\mathbb{C}_i} \mathrm{er}_{P_i}(g) > (4/\gamma)\varepsilon\right) > \gamma. \tag{65}$$

For any given sequence $P_1, \ldots, P_d$, with $P_j \in \mathbb{D}_i$ for each $j \in \{1, \ldots, d\}$, for every $i \in \{1, \ldots, d\}$, denote $\psi_i(P_i, \{P_j\}_{j \neq i}) = \mathbb{P}\left(\mathrm{er}_{P_i}\left(\hat{h}'_{P,i}\right) - \inf_{g \in \mathbb{C}_i} \mathrm{er}_{P_i}(g) > (4/\gamma)\varepsilon\right)$, where $P = \frac{1}{d}\sum_{j=1}^{d} P_j$ as above. Then, by the law of total probability, (65) may be restated as

$$\max_{P_i \in \mathbb{D}_i} \mathbb{E}\left[\psi_i\left(P_i, \left\{\tilde{P}_{j,\Pi^{\setminus i}}\right\}_{j \neq i}\right)\right] > \gamma.$$

Since this holds for every choice of $\Pi^{\setminus i}$, we have that

$$\inf_{\Pi^{\setminus i}} \max_{P_i \in \mathbb{D}_i} \mathbb{E}\left[\psi_i\left(P_i, \left\{\tilde{P}_{j,\Pi^{\setminus i}}\right\}_{j \neq i}\right)\right] \geq \gamma.$$

Since each $\mathbb{D}_j$ is finite, by the minimax theorem (von Neumann, 1928; von Neumann and Morgenstern, 1944), for each $i \in \{1, \ldots, d\}$, there exists a probability measure Π_i over $\mathbb{D}_i$ such that, if $\tilde{P}_i \sim \Pi_i$ (independent from every $\{\tilde{P}_{j,\Pi^{\setminus i}}\}_{j \neq i}$), then

$$\inf_{\Pi^{\setminus i}} \mathbb{E}\left[\psi_i\left(\tilde{P}_i, \left\{\tilde{P}_{j,\Pi^{\setminus i}}\right\}_{j \neq i}\right)\right] = \inf_{\Pi^{\setminus i}} \max_{P_i \in \mathbb{D}_i} \mathbb{E}\left[\psi_i\left(P_i, \left\{\tilde{P}_{j,\Pi^{\setminus i}}\right\}_{j \neq i}\right)\right].$$

In particular, taking these $\{\tilde{P}_i\}_{i=1}^{d}$ to be independent, we have that $\forall i \in \{1, \ldots, d\}$,

$$\mathbb{E}\left[\psi_i\left(\tilde{P}_i, \left\{\tilde{P}_j\right\}_{j \neq i}\right)\right] \geq \inf_{\Pi^{\setminus i}} \mathbb{E}\left[\psi_i\left(\tilde{P}_i, \left\{\tilde{P}_{j,\Pi^{\setminus i}}\right\}_{j \neq i}\right)\right] = \inf_{\Pi^{\setminus i}} \max_{P_i \in \mathbb{D}_i} \mathbb{E}\left[\psi_i\left(P_i, \left\{\tilde{P}_{j,\Pi^{\setminus i}}\right\}_{j \neq i}\right)\right] \geq \gamma.$$

Thus,

$$\sup_{\substack{P_i \in \mathbb{D}_i: \\ i \in \{1,\ldots,d\}}} \sum_{i=1}^{d} \psi_i(P_i, \{P_j\}_{j \neq i}) \geq \mathbb{E}\left[\sum_{i=1}^{d} \psi_i\left(\tilde{P}_i, \left\{\tilde{P}_j\right\}_{j \neq i}\right)\right] = \sum_{i=1}^{d} \mathbb{E}\left[\psi_i\left(\tilde{P}_i, \left\{\tilde{P}_j\right\}_{j \neq i}\right)\right] \geq \gamma d.$$

Altogether, we have that

$$\begin{aligned}
\sup_{\substack{P_i \in \mathbb{D}_i: \\ i \in \{1,\ldots,d\}}} \mathbb{P}\left(\mathrm{er}_P\left(\hat{h}_P\right) - \inf_{h \in \mathbb{C}} \mathrm{er}_P(h) > \varepsilon\right) &\geq -\frac{3\gamma}{4-\gamma} + \frac{4}{4-\gamma}\frac{1}{d} \sup_{\substack{P_i \in \mathbb{D}_i: \\ i \in \{1,\ldots,d\}}} \sum_{i=1}^{d} \psi_i\left(P_i, \{P_j\}_{j \neq i}\right) \\
&\geq -\frac{3\gamma}{4-\gamma} + \frac{4\gamma}{4-\gamma} = \frac{\gamma}{4-\gamma} > \delta.
\end{aligned}$$

Since this holds for any active learning algorithm $\mathcal{A}$ and $n < (\gamma/4) d \min_{i \in \{1,\ldots,d\}} \Lambda_i((4/\gamma)\varepsilon, \gamma)$, the lemma follows. ■

With this lemma in hand, we can now plug in various sets $\mathbb{D}_i$ to obtain lower bounds for learning with this set $\mathbb{C}$ under various noise models. In particular, we can make use of the constructions of lower bounds on $\Lambda_i(\varepsilon, \delta)$ given in the proofs of the theorems in Section 5, noting that the VC dimension of $\mathbb{C}_i$ is 1, and the star number of $\mathbb{C}_i$ is $\lfloor \mathfrak{s}/d \rfloor$. Note that, in the case $d \lesssim 1$, the lower bounds in each of these theorems already match their respective upper bounds up to constant and logarithmic factors (using the lower bound from Theorem 3 as a lower bound on $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \delta)$ for β near 0). We may therefore suppose $d \geq 32$ for the remainder of this subsection.

The realizable case: For the realizable case, for each $i \in \{1, \ldots, d\}$ and $t \in \{1, \ldots, \lfloor \mathfrak{s}/d \rfloor\}$, let $\mathcal{P}_{it}$ be a uniform distribution on $\{\lfloor \mathfrak{s}/d \rfloor(i-1)+1, \ldots, \lfloor \mathfrak{s}/d \rfloor(i-1)+t\} \subseteq \mathcal{X}_i$, and let $\mathbb{D}_i$ denote the set of probability measures P_i in RE having marginal over $\mathcal{X}$ among $\{\mathcal{P}_{it} : 1 \leq t \leq \lfloor \mathfrak{s}/d \rfloor\}$ and having $f^\star_{P_i} \in \mathbb{C}_i$. Noting that the star number of $\mathbb{C}_i$ is $\lfloor \mathfrak{s}/d \rfloor$ and that $\mathcal{X}_i$ is a (maximal) star set for $\mathbb{C}_i$, and recalling that the first term in the "max" in the lower bound of Theorem 3 was proven in Appendix B.1 under the uniform marginal distribution on the first t elements of a maximal star set (for an appropriate value of t, of size at least 1 and at most the star number), we have that for $\varepsilon \in \left(0, \frac{1}{9 \cdot 16}\right)$,

$$\Lambda_i(16\varepsilon, 1/4) \gtrsim \min\left\{\frac{\mathfrak{s}}{d}, \frac{1}{\varepsilon}\right\}.$$

Therefore, Lemma 45 (with $\gamma = 1/4$) implies that for $\mathbb{D} = \left\{\frac{1}{d}\sum_{i=1}^d P_i : \forall i \in \{1, \ldots, d\}, P_i \in \mathbb{D}_i\right\}$, $\forall \delta \in \left(0, \frac{1}{15}\right)$,

$$\Lambda_{\mathbb{D}}(\varepsilon, \delta) \gtrsim \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}.$$

Furthermore, for each choice of $P_1, \ldots, P_d$ (with each $P_i \in \mathbb{D}_i$), by construction, every $i \in \{1, \ldots, d\}$ has at most one $x \in \mathcal{X}_i$ with $P_i(\{(x,+1)\}|\{x\} \times \mathcal{Y}) = 1$, and every other x' in $\mathcal{X}_i$ has $P_i(\{(x',+1)\}|\{x'\} \times \mathcal{Y}) = 0$. Therefore, since $P(\{(x,+1)\}|\{x\} \times \mathcal{Y}) = P_i(\{(x,+1)\}|\{x\} \times \mathcal{Y})$ for every $x \in \mathcal{X}_i$, for $P = \frac{1}{d}\sum_{j=1}^d P_j$, we have that there are at most d points x in $\bigcup_{i=1}^d \mathcal{X}_i$ with $P(\{(x,+1)\}|\{x\} \times \mathcal{Y}) = 1$, and all other points x in $\bigcup_{i=1}^d \mathcal{X}_i$ have $P(\{(x,+1)\}|\{x\} \times \mathcal{Y}) = 0$. In particular, this implies that for $(X,Y) \sim P$, $\mathbb{P}(f^\star_P(X) \neq Y | X \in \bigcup_{i=1}^d \mathcal{X}_i) = 0$. Since we also have that $\forall t \in \mathbb{N} \setminus \bigcup_{i=1}^d \mathcal{X}_i$, $P(\{t\} \times \mathcal{Y}) = 0$, we can take $f^\star_P(x) = -1$ for every $x \in \mathcal{X} \setminus \bigcup_{i=1}^d \mathcal{X}_i$ while guaranteeing $\mathrm{er}_P(f^\star_P) = 0$. Since $\bigcup_{i=1}^d \mathcal{X}_i \subseteq \{1, \ldots, \mathfrak{s}\}$, we also have that $f^\star_P \in \mathbb{C}$. Together, these facts imply $P \in \mathrm{RE}$. Thus, $\mathbb{D} \subseteq \mathrm{RE}$, which implies $\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \geq \Lambda_{\mathbb{D}}(\varepsilon, \delta)$, so that

$$\Lambda_{\mathrm{RE}}(\varepsilon, \delta) \gtrsim \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}$$

as well. Since the upper bound in Theorem 3 is within a factor proportional to $\mathrm{Log}(1/\varepsilon)$ of this,[16] this establishes that the upper bound is sometimes tight to within a factor proportional to $\mathrm{Log}(1/\varepsilon)$.

Bounded noise: In the case of bounded noise, fix any $\beta \in (0, 1/2)$ and $\varepsilon \in (0, (1-2\beta)/(256e))$. Take $\zeta = \frac{32e\varepsilon}{1-2\beta}$ and $k = \min\{\lfloor \mathfrak{s}/d \rfloor - 1, \lfloor 1/\zeta \rfloor\}$, and for each $i \in \{1, \ldots, d\}$, let $\mathbb{D}_i$ be defined as the set $\mathrm{RR}(k, \zeta, \beta)$ in Lemma 26, as applied to the hypothesis class $\mathbb{C}_i$ with $\{x_1, \ldots, x_{k+1}\} = \{\lfloor \mathfrak{s}/d \rfloor(i-1)+1, \ldots, \lfloor \mathfrak{s}/d \rfloor(i-1)+k+1\}$, $h_0 = -1$, and $h_j = 2\mathbb{1}_{\{\lfloor \mathfrak{s}/d \rfloor(i-1)+j\}} - 1$ for each $j \in \{1, \ldots, k\}$. Then Lemma 26 implies

$$\Lambda_i(16e\varepsilon, 1/(4e)) \geq \frac{\beta(k-1)}{3(1-2\beta)^2} \gtrsim \frac{\beta}{(1-2\beta)^2} \min\left\{\frac{\mathfrak{s}}{d}, \frac{1-2\beta}{\varepsilon}\right\}.$$

Furthermore, recall from the definition of $\mathrm{RR}(k, \zeta, \beta)$ in Section A.2 that $\mathbb{D}_i$ is a finite set of probability measures, and every $P_i \in \mathbb{D}_i$ has $P_i((\mathcal{X} \setminus \{x_1, \ldots, x_{k+1}\}) \times \mathcal{Y}) = 0$. In particular, note that $\{x_1, \ldots, x_{k+1}\} \subseteq \mathcal{X}_i$ in this case. Furthermore, every $P_i \in \mathbb{D}_i$ has $\forall x \in \{x_1, \ldots, x_k\}$,

16. Note that, although $\frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})}$ can sometimes be much smaller than $\mathfrak{s} \wedge \frac{d}{\varepsilon}$, we always have $\mathfrak{s} \wedge \frac{d}{\varepsilon} \lesssim \frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$, so that this $\mathfrak{s} \wedge \frac{d}{\varepsilon}$ lower bound does not contradict the $\frac{\mathfrak{s}d}{\mathrm{Log}(\mathfrak{s})}\mathrm{Log}\left(\frac{1}{\varepsilon}\right)$ upper bound.

$P_i(\{(x,+1)\}|\{x\}\times\mathcal{Y}) \in \{\beta, 1-\beta\}$, and at most one $x \in \{x_1,\ldots,x_k\}$ has $P_i(\{(x,+1)\}|\{x\}\times\mathcal{Y}) = 1-\beta$, while $P_i(\{(x_{k+1},+1)\}|\{x_{k+1}\}\times\mathcal{Y}) = 0$. Thus, for any choices of $P_i \in \mathbb{D}_i$ for each $i \in \{1,\ldots,d\}$, the probability measure $P = \frac{1}{d}\sum_{i=1}^{d} P_i$ satisfies the property that, $\forall x \in \mathcal{X}$ with $P(\{x\}\times\mathcal{Y}) > 0$, $P(\{(x,+1)\}|\{x\}\times\mathcal{Y}) \in \{0,\beta,1-\beta\}$, and there are at most d values $x \in \mathcal{X}$ with $P(\{x\}\times\mathcal{Y}) > 0$ and $P(\{(x,+1)\}|\{x\}\times\mathcal{Y}) = 1-\beta$. In particular, this implies that without loss, we can take $f_P^\star \in \mathbb{C}$, and furthermore that $P \in \mathrm{BN}(\beta)$. Thus, for the set $\mathbb{D} = \left\{\frac{1}{d}\sum_{i=1}^{d} P_i : \forall i \in \{1,\ldots,d\}, P_i \in \mathbb{D}_i\right\}$, we have $\mathbb{D} \subseteq \mathrm{BN}(\beta)$. Lemma 45 (with $\gamma = 1/(4e)$) then implies that $\forall \delta \in \left(0, \frac{1}{16e-1}\right)$,

$$\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \geq \Lambda_{\mathbb{D}}(\varepsilon,\delta) \gtrsim d \min_{i\in\{1,\ldots,d\}} \Lambda_i(16e\varepsilon, 1/(4e)) \gtrsim \frac{\beta}{(1-2\beta)^2}\min\left\{\mathfrak{s}, \frac{(1-2\beta)d}{\varepsilon}\right\}.$$

For β bounded away from 0, the upper bound in Theorem 4 is within a $\mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right)$ factor of this, so that this establishes that the upper bound is sometimes tight to within logarithmic factors when β is bounded away from 0. Furthermore, since $\mathrm{RE} \subseteq \mathrm{BN}(\beta)$, the above result for sometimes-tightness of the upper bound in the realizable case implies that the upper bound in Theorem 4 is also sometimes tight to within logarithmic factors for any β near 0.

Tsybakov noise: For the case of Tsybakov noise, the tightness (up to logarithmic factors) of the upper bound for $\alpha \leq 1/2$ is already established by the lower bound for that case in Theorem 5. Thus, it remains only to consider $\alpha \in (1/2, 1)$. Fix any values $a \in [4,\infty)$, $\alpha \in (1/2,1)$, and $\varepsilon \in \left(0, 1/(2^{11}a^{1/\alpha})\right)$, let a' be as in the definition of $\mathrm{TN}(a,\alpha)$, and let

$$k = \min\left\{\left\lfloor\frac{\mathfrak{s}}{d}\right\rfloor - 1, \left\lfloor\frac{(a')^{\frac{\alpha-1}{\alpha}}}{64\varepsilon}\right\rfloor, \left\lfloor\frac{a'}{64\varepsilon}4^{-\frac{1}{1-\alpha}}\right\rfloor\right\},$$

$\beta = \frac{1}{2} - \left(\frac{k64\varepsilon}{a'}\right)^{1-\alpha}$, and $\zeta = \frac{128\varepsilon}{1-2\beta}$. Note that $\zeta \in (0,1)$, $\beta \in [1/4, 1/2)$, and $2 \leq k \leq \min\{\lfloor\mathfrak{s}/d\rfloor - 1, \lfloor 1/\zeta\rfloor\}$ (following the arguments from the proof of Theorem 5, with ε replaced by 64ε). Furthermore, $\forall i \in \{1,\ldots,d\}$, let $\mathbb{D}_i$ be the set $\mathrm{RR}(k,\zeta,\beta)$ in Lemma 26, as applied to the class $\mathbb{C}_i$, with $\{x_1,\ldots,x_{k+1}\} = \{\lfloor\mathfrak{s}/d\rfloor(i-1)+1,\ldots,\lfloor\mathfrak{s}/d\rfloor(i-1)+k+1\}$, $h_0 = -1$, and $h_j = 2\mathbb{1}_{\{\lfloor\mathfrak{s}/d\rfloor(i-1)+j\}} - 1$ for each $j \in \{1,\ldots,k\}$. Thus, by Lemma 26,

$$\begin{aligned}\Lambda_i(64\varepsilon, 1/16) &\geq \frac{\beta(k-1)\ln(4)}{3(1-2\beta)^2} \gtrsim \left(\frac{\varepsilon}{a'}\right)^{2\alpha-2} k^{2\alpha-1} \\ &\gtrsim a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\min\left\{\frac{\mathfrak{s}}{d}, \frac{(a')^{\frac{\alpha-1}{\alpha}}}{\varepsilon}, \frac{a'}{\varepsilon}4^{-\frac{1}{1-\alpha}}\right\}^{2\alpha-1} \gtrsim a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha}\min\left\{\frac{\mathfrak{s}}{d}, \frac{1}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1},\end{aligned}$$

where this last inequality relies on the fact (established in the proof of Theorem 5) that $(a')^{\frac{\alpha-1}{\alpha}} \leq a'4^{-\frac{1}{1-\alpha}}$.

We note that any $P_i \in \mathbb{D}_i$ has $P_i((\mathcal{X}\setminus\{\lfloor\mathfrak{s}/d\rfloor(i-1)+1,\ldots,\lfloor\mathfrak{s}/d\rfloor(i-1)+k+1\})\times\mathcal{Y}) = 0$. Without loss of generality, suppose each $P_i \in \mathbb{D}_i$ has $\eta(x;P_i) = 0$ for every $x \in \mathcal{X}\setminus\{\lfloor\mathfrak{s}/d\rfloor(i-1)+1,\ldots,\lfloor\mathfrak{s}/d\rfloor(i-1)+k+1\}$. As in the proof of the lower bound in Theorem 5, we note that any $P_i \in \mathbb{D}_i$ has $P_i((x,y) : |\eta(x;P_i) - 1/2| \leq t) \leq a't^{\alpha/(1-\alpha)}$ for every $t > 0$, and furthermore that $f_{P_i}^\star(\cdot) = \mathrm{sign}(2\eta(\cdot;P_i) - 1)$, which has at most one x with $f_{P_i}^\star(x_i) = +1$

(by definition of $\mathrm{RR}(k,\zeta,\beta)$ in Section A.2). This further implies that, for any choices of $P_i \in \mathbb{D}_i$ for each $i \in \{1,\ldots,d\}$, the probability measure $P = \frac{1}{d}\sum_{i=1}^{d} P_i$ has support for its marginal over $\mathcal{X}$ only in $\bigcup_{i=1}^{d}\{\lfloor\mathfrak{s}/d\rfloor(i-1)+1,\ldots,\lfloor\mathfrak{s}/d\rfloor(i-1)+k+1\}$, and for each $i \in \{1,\ldots,d\}$, $\forall x \in \{\lfloor\mathfrak{s}/d\rfloor(i-1)+1,\ldots,\lfloor\mathfrak{s}/d\rfloor(i-1)+k+1\}$, $\eta(x;P) = \eta(x;P_i)$, while we may take $\eta(x;P) = 0$ for every $x \notin \bigcup_{i=1}^{d}\{\lfloor\mathfrak{s}/d\rfloor(i-1)+1,\ldots,\lfloor\mathfrak{s}/d\rfloor(i-1)+k+1\}$. Therefore, $f_P^\star$ has at most d points $x \in \bigcup_{i=1}^{d}\mathcal{X}_i$ with $f_P^\star(x) = +1$, and $f_P^\star(x) = -1$ for all other $x \in \mathcal{X}$: that is, $f_P^\star \in \mathbb{C}$. Additionally, since the supports of the marginals of the P_i distributions over $\mathcal{X}$ are disjoint, we have that $\forall t > 0$,

$$
\begin{aligned}
P\left((x,y) : |\eta(x;P) - 1/2| \le t\right) &= \frac{1}{d}\sum_{i=1}^{d} P_i\left((x,y) : |\eta(x;P) - 1/2| \le t\right) \\
&= \frac{1}{d}\sum_{i=1}^{d} P_i\left((x,y) : |\eta(x;P_i) - 1/2| \le t\right) \le \frac{1}{d}\sum_{i=1}^{d} a' t^{\alpha/(1-\alpha)} = a' t^{\alpha/(1-\alpha)}.
\end{aligned}
$$

Thus, the set $\mathbb{D} = \left\{\frac{1}{d}\sum_{i=1}^{d} P_i : \forall i \in \{1,\ldots,d\}, P_i \in \mathbb{D}_i\right\}$ satisfies $\mathbb{D} \subseteq \mathrm{TN}(a,\alpha)$. Combined with the fact that each set $\mathbb{D}_i$ is finite (by the definition of $\mathrm{RR}(k,\zeta,\beta)$ in Section A.2), Lemma 45 (with $\gamma = 1/16$) implies that $\forall \delta \in \left(0, \frac{1}{63}\right)$,

$$
\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta) \ge \Lambda_{\mathbb{D}}(\varepsilon,\delta) \gtrsim d \min_{i\in\{1,\ldots,d\}} \Lambda_i(64\varepsilon, 1/16) \gtrsim a^2\left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \min\left\{\frac{\mathfrak{s}}{d}, \frac{1}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1} d.
$$

Since this is within logarithmic factors of the upper bound of Theorem 5, this establishes that the upper bound is sometimes tight to within logarithmic factors (for sufficiently small values of ε).

Benign noise: We can establish that the upper bound in Theorem 7 is sometimes tight by reduction from the above problems. Specifically, since $\mathrm{RE} \subseteq \mathrm{BE}(\nu)$ for every $\nu \in [0,1/2)$, for the above choice of $\mathbb{C}$ we have that $\forall \nu \in [0,1/2]$, $\forall \varepsilon \in \left(0, \frac{1}{9\cdot 16}\right)$, $\forall \delta \in \left(0, \frac{1}{15}\right)$,

$$
\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \ge \Lambda_{\mathrm{RE}}(\varepsilon,\delta) \gtrsim \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}.
$$

Furthermore, the lower bound in Theorem 7 already implies that $\forall \varepsilon \in \left(0, \frac{1-2\nu}{24}\right)$, $\forall \delta \in \left(0, \frac{1}{24}\right]$,

$$
\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \gtrsim \frac{\nu^2}{\varepsilon^2} d.
$$

Together, we have that $\forall \nu \in [0,1/2)$, $\forall \varepsilon \in \left(0, \frac{1-2\nu}{9\cdot 16}\right)$, $\forall \delta \in \left(0, \frac{1}{24}\right]$,

$$
\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta) \gtrsim \max\left\{\frac{\nu^2}{\varepsilon^2} d, \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}\right\} \gtrsim \frac{\nu^2}{\varepsilon^2} d + \min\left\{\mathfrak{s}, \frac{d}{\varepsilon}\right\}.
$$

Thus, the upper bound in Theorem 7 is sometimes tight to within logarithmic factors.

D.2 The Lower Bounds are Sometimes Tight

We now argue that the lower bounds in Theorems 3, 4, 5, and 7 are sometimes tight (up to logarithmic factors). First we have a general lemma. Let $\mathcal{X}_1 \subset \mathcal{X}$ and $\mathcal{X}_2 = \mathcal{X} \setminus \mathcal{X}_1$, and let

$\mathbb{C}_1, \mathbb{C}_2$ be hypothesis classes such that $\forall i \in \{1,2\}$, $\forall h \in \mathbb{C}_i$, $\forall x \in \mathcal{X} \setminus \mathcal{X}_i$, $h(x) = -1$. Further suppose that $\forall i \in \{1,2\}$, the all-negative classifier $x \mapsto h_-(x) = -1$ is in $\mathbb{C}_i$. For each $i \in \{1,2\}$ and $\gamma \in [0,1]$, let $\mathbb{D}_i(\gamma)$ be a nonempty set of probability measures on $\mathcal{X} \times \mathcal{Y}$ such that $\forall P_i \in \mathbb{D}_i(\gamma)$, $P_i(\mathcal{X}_i \times \mathcal{Y}) = 1$; further suppose $\forall \gamma, \gamma' \in [0,1]$ with $\gamma \leq \gamma'$, $\mathbb{D}_i(\gamma) \supseteq \mathbb{D}_i(\gamma')$. Also, for each $i \in \{1,2\}$, $\gamma, \delta \in [0,1]$, and $\varepsilon > 0$, let $\Lambda_{i,\gamma}(\varepsilon, \delta)$ denote the minimax label complexity under $\mathbb{D}_i(\gamma)$ with respect to $\mathbb{C}_i$ (i.e., the value of $\Lambda_{\mathbb{D}_i(\gamma)}(\varepsilon,\delta)$ when $\mathbb{C} = \mathbb{C}_i$). Let $\mathbb{D} = \{\gamma P_1 + (1-\gamma)P_2 : P_1 \in \mathbb{D}_1(\gamma), P_2 \in \mathbb{D}_2(1-\gamma), \gamma \in [0,1]\}$.

Lemma 46 *For $\mathbb{C} = \mathbb{C}_1 \cup \mathbb{C}_2$, $\forall \varepsilon, \delta \in (0,1)$,*

$$\Lambda_{\mathbb{D}}(\varepsilon,\delta) \leq 2 \sup_{\gamma \in [0,1]} \max\left\{ \Lambda_{1,(\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/8)}, \frac{\delta}{3}\right), \Lambda_{2,(1-\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(1-\gamma+\varepsilon/8)}, \frac{\delta}{3}\right)\right\}.$$

Proof For each $i \in \{1,2\}$ and $\gamma \in [0,1]$, let $\mathcal{A}_{\gamma,i}$ be an active learning algorithm such that, for any integer $n \geq \Lambda_{i,\gamma}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/8)}, \frac{\delta}{3}\right)$, if $\mathcal{P}_{XY} \in \mathbb{D}_i(\gamma)$, then with probability at least $1-\delta/3$, the classifier $\hat{h}$ produced by $\mathcal{A}_{\gamma,i}(n)$ satisfies $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) - \inf_{h \in \mathbb{C}_i} \mathrm{er}_{\mathcal{P}_{XY}}(h) \leq \frac{\varepsilon}{2(\gamma+\varepsilon/8)}$; such an algorithm is guaranteed to exist by the definition of $\Lambda_{i,\gamma}(\cdot,\cdot)$.

Now suppose $\mathcal{P}_{XY} \in \mathbb{D}$, so that $\mathcal{P}_{XY} = \gamma P_1 + (1-\gamma)P_2$ for some $\gamma \in [0,1]$, $P_1 \in \mathbb{D}_1(\gamma)$, and $P_2 \in \mathbb{D}_2(1-\gamma)$. Let $(X_1,Y_1),(X_2,Y_2),\ldots$ be the data sequence, as usual (i.i.d. $\mathcal{P}_{XY}$). Consider an active learning algorithm $\mathcal{A}$ defined as follows. We first split the sequence of indices into three subsequences: $i_{0,k} = 2k-1$ for $k \in \mathbb{N}$, $i_{1,1}, i_{1,2}, \ldots$ is the increasing subsequence of indices i such that $i/2 \in \mathbb{N}$ and $X_i \in \mathcal{X}_1$, and $i_{2,1}, i_{2,2}, \ldots$ is the remaining increasing subsequence (i.e., indices i such that $i/2 \in \mathbb{N}$ and $X_i \in \mathcal{X}_2$). Given a budget $n \in \mathbb{N}$, $\mathcal{A}(n)$ proceeds as follows. First, we let $m = \left\lceil \frac{128}{\varepsilon^2} \ln\left(\frac{12}{\delta}\right)\right\rceil$, $\gamma_1 = \max\left\{\frac{1}{m}\sum_{k=1}^m \mathbb{1}_{\mathcal{X}_1}(X_{i_{0,k}}) - \frac{\varepsilon}{16}, 0\right\}$, and $\gamma_2 = \max\left\{\frac{1}{m}\sum_{k=1}^m \mathbb{1}_{\mathcal{X}_2}(X_{i_{0,k}}) - \frac{\varepsilon}{16}, 0\right\}$. By Hoeffding's inequality and a union bound, with probability at least $1-\delta/3$, $\forall i \in \{1,2\}$,

$$\mathcal{P}_{XY}(\mathcal{X}_i \times \mathcal{Y}) - \frac{\varepsilon}{8} \leq \gamma_i \leq \mathcal{P}_{XY}(\mathcal{X}_i \times \mathcal{Y}). \tag{66}$$

Denote by H this event.

Next, for each $j \in \{1,2\}$, if the subsequence $i_{j,1}, i_{j,2}, \ldots$ is infinite, then we run $\mathcal{A}_{\gamma_j,j}(\lfloor n/2 \rfloor)$ with the data subsequence $\{X_k^{(j)}\}_{k=1}^\infty = \{X_{i_{j,k}}\}_{k=1}^\infty$; if the algorithm $\mathcal{A}_{\gamma_j,j}$ requests the label for an index k (i.e., corresponding to $X_k^{(j)}$), then $\mathcal{A}(n)$ requests the corresponding label $Y_{i_{j,k}}$ and provides this value to $\mathcal{A}_{\gamma_j,j}$ as the label of $X_k^{(j)}$. Let $\hat{h}_j$ denote the classifier returned by this execution of $\mathcal{A}_{\gamma_j,j}(\lfloor n/2 \rfloor)$. On the other hand, if the subsequence $i_{j,1}, i_{j,2}, \ldots$ is finite (or empty), then we let $\hat{h}_j$ denote an arbitrary classifier. Finally, let $\mathcal{A}(n)$ return the classifier $\hat{h} = \hat{h}_1 \mathbb{1}_{\mathcal{X}_1} + \hat{h}_2 \mathbb{1}_{\mathcal{X}_2}$. In particular, note that this method requests at most n labels, since all labels are requested by one of the $\mathcal{A}_{\gamma_j,j}$ algorithms, each of which requests at most $\lfloor n/2 \rfloor$ labels.

For this method, we have that

$$\begin{aligned}\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) - \inf_{h \in \mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) &= \gamma \mathrm{er}_{P_1}(\hat{h}_1) + (1-\gamma)\mathrm{er}_{P_2}(\hat{h}_2) - \inf_{h \in \mathbb{C}} \left(\gamma \mathrm{er}_{P_1}(h) + (1-\gamma)\mathrm{er}_{P_2}(h)\right) \\ &\leq \gamma\left(\mathrm{er}_{P_1}(\hat{h}_1) - \inf_{h \in \mathbb{C}} \mathrm{er}_{P_1}(h)\right) + (1-\gamma)\left(\mathrm{er}_{P_2}(\hat{h}_2) - \inf_{h \in \mathbb{C}} \mathrm{er}_{P_2}(h)\right).\end{aligned}$$

For each $j \in \{1,2\}$, since every $h \in \mathbb{C} \setminus \mathbb{C}_j$ has $h(x) = h_-(x)$ for every $x \in \mathcal{X}_j$, and $h_- \in \mathbb{C}_j$, we have that $\inf_{h \in \mathbb{C}} \mathrm{er}_{P_j}(h) = \inf_{h \in \mathbb{C}_j} \mathrm{er}_{P_j}(h)$. Thus, the above implies

$$\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) - \inf_{h \in \mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) \leq \gamma \left(\mathrm{er}_{P_1}(\hat{h}_1) - \inf_{h \in \mathbb{C}_1} \mathrm{er}_{P_1}(h) \right) + (1-\gamma) \left(\mathrm{er}_{P_2}(\hat{h}_2) - \inf_{h \in \mathbb{C}_2} \mathrm{er}_{P_2}(h) \right). \tag{67}$$

If $\gamma = 0$, then with probability one, every $X_i \in \mathcal{X}_2$, and $\{(X_{i_{2,k}}, Y_{i_{2,k}})\}_{k=1}^{\infty}$ is an infinite i.i.d. P_2-distributed sequence. Furthermore, $1 - \varepsilon/8 < \gamma_2 = 1 - \varepsilon/16 < 1$, so that $\mathcal{P}_{XY} \in \mathbb{D}_2(\gamma_2)$. Thus, if $n \geq 2\Lambda_{2,1-\varepsilon/8}\left(\frac{\varepsilon}{2(1+\varepsilon/8)}, \frac{\delta}{3}\right)$, then we also have $n \geq \Lambda_{2,\gamma_2}\left(\frac{\varepsilon}{2(\gamma_2+\varepsilon/8)}, \frac{\delta}{3}\right)$ (by monotonicity of $\mathbb{D}_2(\cdot)$ and the label complexity), so that with probability at least $1 - \delta/3$, $\mathrm{er}_{P_2}(\hat{h}_2) - \inf_{h \in \mathbb{C}_2} \mathrm{er}_{P_2}(h) \leq \frac{\varepsilon}{2(\gamma_2+\varepsilon/8)} = \frac{\varepsilon}{2(1+\varepsilon/16)} < \frac{\varepsilon}{2}$ (here we are evaluating the label complexity guarantee of $\mathcal{A}_{\gamma_2,2}$ under the conditional distribution given γ_2, and then invoking the law of total probability and intersecting with the above probability-one event). Combined with (67), this implies $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) - \inf_{h \in \mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) < \frac{\varepsilon}{2}$. If $\gamma = 1$, then a symmetric argument implies that if $n \geq 2\Lambda_{1,1-\varepsilon/8}\left(\frac{\varepsilon}{2(1+\varepsilon/8)}, \frac{\delta}{3}\right)$, then with probability at least $1 - \delta/3$, $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) - \inf_{h \in \mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) < \frac{\varepsilon}{2}$.

Otherwise, suppose $0 < \gamma < 1$. Note that, on the event H, $\gamma - \varepsilon/8 \leq \gamma_1 \leq \gamma$ and $1 - \gamma - \varepsilon/8 \leq \gamma_2 \leq 1 - \gamma$, so that $\mathbb{D}_1(\gamma_1) \subseteq \mathbb{D}_1((\gamma - \varepsilon/8) \vee 0)$ and $\mathbb{D}_2(\gamma_2) \subseteq \mathbb{D}_2((1 - \gamma - \varepsilon/8) \vee 0)$, and hence that

$$\Lambda_{1,\gamma_1}\left(\frac{\varepsilon}{2(\gamma_1 + \varepsilon/8)}, \frac{\delta}{3}\right) \leq \Lambda_{1,(\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(\gamma + \varepsilon/8)}, \frac{\delta}{3}\right)$$

and

$$\Lambda_{2,\gamma_2}\left(\frac{\varepsilon}{2(\gamma_2 + \varepsilon/8)}, \frac{\delta}{3}\right) \leq \Lambda_{2,(1-\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(1 - \gamma + \varepsilon/8)}, \frac{\delta}{3}\right).$$

In this case, by the strong law of large numbers, with probability one, $\forall j \in \{1,2\}$, the sequence $i_{j,1}, i_{j,2}, \ldots$ exists and is infinite. Since the support of the marginal of P_j over $\mathcal{X}$ is contained within $\mathcal{X}_j$, and $\mathcal{X}_1$ and $\mathcal{X}_2$ are disjoint, we may note that $(X_{i_{j,1}}, Y_{i_{j,1}}), (X_{i_{j,2}}, Y_{i_{j,2}}), \ldots$ are independent P_j-distributed random variables. In particular, if

$$n \geq 2\max\left\{\Lambda_{1,(\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(\gamma + \varepsilon/8)}, \frac{\delta}{3}\right), \Lambda_{2,(1-\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(1 - \gamma + \varepsilon/8)}, \frac{\delta}{3}\right)\right\},$$

then (by the label complexity guarantee of $\mathcal{A}_{\gamma_j,j}$ applied under the conditional distribution given γ_j, combined with the law of total probability, and intersecting with the above probability-one event) there are events H_1 and H_2, each of probability at least $1 - \delta/3$, such that on the event $H \cap H_1$, $\mathrm{er}_{P_1}(\hat{h}_1) - \inf_{h \in \mathbb{C}_1} \mathrm{er}_{P_1}(h) \leq \frac{\varepsilon}{2(\gamma_1+\varepsilon/8)} \leq \frac{\varepsilon}{2\gamma}$, and on the event $H \cap H_2$, $\mathrm{er}_{P_2}(\hat{h}_2) - \inf_{h \in \mathbb{C}_2} \mathrm{er}_{P_2}(h) \leq \frac{\varepsilon}{2(\gamma_2+\varepsilon/8)} \leq \frac{\varepsilon}{2(1-\gamma)}$. Therefore, on the event $H \cap H_1 \cap H_2$, the right hand side of (67) is at most $\gamma\frac{\varepsilon}{2\gamma} + (1-\gamma)\frac{\varepsilon}{2(1-\gamma)} = \varepsilon$, so that $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) - \inf_{h \in \mathbb{C}} \mathrm{er}_{\mathcal{P}_{XY}}(h) \leq \varepsilon$. By a union bound, the probability of $H \cap H_1 \cap H_2$ is at least $1 - \delta$. Since this holds for any $\mathcal{P}_{XY} \in \mathbb{D}$, the result follows. ■

We can now apply this result with various choices of the sets $\mathbb{D}_1(\gamma)$ and $\mathbb{D}_2(\gamma)$ to obtain upper bounds for the above space $\mathbb{C}$, matching the lower bounds proven above for various noise models. Specifically, consider $\mathcal{X} = \mathbb{N}$, $\mathcal{X}_1 = \{1, \ldots, d\}$, $\mathcal{X}_2 = \{d+1, d+2, \ldots\}$, $\mathbb{C}_1 = \{x \mapsto 2\mathbb{1}_S(x) - 1 : S \subseteq \{1, \ldots, d\}\}$, and $\mathbb{C}_2 = \left\{x \mapsto 2\mathbb{1}_{\{t\}}(x) - 1 : t \in \{d+1, d+2, \ldots, \mathfrak{s}\}\right\} \cup$

$\{x \mapsto -1\}$. Note that $\mathbb{C}_1$ and $\mathbb{C}_2$ satisfy the requirements specified above, and also that the VC dimension of $\mathbb{C}_1$ is d and the star number of $\mathbb{C}_1$ is d, while the VC dimension of $\mathbb{C}_2$ is 1 and the star number of $\mathbb{C}_2$ is $\mathfrak{s} - d$. Furthermore, take $\mathbb{C} = \{x \mapsto 2\mathbb{1}_S(x) - 1 : S \in 2^{\{1,\ldots,d\}} \cup \{\{i\} : d+1 \leq i \leq \mathfrak{s}\}\}$, and note that this satisfies $\mathbb{C} = \mathbb{C}_1 \cup \mathbb{C}_2$, and $\mathbb{C}$ has VC dimension d and star number $\mathfrak{s}$.

The realizable case: For the realizable case, we can in fact show that that lower bound in Theorem 3 is sometimes tight up to *universal constant* factors. Specifically, let $\mathbb{D}_i$ denote the set of all $P_i \in \mathrm{RE}$ with $P_i(\mathcal{X}_i \times \mathcal{Y}) = 1$, for each $i \in \{1,2\}$. For every $\gamma \in [0,1]$ and $i \in \{1,2\}$, define $\mathbb{D}_i(\gamma) = \mathbb{D}_i$. In particular, note that for any $P \in \mathrm{RE}$, for any measurable $A \subseteq \mathcal{X} \times \mathcal{Y}$, $P(A) = P(\mathcal{X}_1 \times \mathcal{Y})P(A|\mathcal{X}_1 \times \mathcal{Y}) + P(\mathcal{X}_2 \times \mathcal{Y})P(A|\mathcal{X}_2 \times \mathcal{Y})$. Furthermore, note that any $i \in \{1,2\}$ with $P(\mathcal{X}_i \times \mathcal{Y}) > 0$ has $P(\cdot \times \mathcal{Y}|\mathcal{X}_i \times \mathcal{Y})$ supported only in $\mathcal{X}_i$, and has $P(\cdot|\mathcal{X}_i \times \mathcal{Y}) \in \mathrm{RE}$, so that $P(\cdot|\mathcal{X}_i \times \mathcal{Y}) \in \mathbb{D}_i$. Thus, $P \in \mathbb{D} = \{\gamma P_1 + (1-\gamma)P_2 : P_1 \in \mathbb{D}_1, P_2 \in \mathbb{D}_2, \gamma \in [0,1]\}$. Therefore, $\mathrm{RE} \subseteq \mathbb{D}$. Together with Lemma 46, this implies $\forall \varepsilon, \delta \in (0,1)$,

$$\begin{aligned}\Lambda_{\mathrm{RE}}(\varepsilon,\delta) \leq \Lambda_{\mathbb{D}}(\varepsilon,\delta) &\leq 2\max\left\{\Lambda_{1,0}\left(\frac{\varepsilon}{2(1+\varepsilon/8)}, \frac{\delta}{3}\right), \Lambda_{2,0}\left(\frac{\varepsilon}{2(1+\varepsilon/8)}, \frac{\delta}{2}\right)\right\} \\ &\leq 2\max\left\{\Lambda_{1,0}\left(\frac{\varepsilon}{3}, \frac{\delta}{3}\right), \Lambda_{2,0}\left(\frac{\varepsilon}{3}, \frac{\delta}{2}\right)\right\},\end{aligned}$$

for $\Lambda_{i,0}(\cdot,\cdot)$ defined as above.

Now note that, since every $P_1 \in \mathbb{D}_1$ has $P_1(\cdot \times \mathcal{Y})$ supported only in $\mathcal{X}_1$, and $P_1 \in \mathrm{RE}$, and since $\mathbb{C}_1$ contains classifiers realizing all 2^d distinct classifications of $\mathcal{X}_1$, $\exists h_{P_1} \in \mathbb{C}_1$ with $\mathrm{er}_{P_1}(h_{P_1}) = 0$; thus, without loss, we can take $f^\star_{P_1} = h_{P_1}$, so that P_1 is in the realizable case with respect to $\mathbb{C}_1$. In particular, since there are only d points in $\mathcal{X}_1$, if we consider the active learning algorithm that (given a budget $n \geq d$) simply requests Y_i for exactly one i s.t. $X_i = x$, for each $x \in \mathcal{X}_1$ for which $\exists X_i = x$, and then returns any classifier $\hat{h}$ consistent with these labels, if $\mathcal{P}_{XY} \in \mathbb{D}_1$, with probability one every $x \in \mathcal{X}_1$ with $\mathcal{P}_{XY}(\{x\} \times \mathcal{Y}) > 0$ has some $X_i = x$, so that $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) = 0$. Noting that this algorithm requests at most d labels, we have that $\forall \varepsilon, \delta \in (0,1)$,

$$\Lambda_{1,0}\left(\frac{\varepsilon}{3}, \frac{\delta}{3}\right) \leq d.$$

Similarly, since every $P_2 \in \mathbb{D}_2$ has $P_2(\cdot \times \mathcal{Y})$ supported only in $\mathcal{X}_2$, and $P_2 \in \mathrm{RE}$, $f^\star_{P_2}$ is either equal -1 with P_2-probability one, or else $\exists x \in \{d+1,\ldots,\mathfrak{s}\}$ with $f^\star_{P_2}(x) = +1$; in either case, $\exists h_{P_2} \in \mathbb{C}_2$ with $\mathrm{er}_{P_2}(h_{P_2}) = 0$; thus, without loss, we can take $f^\star_{P_2} = h_{P_2}$, so that P_2 is in the realizable case with respect to $\mathbb{C}_2$. Now consider an active learning algorithm that first calculates the empirical frequency $\hat{\mathcal{P}}(\{x\}) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}[X_i = x]$ for each $x \in \{d+1,\ldots,\mathfrak{s}\}$ among the first $m = \left\lceil \frac{3^4}{2\varepsilon^4} \ln\left(\frac{3(\mathfrak{s}-d)}{\delta}\right)\right\rceil$ unlabeled data points. Then, for each $x \in \{d+1,\ldots,\mathfrak{s}\}$, if $\hat{\mathcal{P}}(\{x\}) > (1-\varepsilon/3)\varepsilon/3$, the algorithm requests Y_i for the first $i \in \mathbb{N}$ with $X_i = x$ (supposing the budget n has not yet been reached). If any requested value Y_i equals $+1$, then for the $x \in \{d+1,\ldots,\mathfrak{s}\}$ with $X_i = x$, the algorithm returns the classifier $x' \mapsto 2\mathbb{1}_{\{x\}}(x') - 1$. Otherwise, the algorithm returns the all-negative classifier: $x' \mapsto -1$. Denote by $\hat{h}$ the classifier returned by the algorithm. By Hoeffding's inequality and a union bound, with probability at least $1-\delta/3$, every $x \in \{d+1,\ldots,\mathfrak{s}\}$ has $\hat{\mathcal{P}}(\{x\}) \geq \mathcal{P}_{XY}(\{x\} \times \mathcal{Y}) - (\varepsilon/3)^2$. Also, if $\mathcal{P}_{XY} \in \mathrm{RE}$, then with probability one, every $Y_i = f^\star_{\mathcal{P}_{XY}}(X_i)$. Therefore, if $\mathcal{P}_{XY} \in \mathbb{D}_2$, on these events, every $x \in \{d+1,\ldots,\mathfrak{s}\}$ with $\mathcal{P}_{XY}(\{x\} \times \mathcal{Y}) > \varepsilon/3$ will have a label Y_i with $X_i = x$ requested by the algorithm (supposing sufficiently large n), which implies $\hat{h}(x) = f^{\star}_{\mathcal{P}_{XY}}(x)$. Since $f^{\star}_{\mathcal{P}_{XY}}$ has at most one $x \in \mathcal{X}_2$ with $f^{\star}_{\mathcal{P}_{XY}}(x) = +1$, and if such an x exists it must be in $\{d+1,\ldots,\mathfrak{s}\}$, if any requested $Y_i = +1$, we have $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) = 0$, and otherwise either no $x \in \mathcal{X}_2$ has $f^{\star}_{\mathcal{P}_{XY}}(x) = +1$ or else the one such x has $\mathcal{P}_{XY}(\{x\}\times\mathcal{Y}) \leq \varepsilon/3$; in either case, we have $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) = \mathcal{P}_{XY}(\{x : f^{\star}_{\mathcal{P}_{XY}}(x) = +1\}\times\mathcal{Y}) \leq \varepsilon/3$. Thus, regardless of whether the algorithm requests a Y_i with value $+1$, we have $\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) \leq \varepsilon/3$. By a union bound for the two events, we have that $\mathbb{P}(\mathrm{er}_{\mathcal{P}_{XY}}(\hat{h}) > \varepsilon/3) \leq \delta/3$ (given a sufficiently large n). Furthermore, there are at most $\min\left\{\mathfrak{s}-d, \frac{1}{(1-\varepsilon/3)\varepsilon/3}\right\}$ points $x \in \{d+1,\ldots,\mathfrak{s}\}$ with $\hat{\mathcal{P}}(\{x\}) > (1-\varepsilon/3)\varepsilon/3$, and therefore at most this many labels Y_i are requested by the algorithm. Thus, a budget n of at least this size suffices for this guarantee. Since this holds for every $\mathcal{P}_{XY} \in \mathbb{D}_2$, we have that

$$\Lambda_{2,0}\left(\frac{\varepsilon}{3},\frac{\delta}{3}\right) \leq \min\left\{\mathfrak{s}-d, \frac{1}{(1-\varepsilon/3)\varepsilon/3}\right\} \lesssim \min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}.$$

Altogether, we have that $\forall \varepsilon,\delta \in (0,1)$,

$$\Lambda_{\mathrm{RE}}(\varepsilon,\delta) \lesssim \max\left\{\min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\}, d\right\}.$$

Thus, the lower bound in Theorem 3 is tight up to universal constant factors in this case.[17]

Bounded noise: To prove that the lower bound in Theorem 4 is sometimes tight, fix any $\beta \in (0,1/2)$, and let $\mathbb{D}_i$ denote the set of all $P_i \in \mathrm{BN}(\beta)$ with $P_i(\mathcal{X}_i \times \mathcal{Y}) = 1$, for each $i \in \{1,2\}$. For all $\gamma \in [0,1]$ and $i \in \{1,2\}$, define $\mathbb{D}_i(\gamma) = \mathbb{D}_i$. As above, note that for any $P \in \mathrm{BN}(\beta)$, for any measurable $A \subseteq \mathcal{X}\times\mathcal{Y}$, $P(A) = P(\mathcal{X}_1\times\mathcal{Y})P(A|\mathcal{X}_1\times\mathcal{Y}) + P(\mathcal{X}_2\times\mathcal{Y})P(A|\mathcal{X}_2\times\mathcal{Y})$. Furthermore, any $i \in \{1,2\}$ with $P(\mathcal{X}_i\times\mathcal{Y}) > 0$ has $P(\cdot\times\mathcal{Y}|\mathcal{X}_i\times\mathcal{Y})$ supported only on $\mathcal{X}_i$, and since $\eta(x; P(\cdot|\mathcal{X}_i\times\mathcal{Y})) = \eta(x;P)$ for every $x \in \mathcal{X}_i$, we have $P(\cdot|\mathcal{X}_i\times\mathcal{Y}) \in \mathrm{BN}(\beta)$, so that $P(\cdot|\mathcal{X}_i\times\mathcal{Y}) \in \mathbb{D}_i$. Thus, $P \in \mathbb{D} = \{\gamma P_1 + (1-\gamma)P_2 : P_1 \in \mathbb{D}_1, P_2 \in \mathbb{D}_2, \gamma \in [0,1]\}$. Therefore, $\mathrm{BN}(\beta) \subseteq \mathbb{D}$. Together with Lemma 46, this implies $\forall \varepsilon,\delta \in (0,1)$,

$$\Lambda_{\mathrm{BN}(\beta)}(\varepsilon,\delta) \leq \Lambda_{\mathbb{D}}(\varepsilon,\delta) \leq 2\max\left\{\Lambda_{1,0}\left(\frac{\varepsilon}{3},\frac{\delta}{3}\right), \Lambda_{2,0}\left(\frac{\varepsilon}{3},\frac{\delta}{3}\right)\right\},$$

for $\Lambda_{i,0}(\cdot,\cdot)$ defined as above.

Now note that, for each $i \in \{1,2\}$, since every $P_i \in \mathbb{D}_i$ has $P_i \in \mathrm{BN}(\beta)$, we have $f^{\star}_{P_i} \in \mathbb{C}$. Furthermore, since every $h \in \mathbb{C}\setminus\mathbb{C}_i$ has $h(x) = -1$ for every $x \in \mathcal{X}_i$, and the all-negative function $x \mapsto -1$ is contained in $\mathbb{C}_i$, and since $P_i(\mathcal{X}_i\times\mathcal{Y}) = 1$, without loss we can take $f^{\star}_{P_i} \in \mathbb{C}_i$ (i.e., there is a version of $f^{\star}_{P_i}$ contained in $\mathbb{C}_i$). Together with the condition on $\eta(\cdot;P_i)$ from the definition of $\mathrm{BN}(\beta)$, this implies each P_i satisfies the bounded noise condition (with parameter β) with respect to $\mathbb{C}_i$.

Since this is true of every $P_1 \in \mathbb{D}_1$, and the star number and VC dimension of $\mathbb{C}_1$ are both equal d, the upper bound in Theorem 4 implies $\forall \varepsilon \in (0,(1-2\beta)/8)$, $\delta \in (0,1/8]$,

$$\Lambda_{1,0}\left(\frac{\varepsilon}{3},\frac{\delta}{3}\right) \lesssim \frac{1}{(1-2\beta)^2} d \cdot \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

17. The term $\mathrm{Log}\left(\min\left\{\frac{1}{\varepsilon}, |\mathbb{C}|\right\}\right)$ in the lower bound is dominated by the other terms in this example, so that this upper bound is still consistent with the existence of this term in the lower bound.

Similarly, since every $P_2 \in \mathbb{D}_2$ satisfies the bounded noise condition (with parameter β) with respect to $\mathbb{C}_2$, and the star number of $\mathbb{C}_2$ is $\mathfrak{s} - d \leq \mathfrak{s}$ while the VC dimension of $\mathbb{C}_2$ is 1, the upper bound in Theorem 4 implies $\forall \varepsilon \in (0, (1-2\beta)/8)$, $\delta \in (0, 1/8]$,

$$\Lambda_{2,0}\left(\frac{\varepsilon}{3}, \frac{\delta}{3}\right) \lesssim \frac{1}{(1-2\beta)^2} \min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon}\right\} \operatorname{polylog}\left(\frac{1}{\varepsilon\delta}\right).$$

Altogether, we have that

$$\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \delta) \lesssim \frac{1}{(1-2\beta)^2} \max\left\{\min\left\{\mathfrak{s}, \frac{1-2\beta}{\varepsilon}\right\}, d\right\} \operatorname{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

For β bounded away from 0, this is within logarithmic factors of the lower bound in Theorem 4, so that we may conclude that the lower bound is sometimes tight to within logarithmic factors in this case. Furthermore, when β is near 0, it is within logarithmic factors of the lower bound in Theorem 3, which is also a lower bound on $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \delta)$ since $\mathrm{RE} \subseteq \mathrm{BN}(\beta)$; thus, this inherited lower bound on $\Lambda_{\mathrm{BN}(\beta)}(\varepsilon, \delta)$ is also sometimes tight to within logarithmic factors when β is near 0.

Tsybakov noise: The case of Tsybakov noise is slightly more involved than the above. In this case, fix any $a \in [1, \infty)$, $\alpha \in (0, 1)$. Since the upper bound in Theorem 5 already matches the lower bound up to logarithmic factors when $\alpha \in (0, 1/2]$, it suffices to focus on the case $\alpha \in (1/2, 1)$. In this case, for $\gamma \in (0, 1]$, let $\mathbb{D}_i(\gamma)$ denote the set of all $P_i \in \mathrm{TN}(a/\gamma^{1-\alpha}, \alpha)$ with $P_i(\mathcal{X}_i \times \mathcal{Y}) = 1$, for each $i \in \{1, 2\}$. Also let $\mathbb{D}_i(0)$ denote the set of all probability measures P_i with $P_i(\mathcal{X}_i \times \mathcal{Y}) = 1$, for each $i \in \{1, 2\}$. Again, for any $P \in \mathrm{TN}(a, \alpha)$, $P(\cdot) = P(\mathcal{X}_1 \times \mathcal{Y})P(\cdot|\mathcal{X}_1 \times \mathcal{Y}) + P(\mathcal{X}_2 \times \mathcal{Y})P(\cdot|\mathcal{X}_2 \times \mathcal{Y})$, and for any $i \in \{1, 2\}$ with $P(\mathcal{X}_i \times \mathcal{Y}) > 0$, $P(\cdot \times \mathcal{Y}|\mathcal{X}_i \times \mathcal{Y})$ is supported only in $\mathcal{X}_i$, and $\eta(\cdot; P(\cdot|\mathcal{X}_i \times \mathcal{Y})) = \eta(\cdot; P)$ on $\mathcal{X}_i$, so that for any $t > 0$,

$$\begin{aligned}
&P\left(\{x : |\eta(x; P(\cdot|\mathcal{X}_i \times \mathcal{Y})) - 1/2| \leq t\} \times \mathcal{Y} \middle| \mathcal{X}_i \times \mathcal{Y}\right) \\
&= \frac{1}{P(\mathcal{X}_i \times \mathcal{Y})} P\left(\{x \in \mathcal{X}_i : |\eta(x; P) - 1/2| \leq t\} \times \mathcal{Y}\right) \\
&\leq \frac{1}{P(\mathcal{X}_i \times \mathcal{Y})} a' t^{\alpha/(1-\alpha)} = (1-\alpha)(2\alpha)^{\alpha/(1-\alpha)} \left(\frac{a}{P(\mathcal{X}_i \times \mathcal{Y})^{1-\alpha}}\right)^{1/(1-\alpha)} t^{\alpha/(1-\alpha)}.
\end{aligned}$$

Also, since $f^\star_P \in \mathbb{C}$, and $\eta(\cdot; P(\cdot|\mathcal{X}_i \times \mathcal{Y})) = \eta(\cdot; P)$ on $\mathcal{X}_i$, we can take $f^\star_{P(\cdot|\mathcal{X}_i \times \mathcal{Y})}(x) = f^\star_P(x)$ for every $x \in \mathcal{X}_i$, so that there exists a version of $f^\star_{P(\cdot|\mathcal{X}_i \times \mathcal{Y})}$ contained in $\mathbb{C}$. Together, these imply that $P(\cdot|\mathcal{X}_i \times \mathcal{Y}) \in \mathbb{D}_i(P(\mathcal{X}_i \times \mathcal{Y}))$. We therefore have that $\forall P \in \mathrm{TN}(a, \alpha)$, $P = \gamma P_1 + (1-\gamma)P_2$ for some $\gamma \in [0, 1]$, $P_1 \in \mathbb{D}_1(\gamma)$, and $P_2 \in \mathbb{D}_2(1-\gamma)$: that is, $\mathrm{TN}(a, \alpha) \subseteq \mathbb{D}$, for $\mathbb{D}$ as in Lemma 46 (with respect to these definitions of $\mathbb{D}_i(\cdot)$). Therefore, Lemma 46 implies that $\forall \varepsilon, \delta \in (0, 1)$,

$$\begin{aligned}
&\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon, \delta) \leq \Lambda_{\mathbb{D}}(\varepsilon, \delta) \\
&\lesssim \sup_{\gamma \in [0,1]} \max\left\{\Lambda_{1,(\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/8)}, \frac{\delta}{3}\right), \Lambda_{2,(1-\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(1-\gamma+\varepsilon/8)}, \frac{\delta}{3}\right)\right\}. \quad (68)
\end{aligned}$$

First note that, for the case $\gamma \leq \varepsilon/4$, we trivially have

$$\Lambda_{1,(\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/8)}, \frac{\delta}{3}\right) \leq \Lambda_{1,0}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/4)}, \frac{\delta}{3}\right) \leq \Lambda_{1,0}\left(1, \frac{\delta}{3}\right) = 0,$$

and similarly for the case $\gamma \geq 1-\varepsilon/4$, we have $\Lambda_{2,(1-\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(1-\gamma+\varepsilon/8)}, \frac{\delta}{3}\right) = 0$.

For the remaining cases, for any $\gamma \in (0,1]$, since every $P_i \in \mathbb{D}_i(\gamma)$ has $f^{\star}_{P_i} \in \mathbb{C}$, and every $h \in \mathbb{C} \setminus \mathbb{C}_i$ has $h(x) = -1$ for every $x \in \mathcal{X}_i$, and the all-negative function $x \mapsto -1$ is contained in $\mathbb{C}_i$, and $P_i(\mathcal{X}_i \times \mathcal{Y}) = 1$, without loss we can take $f^{\star}_{P_i} \in \mathbb{C}_i$. Together with the definition of $\mathbb{D}_i(\gamma)$, we have that $\mathbb{D}_i(\gamma)$ is contained in the set of probability measures P_i satisfying the Tsybakov noise condition with respect to the hypothesis class $\mathbb{C}_i$, with parameters $\frac{a}{\gamma^{1-\alpha}}$ and α. Therefore, since the star number and VC dimension of $\mathbb{C}_1$ are both d, Theorem 5 implies that for any $\gamma \in (\varepsilon/4, 1]$,[18]

$$\begin{aligned}
&\Lambda_{1,\gamma-\varepsilon/8}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/8)}, \frac{\delta}{3}\right) \leq \Lambda_{1,\gamma/2}\left(\frac{\varepsilon}{3\gamma}, \frac{\delta}{3}\right) \\
&\lesssim \left(\frac{a}{\gamma^{1-\alpha}}\right)^2 \left(\frac{\gamma}{\varepsilon}\right)^{2-2\alpha} d \cdot \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right) = a^2 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha} d \cdot \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).
\end{aligned}$$

Similarly, since the star number of $\mathbb{C}_2$ is $\mathfrak{s} - d$ and the VC dimension of $\mathbb{C}_2$ is 1, Theorem 5 implies that for any $\gamma \in [0, 1-\varepsilon/4)$,

$$\begin{aligned}
&\Lambda_{2,1-\gamma-\varepsilon/8}\left(\frac{\varepsilon}{2(1-\gamma+\varepsilon/8)}, \frac{\delta}{3}\right) \leq \Lambda_{2,(1-\gamma)/2}\left(\frac{\varepsilon}{3(1-\gamma)}, \frac{\delta}{3}\right) \\
&\lesssim \left(\frac{a}{(1-\gamma)^{1-\alpha}}\right)^2 \left(\frac{1-\gamma}{\varepsilon}\right)^{2-2\alpha} \min\left\{\mathfrak{s} - d, \frac{(1-\gamma)^{1/\alpha}(1-\gamma)}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1} \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right) \\
&\leq a^2 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \min\left\{\mathfrak{s}, \frac{1}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1} \mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right).
\end{aligned}$$

Plugging this into (68), we have that

$$\Lambda_{\mathrm{TN}(a,\alpha)}(\varepsilon,\delta) \lesssim a^2 \left(\frac{1}{\varepsilon}\right)^{2-2\alpha} \max\left\{\min\left\{\mathfrak{s}, \frac{1}{a^{1/\alpha}\varepsilon}\right\}^{2\alpha-1}, d\right\} \mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

As claimed, this is within logarithmic factors of the lower bound in Theorem 5 (for $1/2 < \alpha < 1$, $a \geq 4$, $\varepsilon \in (0, 1/(24a^{1/\alpha}))$, and $\delta \in (0, 1/24]$), so that, combined with the tightness (always) for the case $0 < \alpha \leq 1/2$, we may conclude that the lower bounds in Theorem 5 are sometimes tight to within logarithmic factors.

Benign noise: The case of benign noise proceeds analogously to the above. Since $\mathrm{BE}(0) = \mathrm{RE}$, tightness of the lower bound for the case $\nu = 0$ (up to constant factors) has already been addressed above (supposing we include the lower bound from Theorem 3 as a lower bound on $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta)$ to strengthen the lower bound in Theorem 7). For the remainder, we suppose $\nu \in (0, 1/2)$. For $\gamma \in [0,1]$, let $\mathbb{D}_i(\gamma)$ denote the set of all $P_i \in \mathrm{BE}(\nu/(\gamma \vee 2\nu))$ with $P_i(\mathcal{X}_i \times \mathcal{Y}) = 1$, for each $i \in \{1,2\}$. Again, for any $P \in \mathrm{BE}(\nu)$, $P(\cdot) = P(\mathcal{X}_1 \times \mathcal{Y})P(\cdot|\mathcal{X}_1 \times \mathcal{Y}) + P(\mathcal{X}_2 \times \mathcal{Y})P(\cdot|\mathcal{X}_2 \times \mathcal{Y})$, and for any $i \in \{1,2\}$ with $P(\mathcal{X}_i \times \mathcal{Y}) > 0$, $P(\cdot \times \mathcal{Y}|\mathcal{X}_i \times \mathcal{Y})$ is supported only in $\mathcal{X}_i$, and $\eta(\cdot; P(\cdot|\mathcal{X}_i \times \mathcal{Y})) = \eta(\cdot; P)$ on $\mathcal{X}_i$, so that we can take $f^{\star}_{P(\cdot|\mathcal{X}_i \times \mathcal{Y})}(x) = f^{\star}_P(x)$ for every $x \in \mathcal{X}_i$;

18. Recall that, as mentioned in Section 5, the upper bounds on the label complexities stated in Section 5 hold without the stated restrictions on the values $\varepsilon, \delta \in (0,1)$ and a.

thus, there is a version of $f^{\star}_{P(\cdot|\mathcal{X}_i\times\mathcal{Y})}$ contained in $\mathbb{C}$. Furthermore,

$$\begin{aligned}\mathrm{er}_{P(\cdot|\mathcal{X}_i\times\mathcal{Y})}(f^{\star}_{P(\cdot|\mathcal{X}_i\times\mathcal{Y})}) &= \frac{1}{P(\mathcal{X}_i\times\mathcal{Y})}P\left((x,y): f^{\star}_P(x)\neq y \text{ and } x\in\mathcal{X}_i\right)\\ &\leq \frac{1}{P(\mathcal{X}_i\times\mathcal{Y})}P\left((x,y): f^{\star}_P(x)\neq y\right) \leq \frac{\nu}{P(\mathcal{X}_i\times\mathcal{Y})}.\end{aligned}$$

Also, since every $x\in\mathcal{X}_i$ has $f^{\star}_{P(\cdot|\mathcal{X}_i\times\mathcal{Y})}(x) = f^{\star}_P(x) = \mathrm{sign}(2\eta(x;P)-1) = \mathrm{sign}(2\eta(x;P(\cdot|\mathcal{X}_i\times\mathcal{Y}))-1)$, we have $P((x,y): f^{\star}_{P(\cdot|\mathcal{X}_i\times\mathcal{Y})}(x) = y | x\in\mathcal{X}_i)\geq 1/2$, so that $\mathrm{er}_{P(\cdot|\mathcal{X}_i\times\mathcal{Y})}(f^{\star}_{P(\cdot|\mathcal{X}_i\times\mathcal{Y})})\leq 1/2$. Together, these imply that $P(\cdot|\mathcal{X}_i\times\mathcal{Y})\in\mathbb{D}_i(P(\mathcal{X}_i\times\mathcal{Y}))$. We therefore have that $\forall P\in\mathrm{BE}(\nu)$, $P=\gamma P_1+(1-\gamma)P_2$ for some $\gamma\in[0,1]$, $P_1\in\mathbb{D}_1(\gamma)$, and $P_2\in\mathbb{D}_2(1-\gamma)$: that is, $\mathrm{BE}(\nu)\subseteq\mathbb{D}$, for $\mathbb{D}$ as in Lemma 46 (with respect to these definitions of $\mathbb{D}_i(\cdot)$). Therefore, Lemma 46 implies that $\forall\varepsilon,\delta\in(0,1)$,

$$\begin{aligned}&\Lambda_{\mathrm{BE}(\nu)}(\varepsilon,\delta)\leq\Lambda_{\mathbb{D}}(\varepsilon,\delta)\\ &\lesssim \sup_{\gamma\in[0,1]}\max\left\{\Lambda_{1,(\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/8)},\frac{\delta}{3}\right),\Lambda_{2,(1-\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(1-\gamma+\varepsilon/8)},\frac{\delta}{3}\right)\right\}. \quad (69)\end{aligned}$$

First note that, as above, for the case $\gamma\leq\varepsilon/4$, we trivially have

$$\Lambda_{1,(\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/8)},\frac{\delta}{3}\right)\leq\Lambda_{1,0}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/4)},\frac{\delta}{3}\right)\leq\Lambda_{1,0}\left(1,\frac{\delta}{3}\right)=0,$$

and similarly for the case $\gamma\geq 1-\varepsilon/4$, we have $\Lambda_{2,(1-\gamma-\varepsilon/8)\vee 0}\left(\frac{\varepsilon}{2(1-\gamma+\varepsilon/8)},\frac{\delta}{3}\right)=0$.

For the remaining cases, for any $\gamma\in(0,1]$, since every $P_i\in\mathbb{D}_i(\gamma)$ has $f^{\star}_{P_i}\in\mathbb{C}$, and every $h\in\mathbb{C}\setminus\mathbb{C}_i$ has $h(x)=-1$ for every $x\in\mathcal{X}_i$, and the all-negative function $x\mapsto -1$ is contained in $\mathbb{C}_i$, and $P_i(\mathcal{X}_i\times\mathcal{Y})=1$, without loss we can take $f^{\star}_{P_i}\in\mathbb{C}_i$. Together with the definition of $\mathbb{D}_i(\gamma)$, we have that $\mathbb{D}_i(\gamma)$ is contained in the set of probability measures P_i satisfying the benign noise condition with respect to the hypothesis class $\mathbb{C}_i$, with parameter $\frac{\nu}{\gamma}\wedge\frac{1}{2}$. Therefore, since the star number and VC dimension of $\mathbb{C}_1$ are both d, Theorem 7 implies that for any $\gamma\in(\varepsilon/4,1]$,[19]

$$\begin{aligned}\Lambda_{1,\gamma-\varepsilon/8}\left(\frac{\varepsilon}{2(\gamma+\varepsilon/8)},\frac{\delta}{3}\right)\leq\Lambda_{1,\gamma/2}\left(\frac{\varepsilon}{3\gamma},\frac{\delta}{3}\right)&\lesssim\left(\frac{(\nu/\gamma)^2}{(\varepsilon/\gamma)^2}d+d\right)\mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right)\\ &\lesssim\left(\frac{\nu^2}{\varepsilon^2}\vee 1\right)d\cdot\mathrm{polylog}\left(\frac{d}{\varepsilon\delta}\right).\end{aligned}$$

Similarly, since the star number of $\mathbb{C}_2$ is $\mathfrak{s}-d$ and the VC dimension of $\mathbb{C}_2$ is 1, Theorem 7 implies that for any $\gamma\in[0,1-\varepsilon/4)$,

$$\begin{aligned}&\Lambda_{2,1-\gamma-\varepsilon/8}\left(\frac{\varepsilon}{2(1-\gamma+\varepsilon/8)},\frac{\delta}{3}\right)\leq\Lambda_{2,(1-\gamma)/2}\left(\frac{\varepsilon}{3(1-\gamma)},\frac{\delta}{3}\right)\\ &\lesssim\left(\frac{(\nu/(1-\gamma))^2}{(\varepsilon/(1-\gamma))^2}+\min\left\{\mathfrak{s}-d,\frac{1}{\varepsilon}\right\}\right)\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)\lesssim\left(\frac{\nu^2}{\varepsilon^2}\vee\min\left\{\mathfrak{s},\frac{1}{\varepsilon}\right\}\right)\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right).\end{aligned}$$

19. Again, as mentioned in Section 5, the restrictions on ε,δ stated in Theorem 7 are only required for the lower bounds.

Plugging these into (69), we have that for $\varepsilon \in (0, \nu)$,

$$\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta) \lesssim \left(\frac{\nu^2}{\varepsilon^2} d + \min\left\{ \mathfrak{s}, \frac{1}{\varepsilon} \right\} \right) \mathrm{polylog}\left(\frac{d}{\varepsilon\delta} \right).$$

Again, this is within logarithmic factors of the lower bound in Theorem 7 (for $\varepsilon \in (0, (1-2\nu)/24)$ and $\delta \in (0, 1/24]$), so that we may conclude that this lower bound is sometimes tight to within logarithmic factors when ν is not near 0 (specifically, when $\varepsilon < \nu$). For $\nu \le \varepsilon$, the above implies

$$\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta) \lesssim \max\left\{ d, \min\left\{ \mathfrak{s}, \frac{1}{\varepsilon} \right\} \right\} \mathrm{polylog}\left(\frac{d}{\varepsilon\delta} \right),$$

which is within logarithmic factors of the lower bound in Theorem 3 (for $\varepsilon \in (0, 1/9)$ and $\delta \in (0, 1/3)$). Since $\mathrm{RE} \subseteq \mathrm{BE}(\nu)$, this is also a lower bound on $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta)$. Thus, in this case, we may conclude that this inherited lower bound on $\Lambda_{\mathrm{BE}(\nu)}(\varepsilon, \delta)$ is sometimes tight to within logarithmic factors, for ν near 0 (specifically, when $\varepsilon \ge \nu$).

References

T. M. Adams and A. B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *Annals of Probability*, 38(4):1345–1367, 2010. A.3

T. M. Adams and A. B. Nobel. Uniform approximation and bracketing properties of VC classes. *Bernoulli*, 18:1310–1319, 2012. 5.1, A.3

N. Ailon, R. Begleiter, and E. Ezra. Active learning using smooth relative regret approximations with applications. In *Proceedings of the* 25^{th} *Conference on Learning Theory*, 2012. 1.1

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

P. Awasthi, M.-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the* 46^{th} *ACM Symposium on the Theory of Computing*, 2014. 3

M.-F. Balcan and S. Hanneke. Robust interactive learning. In *Proceedings of the* 25^{th} *Conference on Learning Theory*, 2012. 7.2, B.2

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. 1.1, 7.1

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the 23*$^{\text{rd}}$ *International Conference on Machine Learning*, 2006. 1.1, 3, 7.1, B.2, B.2

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the* 20^{th} *Conference on Learning Theory*, 2007. 1.1, 1.2, 3

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009. 1.1, 7.1, 7.4, B.2, B.2

M.-F. Balcan, S. Hanneke, and J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, 2010. 1.1, 7.1, 8

P. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006. 3

P. L. Bartlett, S. Mendelson, and P. Philips. Local complexities for empirical risk minimization. In *Proceedings of the* 17^{th} *Conference on Learning Theory*, 2004. 3

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the* 26^{th} *International Conference on Machine Learning*, 2009. 1.1, 1.2, B.2, B.2, B.2, B.2

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems 23*, 2010. 1.1

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. 2, 5.1, 6, 7.3.1, A.1

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, November 2005. 3

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence*, 3176:169–207, 2004. A.1

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. 7.4, 7.4, 7.4

R. Castro and R. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, July 2008. 1.1

R.M. Castro and R.D. Nowak. Upper and lower error bounds for active learning. In *The 44*$^{\text{th}}$ *Annual Allerton Conference on Communication, Control and Computing*, 2006. 1.1

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994. 1.1, 1.2, 5.1, 7.1, 7.3.1, B.2

S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems 17*, 2004. 1.1, 4, B.1

S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, 2005. 1.1, 1.2, 4, 4, 4, 7, 7.2, 7.2, 7.2, 7.2, 7.4, 8, B.1, C.2

S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the* 18^{th} *Conference on Learning Theory*, 2005. 1.1

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems 20*, 2007. 1.1, 5, 7.1, B.2, B.2

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989. 6

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. 1.1, 4, 7, 7.3.1, B.1

R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13:255–279, 2012. 1.1, 4, 7.3.1, B.1

G. Fan. A graph-theoretic view of teaching. Master's thesis, Department of Computer Science, University of Regina, 2012. 5, 7.3

S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21:269–304, 1995. A.1

Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997. 1.1

E. Friedman. Active learning for smooth problems. In *Proceedings of the* 22^{nd} *Conference on Learning Theory*, 2009. 7.1

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006. 3, 6, 7.4, B.2

S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995. 7.3

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. 7.4

S. Hanneke. The cost complexity of interactive learning, 2006. 7.3, 7.3, 7.3

S. Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of the* 20^{th} *Conference on Learning Theory*, 2007a. 1.1, 3, 4, 4, 5, 7, 7.3, 10, 7.3, 7.3, 7.3.1, B.1, B.2

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the* 24^{th} *International Conference on Machine Learning*, 2007b. 1.1, 3, 4, 7, 7.1, 7.1, 7.4, B.2

S. Hanneke. Adaptive rates of convergence in active learning. In *Proceedings of the* 22^{nd} *Conference on Learning Theory*, 2009a. 1.1, 7.1, B.2, B.2

S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009b. 1.1, 3, 7.1, 7.1, 7.3, B.2

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011. 1.1, 1.2, 3, 7.1, B.2, B.2, B.2, B.2, B.2

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13(5):1469–1587, 2012. 1.1, 3, 9, 7.1, 8, B.2, B.2

S. Hanneke. Theory of Active Learning. Version 1.1. `http://www.stevehanneke.com`, 2014. 1, 1.1, 1.2, 3, 4, 4, 5, 6, 7.1, 7.2, B.2, B.2, B.2, B.2, B.2, C.1

S. Hanneke and L. Yang. Surrogate losses in passive and active learning. *arXiv:1207.3772*, 2012. 1.1, 3, 5, 6, 7.1, B.2

D. Haussler and E. Welzl. ε-nets and simplex range queries. *Discrete Computational Geometry*, 2: 127–151, 1987. A.1, A.1

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115:248–292, 1994. 4, 5, 6

T. Hegedüs. Generalized teaching dimension and the query complexity of learning. In *Proceedings of the 8th Conference on Computational Learning Theory*, 1995. 7.3, B.1, 14

L. Hellerstein, K. Pillaipakkamnatt, V. Raghavan, and D. Wilkins. How many queries are needed to learn? *Journal of the Association for Computing Machinery*, 43(5):840–862, 1996. 7.3, B.1

D. Hsu. *Algorithms for Active Learning*. PhD thesis, Department of Computer Science and Engineering, School of Engineering, University of California, San Diego, 2010. 1.1

M. Kääriäinen. Active learning in the non-realizable case. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory*, 2006. 1.2, 3, 5, 5.1, 6, B.2, B.2, B.2, B.2

A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proceedings of the* 46^{th} *Annual IEEE Symposium on Foundations of Computer Science*, 2005. 3

O. Kallenberg. *Foundations of Modern Probability, 2nd Edition*. Springer Verlag, New York, 2002. B.1

M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994. 3

A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959. C.4

A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961. C.4

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. 3

V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010. 1.1, 3, 7.1, B.2, B.2, B.2

S. R. Kulkarni. On metric entropy, Vapnik-Chervonenkis dimension, and learnability for a class of distributions. Technical Report CICS-P-160, Center for Intelligent Control Systems, 1989. B.1, B.1, C.4

S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11:23–35, 1993. B.1, B.1

L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1 (1):38–53, 1973. 7.4

Y. Li and P. M. Long. Learnability and the doubling dimension. In *Advances in Neural Information Processing 20*, 2007. 7, 7.4

N. Littlestone and M. Warmuth. Relating data compression and learnability, 1986. A.1

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27: 1808–1829, 1999. 3, B.2

P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5): 2326–2366, 2006. 3, 6

C. McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer-Verlag, 1998. B.2

S. Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(1): 67–90, 2012. 1.1

T. Mitchell. Version spaces: a candidate elimination approach to rule learning. In *IJCAI'77: Proceedings of the 5th international joint conference on Artificial intelligence*, pages 305–310, 1977. 7.3

M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems 24*, 2011. 1.2, 3, 5, 6, 7.4, A.2, 13

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13:145–147, 1972. A.1, B.1, C.4

B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2012. 1

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. 3, B.2

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984. 3

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996. 2

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

R. van Handel. The universal Glivenko-Cantelli property. *Probability and Related Fields*, 155: 911–934, 2013. A.3

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982. 3, 5.1, 6, 7.3.1, A.1, A.1, B.2

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998. 3, 5.1, 6, 7.3.1, B.2

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971. 2, 3, 6, A.1, B.1, B.2, C.4

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. A.1, A.1

M. Vidyasagar. *Learning and Generalization with Applications to Neural Networks, 2$^{\text{nd}}$ Edition*. Springer-Verlag, 2003. A.1, C.4

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. D.1

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944. D.1

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945. 5.1

A. Wald. *Sequential Analysis*. John Wiley & Sons, Inc., New York, 1947. 5.1

L. Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12:2269–2292, 2011. 7.1

Y. Wiener, S. Hanneke, and R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *arXiv:1404.1504*, 2014. 1.1, 1.2, 5.1, 7.3, 7.3.1, 7.3.1, B.1

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. 7.4