Agnostic Active Learning Is Always Better Than Passive Learning

Steve Hanneke
Department of Computer Science
Purdue University
steve.hanneke@gmail.com

Abstract

We sharply characterize the optimal first-order query complexity of agnostic active learning for all concept classes, and propose a new general active learning algorithm which achieves it. Remarkably, the optimal query complexity admits a leading term which is *always* strictly smaller than the sample complexity of passive supervised learning (by a factor proportional to the best-in-class error rate). This was not previously known to be possible in the agnostic setting. For comparison, in all previous general analyses, the leading term exhibits an additional factor, such as the disagreement coefficient or related complexity measure, and therefore only provides improvements over passive learning in restricted cases. The present work completely removes such factors from the leading term, implying that *every* concept class benefits from active learning in the non-realizable case. The results established in this work resolve an important long-standing open question central to the past two decades of research on the theory of agnostic active learning.

1 Introduction

Active learning is a well-known powerful variant of supervised learning, in which the learning algorithm interactively participates in the process of labeling the training examples. In this setting, there is a pool (or stream) of unlabeled examples, and the learning algorithm selects individual examples and queries an oracle (typically a human labeler) to observe their labels. This happens sequentially, so that the learner has observed previously-queried labels before deciding which example to query next. The intended purpose of active learning is to reduce the overall number of labels necessary for learning to a given accuracy, called the *query complexity*. We are therefore particularly interested in using active learning in scenarios where its query complexity is significantly smaller than the number of randomly-sampled training examples which would be needed to achieve the same accuracy, called the *sample complexity* of *passive* supervised learning.

Active learning has not only been incredibly useful for many practical machine learning problems (e.g., Cohn, Ghahramani, and Jordan, 1996; Tong and Koller, 2001; Zhu, Lafferty, and Ghahramani, 2003; Olsson, 2009; Settles, 2012; Ren, Xiao, Chang, Huang, Li, Gupta, Chen, and Wang, 2021; Mosqueira-Rey, Hernández-Pereira, Alonso-Ríos, Bobes-Bascarán, and Fernández-Leal, 2023) but has also given rise to a rich and nuanced theoretical literature (see e.g., Dasgupta, 2005, 2011; Balcan, Beygelzimer, and Langford, 2009; Hanneke, 2007b, 2014; Zhang and Chaudhuri, 2014; Hanneke and Yang, 2015; see Appendix A for a detailed summary of related work). Moreover, the insights and techniques discovered in this literature have had tremendous influence on other branches of the learning theory literature (e.g., Awasthi, Balcan, and Long, 2014; Foster, Rakhlin, Simchi-Levi, and Xu, 2021; Hanneke, 2009b, 2016a,b, 2024; Zhivotovskiy and Hanneke, 2018; Simon, 2015; Balcan and Long, 2013; El-Yaniv and Wiener, 2010; Balcan, Blum, Hanneke, and Sharma, 2022).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Within the literature on the theory of active learning, a central topic which has garnered by-far the most interest is that of *agnostic* active learning: that is, the study of active learning algorithms capable of providing performance guarantees even in noisy or otherwise non-realizable learning problems, without assumptions on the form of the noise. This line of work was initiated by the groundbreaking A^2 algorithm (Agnostic Active) of Balcan, Beygelzimer, and Langford (2005, 2006, 2009) (with its general analysis later given by Hanneke, 2007b) and concurrently a lower bound analysis of Kääriäinen (2005, 2006) (later strengthened by Beygelzimer, Dasgupta, and Langford, 2009). These results were later refined and extended in numerous ways. However, throughout this two-decades long history, there has persisted a significant gap between the sharpest known upper and lower bounds on the optimal query complexity. Moreover, this gap represents an important *qualitative* distinction: while the lower bound is always smaller than the sample complexity of passive learning, the existing upper bounds only reflect such improvements under further restrictive conditions (e.g., bounded disagreement coefficient). Thus, the issue of resolving this gap is of central importance to this subject, since it has implications for answering the question of whether the query complexity of active learning can *always* offer improvements over the sample complexity of passive learning in the non-realizable case: i.e., for *every* concept class, with *no* distributional assumptions.

The main contribution of the present work is to establish that this is indeed possible, and in fact the lower bound is always attainable. To achieve this, we introduce new algorithmic principles for active learning, improving concentration of error estimates via adaptively isolating regions where the error estimates have high variance and allocating more queries to such regions.

2 Background and Summary of the Main Result

Let $\mathbb{C}$ be any concept class[1] (a set of functions $\mathcal{X} \to \{0,1\}$ on a set $\mathcal{X}$ called the *instance space*) and denote by $\mathsf{d} = \mathrm{VC}(\mathbb{C})$ the VC dimension of $\mathbb{C}$ (Vapnik and Chervonenkis, 1971; see Definition 4). Let P be an (unknown) joint distribution on $\mathcal{X} \times \{0,1\}$, and define the *error rate* of any *classifier* $h : \mathcal{X} \to \{0,1\}$ as $\mathrm{er}_P(h) := P((x,y) : h(x) \neq y)$. In the *active learning* problem, there is a sequence $(X_1, Y_1), \ldots, (X_m, Y_m)$ of i.i.d. samples from P, but the learner initially only observes the X_i values (the *unlabeled* examples). It then has the capability to *query* any example X_i, which reveals the corresponding true label Y_i, in a *sequential* manner (i.e., it chooses its next query $X_{i'}$ after observing the label Y_i of its previous query point X_i). After a number of such queries, the learner returns a classifier $\hat{h}$. The goal is to achieve a small *excess* error rate $\mathrm{er}_P(\hat{h}) \leq \inf_{h \in \mathbb{C}} \mathrm{er}_P(h) + \varepsilon$ while making as few queries as possible. We are particularly interested in quantifying the number of queries sufficient to achieve this, as a function of ε and the value of the *best-in-class* error rate $\inf_{h \in \mathbb{C}} \mathrm{er}_P(h)$, known as a *first-order* query complexity bound.

Specifically, for any $\varepsilon, \delta, \beta \in (0,1)$, the *optimal query complexity*, $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C})$, is defined as the minimal $Q \in \mathbb{N}$ for which there exists an active learner $\mathbb{A}_a$ such that (for a sufficiently large number m of unlabeled examples), for every P with $\inf_{h \in \mathbb{C}} \mathrm{er}_P(h) \leq \beta$, with probability at least $1 - \delta$, $\mathbb{A}_a$ makes at most Q queries and returns $\hat{h}$ satisfying $\mathrm{er}_P(\hat{h}) \leq \inf_{h \in \mathbb{C}} \mathrm{er}_P(h) + \varepsilon$. The main quantity for comparison is the *sample complexity* of supervised *passive learning*. A passive learner $\mathbb{A}_p$ simply trains on n *labeled* training examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ sampled i.i.d. from P to produce a classifier $\hat{h}$. For $\varepsilon, \delta, \beta \in (0,1)$, the *optimal sample complexity* of passive learning, $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$, is defined as the minimal size $n \in \mathbb{N}$ of such a training sample for which there exists a passive learner $\mathbb{A}_p$ such that, for every P with $\inf_{h \in \mathbb{C}} \mathrm{er}_P(h) \leq \beta$, with probability at least $1 - \delta$, $\mathbb{A}_p$ returns $\hat{h}$ satisfying $\mathrm{er}_P(\hat{h}) \leq \inf_{h \in \mathbb{C}} \mathrm{er}_P(h) + \varepsilon$. We remark that, in both the active and passive cases, these definitions place no restrictions on the computational efficiency of the learning algorithms, but rather focus on the *data efficiency*, which is our primary interest in this work (see Section G).

Since both the query complexity and sample complexity concern the number of *labels* sufficient for learning, it is natural to compare $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C})$ with $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$ to quantify the benefits of active learning. Thus, the primary interest in the theory of agnostic active learning is quantifying how much smaller $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C})$ is compared to $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$. Since our interest is *agnostic* learning, it is most interesting to focus on the regime where P is far-from-realizable: that is, where β is much larger than ε. In this regime, it is well known from the works of Vapnik and Chervonenkis (1974);

[1] To focus on non-trivial cases, we suppose $|\mathbb{C}| \geq 3$. We also suppose $\mathcal{X}$ is equipped with a σ-algebra specifying its measurable subsets, and we adopt the standard mild measure-theoretic restrictions on the σ-algebra and the class $\mathbb{C}$ from empirical process theory: namely, the image-admissible Suslin property (Dudley, 1999).

Devroye and Lugosi (1995); Hanneke, Larsen, and Zhivotovskiy (2024b) that the optimal sample complexity of passive learning satisfies $\mathcal{M}_p(\varepsilon,\delta;\beta,\mathbb{C}) = \Theta\left(\frac{\beta}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)\right)$. In comparison, the known lower bound for active learning is $\mathrm{QC}_a(\varepsilon,\delta;\beta,\mathbb{C}) = \Omega\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)\right)$ (Kääriäinen, 2006; Beygelzimer, Dasgupta, and Langford, 2009). Thus, the strongest improvement we might hope from active learning is a factor of β (representing the *best-in-class* error rate).

However, in the prior literature, this β-factor improvement has only been demonstrated in upper bounds under *restrictions* to $\mathbb{C}$ or P. Specifically, every general upper bound on $\mathrm{QC}_a(\varepsilon,\delta;\beta,\mathbb{C})$ in the literature has the form $c(\beta)\mathsf{d}\frac{\beta^2}{\varepsilon^2}$ (ignoring logs), where $c(\beta)$ is a $(\mathbb{C},P)$-dependent quantity. For instance, one commonly appearing such quantity $c(\beta)$ is the *disagreement coefficient* $\theta(\beta)$ of Hanneke (2007b). We refer the reader to Appendix A for a detailed survey of such quantities $c(\beta)$ which have appeared in the literature. Importantly, for all such upper bounds in the literature, the corresponding factor $c(\beta)$ has the property that there exist simple classes $\mathbb{C}$ and distributions P for which $c(\beta) \geq \frac{1}{\beta}$ (see Hanneke and Yang, 2015; Hanneke, 2016b, 2024): for instance, even for linear classifiers on $\mathbb{R}^2$ or singletons on $\mathbb{N}$. Note that when $c(\beta) \geq \frac{1}{\beta}$, a query complexity $c(\beta)\mathsf{d}\frac{\beta^2}{\varepsilon^2}$ becomes *no smaller* than $\mathsf{d}\frac{\beta}{\varepsilon^2}$, the sample complexity of *passive* learning. Moreover, one can show that avoiding such $\mathsf{d}\frac{\beta}{\varepsilon^2}$ query complexities would require new algorithmic techniques (see Appendix A).

Naturally, the question of refining such $c(\beta)$ factors has been a subject of much interest for many years. In particular, it has remained open whether such factors might even be *avoided entirely*, so that the β-factor improvement might *always* be achievable. In a series of talks, I conjectured that the lower bound $\Omega\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)\right)$ is *always* sharp (in the far-from-realizable regime), and even offered a sizable prize for a resolution of this question (along with lower-order terms): for instance, see our 2019 ICML tutorial (Hanneke and Nowak, 2019).

Contributions of this Work: In the present work, we completely resolve this question. We prove that (in the above regime) $\mathrm{QC}_a(\varepsilon,\delta;\beta,\mathbb{C}) = \Theta\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)\right)$. In other words, the β-factor improvement is *always* achievable, the known lower bound is *sharp*, and there is *no need* for restrictions on $(\mathbb{C},P)$ or additional factors $c(\beta)$ as appear in all prior works.

Extending to the *full range* of β, the more-general form of the bound we prove also includes an additive *lower-order* term to account for the small-β regime. In the simplest such bound (Theorem 1), this lower-order term is simply $\tilde{O}\left(\frac{\mathsf{d}}{\varepsilon}\right)$, so that the general form is $\mathrm{QC}_a(\varepsilon,\delta;\beta,\mathbb{C}) = \tilde{O}\left(\mathsf{d}\frac{\beta^2}{\varepsilon^2}+\frac{\mathsf{d}}{\varepsilon}\right)$ (Theorem 3 and Appendix F refine this lower-order term for some classes). For comparison, the general form of the passive sample complexity is $\mathcal{M}_p(\varepsilon,\delta;\beta,\mathbb{C}) = \tilde{\Theta}\left(\mathsf{d}\frac{\beta}{\varepsilon^2}+\frac{\mathsf{d}}{\varepsilon}\right)$. We note that, even in the *nearly-realizable* regime ($\beta = \tilde{O}(\varepsilon)$), it is known that $\frac{\mathsf{d}}{\varepsilon}$ is a lower bound on the query complexity for many classes $\mathbb{C}$ (Dasgupta, 2005; Hanneke, 2014; see Appendix D of Hanneke and Yang, 2015), so that this term is sometimes unavoidable, and hence the benefits of active learning can wane in the nearly-realizable regime. Likewise, the lower bound $\mathsf{d}\frac{\beta^2}{\varepsilon^2}$ implies the benefits can also diminish in the very-high-noise regime ($\beta = \Omega(1)$). In contrast, as discussed above, in the *far-from-realizable* regime ($\sqrt{\varepsilon} \leq \beta \ll 1$), the bound is of order $\mathsf{d}\frac{\beta^2}{\varepsilon^2}$, reflecting a β-factor improvement over the sample complexity of passive learning $\mathsf{d}\frac{\beta}{\varepsilon^2}$. Additionally, the intermediate regime of *moderate-size* β (i.e., $\varepsilon \ll \beta < \sqrt{\varepsilon}$) *also* exhibits improvements over passive learning for all $\mathbb{C}$: in this regime, $\mathcal{M}_p(\varepsilon,\delta;\beta,\mathbb{C}) = \Omega\left(\mathsf{d}\frac{\beta}{\varepsilon^2}\right)$, whereas $\mathrm{QC}_a(\varepsilon,\delta;\beta,\mathbb{C}) = \tilde{O}\left(\frac{\mathsf{d}}{\varepsilon}\right) \ll \mathsf{d}\frac{\beta}{\varepsilon^2}$, reflecting an improvement by a factor $\tilde{O}(\frac{\varepsilon}{\beta})$. Altogether, this result reveals a previously-unknown and truly remarkable fact: $\mathrm{QC}_a(\varepsilon,\delta;\beta,\mathbb{C}) \ll \mathcal{M}_p(\varepsilon,\delta;\beta,\mathbb{C})$ in all regimes $\varepsilon \ll \beta \ll 1$, or in other words, in all regimes outside the nearly-realizable and very-high-noise cases, the following is true:

For every concept class $\mathbb{C}$, the optimal query complexity of agnostic active learning is strictly smaller than the optimal sample complexity of agnostic passive learning.

This result resolves an important long-standing open question central to the past two decades of research on the theory of agnostic active learning.

3 Main Results

Formally, the following theorem expresses the new upper bound, together with known lower bounds for comparison (Kääriäinen, 2006; Beygelzimer, Dasgupta, and Langford, 2009; Hanneke, 2014; Hanneke and Yang, 2015). A more-detailed version of the result appears in Theorem 5 (Appendix C).

Theorem 1. *For every concept class* $\mathbb{C}$*, letting* $\mathsf{d} = \mathrm{VC}(\mathbb{C})$, $\forall \varepsilon, \delta \in (0, 1/8)$, $\forall \beta \in [0, 1]$,

$$\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}) = O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\frac{\mathsf{d}}{\varepsilon}\right)$$

and $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}) = \Omega\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right)$*. Moreover, for every* $\mathsf{d} \in \mathbb{N}$ *there exists* $\mathbb{C}$ *with* $\mathrm{VC}(\mathbb{C}) = \mathsf{d}$ *such that* $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}) = \Omega\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right) + \frac{\mathsf{d}}{\varepsilon}\right)$.

We provide a new general active learning algorithm $\mathbb{A}_{\mathrm{avid}}$ achieving this upper bound in Section 4. Importantly, the algorithm *does not need to know* β (or anything else about P) to achieve this guarantee: i.e., it is completely *adaptive* to the value β. Moreover, the number of *unlabeled* examples the algorithm requires is only $\tilde{\Theta}\left(\mathsf{d}\frac{\beta}{\varepsilon^2} + \frac{\mathsf{d}}{\varepsilon}\right)$, of the same order as the sample complexity of passive learning; it can also adaptively determine how many unlabeled examples to use without knowing β.

The AVID Principle: The main innovation underlying the algorithm, which enables it to achieve this query complexity, represents a new principle for the design of active learning learning algorithms, which we call *adaptive localized variance isolation by disagreements* (AVID). The algorithm adaptively partitions the instance space $\mathcal{X}$ into *regions*, with the aim of *isolating* a region $\Delta \subseteq \mathcal{X}$ where it is most challenging to learn, due to exceptionally high *variance* in the error estimation problem in the Δ region (where Δ will be defined as a union of pairwise *disagreement* regions witnessing the high variance, carefully selected to ensure $P_X(\Delta) = O(\beta)$). It then allocates disproportionately *more* queries to this challenging region Δ compared to the (considerably-easier) remaining region $\mathcal{X} \setminus \Delta$. This idea has interesting connections to techniques explored in other branches of the literature (e.g., Hanneke, Larsen, and Zhivotovskiy, 2024b; Bousquet and Zhivotovskiy, 2021; Puchkin and Zhivotovskiy, 2022), discussed in Appendix A.

3.1 Refinement of the Lower-order Term for Some Classes

The AVID principle already suffices to achieve the query complexity bound in Theorem 1. Moreover, for *most* concept classes of interest, the query complexity bound in Theorem 1 is already *optimal*, matching a lower bound (up to log factors in the lower-order term): e.g., linear classifiers in $\mathbb{R}^k$, $k \geq 2$ (Dasgupta, 2005; Hanneke, 2014; Hanneke and Yang, 2015). However, while the lead term $\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)$ is already optimal for every concept class $\mathbb{C}$, there do exist some special classes $\mathbb{C}$ for which a further refinement of the *lower-order* term $\frac{\mathsf{d}}{\varepsilon}$ is possible (e.g., threshold classifiers $\mathbb{1}_{[a,\infty)}$ on $\mathbb{R}$). As our second main result, we provide a refinement of the upper bound in Theorem 1 to capture such special classes, thereby establishing a query complexity bound which is nearly optimal for *every* concept class.

Since such refinements are only possible for some concept classes, the expression of this refinement necessarily depends on an additional complexity measure of the class $\mathbb{C}$. We prove that the *optimal* lower-order term in the query complexity is well-captured by a quantity known as the *star number* of $\mathbb{C}$, introduced by Hanneke and Yang (2015). In particular, Hanneke and Yang (2015) showed that the star number precisely characterizes the optimal query complexity in the *realizable case* ($\beta = 0$); since this is a limiting case of agnostic learning, it is natural that this quantity plays a crucial role in characterizing the optimal lower-order term. The formal definition is as follows.

Definition 2. *For any concept class* $\mathbb{C}$*, the* star number $\mathfrak{s} = \mathfrak{s}(\mathbb{C})$ *is the supremum* $n \in \mathbb{N}$ *for which* $\exists x_1, \ldots, x_n \in \mathcal{X}$ *and* $h_0, h_1, \ldots, h_n \in \mathbb{C}$ *such that* $\forall i, j \in \{1, \ldots, n\}$, $h_i(x_j) \neq h_0(x_j) \Leftrightarrow i = j$.

The star number essentially describes a scenario which is intuitively challenging for active learners in the realizable case, wherein there is a set of instances x_j and a *default* labeling $h_0(x_j)$, but the *target* concept is some h_i which differs from h_0 at just *one* instance x_i, unknown to the learner (which must therefore query nearly all of these x_j instances, searching for the special point x_i, in order to

identify the target concept h_i). Hanneke and Yang (2015) provide numerous examples calculating $\mathfrak{s}$ for various concept classes. For instance, thresholds on $\mathbb{R}$ have $\mathfrak{s} = 2$ and decision stumps on $\mathbb{R}^k$ have $\mathfrak{s} = 2k$. However, it is worth noting that $\mathfrak{s}$ is typically large (or infinite) for most concept classes of interest in learning theory (e.g., $\mathfrak{s} = \infty$ for linear classifiers on $\mathbb{R}^k$, $k \geq 2$). This fact is important to the present work, since Hanneke and Yang (2015); Hanneke (2016b, 2024) have shown that the $c(\beta)$ factors (discussed in Section 2 above) appearing in all previous general upper bounds *all* become no smaller than $\mathfrak{s} \wedge \frac{1}{\beta}$ in the worst case over distributions (subject to the β constraint). Thus, all general upper bounds $c(\beta)\mathsf{d}\frac{\beta^2}{\varepsilon^2}$ from the prior literature become no smaller than $\mathsf{d}\frac{\beta}{\varepsilon^2}$ in the worst case when $\mathfrak{s} = \infty$. In a sense, this means Theorem 1 is actually *most* interesting in the (typical) case of $\mathfrak{s} = \infty$, since *no* previously known upper bounds offer any improvements over passive learning in this case (without further restrictions to P), in stark contrast to Theorem 1 which has *no dependence* on $\mathfrak{s}$ and provides improvements over passive learning in the lead term for *every* concept class.

Nonetheless, the special structure of classes with $\mathfrak{s} < \infty$ turns out to provide some additional advantages for active learning, so that in order to state a general query complexity bound which is optimal for *every* concept class $\mathbb{C}$, we need to account for this structure, via a dependence on $\mathfrak{s}$ in the lower-order term. Specifically, by combining the AVID principle with existing principles for active learning (namely, *disagreement-based* queries), we can take further advantage of the power of active learning, thereby enabling a refinement of the lower-order term for classes with $\mathfrak{s} < \infty$. The following result presents a new general query complexity bound reflecting such refinements, together with a known lower bound for comparison (due to Kääriäinen, 2006; Beygelzimer, Dasgupta, and Langford, 2009; Hanneke and Yang, 2015). In particular, the result shows that this new upper bound is nearly optimal for *every* concept class $\mathbb{C}$ (including the lower-order term, up to a factor of d, which we discuss below). A more-detailed version of the result appears in Theorem 5 of Appendix C (and distribution-dependent variants are presented in Appendix F, replacing $\mathfrak{s}$ with variants of the *disagreement coefficient*).

Theorem 3. *For every* $\mathbb{C}$, *letting* $\mathsf{d} = \mathrm{VC}(\mathbb{C})$ *and* $\mathfrak{s} = \mathfrak{s}(\mathbb{C})$, $\forall \varepsilon, \delta \in (0, 1/8)$, $\forall \beta \in [0, 1]$,

$$\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}) = O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right)\mathsf{d}\right),$$

and $$\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}) = \Omega\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right) + \mathfrak{s} \wedge \frac{1}{\varepsilon}\right).$$

We may note that the upper bound in Theorem 1 is an immediate implication of Theorem 3 (we have stated Theorem 1 separately merely to emphasize that the improvements over passive learning are available without any special properties of $\mathbb{C}$ such as finite star number). Theorem 3 provides a refinement in the lower-order term compared to Theorem 1 when $\mathfrak{s} < \frac{1}{\varepsilon}$. In particular, for $\mathfrak{s} < \infty$, the asymptotic dependence on ε in the lower-order term is $\log^2\left(\frac{1}{\varepsilon}\right)$. We leave open the question of whether this can be further refined to $\log\left(\frac{1}{\varepsilon}\right)$, which would match a known lower bound on this dependence for all infinite classes (Kulkarni, Mitter, and Tsitsiklis, 1993; Hanneke and Yang, 2015). The only significant difference between the upper and lower bounds in Theorem 3 is the factor of d in the lower-order term. I conjecture this term can be further refined to $\tilde{O}\left(\mathfrak{s} \wedge \frac{\mathsf{d}}{\varepsilon}\right)$, which is known to be sharp for some classes (Hanneke and Yang, 2015), and would fully answer a question posed by Hanneke and Nowak (2019). Beyond this, it is known that a gap between such lower-order terms in general upper and lower bounds is unavoidable if the only dependence on $\mathbb{C}$ is via d and $\mathfrak{s}$. Specifically, it follows from arguments in Appendix D of Hanneke and Yang (2015) that for some classes $\mathbb{C}$ this term should be $\tilde{\Theta}\left(\mathfrak{s} \wedge \frac{\mathsf{d}}{\varepsilon}\right)$ while for other classes $\mathbb{C}$ the term should be $\tilde{\Theta}\left(\mathfrak{s} \wedge \frac{1}{\varepsilon} + \mathsf{d}\right)$. Thus, obtaining matching (big-Θ) upper and lower bounds would require introducing a new complexity measure reflecting the distinctions between these types of classes, which we leave as an open question.

4 Algorithm and Outline of the Analysis

We next present the algorithm achieving Theorems 1 and 3 and a sketch of its analysis (the complete formal proof is given in Section E). Before stating the algorithm, we first introduce a few additional definitions and convenient notational conventions.

Error and disagreement regions: For any function $h : \mathcal{X} \to \{0,1\}$, define its *error region* $\mathrm{ER}(h) := \{(x,y) \in \mathcal{X} \times \{0,1\} : h(x) \neq y\}$. In particular, note that $\mathrm{er}_P(h) = P(\mathrm{ER}(h))$. For any non-empty set $\mathbb{C}'$ of functions $\mathcal{X} \to \{0,1\}$, define the *region of disagreement*:

$$\mathrm{DIS}(\mathbb{C}') := \{x \in \mathcal{X} : \exists f, g \in \mathbb{C}', f(x) \neq g(x)\}.$$

Also, for any two functions $f, g : \mathcal{X} \to \{0,1\}$, abbreviate by $\{f \neq g\} := \{x \in \mathcal{X} : f(x) \neq g(x)\}$ their *pairwise disagreement region*.

Overloaded set notation: For convenience, we adopt a convention of treating sets $A \subseteq \mathcal{X}$ as notationally interchangeable with their *labeled extension* $A \times \{0,1\} \subseteq \mathcal{X} \times \{0,1\}$. For instance, for functions $f, g, h : \mathcal{X} \to \{0,1\}$, we may write $\mathrm{ER}(h) \cap \{f \neq g\}$, which, by the above convention, is interpreted as $\mathrm{ER}(h) \cap (\{f \neq g\} \times \{0,1\})$. We also overload notation for set intersections to allow for intersections of sets with *sequences*: that is, for any set $\mathcal{Z}$, sequence $S = \{z_1, \ldots, z_m\} \in \mathcal{Z}^m$, and set $A \subseteq \mathcal{Z}$, we define $S \cap A$ as the *subsequence* $\{z_i : i \leq m, z_i \in A\}$, and likewise $S \setminus A := S \cap (\mathcal{Z} \setminus A)$. We also apply these conventions in combination: i.e., for a sequence $S \in (\mathcal{X} \times \{0,1\})^m$ and a set $\Delta \subseteq \mathcal{X}$, we define $S \cap \Delta := S \cap (\Delta \times \{0,1\})$ and $S \setminus \Delta := S \cap ((\mathcal{X} \setminus \Delta) \times \{0,1\})$.

Empirical estimates: We will make use of *empirical estimates* of quantities such as $\mathrm{er}_P(h)$ and $P_X(f \neq g)$. For any set $\mathcal{Z}$ and sequence $S = \{z_1, \ldots, z_m\} \in \mathcal{Z}^m$, for any set $A \subseteq \mathcal{Z}$, define the *empirical measure*: $\hat{P}_S(A) := \frac{1}{m}|S \cap A| = \frac{1}{m}\sum_{i=1}^m \mathbb{1}[z_i \in A]$. Again, we also apply these conventions in combination: i.e., for $S \in (\mathcal{X} \times \{0,1\})^*$ and $\Delta \subseteq \mathcal{X}$, we define $\hat{P}_S(\Delta) := \hat{P}_S(\Delta \times \{0,1\})$. For any sequence $S \in (\mathcal{X} \times \{0,1\})^*$ and function $h : \mathcal{X} \to \{0,1\}$, define its *empirical error rate* (or *empirical risk*): $\hat{\mathrm{er}}_S(h) := \hat{P}_S(\mathrm{ER}(h))$.

Decision lists: We will often express *decision-list* aggregations of functions $f, g : \mathcal{X} \to \{0,1\}$. For instance, for any set $\Delta \subseteq \mathcal{X}$, we may write $h = f\mathbb{1}_{\mathcal{X}\setminus\Delta} + g\mathbb{1}_\Delta$ to express a function h with $h(x) = f(x)$ for $x \notin \Delta$ and $h(x) = g(x)$ for $x \in \Delta$.

4.1 The AVID Agnostic Algorithm: Adaptive Localized Variance Isolation by Disagreements

We are now ready to describe the algorithm achieving the upper bounds in Theorems 1 and 3 (for full formality, some additional technical minutiae for the definition are given in Section C). Fix any values $\varepsilon, \delta \in (0,1)$ (the error and confidence parameters input to the learner). Fix any distribution P (unknown to the learner) and let $(X_1, Y_1), \ldots, (X_m, Y_m)$ be independent P-distributed random variables (for any sufficiently large m, quantified explicitly in Theorem 5). The algorithm is stated in Figure 1, expressed in terms of certain quantities and data subsets defined as follows.[2] Let $C := \frac{11}{10}$, $N := \left\lceil \log_C\left(\frac{2}{\varepsilon}\right)\right\rceil$, and for each $k \in \mathbb{N}$ define $\varepsilon_k := C^{1-k}$ and $m_k := \Theta\left(\frac{1}{\varepsilon_k}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon_k}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$ (see Section C for the precise value). In Step 3, C' denotes an appropriate universal constant (see Section C). As defined in Figure 1, the algorithm makes use of different portions of the data (S_k^1, S_k^2, $S_{k,i}^3$, S_k^4) for different purposes, and to complete the definition of the algorithm we next specify how these data subsets are defined in the algorithm. We first split the initial $2M_1 := 2\sum_{k=1}^{N+1} m_k$ examples $\{(X_1, Y_1), \ldots, (X_{2M_1}, Y_{2M_1})\}$ into consecutive disjoint contiguous segments $S_1^1, \ldots, S_{N+1}^1, S_1^4, \ldots, S_{N+1}^4$, with the segments S_k^1 and S_k^4 being of size m_k. The algorithm also allocates disjoint segments (S_k^2, $S_{k,i}^3$) of the remaining data $\{(X_i, Y_i) : 2M_1 < i \leq m\}$, but does so *adaptively* during its execution. Specifically, if and when the algorithm reaches Step 2 with a value k, or reaches Step 9 (in which case let $k = N+1$), for the value i_k and the set Δ_{i_k} as defined at that time in the algorithm, it constructs a data subset S_k^2, allocating to S_k^2 the next m_k' consecutive examples which have not yet been allocated to any data subset $S_{k'}^1$, $S_{k'}^2$, $S_{k',i'}^3$, $S_{k'}^4$ (i.e., *fresh*, previously-unused, examples), where, letting $\hat{p}_k := 2\hat{P}_{S_k^4}(\Delta_{i_k})$, we define $m_k' := \Theta\left(\frac{\hat{p}_k}{\varepsilon_k^2}\left(\mathsf{d} + \log\left(\frac{3+N-k}{\delta}\right)\right)\right)$ (see Section C for the precise value of m_k'). Similarly, if and when the algorithm reaches Step 5 with some values of (k,i), it constructs a data subset $S_{k,i}^3$, allocating to $S_{k,i}^3$ the next m_k consecutive examples which have not yet been allocated.

[2] For simplicity, we have expressed the algorithm as representing a set of surviving concepts $V_k \subseteq \mathbb{C}$. However, it should be clear from the definition that running the algorithm does not require explicitly storing V_k. Rather, the various uses of this set can be implemented as constrained optimization problems (in Steps 4-6 and

Algorithm $\mathbb{A}_{\text{avid}}$
Input: Error parameter ε, Confidence parameter δ, Unlabeled data $X_1, \ldots, X_m$
Output: Classifier $\hat{h}$
0. Initialize $i = i_1 = 0$, $\Delta_0 = \emptyset$, $V_0 = \mathbb{C}$
1. For $k = 1, \ldots, N$
2. Query all examples in $S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$ and $S_k^2 \cap \Delta_{i_k}$
3. $V_k \leftarrow \left\{h \in V_{k-1} : \hat{\text{er}}_k^{1,2}(h) \leq \hat{\text{er}}_k^{1,2}(\hat{h}_k) + \frac{\varepsilon_k}{C'}\right\}$
4. If $V_k = \emptyset$ or $\hat{\text{er}}_k^{1,2}(\hat{h}_k) < \min_{h \in V_k} \hat{\text{er}}_k^{1,2}(h) - \frac{\varepsilon_k}{4C'}$, Then Return $\hat{h} := \hat{h}_k$
5. While $\max_{f,g \in V_k} \hat{P}_{S_{k,i}^3}(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+2}$
6. $(f, g) \leftarrow \text{argmax}_{(f',g') \in V_k^2} \hat{P}_{S_{k,i}^3}(\{f' \neq g'\} \setminus \Delta_i)$
7. $\Delta_{i+1} \leftarrow \Delta_i \cup \{f \neq g\}$, and update $i \leftarrow i + 1$
8. $i_{k+1} \leftarrow i$
9. Query all examples in $S_{N+1}^1 \cap D_N \setminus \Delta_{i_{N+1}}$ and $S_{N+1}^2 \cap \Delta_{i_{N+1}}$ and Return $\hat{h} := \hat{h}_{N+1}$

Figure 1: The AVID Agnostic algorithm. Notations N, D_{k-1}, ε_k, $\hat{h}_k$, S_k^1, S_k^2, $S_{k,i}^3$, $\hat{\text{er}}_k^{1,2}$ defined in the text.

To complete the definition of the algorithm, we define D_{k-1}, $\hat{\text{er}}_k^{1,2}$, and $\hat{h}_k$, appearing in the algorithm, as follows. For each value of k encountered in the '*For*' loop, as well as for $k = N + 1$ in the case the algorithm reaches Step 9, define (where V_{k-1} and Δ_{i_k} are as defined in the algorithm):

$$D_{k-1} := \text{DIS}(V_{k-1}),$$

$$\forall h,\ \hat{\text{er}}_k^{1,2}(h) := \hat{P}_{S_k^1}(\text{ER}(h) \cap D_{k-1} \setminus \Delta_{i_k}) + \hat{P}_{S_k^2}(\text{ER}(h) \cap \Delta_{i_k}), \tag{1}$$

$$V_{k-1}^{(4)} := \{f \mathbb{1}_{\{f=g\} \setminus \Delta_{i_k}} + h_1 \mathbb{1}_{\{f \neq g\} \setminus \Delta_{i_k}} + h_2 \mathbb{1}_{\Delta_{i_k}} : f, g \in V_{k-1}, h_1, h_2 \in \mathbb{C}\}, \tag{2}$$

$$\text{and} \quad \hat{h}_k := \operatorname*{argmin}_{h \in V_{k-1}^{(4)}} \hat{\text{er}}_k^{1,2}(h). \tag{3}$$

This completes the definition of the $\mathbb{A}_{\text{avid}}$ algorithm.

We remark that the examples in $S_{k,i}^3$ and S_k^4 are *never queried* in the algorithm, and thus the algorithm (necessarily) only uses the unlabeled X_i values in these data subsets (to estimate certain marginal P_X probabilities), so in fact these can be regarded as *unlabeled* data subsets. Similarly, the algorithm only queries a *portion* of S_k^1 and S_k^2, and the remaining unqueried portions are in fact *never used* by the algorithm. For notational simplicity, we do not make these facts explicit in the notation.

Description of the algorithm: We briefly summarize the behavior of the algorithm, as follows (with explanations following in Section 4.2). As the algorithm iterates over rounds k of the '*For*' loop, it maintains a partition of the space into a region Δ_{i_k} and its complement $\mathcal{X} \setminus \Delta_{i_k}$. In each round, the algorithm refines a set V_k of surviving concepts from $\mathbb{C}$, by pruning out concepts h having large estimated difference of error rate compared to a function $\hat{h}_k$ (Step 3). There are several crucial aspects of this, both in how these estimates of $\text{er}_P(h) - \text{er}_P(\hat{h}_k)$ are defined, and in the choice of function $\hat{h}_k$ we compare to. For the purpose of estimating the error differences $\text{er}_P(h) - \text{er}_P(\hat{h}_k)$, in Step 2 it queries a number of random examples in $\mathcal{X} \setminus \Delta_{i_k}$ (or rather, the slightly smaller region $D_{k-1} \setminus \Delta_{i_k}$, since examples in $\mathcal{X} \setminus D_{k-1}$ are uninformative for estimating these error differences) and a number of random examples in Δ_{i_k}. It then prunes suboptimal concepts from V_{k-1} using estimates of the error differences $\text{er}_P(h) - \text{er}_P(\hat{h}_k)$ defined by the empirical difference $\hat{\text{er}}_k^{1,2}(h) - \hat{\text{er}}_k^{1,2}(\hat{h}_k)$ (Step 3), where $\hat{\text{er}}_k^{1,2}(h)$ uses the examples we queried in each of the two regions to estimate the error rate in that respective region, as defined in (1). The reason for this definition is that, as it will turn out, we require a disproportionately larger number of samples to accurately estimate the difference of error rates in the region Δ_{i_k} compared to the region $D_{k-1} \setminus \Delta_{i_k}$: namely, for the latter, we use the samples in $S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$ (queried in Step 2), where S_k^1 has a modest size $m_k = \tilde{\Theta}\left(\frac{\text{d}}{\varepsilon_k}\right)$, while for the

$\hat{h}_k$), where the constraints are merely the inequalities which would define the sets $V_{k'}$, $k' \leq k$, and Step 3 is then replaced by simply adding one more constraint to the constraint set.

former we use the samples in $S_k^2 \cap \Delta_{i_k}$ (also queried in Step 2), where S_k^2 has a potentially larger size m'_k which is roughly $\tilde{\Theta}\left(\frac{P_X(\Delta_{i_k})\mathsf{d}}{\varepsilon_k^2}\right)$. The other crucial aspect is how we define the function $\hat{h}_k$ we compare to. For this, rather than simply comparing to the best $\hat{\mathrm{er}}_k^{1,2}(h)$ among $h \in V_{k-1}$, we instead compare to an even smaller value: the best $\hat{\mathrm{er}}_k^{1,2}(h)$ among a more-complex class of functions $V_{k-1}^{(4)}$, comprised of *decision list* functions which use some concept h_2 for predictions in Δ_{i_k}, and use a separate function for predictions in $\mathcal{X} \setminus \Delta_{i_k}$, where the latter is specified by three concepts f, g, h_1, as described in (2) (equivalently, it predicts with a majority vote of f, g, h_1 in $\mathcal{X} \setminus \Delta_{i_k}$). $\hat{h}_k$ is defined as a minimizer of $\hat{\mathrm{er}}_k^{1,2}(h)$ among $h \in V_{k-1}^{(4)}$, as in (3). Given that $\hat{h}_k$ is chosen from a more-complex class, it is possible that, after the update in Step 3, there may be *no* surviving concepts in V_k. In this event (or if $\hat{h}_k$ is at least somewhat better than all surviving concepts in V_k), the algorithm terminates early and returns $\hat{h}_k$ (Step 4). Otherwise, if it makes it past this early-stopping case, its next objective is to define the region $\Delta_{i_{k+1}}$ for use in the next iteration. This occurs in the '*While*' loop (Steps 5-7). On each round of this loop, it uses a fresh data set $S_{k,i}^3$ of size $m_k = \tilde{\Theta}\left(\frac{\mathsf{d}}{\varepsilon_k}\right)$ to check whether there exist $f, g \in V_k$ significantly *distant* from each other in the region $\mathcal{X} \setminus \Delta_i$ (Step 5). If so, it adds their pairwise disagreement region $\{f \neq g\}$ to the Δ_i region to define Δ_{i+1} and increments i (Step 7). It repeats this until no such pair f, g exists, at which time it defines $i_{k+1} = i$ (Step 8) and proceeds to the next iteration of the '*For*' loop. After $N = O(\log(1/\varepsilon))$ such iterations, it returns $\hat{h}_{N+1}$ (Step 9).

We note that the algorithm's returned classifier $\hat{h}$ might *not* be an element of $\mathbb{C}$ (known as an *improper* learner), but rather can be represented as a (shallow) *decision list* of concepts from $\mathbb{C}$. This aspect is quite important to certain parts of the proof, and we leave open the question of whether Theorems 1 and 3 are achievable by a proper learner (see Appendix G). We also remark that the D_{k-1} set is *only* needed for establishing Theorem 3: the algorithm achieves the query complexity bound in Theorem 1 even if we replace D_{k-1} with the full space $\mathcal{X}$ everywhere.

4.2 Principles and Outline of the Proof

Next we explain the high-level principles underlying the design of the algorithm, highlighting the *two key innovations* compared to previous approaches, which enable the improved query complexity guarantee (namely, separating out the Δ_{i_k} regions, and the definition of $\hat{h}_k$).

Empirical localization: The principles underlying the design of the algorithm begin with a familiar principle from statistical learning: *empirical localization* (Koltchinskii, 2006; Bartlett, Bousquet, and Mendelson, 2005). Specifically, the uniform Bernstein inequality (Lemma 7) implies that for an i.i.d. data set S, the sample complexity of uniform concentration of differences $|(\hat{\mathrm{er}}_S(f) - \hat{\mathrm{er}}_S(g)) - (\mathrm{er}_P(f) - \mathrm{er}_P(g))|$ becomes smaller when the *diameter* $\mathrm{diam}(\mathbb{C})$ of the concept class is small, measured under the pairwise-disagreement pseudo-metric $P_X(f \neq g)$ (bounding the *variance* of loss differences $\mathbb{1}[f(x) \neq y] - \mathbb{1}[g(x) \neq y]$). Quantitatively (for $0 < \varepsilon < \mathrm{diam}(\mathbb{C})$), $\tilde{\Theta}\left(\mathsf{d}\frac{\mathrm{diam}(\mathbb{C})}{\varepsilon^2}\right)$ samples suffice to guarantee $|(\hat{\mathrm{er}}_S(f) - \hat{\mathrm{er}}_S(g)) - (\mathrm{er}_P(f) - \mathrm{er}_P(g))| < \varepsilon$. This fact leads to a natural well-known algorithmic principle, wherein we can *prune* from $\mathbb{C}$ concepts h having $\hat{\mathrm{er}}_S(h) - \min_{h' \in \mathbb{C}} \hat{\mathrm{er}}_S(h') > \varepsilon$ (as the above inequality implies these verifiably have suboptimal error rates), leaving a subset V'_1 of surviving concepts, while preserving $h^\star \in V'_1$, where $h^\star := \mathrm{argmin}_{h \in \mathbb{C}} \mathrm{er}_P(h)$. If $\mathrm{diam}(V'_1)$ is *smaller* than $\mathrm{diam}(\mathbb{C})$, this *improves* the resolution of the uniform Bernstein inequality, which enables us to prune *even more* concepts from V'_1, leaving a set V'_2 of surviving concepts, and so on for $V'_3, V'_4, \ldots$. Quantitatively, we can combine this with a schedule of resolutions ε_k, so that as long as $h^\star \in V'_{k-1}$ and $\mathrm{diam}(V'_{k-1}) \leq \varepsilon_k$, an i.i.d. data set S_k^1 of size $m_k = \tilde{\Theta}\left(\frac{\mathsf{d}}{\varepsilon_k}\right)$ suffices to reduce to a subset $V'_k = \left\{h \in V'_{k-1} : \hat{\mathrm{er}}_{S_k^1}(h) \leq \min_{h' \in V'_{k-1}} \hat{\mathrm{er}}_{S_k^1}(h') + \frac{\varepsilon_k}{C'}\right\}$ for which all $h \in V'_k$ have $\mathrm{er}_P(h) \leq \mathrm{er}_P(h^\star) + 2\frac{\varepsilon_k}{C'}$, while preserving $h^\star \in V'_k$. Iterating this $N = \Theta\left(\log_C\left(\frac{1}{\varepsilon}\right)\right)$ times results in a subset V'_N of concepts with excess error ε (recalling $\varepsilon_k = C^{1-k}$).

Disagreement-based active learning: An additional observation, underlying many active learning algorithms (*disagreement-based* methods), is that the above argument still holds while replacing $\hat{\mathrm{er}}_{S_k^1}(h)$ with $\hat{P}_{S_k^1}(\mathrm{ER}(h) \cap D'_{k-1})$, where $D'_{k-1} := \mathrm{DIS}(V'_{k-1})$. To see this, note that $\forall h, h' \in V'_{k-1}$, $\hat{P}_{S_k^1}(\mathrm{ER}(h) \cap D'_{k-1}) - \hat{P}_{S_k^1}(\mathrm{ER}(h') \cap D'_{k-1}) = \hat{\mathrm{er}}_{S_k^1}(h) - \hat{\mathrm{er}}_{S_k^1}(h')$. Thus, we may equivalently define $V'_k = \left\{h \in V'_{k-1} : \hat{P}_{S_k^1}(\mathrm{ER}(h) \cap D'_{k-1}) \leq \min_{h' \in V'_{k-1}} \hat{P}_{S_k^1}(\mathrm{ER}(h') \cap D'_{k-1}) + \frac{\varepsilon_k}{C'}\right\}$. Moreover,

as long as $\operatorname{diam}(V'_{k-1}) \le \varepsilon_k$, we have $P_X(D'_{k-1}) \le \mathfrak{s}\varepsilon_k$ (Hanneke and Yang, 2015). Since the quantities in V'_k only rely on the labels of examples in $D'_{k-1} \cap S^1_k$, constructing V'_k only requires a number of queries $O(\mathfrak{s}\varepsilon_k m_k) \wedge m_k$. Summing over k, these queries total to at most the claimed lower-order term in Theorem 3 (though note that even without this D'_{k-1} refinement we still recover the lower-order term from Theorem 1). So far, this is all essentially standard reasoning commonly followed in the prior literature on active learning (e.g., Hanneke, 2009b, 2014; Koltchinskii, 2010).

Handling non-shrinking diameter: However, the above algorithmic principle breaks down if we reach a k with $\operatorname{diam}(V'_{k-1}) \neq O(\varepsilon_k)$. This failure can easily occur in the agnostic setting, where it is possible for the set V'_{k-1} above to contain multiple relatively-good functions f, g which are nevertheless *far* from each other.[3] This is the motivation for the *first key innovation* in $\mathbb{A}_{\text{avid}}$: namely, if we ever reach such a k, where the V_k set does not naturally have $\operatorname{diam}(V_k) \le \varepsilon_{k+1}$ (as tested in Step 5), the algorithm *removes* a portion of the space $\mathcal{X}$ to *artificially* reduce the diameter. Specifically, it identifies a pair $f, g \in V_k$ with $P_X(f \neq g) > \varepsilon_{k+1}$ (intuitively, an *obstruction* to having low diameter) and *separates out* their pairwise disagreement region $\{f \neq g\}$ from the region of focus of the algorithm (Steps 5-7).[4] Having set aside this region, the algorithm continues, focusing on the remaining set $\mathcal{X} \setminus \{f \neq g\}$. This step is repeated, and these set-aside regions $\{f \neq g\}$ are altogether captured in the set Δ_i (Step 7). Thus, we repeatedly find pairs $f, g \in V_k$ with $P_X(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+1}$ (Steps 5-6) and add $\{f \neq g\}$ to Δ_i (Step 7) until the diameter of V_k on $\mathcal{X} \setminus \Delta_i$ is reduced below ε_{k+1}. At that point, the algorithm proceeds to the next round ($k \leftarrow k+1$). On the next round k, since we have (artificially) ensured the diameter of V_{k-1} is at most ε_k in the region $\mathcal{X} \setminus \Delta_{i_k}$, the uniform Bernstein argument implies m_k examples S^1_k suffice to guarantee every $f, g \in V_{k-1}$ have $\hat{P}_{S^1_k}(\mathrm{ER}(f) \cap D_{k-1} \setminus \Delta_{i_k}) - \hat{P}_{S^1_k}(\mathrm{ER}(g) \cap D_{k-1} \setminus \Delta_{i_k})$ within $\pm\frac{\varepsilon_k}{2C'}$ of $P(\mathrm{ER}(f) \setminus \Delta_{i_k}) - P(\mathrm{ER}(g) \setminus \Delta_{i_k})\big)$.

Error in the $\boldsymbol{\Delta_{i_k}}$ region: There remains the issue of estimating error rates in the Δ_{i_k} *isolated* region. For this, the algorithm uses a data set S^2_k of size $m'_k \approx \mathsf{d}\frac{P_X(\Delta_{i_k})}{\varepsilon_k^2}$, queries all examples in $S^2_k \cap \Delta_{i_k}$, and uses these to estimate the error rates $P(\mathrm{ER}(h) \cap \Delta_{i_k})$ in the Δ_{i_k} region. By a refinement of the uniform convergence bound of Talagrand (1994) which accounts for an *envelope* set Δ_{i_k} (Lemma 8), this number m'_k of examples suffices to ensure $\left|\hat{P}_{S^2_k}(\mathrm{ER}(h) \cap \Delta_{i_k}) - P(\mathrm{ER}(h) \cap \Delta_{i_k})\right| \le \frac{\varepsilon_k}{4C'}$ for every $h \in \mathbb{C}$. Combining this with the above error-differences estimates in the $\mathcal{X} \setminus \Delta_{i_k}$ region, we can guarantee that the functions $f, g \in V_{k-1}$ have $\left|\big(\hat{\mathrm{er}}^{1,2}_k(f) - \hat{\mathrm{er}}^{1,2}_k(g)\big) - \big(\mathrm{er}_P(f) - \mathrm{er}_P(g)\big)\right| \le \frac{\varepsilon_k}{C'}$, recalling the definition of $\hat{\mathrm{er}}^{1,2}_k$ from (1). Altogether, we conclude that a set $V''_k = \left\{h \in V_{k-1} : \hat{\mathrm{er}}^{1,2}_k(h) \le \min_{h' \in V'_{k-1}} \hat{\mathrm{er}}^{1,2}_k(h') + \frac{\varepsilon_k}{C'}\right\}$ would contain only functions h with $\mathrm{er}_P(h) - \mathrm{er}_P(h^\star) \le 2\frac{\varepsilon_k}{C'}$ while preserving $h^\star \in V''_k$. The actual definition of V_k in Step 3 is only slightly different from this, for reasons we discuss next.

Bounding the size of $\boldsymbol{\Delta_{i_k}}$: Since the number of queries in $S^2_k \cap \Delta_{i_k}$ is $\approx \mathsf{d}\frac{P_X(\Delta_{i_k})^2}{\varepsilon_k^2}$, if we hope to achieve a small query complexity it is crucial to bound the size $P_X(\Delta_{i_k})$ of the Δ_{i_k} region. This is the motivation for the *second key innovation* in $\mathbb{A}_{\text{avid}}$: defining the update in V_k by comparison to the function $\hat{h}_k$ in (3), rather than the best $h \in V_{k-1}$. This turns out to be the most subtle part of the argument, requiring precise choices in the design of the algorithm. The essential argument is as follows. Suppose the algorithm reaches Step 6 for some (k, i), so that it will add $\{f \neq g\}$ to the Δ_i region. We then want to argue that $P(\mathrm{ER}(h^\star) \cap \{f \neq g\} \setminus \Delta_i) = \Omega(P(\{f \neq g\} \setminus \Delta_i))$: that is, each time we add to Δ_i, we *chop off* a portion of $\mathrm{ER}(h^\star)$ of size proportional to the increase in Δ_i. Clearly if we can show this is always the case, we will inductively maintain $P_X(\Delta_i) = O(\beta)$, resulting in the claimed leading term in the query complexity bounds in Theorems 1 and 3. Now, to show this indeed occurs, we first note that one of f, g must err on at least *half* of $\{f \neq g\} \setminus \Delta_{i_k}$; w.l.o.g. suppose it is f: that is, $P(\mathrm{ER}(f) \cap \{f \neq g\} \setminus \Delta_{i_k}) \ge \frac{1}{2}P_X(\{f \neq g\} \setminus \Delta_{i_k})$. Now consider a function

[3] For instance, for $\mathbb{C}$ the class of *intervals* $\mathbb{1}_{[a,b]}$ on $\mathbb{R}$, with $P_X = \mathrm{Uniform}([0,1])$ and $P(Y=1|X) = \mathbb{1}_{[0,1/4]\cup[3/4,1]}(X)$, the concepts $\mathbb{1}_{[0,1/4]}$ and $\mathbb{1}_{[3/4,1]}$ are both optimal among $\mathbb{C}$, yet distance $1/2$ apart.

[4] This reasoning is somewhat reminiscent of the motivation for the *splitting* approach to active learning (Dasgupta, 2005), differing only in how we resolve the obstruction: whereas splitting would resolve it with queries to eliminate one element from each obstructing pair, here we resolve it by subtracting the pairwise disagreement region from the region of focus $\mathcal{X} \setminus \Delta_i$. This idea is also related to a technique of Hanneke, Larsen, and Zhivotovskiy (2024b) for agnostic passive learning, discussed in Appendix A.

$f^\star = f\mathbb{1}_{\{f=g\}\setminus\Delta_{i_k}} + h^\star\mathbb{1}_{\{f\neq g\}\setminus\Delta_{i_k}} + f\mathbb{1}_{\Delta_{i_k}}$ which replaces f by $h^\star$ in the region $\{f \neq g\} \setminus \Delta_{i_k}$. Note that $f^\star \in V_{k-1}^{(4)}$ defined in (2). Since $\hat{h}_k$ has minimal $\hat{\mathrm{er}}_k^{1,2}$ among $V_{k-1}^{(4)}$, and $f \in V_k$ implies $\hat{\mathrm{er}}_k^{1,2}(f) \leq \hat{\mathrm{er}}_k^{1,2}(\hat{h}_k) + \frac{\varepsilon_k}{C'}$, extending the above concentration of $\hat{\mathrm{er}}_k^{1,2}$ differences to functions in $V_{k-1}^{(4)}$ (with appropriate adjustment of constants in m_k, m_k') implies $\mathrm{er}_P(f) - \mathrm{er}_P(f^\star) \leq 2\frac{\varepsilon_k}{C'}$. Thus,

$$\begin{aligned}
&\frac{1}{2}P_X(\{f \neq g\} \setminus \Delta_{i_k}) - P(\mathrm{ER}(h^\star) \cap \{f \neq g\} \setminus \Delta_{i_k}) \\
&\leq P(\mathrm{ER}(f) \cap \{f \neq g\} \setminus \Delta_{i_k}) - P(\mathrm{ER}(h^\star) \cap \{f \neq g\} \setminus \Delta_{i_k}) = \mathrm{er}_P(f) - \mathrm{er}_P(f^\star) \leq 2\frac{\varepsilon_k}{C'}.
\end{aligned}$$

In other words, $P(\mathrm{ER}(h^\star) \cap \{f \neq g\} \setminus \Delta_{i_k}) \geq \frac{1}{2}P_X(\{f \neq g\} \setminus \Delta_{i_k}) - 2\frac{\varepsilon_k}{C'}$. This is almost what we wanted, aside from having Δ_{i_k} in place of Δ_i. We then argue $P(\mathrm{ER}(h^\star) \cap \{f \neq g\} \setminus \Delta_i) \geq P(\mathrm{ER}(h^\star) \cap \{f \neq g\} \setminus \Delta_{i_k}) - P_X(\{f \neq g\} \setminus \Delta_{i_k}) + P_X(\{f \neq g\} \setminus \Delta_i)$, which (by the above) is at least $P_X(\{f \neq g\} \setminus \Delta_i) - \frac{1}{2}P_X(\{f \neq g\} \setminus \Delta_{i_k}) - 2\frac{\varepsilon_k}{C'}$. Since $f, g \in V_{k-1}$, we know $P_X(\{f \neq g\} \setminus \Delta_{i_k}) \leq \varepsilon_k$, so that this lower-bound is at least $P_X(\{f \neq g\} \setminus \Delta_i) - \frac{\varepsilon_k}{2} - 2\frac{\varepsilon_k}{C'}$. On the other hand, the condition in Step 5 ensures $P_X(\{f \neq g\} \setminus \Delta_i) \geq \frac{\varepsilon_k}{c}$ for a constant $c > C^2$. Thus, we have $P(\mathrm{ER}(h^\star) \cap \{f \neq g\} \setminus \Delta_i) \geq \left(1 - \frac{c}{2} - 2\frac{c}{C'}\right) P_X(\{f \neq g\} \setminus \Delta_i)$. For appropriately large C' and constants in m_k, m_k', we can obtain this with $c \leq \frac{3}{2} \wedge \frac{C'}{9}$, so that the factor $1 - \frac{c}{2} - 2\frac{c}{C'} > 0$.

The early stopping case: Since $\hat{h}_k$ is a more-complex function than the functions in V_{k-1}, there is a chance that V_{k-1} might simply not have *any* relatively good functions in it. For this reason, we have added the early stopping case in Step 4. In this event (using slightly tighter concentration inequalities), we have effectively verified that $\hat{h}_k$ is even *better* than $h^\star$, and we can safely return $\hat{h}_k$.

Overall behavior: The effective overall behavior of the algorithm is to *isolate* in the region Δ_{i_k} the most-challenging part of the error estimation problem, due to the high variance (diameter) of the error differences in that region. It then allocates a disproportionately larger number of queries $S_k^2 \cap \Delta_{i_k}$ to this region, toward estimating the error rates there. By comparing with the function $\hat{h}_k$ (which separately optimizes errors in pairwise difference regions $\{f \neq g\} \setminus \Delta_{i_k}$) in the definition of V_k, we can maintain that Δ_{i_k} never grows larger than $O(\beta)$, so that the number of queries in $S_k^2 \cap \Delta_{i_k}$ does not grow excessively large. The remaining region $\mathcal{X} \setminus \Delta_{i_k}$ enjoys the property that the set V_{k-1} has diameter $\leq \varepsilon_k$, so that we can easily estimate error differences in this region by a uniform Bernstein inequality. Altogether, after at most $N = O\left(\log\left(\frac{1}{\varepsilon}\right)\right)$ rounds, this achieves the objective of ε excess error rate, while using a number of queries as stated in the query complexity bound in Theorem 3.

The formal proof is given in Appendix E.

5 Conclusions and Summary of the Appendices

This work establishes a new sharp upper bound on the first-order query complexity of agnostic active learning. The leading term is smaller than that of passive learning by a factor proportional to the best-in-class error rate. This reveals a heretofore unknown fact that *every* concept class benefits from active learning in the non-realizable case.

The appendices include the formal proofs, along with additional contents. Appendix A presents a thorough summary of related work and background on the theory of active learning, as well as other works with techniques related to those used here. Appendix B introduces additional definitions and notation needed for the formal proofs. Appendix C presents remaining minutiae for the definition of $\mathbb{A}_{\mathrm{avid}}$, along with a more-detailed version of Theorem 3, including formal claims regarding the number of *unlabeled* examples. Appendix D presents useful concentration inequalities important to the proof. Appendix E presents the formal proof of Theorem 3. Appendix F presents distribution-dependent refinements of Theorem 3, which replace the star number $\mathfrak{s}$ with certain P-dependent complexity measures: variants of the disagreement coefficient. We further argue that the disagreement coefficient $\theta_P(\varepsilon)$, as originally defined by Hanneke (2007b), provably *cannot* be attained as a replacement for $\mathfrak{s}$ in the lower-order term (by any algorithm), while on the other hand $\mathbb{A}_{\mathrm{avid}}$ *does* achieve a lower-order term $\tilde{O}(\theta_P(\beta + \varepsilon)^2\mathsf{d})$. We also present subregion-based refinements of the algorithm and analysis, based on techniques of Zhang and Chaudhuri (2014). Appendix G presents extensions (*multiclass* classification, *stream-based* active learning), along with several open questions and future directions.

References

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

Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1987. doi: 10.1007/BF00116828. URL https://doi.org/10.1007/BF00116828.

J. Asilis, S. Devic, S. Dughmi V. Sharan, and S.-H. Teng. Proper learnability and the role of unlabeled data. In *Proceedings of the* 36^{th} *International Conference on Algorithmic Learning Theory*, 2025a.

J. Asilis, M. M. Høgsgaard, and G. Velegkas. Understanding aggregations of proper learners in multiclass classification. In *Proceedings of the* 36^{th} *International Conference on Algorithmic Learning Theory*, 2025b.

P. Awasthi, V. Feldman, and V. Kanade. Learning using local membership queries. In *Proceedings of the* 26^{th} *Conference on Learning Theory*, 2013.

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.

M.-F. Balcan and A. Blum. A discriminative model for semi-supervised learning. *Journal of the ACM*, 57(3):1–46, 2010.

M.-F. Balcan and S. Hanneke. Robust interactive learning. In *Proceedings of the* 25^{th} *Conference on Learning Theory*, 2012.

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.

M.-F. Balcan and H. Zhang. Sample and computationally efficient learning algorithms under s-concave distributions. 2017.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *NIPS Workshop on Foundations of Active Learning*, 2005.

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

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

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

M.-F. Balcan, S. Hanneke, and J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, 2010.

M.-F. Balcan, A. Blum, S. Hanneke, and D. Sharma. Robustly-reliable learners under poisoning attacks. In *Proceedings of the* 35^{th} *Conference on Learning Theory*, 2022.

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

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

E. Baum. Neural net algorithms that learn in polynomial time from examples and queries. *IEEE Transactions on Neural Networks*, 2(1):5–19, 1991.

E. Baum and K. Lang. Query learning can work poorly when a human oracle is used. In *Proceedings of the International Joint Conference in Neural Networks*, 1992.

G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

S. Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Annales Scientifiques de l'Institut de la Société des Savants d'Ukraine, Section de Mathématiques*, 1(4):38–49, 1924.

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the* 26^{th} *International Conference on Machine Learning*, 2009.

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems 23*, 2010.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

O. Bousquet and N. Zhivotovskiy. Fast classification rates without standard margin assumptions. *Information and Inference: A Journal of the IMA*, 10(4):1389–1421, 2021.

O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Proceedings of the* 33^{rd} *Conference on Learning Theory*, 2020.

N. Brukhim, D. Carmon, I. Dinur, S. Moran, and A. Yehudayoff. A characterization of multiclass learnability. In *Proceedings of the* 63^{rd} *Annual IEEE Symposium on Foundations of Computer Science*, 2022.

G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Learning noisy linear classifiers via adaptive and selective sampling. *Machine Learning*, 83:71–102, 2011.

O. Chapelle, B. Scholkopf, and A. Zien. *Semi-Supervised Learning*. Adaptive Computation and Machine Learning Series. MIT Press, 2006.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

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

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and N. Zhang. Active learning with region graphs. In *Proceedings of the* 36^{th} *International Conference on Machine Learning*, 2019a.

C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and N. Zhang. Region-based active learning. In *Proceedings of the* 22^{nd} *International Conference on Artificial Intelligence and Statistics*, 2019b.

C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and N. Zhang. Active learning with disagreement graphs. In *Proceedings of the* 36^{th} *International Conference on Machine Learning*, 2019c.

C. Cortes, G. DeSalvo, C. Gentile, M. Mohri, and N. Zhang. Adaptive region-based active learning. In *Proceedings of the* 37^{th} *International Conference on Machine Learning*, 2020.

A. Daniely and S. Shalev-Shwartz. Optimal learners for multiclass problems. In *Proceedings of the* 27^{th} *Conference on Learning Theory*, 2014.

S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems 17*, 2004.

S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, 2005.

S. Dasgupta. The two faces of active learning. *Theoretical Computer Science*, 412(19), 2011.

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

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

O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13(9):2655–2697, 2012.

G. DeSalvo, C. Gentile, and T. S. Thune. Online active learning with surrogate loss functions. *Advances in Neural Information Processing Systems 34*, 2021.

L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28:1011–1018, 1995.

I. Diakonikolas, D. Kane, and M. Ma. Active learning of general halfspaces: Label queries vs membership queries. In *Advances in Neural Information Processing Systems 37*, 2024.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

S. Efromovich. Sequential design and estimation in heteroscedastic nonparametric regression. *Sequential Analysis*, 26(1):3–25, 2007.

B. Eisenberg. *On the Sample Complexity of PAC-Learning using Random and Chosen Examples*. PhD thesis, Massachusetts Institute of Technology, 1992.

B. Eisenberg and R. Rivest. On the sample complexity of PAC-learning using random and chosen examples. In *Proceedings of the 3*$^{\mathrm{rd}}$ *Annual Workshop on Computational Learning Theory*, 1990.

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

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

D. J. Foster, A. Rakhlin, D. Simchi-Levi, and Y. Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Proceedings of the* 34^{th} *Conference on Learning Theory*, 2021.

Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

E. Friedman. Active learning for smooth problems. In *Proceedings of the* 22^{nd} *Conference on Learning Theory*, 2009.

R. Gelbhart and R. El-Yaniv. The relationship between agnostic selective classification, active learning and the disagreement coefficient. *Journal of Machine Learning Research*, 20(33):1–38, 2019.

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

A. Gonen, S. Sabato, and S. Shalev-Shwartz. Efficient active learning of halfspaces: An aggressive approach. *The Journal of Machine Learning Research*, 14(1):2583–2615, 2013.

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

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

S. Hanneke. Adaptive rates of convergence in active learning. In *Proceedings of the* 22^{nd} *Conference on Learning Theory*, 2009a.

S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009b.

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

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

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

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

S. Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 17(135):1–55, 2016b.

S. Hanneke. The star number and eluder dimension: Elementary observations about the dimensions of disagreement. In *Proceedings of the* 37^{th} *Conference on Learning Theory*, 2024.

S. Hanneke and S. Kpotufe. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

S. Hanneke and R. Nowak. Tutorial on Active Learning: From Theory to Practice. In *The* 36^{th} *International Conference on Machine Learning*, 2019. URL https://youtu.be/0TADiY7iPAc?t=5865.

S. Hanneke and L. Yang. Negative results for active learning with convex losses. In *Proceedings of the* 13^{th} *International Conference on Artificial Intelligence and Statistics*, 2010.

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

S. Hanneke and L. Yang. Surrogate losses in passive and active learning. *Electronic Journal of Statistics*, 13(2):4646–4708, 2019.

S. Hanneke, A. Karbasi, S. Moran, and G. Velegkas. Universal rates for active learning. In *Advances in Neural Information Processing Systems 37*, 2024a.

S. Hanneke, K. G. Larsen, and N. Zhivotovskiy. Revisiting agnostic PAC learning. In *Proceedings of the* 65^{th} *IEEE Symposium on Foundations of Computer Science*, 2024b.

S. Har-Peled, D. Roth, and D. Zimak. Maximum margin coresets for active and noise tolerant learning. In *Proceedings of the* 35^{th} *International Joint Conference on Artificial Intelligence*, 2007.

D. Haussler and P. M. Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

T. Hegedüs. Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the 8th Conference on Computational Learning Theory*, 1995.

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.

M. Hopkins, D. Kane, S. Lovett, and G. Mahajan. Point location and active learning: Learning halfspaces almost optimally. In *Proceedings of the* 61^{st} *Annual IEEE Symposium on Foundations of Computer Science*, 2020.

D. Hsu. *Algorithms for Active Learning*. PhD thesis, Department of Computer Science and Engineering, School of Engineering, University of California, San Diego, 2010.

T.-K. Huang, A. Agarwal, D. J. Hsu, J. Langford, and R. E. Schapire. Efficient and parsimonious agnostic active learning. In *Advances in Neural Information Processing Systems 28*, 2015.

M. Kääriäinen. On active learning in the non-realizable case. In *NIPS Workshop on Foundations of Active Learning*, 2005.

M. Kääriäinen. Active learning in the non-realizable case. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory*, 2006.

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

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

S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11:23–35, 1993.

S. Mahalanabis. A note on active learning for smooth problems. arXiv*:1103.3095*, 2011.

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

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

T. Mitchell. *Version Spaces: An Approach to Concept Learning*. PhD thesis, Stanford University, 1979.

O. Montasser, S. Hanneke, and N. Srebro. VC classes are adversarially robustly learnable, but only improperly. In *Proceedings of the* 32^{nd} *Conference on Learning Theory*, 2019.

E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4): 3005–3054, 2023.

B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

R. D. Nowak. Generalized binary search. In *Proceedings of the* 46^{th} *Allerton Conference on Communication, Control, and Computing*, 2008.

R. D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12), 2011.

F. Olsson. A literature survey of active machine learning in the context of natural language processing. 2009.

N. Puchkin and N. Zhivotovskiy. Exponential savings in agnostic active learning through abstention. *IEEE Transactions on Information Theory*, 68(7):4651–4665, 2022.

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

P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021.

B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2012.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the* 5^{th} *Annual Workshop on Computational Learning Theory*, 1992.

H. Shayestehmanesh. Active learning under the Bernstein condition for general losses. Master's thesis, University of Victoria, 2020.

H. Simon. An almost optimal PAC algorithm. In *Proceedings of the* 28^{th} *Conference on Learning Theory*, 2015.

M. Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22: 28–76, 1994.

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(11):45–66, 2001.

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

G. Turán. Lower bounds for PAC learning with queries. In *Proceedings of the 6$^{\text{th}}$ Annual Conference on Computational Learning Theory*, 1993.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

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

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

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

M. Vidyasagar. *Learning and Generalization with Applications to Neural Networks*. Springer-Verlag, 2$^{\text{nd}}$ edition, 2003.

L. Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12(7):2269–2292, 2011.

Y. Wang and A. Singh. Noise-adaptive margin-based active learning and lower bounds under Tsybakov noise condition. In *Proceedings of the* 30$^{\text{th}}$ *AAAI Conference on Artificial Intelligence*, 2016.

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

S. Yan, K. Chaudhuri, and T. Javidi. Active learning with logged data. In *Proceedings of the 35$^{\text{th}}$ International Conference on Machine Learning*, 2018.

S. Yan, K. Chaudhuri, and T. Javidi. The label complexity of active learning from observational data. In *Advances in Neural Information Processing Systems 32*, 2019.

C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems 27*, 2014.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

T. Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023.

N. Zhivotovskiy and S. Hanneke. Localization of VC classes: Beyond local Rademacher complexities. *Theoretical Computer Science*, 742:27–49, 2018.

X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *ICML workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, 2003.

Y. Zhu and R. Nowak. Efficient active learning with abstention. In *Advances in Neural Information Processing Systems 35*, 2022.

A Survey of the Theory of Active Learning and Other Related Work

There is at this time quite an extensive literature on the theory of active learning. We refer the interested reader to the surveys of Hanneke (2014), Dasgupta (2011), and the 2019 ICML tutorial of Hanneke and Nowak (2019) for detailed discussions of classic works in this literature. In this section, we present a brief survey of the subject, with particular emphasis on the parts most-closely related to the present work.

A.1 A Brief Historical Overview

The literature on active learning has a long history, dating back at least to the classical works on *experiment design* in statistics, wherein the analogous setting to active learning is referred to as *sequential design*. Active learning has also been an important subject within the machine learning literature from the very beginning (see Mitchell, 1979 and references therein). Below we briefly mention some of the background of the subject in the *learning theory* literature, before giving detailed background of the literature on agnostic active learning.

Membership Queries: In the learning theory literature, the idea of active learning also appeared as a natural variant of the problem of *Exact learning with queries*. Specifically, in this setting, supposing there is an unknown *target concept* $h^\star \in \mathbb{C}$, the objective of the learner is to *exactly identify* $h^\star$. To achieve this goal, the learner has access to an oracle (who knows $h^\star$), to which it may pose queries of a given type. The most relevant such queries (to the present work) are *membership queries*: namely, it may construct any $x \in \mathcal{X}$ and query for the value $h^\star(x)$ (in later works in machine learning, this is sometimes known as *query synthesis*). Early discussion of this framework and corresponding algorithmic principles appear in the seminal work of Mitchell (1979). General analyses of the number of queries necessary and sufficient to identify $h^\star$ (i.e., the *query complexity*) were developed in the works of Angluin (1987); Hegedüs (1995); Hellerstein, Pillaipakkamnatt, Raghavan, and Wilkins (1996); Nowak (2008, 2011); Hopkins, Kane, Lovett, and Mahajan (2020), and a related average-case analysis was developed by Dasgupta (2004).

Closer to the setting considered in the present work, the idea of learning with membership queries has also been extensively studied in the context of PAC learning in the realizable case. In that setting, the learner observes i.i.d. samples (X_i, Y_i) with unknown distribution P, under the assumption that there exists an unknown target concept $h^\star \in \mathbb{C}$ with $\mathrm{er}_P(h^\star) = 0$. The learner is additionally permitted to make membership queries relative to this concept $h^\star$, with the goal of producing a predictor $\hat{h}$ having $\mathrm{er}_P(\hat{h}) \leq \varepsilon$ with high probability. The literature contains numerous works on learning strategies and sample complexity analysis for this setting (e.g., Valiant, 1984; Eisenberg and Rivest, 1990; Eisenberg, 1992; Turán, 1993; Kulkarni, Mitter, and Tsitsiklis, 1993; Baum, 1991; Diakonikolas, Kane, and Ma, 2024).

Modern Active Learning with Label Queries: While the early literature on PAC learning with membership queries included several strong positive results (exhibiting advantages in both query complexity and computational complexity compared to learning from i.i.d. samples alone), when researchers implemented these algorithms and tried to use them for practical machine learning with a human labeler used to answer the queries, they found that the instances $x \in \mathcal{X}$ queried by the learner often turned out to be rather nonsensical, unnatural, or borderline cases between two labels (e.g., Baum and Lang, 1992). As such, human labelers were unable to provide useful answers to the queries, leading to poor performance of the learning algorithm. To address this issue, researchers turned to studying algorithms whose queries are restricted to only *natural* instances $x \in \mathcal{X}$, which in most works (with a few notable exceptions, such as Awasthi, Feldman, and Kanade, 2013) essentially means x in the support of the marginal distribution P_X: i.e., the types of examples that might occur naturally in the population. To actualize this restriction, researchers proposed a simple variant of active learning (which has become the standard framework in the literature, and is now essentially synonymous with the term *active learning*), in which there are i.i.d. samples (X_i, Y_i) from an unknown distribution P, but the learner initially only observes the *unlabeled* examples X_i, and can *query* to observe individual labels Y_i (in a *sequential* fashion, so that it observes the label Y_i of its previous query before selecting the $X_{i'}$ to query next) (Mitchell, 1979; Cohn, Atlas, and Ladner, 1994; Freund, Seung, Shamir, and Tishby, 1997; Tong and Koller, 2001). Such queries can typically be answered by human experts, being of the same type as used for data annotation in standard supervised machine learning. In this scenario, the unlabeled examples X_i are typically assumed to be available in abundance, so that the primary objective is to minimize the number of *label queries* needed to achieve a given accuracy of a learned predictor $\hat{h}$.

The theoretical literature on this subject has origins in early works discussing algorithmic principles based on version spaces (Mitchell, 1979; Cohn, Atlas, and Ladner, 1994). Many of the early works providing actual bounds on the query complexity focused on showing improvements over passive learning in the realizable case for special scenarios, such as linear classifiers under distribution assumptions (e.g., Seung, Opper, and Sompolinsky, 1992; Freund, Seung, Shamir, and Tishby, 1997;

Dasgupta, 2004; Dasgupta, Kalai, and Monteleoni, 2005; Har-Peled, Roth, and Zimak, 2007; Balcan, Beygelzimer, and Langford, 2006; Balcan, Broder, and Zhang, 2007; Balcan, Hanneke, and Vaughan, 2010; Balcan and Long, 2013; Gonen, Sabato, and Shalev-Shwartz, 2013; Wang and Singh, 2016; Cavallanti, Cesa-Bianchi, and Gentile, 2011; Dekel, Gentile, and Sridharan, 2012). This was followed by a boom of general-case analyses, providing general theories analyzing the query complexity for any concept class (e.g., Dasgupta, 2005; Hanneke, 2007a,b, 2009b,a, 2011, 2012, 2014; Dasgupta, Hsu, and Monteleoni, 2007; Balcan, Hanneke, and Vaughan, 2010; Beygelzimer, Dasgupta, and Langford, 2009; Koltchinskii, 2010; Zhang and Chaudhuri, 2014; El-Yaniv and Wiener, 2012; Wiener, Hanneke, and El-Yaniv, 2015; Hanneke and Yang, 2015; Hanneke, Karbasi, Moran, and Velegkas, 2024a), some of which are discussed in more detail below.

The Passive Learning Baseline: Since the predictor $\hat{h}$ produced by an active learning algorithm is based on a selected subset of a given set of i.i.d. examples (X_i, Y_i), the natural quantity for comparison is the number of *i.i.d. labeled* examples necessary to obtain the same accuracy: i.e., the sample complexity of standard supervised learning, which in this literature is termed *passive learning*.[5] Recall from Section 2 that we denote by $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$ the minimax optimal sample complexity of passive learning: i.e., the minimal n such that there exists a passive learning algorithm $\mathbb{A}_p$ that, for every P with $\inf_{h\in\mathbb{C}} \mathrm{er}_P(h) \leq \beta$, for $S \sim P^n$ and $\hat{h}_n = \mathbb{A}_p(S)$, guarantees with probability at least $1-\delta$ that $\mathrm{er}_P(\hat{h}_n) \leq \inf_{h\in\mathbb{C}} \mathrm{er}_P(h) + \varepsilon$. Lower bounds of Vapnik and Chervonenkis (1974); Devroye and Lugosi (1995) establish that $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C}) = \Omega\left(\frac{\beta}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{\varepsilon}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right)$. The classic analysis of Vapnik and Chervonenkis (1974) further established this lower bound can nearly be achieved by the simple method of *empirical risk minimization*, i.e., $\hat{h}_n = \mathrm{argmin}_{h\in\mathbb{C}} \hat{\mathrm{er}}_S(h)$, providing an upper bound $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C}) = O\left(\frac{\beta}{\varepsilon^2}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{\varepsilon}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$. This has since been refined in various ways, such as via localized chaining arguments (e.g., Giné and Koltchinskii, 2006). Most recently, Hanneke, Larsen, and Zhivotovskiy (2024b) proved an upper bound $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C}) = O\left(\frac{\beta}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\frac{1}{\varepsilon}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right)$, matching the lower bound up to log factors in the lower-order term (the problem of removing these remaining log factors remains open at this time). The algorithm achieving this is *improper*, meaning its returned $\hat{h}_n$ is not necessarily an element of $\mathbb{C}$, and Hanneke, Larsen, and Zhivotovskiy (2024b) in fact show improperness is *necessary* to match the lower bound in the lead term without a log factor gap. In the special case of $\beta = 0$ (the *realizable case*), the lower bound was shown to be achievable by Hanneke (2016a) (also necessarily via an improper learner), so that $\mathcal{M}_p(\varepsilon, \delta; 0, \mathbb{C}) = \Theta\left(\frac{1}{\varepsilon}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right)$. The lower bound $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C}) = \Omega\left(\frac{\beta}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{\varepsilon}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right)$ will therefore serve as a suitable baseline for gauging whether the query complexity $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C})$ of active learning is smaller than the sample complexity $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$ of passive learning.

The Need for Distribution-dependent Analysis in Realizable Active Learning: Much of the early work on active learning focused on the *realizable case*, i.e., the special case $\beta = \inf_{h\in\mathbb{C}} \mathrm{er}_P(h) = 0$. In this special case, it was quickly observed by Dasgupta (2004, 2005) that there are some concept classes (e.g., *thresholds* $\mathbb{1}_{[a,\infty)}$ on $\mathbb{R}$) where active learning offers strong improvements over passive learning, and other concept classes (e.g., *intervals* $\mathbb{1}_{[a,b]}$ on $\mathbb{R}$) where the minimax query complexity $\mathrm{QC}_a(\varepsilon, \delta; 0, \mathbb{C})$ of active learning offers *no significant advantages* over passive learning. The essential challenge in the latter case is the problem of "searching in the dark" for a small-but-important region (e.g., $P_X = \mathrm{Uniform}(\{x_1, \ldots, x_{1/\varepsilon}\})$ and the optimal concept labels exactly one x_i as 1). This challenging scenario is embedded in *most* concept classes of interest in learning theory,

[5] Since the active learner also has access to the remaining i.i.d. *unlabeled examples* X_i it does not query, it is also natural to compare to the related framework of *semi-supervised* learning, in which a learner has access to some number n of i.i.d. labeled examples with distribution P and additionally some larger number m of i.i.d. *unlabeled* examples with distribution P_X (Chapelle, Scholkopf, and Zien, 2006). While, under some favorable conditions, the labeled sample complexity n of semi-supervised learning can be smaller than that of strictly-supervised passive learning (see Balcan and Blum, 2010), the lower bounds on the (distribution-free) sample complexity of passive learning discussed in this work remain valid for the labeled sample complexity of semi-supervised learning, so that for the purpose of comparison in the present work, the distinction between supervised and semi-supervised passive learning as a baseline is not important, and we will simply compare to passive supervised learning for simplicity.

a fact which was formalized abstractly in the *star number* complexity measure $\mathfrak{s}$ (Definition 2) by Hanneke and Yang (2015), who also show it sharply characterizes the optimal distribution-free query complexity in the realizable case: in particular, they showed $\mathrm{QC}_a(\varepsilon,\delta;0,\mathbb{C})$ admits an improved dependence on ε compared to $\mathcal{M}_p(\varepsilon,\delta;0,\mathbb{C})$ *if and only if* $\mathfrak{s}<\infty$ (whereas most commonly-studied classes $\mathbb{C}$ have $\mathfrak{s}=\infty$). They also show there exist classes $\mathbb{C}$ of any given $\mathtt{d}=\mathrm{VC}(\mathbb{C})$ where $\mathrm{QC}_a(\varepsilon,\delta;0,\mathbb{C})=\Theta\left(\frac{\mathtt{d}}{\varepsilon}\right)$, indicating that $\mathrm{QC}_a(\varepsilon,\delta;0,\mathbb{C})$ is not significantly smaller than $\mathcal{M}_p(\varepsilon,\delta;0,\mathbb{C})$ (only offering an improvement in the dependence on δ).

Motivated by the fact that this "searching in the dark" scenario is embedded in *most* concept classes of interest, Dasgupta (2005) suggested that, for such concept classes, the only viable way to understand the potential advantages of active learning is to focus on *distribution-dependent* analysis, toward identifying special scenarios where active learning algorithms offer improvements over passive learning under appropriate assumptions on the distribution P. This narrative quickly caught on in the literature, with a variety of distribution-dependent analyses and general P-dependent complexity measures being proposed to analyze certain active learning strategies under various restrictions on the realizable distribution P (Dasgupta, 2005; Hanneke, 2007a,b, 2014; Balcan, Broder, and Zhang, 2007; Balcan and Long, 2013; Zhang and Chaudhuri, 2014; El-Yaniv and Wiener, 2012; Wiener, Hanneke, and El-Yaniv, 2015), discussed in more detail below.

Active Learning in the Non-realizable Case: Given the above narrative, when approaching the analysis of active learning in the *non-realizable* case ($\beta>0$), it might at first seem perfectly reasonable to expect that for many concept classes $\mathbb{C}$ the query complexity $\mathrm{QC}_a(\varepsilon,\delta;\beta,\mathbb{C})$ might be not be much smaller than the sample complexity of passive learning $\mathcal{M}_p(\varepsilon,\delta;\beta,\mathbb{C})$. As such, the literature on agnostic active learning largely focused on *extending* the distribution-dependent analyses from the realizable case to the agnostic setting (Balcan, Beygelzimer, and Langford, 2006, 2009; Hanneke, 2007b,a, 2009b, 2011, 2014; Balcan and Hanneke, 2012; Zhang and Chaudhuri, 2014). These upper bounds conformed to the accepted narrative, in that they offer improvements under some distributions, but for most classes $\mathbb{C}$, in the worst case over distributions P (with $\inf_{h\in\mathbb{C}}\mathrm{er}_P(h)\leq\beta$) they revert to the passive sample complexity $\mathcal{M}_p(\varepsilon,\delta;\beta,\mathbb{C})$.

However, since this narrative was born from analysis of the *realizable case*, there was no actual reason to believe it should remain valid in the non-realizable case. In particular, there remained an intriguing possibility that there could perhaps be *other advantages* of active learning *specific to the non-realizable case*: that is, beyond the "binary search" type advantages known from the realizable case (as captured by the complexity measures proposed for realizable-case analysis). Some hints that such additional advantages may exist appear in the works of Efromovich (2007); Hanneke and Yang (2015) studying certain special scenarios (e.g., noise models), where they find that active learning can be useful for *adaptively identifying noisy regions* (i.e., regions of X's where $P(Y|X)$ is close to $\frac{1}{2}$) and allocating queries appropriately to compensate for this noisiness without wasting excessive queries in less-noisy regions (as passive learning would). This additional advantage, specific to the non-realizable case, offered quantitative advantages over passive learning under the specific conditions studied in those works (e.g., Hanneke and Yang, 2015 showed improvements in query complexity under certain noise models, namely Tsybakov noise and Benign noise, for all classes $\mathbb{C}$, including those with the "searching in the dark" scenario embedded in them). However, these works left open the question of whether such advantages can be observed also in the more-challenging *agnostic* setting. The extension to the agnostic case is not at all clear, since in this setting (unlike the special noise models in the works above) the source of non-realizability is not only the noisiness of the $P(Y|X)$ label distribution, but also *model misspecification*: i.e., it is possible to have $\beta>0$ even when $P(Y|X)\in\{0,1\}$, if the *Bayes classifier* is not in $\mathbb{C}$, in which case the idea of adapting to "noisiness" of labels is no longer a useful framing of the problem.

Quantifying the Query Complexity: Quantitatively, in the realizable case ($\beta=0$), the above analyses produced P-dependent query complexity bounds of the form $c_P(\varepsilon)\cdot\mathtt{d}\cdot\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$ for some P-dependent complexity measure $c_P(\varepsilon)$. Examples of such complexity measures include the splitting index (Dasgupta, 2005), disagreement coefficient (Hanneke, 2007b, 2009b, 2011), empirical extended teaching dimension (Hanneke, 2007a), and a subregion variant of the disagreement coefficient (Zhang and Chaudhuri, 2014), among others (e.g., El-Yaniv and Wiener, 2012; Hanneke, 2012; Hanneke and Yang, 2015; Hanneke, 2014; Wiener, Hanneke, and El-Yaniv, 2015). Some of these were accompanied by related minimax lower bounds holding for any fixed P_X marginal distribution (Dasgupta, 2005; Hanneke, 2007a; Balcan and Hanneke, 2012). The works of Hanneke and Yang (2015); Hanneke

(2016b, 2024) later showed that all of these proposed complexity measures $c_P(\varepsilon)$ have worst-case values precisely equal the *star number* $\mathfrak{s}$ (Definition 2), a quantity which effectively formalizes and quantifies the above discussion of whether the class has within it an embedded "searching in the dark" problem. In particular, they showed that in each case, $\sup_P c_P(\varepsilon) = \mathfrak{s} \wedge \frac{1}{\varepsilon}$. Thus, the star number unifies all of these complexity measures in the case of distribution-free analysis. Moreover, Hanneke and Yang (2015) also established upper and lower bounds on $\mathrm{QC}_a(\varepsilon, \delta; 0, \mathbb{C})$ itself, showing it is $\Omega\big(\mathfrak{s} \wedge \frac{1}{\varepsilon}\big)$ and $O\Big(\big(\mathfrak{s} \wedge \frac{\mathsf{d}}{\varepsilon}\big) \log\big(\frac{1}{\varepsilon}\big)\Big)$. As mentioned, they also show there exist classes $\mathbb{C}$ of any given $\mathsf{d} = \mathrm{VC}(\mathbb{C})$ where $\mathrm{QC}_a(\varepsilon, \delta; 0, \mathbb{C}) = \Theta\big(\frac{\mathsf{d}}{\varepsilon}\big)$. Thus, any general upper bound on $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C})$ can be no smaller than $\mathfrak{s} \wedge \frac{1}{\varepsilon}$, and any such bound depending on $\mathbb{C}$ only via d can be no smaller than $\frac{\mathsf{d}}{\varepsilon}$.

Turning to the agnostic case ($\beta \geq 0$), Kääriäinen (2006) established a general lower bound $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}) = \Omega\Big(\frac{\beta^2}{\varepsilon^2} \log\big(\frac{1}{\delta}\big)\Big)$, later strengthened by Beygelzimer, Dasgupta, and Langford (2009) to $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}) = \Omega\Big(\frac{\beta^2}{\varepsilon^2} \big(\mathsf{d} + \log\big(\frac{1}{\delta}\big)\big)\Big)$. Comparing this to the sample complexity of passive learning, namely $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C}) = \Theta\Big(\frac{\beta}{\varepsilon^2} \big(\mathsf{d} + \log\big(\frac{1}{\delta}\big)\big)\Big) + \tilde{\Theta}\big(\frac{1}{\varepsilon} \big(\mathsf{d} + \log\big(\frac{1}{\delta}\big)\big)\big)$, we see that in the regime $\beta \gg \sqrt{\varepsilon}$, the best improvement we can hope for from active learning would be to replace the factor β with β^2: i.e., squaring the dependence on the best-in-class error rate.

The work of Balcan, Beygelzimer, and Langford (2006) initiated the study of upper bounds on the query complexity in the agnostic case, showing that the above lower bound can be matched in the special cases of *threshold* classifiers (concepts $\mathbb{1}_{[a,\infty)}$ on $\mathcal{X} = \mathbb{R}$), and (in the regime $\beta \lesssim \varepsilon/\sqrt{\mathsf{d}}$) matched up to a factor d for homogeneous linear classifiers under P_X uniform in an origin-centered ball, extending these well-known examples from the realizable case This analysis was generalized to *all* concept classes by Hanneke (2007b), expressing a query complexity bound of the form $\tilde{O}\Big(c_P(\beta + \varepsilon)\mathsf{d}\left(\frac{\beta^2}{\varepsilon^2} + 1\right)\Big)$, where the factor $c_P(\beta + \varepsilon)$ is based on a P-dependent complexity measure $\theta_P(\beta + \varepsilon)$ therein termed the *disagreement coefficient* (Definition 25). In particular, the bound of Hanneke (2007b) matches the lower bound of Kääriäinen (2006); Beygelzimer, Dasgupta, and Langford (2009) up to logs only when $\theta_P(\beta + \varepsilon) = \tilde{O}(1)$. The latter holds for threshold classifiers, and for other classes under restrictions on P, but in many other cases $\theta_P(\beta + \varepsilon)$ can be as large as $\frac{1}{\beta+\varepsilon}$ due to the "searching in the dark" problem discussed above; in such cases, the query complexity upper bound of Hanneke (2007b) is no smaller than the sample complexity of passive learning $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$. Numerous later works (some discussed in detail below) discovered refinements and alternative P-dependent complexity measures used to express upper bounds on the query complexity (Hanneke, 2007a; Dasgupta, Hsu, and Monteleoni, 2007; Hanneke, 2009b, 2011, 2014; Zhang and Chaudhuri, 2014). However, like the bound of Hanneke (2007b), all of these results establish query complexity upper bounds of the form $\tilde{O}\Big(c_P(\beta + \varepsilon)\mathsf{d}\left(\frac{\beta^2}{\varepsilon^2} + 1\right)\Big)$ for some P-dependent complexity measure $c_P(\beta + \varepsilon)$, all of which have the property that, in the "searching in the dark" type scenarios discussed above, the value $c_P(\beta + \varepsilon) \geq \frac{1}{\beta+\varepsilon}$, so that in such scenarios these upper bounds are all no smaller than the sample complexity of passive learning $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$.

As in the realizable case, these various analysis were later all unified under worst-case analysis over P by the *star number* in the work of Hanneke and Yang (2015). Indeed, these complexity measures $c_P(\beta + \varepsilon)$ are in fact the same family of complexity measures alluded to above for the realizable case. As such, by the aforementioned result of Hanneke and Yang (2015), they all satisfy $\sup_P c_P(\beta + \varepsilon) = \mathfrak{s} \wedge \frac{1}{\beta+\varepsilon}$. Thus, the upper bounds established by these works, all being of the form $\tilde{O}\Big(c_P(\beta + \varepsilon)\mathsf{d}\left(\frac{\beta^2}{\varepsilon^2} + 1\right)\Big)$, unify to a single upper bound of the form $\tilde{O}\Big(\Big(\mathfrak{s} \wedge \frac{1}{\beta+\varepsilon}\Big)\,\mathsf{d}\left(\frac{\beta^2}{\varepsilon^2} + 1\right)\Big)$ in the worst case over distribution P (subject to $\inf_{h\in\mathbb{C}} \mathrm{er}_P(h) \leq \beta$). In particular, this also means they all fail to imply any improvements over the sample complexity of passive learning $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$ in the worst case over such distributions P, for any concept class $\mathbb{C}$ with $\mathfrak{s} = \infty$.[6] This is particularly significant, since most commonly-studied concept classes have $\mathfrak{s} = \infty$, including, for instance, linear classifiers in $\mathbb{R}^p$, $p \geq 2$. On the other hand, the lower bound of Kääriäinen (2006); Beygelzimer,

[6] One can also show that this is not merely a result of loose analysis. The algorithms (prior to the present work) can be made to behave similarly to passive learners (meaning they query almost indiscriminately) in some scenarios constructed on large star sets, resulting in a number of queries $\frac{\beta}{\varepsilon^2}$.

Dasgupta, and Langford (2009), of the form $\Omega\Big(\frac{\beta^2}{\varepsilon^2}\left(\mathtt{d}+\log\left(\frac{1}{\delta}\right)\right)\Big)$, has no such factor $\mathfrak{s}\wedge\frac{1}{\beta+\varepsilon}$. The natural question is therefore which of these can be strengthened: the upper bound or lower bound.

The above gap has a qualitative significance. If the lower bound could be strengthened to match the upper bound, it would mean that (as in the realizable case) there are classes where active learning offers no advantage in its minimax query complexity compared to passive learning. On the other hand, if the upper bound can be strengthened to match the lower bound, it would mean that (unlike the realizable case) the query complexity of active learning is always smaller than the sample complexity of passive learning in the agnostic setting. The problem of resolving this gap has remained open until now. In the present work, we completely resolve this question, strengthening the upper bound to match the above lower bound, and thereby establishing that active learning is always better than passive learning in the agnostic case, providing an improvement by squaring the dependence on the best-in-class error rate: i.e., replacing β with β^2. Establishing this upper bound requires a new principle for active learning, specific to the agnostic setting, which we develop in this work (termed AVID, for *adaptive localized variance isolation by disagreements*).

Before proceeding with the presentation of our results, we first provide, in the next subsection, a detailed survey of several of the prior works mentioned in the above brief historical summary.

A.2 Detailed Description of Relevant Techniques in the Prior Literature

In this subsection, we provide further details of relevant works in the literature. Due to the vastness and diversity of the literature on the theory of active learning, we will not provide an exhaustive survey here, instead focusing on the techniques and results most-relevant to the present work.

Disagreement-based Active Learning: By-far the most well-studied technique in the literature on the theory of active learning is known as *disagreement-based* active learning. A disagreement-based active learner is given as input the sequence $X_1, X_2, \ldots, X_m$ of unlabeled examples. It maintains (either explicitly or implicitly) a set $V \subseteq \mathbb{C}$ of *surviving concepts* (known as a *version space*), with a guarantee that the *best-in-class* concept[7] $h^\star$ is retained in V. To choose its query points, it finds the next unlabeled example X_i in the sequence for which $\exists f, g \in V$ with $f(X_i) \neq g(X_i)$, and queries for the label Y_i: or more succinctly, it queries the next $X_i \in \mathrm{DIS}(V)$, where

$$\mathrm{DIS}(V) := \{x \in \mathcal{X} : \exists f, g \in V, f(x) \neq g(x)\},$$

denotes the *region of disagreement* of V. It then updates the set V of surviving concepts based on this new information (or, in some variants, it performs this update only periodically, rather than after every query). This is abstractly summarized in the following skeleton.

Algorithm Outline: Disagreement-based Active Learning
Input: Unlabeled data $X_1, \ldots, X_m$
Output: Classifier $\hat{h}$
0. Initialize $V = \mathbb{C}$
1. For $i = 1, 2, \ldots, m$
2. If $X_i \in \mathrm{DIS}(V)$
3. Query for label Y_i
4. Update V
5. Return any $\hat{h} \in V$

The idea is that, if we seek to return a concept $\hat{h} \in V$ with small $\mathrm{er}_P(\hat{h}) - \mathrm{er}_P(h^\star)$, then for any $X_i \notin \mathrm{DIS}(V)$, since all surviving concepts *agree* on the classification of X_i, the label Y_i would provide no information that would help with this goal, so we do not bother querying for this label: that is, such a Y_i cannot help to estimate the *relative* performances $\mathrm{er}_P(f) - \mathrm{er}_P(g)$ of concepts $f, g \in V$, since regardless of Y_i, we have $\mathbb{1}[f(X_i) \neq Y_i] - \mathbb{1}[g(X_i) \neq Y_i] = 0$. In contrast, the next $X_i \in \mathrm{DIS}(V)$ in the sequence is a random sample from $P_X(\cdot|\mathrm{DIS}(V))$, so that for any $f, g \in V$, $\mathbb{1}[f(X_i) \neq Y_i] - \mathbb{1}[g(X_i) \neq Y_i]$ is an unbiased estimate of the difference of error rates under the conditional distribution $P(\cdot|\mathrm{DIS}(V) \times \{0,1\})$, which (again since f, g agree outside $\mathrm{DIS}(V)$) is

[7] Suppose, for simplicity, such a $h^\star$ exists in $\mathbb{C}$: i.e., $\mathrm{er}_P(h^\star) = \inf_{h\in\mathbb{C}} \mathrm{er}_P(h)$. The theory easily generalizes to cases where the infimum is not attained.

proportional to $\mathrm{er}_P(f) - \mathrm{er}_P(g)$. By reasoning about uniform concentration of these estimates, we can define an update rule for V in Step 4 that never removes the best-in-class concept $h^\star$ while pruning sub-optimal concepts from V (where the resolution of this pruning improves with m).

The algorithmic principle underlying disagreement-based active learning, and corresponding reasoning about correctness and potential advantages, was already identified in the early work of Mitchell (1979) for the realizable case,[8] where (since we always have $h^\star(X_i) = Y_i$ in this case) the update to the version space (Step 4) simply removes any concepts incorrect on (X_i, Y_i): that is, $V \leftarrow \{h \in V : h(X_i) = Y_i\}$. The precise form expressed above was first explicitly studied by Cohn, Atlas, and Ladner (1994), in the realizable case (again with the above simple rule for updating V in Step 4). In their honor, this realizable-case technique is referred to as CAL in the literature. While the original works of Mitchell (1979); Cohn, Atlas, and Ladner (1994) include the observation that $h^\star$ is retained in V, and Cohn, Atlas, and Ladner (1994) include some discussion of generalization, the formal analysis of the query complexity of this technique only began with the later work of Balcan, Beygelzimer, and Langford (2006) (bounding the query complexity for some specific concept classes), and the general analysis of the technique (applicable to any concept class) began with the works of Hanneke (2007b, 2009b, 2011); Dasgupta, Hsu, and Monteleoni (2007).

The idea of disagreement-based active learning was first extended to the agnostic setting ($\beta \geq 0$) by Balcan, Beygelzimer, and Langford (2006), with an instantiation of the above outline they called the A^2 algorithm (for *Agnostic Active*). The main idea was to instantiate the update to V in Step 4 using uniform concentration inequalities. In their original version, they specifically define $\mathrm{UB}(h)$ and $\mathrm{LB}(h)$ as high-probability uniform upper and lower bounds on $\mathrm{er}_{P(\cdot|\mathrm{DIS}(V)\times\{0,1\})}(h)$ based on the queries from $\mathrm{DIS}(V)$ since the last update to V (where they only update V periodically in their algorithm). They then define the update as $V \leftarrow \{h \in V : \mathrm{LB}(h) \leq \min_{h' \in V} \mathrm{UB}(h')\}$. The idea is that they wish to remove a concept h from V if there is another concept $h' \in V$ whose *upper bound* $\mathrm{UB}(h')$ on its error rate is smaller than the *lower bound* $\mathrm{LB}(h)$ on the error rate of h. In particular, since h, h' agree on all $x \notin \mathrm{DIS}(V)$, assuming the validity of these upper and lower bounds, a concept h can be removed from V *only if* $\mathrm{er}_P(h) \geq \mathrm{LB}(h) > \mathrm{UB}(h') \geq \mathrm{er}_P(h')$, so that h is verifiably suboptimal (hence, necessarily, $h \neq h^\star$).

Balcan, Beygelzimer, and Langford (2006) included the above correctness guarantee (i.e., the algorithm maintains $h^\star \in V$) and argue the A^2 algorithm returns an $\hat{h}$ with $\mathrm{er}_P(\hat{h}) \leq \mathrm{er}_P(h^\star) + \varepsilon$, with a number of queries never significantly worse than that of passive learning. Also, as a sort of *proof of concept* illustrating the potential benefits of A^2 in a simple example, they also quantified the query complexity advantages in the special case of *threshold* classifiers (concepts $\mathbb{1}_{[a,\infty)}$ on $\mathcal{X} = \mathbb{R}$), showing a bound $\tilde{O}\left(\frac{\beta^2}{\varepsilon^2}\right)$ for that class (matching the lower bound of Kääriäinen, 2006). They also studied the special case of learning homogeneous linear classifiers under a uniform distribution in an origin-centered ball in $\mathbb{R}^{\mathsf{d}}$, focusing on the regime $\beta \lesssim \varepsilon/\sqrt{\mathsf{d}}$, for which they showed the query complexity is $\tilde{O}\big(\mathsf{d}^2 \log\big(\frac{1}{\varepsilon}\big) \log\big(\frac{1}{\delta}\big)\big)$.

The first *general* analyses (i.e., applicable to *any* concept class) of the query complexity of active learning in the agnostic setting were given in the works of Hanneke (2007b,a). In particular, Hanneke (2007b) analyzed the A^2 disagreement-based active learning algorithm, providing a general query complexity bound expressed in terms of a new complexity measure therein termed the *disagreement coefficient*. Specifically, for $r > 0$, denoting by $\mathrm{B}_{P_X}(h^\star, r) = \{h \in \mathbb{C} : P_X(x : h(x) \neq h^\star(x)) \leq r\}$ the $h^\star$-centered r-ball (under $\mathrm{L}_1(P_X)$), the disagreement coefficient is defined as

$$\theta_P(\beta + \varepsilon) := \sup_{r > \beta + \varepsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}_{P_X}(h^\star, r)))}{r} \vee 1.$$

The intuitive interpretation of the relevance of this quantity is that, as the algorithm progresses, the set V of surviving concepts will become closer and closer to $h^\star$ (up until a distance $O(\beta + \varepsilon)$), so that the probability of querying decreases as $P_X(\mathrm{DIS}(V)) \leq P_X(\mathrm{DIS}(\mathrm{B}_{P_X}(h^\star, r)))$ for an appropriate r decreasing as the number of queries grows.

Hanneke (2007b) proves that, for any $\mathbb{C}$ and P, for $\beta = \mathrm{er}_P(h^\star)$, the A^2 algorithm succeeds after a number of queries $\tilde{O}\left(\theta_P(\beta + \varepsilon)^2 \mathsf{d} \left(\frac{\beta^2}{\varepsilon^2} + 1\right)\right)$. This matches the lower bound of Kääriäinen (2006); Beygelzimer, Dasgupta, and Langford (2009) up to logs whenever $\theta_P(\beta + \varepsilon) = \tilde{O}(1)$. In particular,

[8] Mitchell (1979) also presents some discussion of reasonable extensions to the non-realizable case.

Hanneke (2007b) bounds $\theta_P(\beta+\varepsilon)$ for a number of examples, including showing that this general query complexity upper bound recovers the examples of Balcan, Beygelzimer, and Langford (2006): $\theta_P(\beta+\varepsilon) \leq 2$ for threshold classifiers, and $\theta_P(\beta+\varepsilon) = O(\sqrt{\mathsf{d}})$ for homogeneous linear classifiers under a uniform distribution on an origin-centered sphere (thus also removing the constraints on β, ε from the result of Balcan, Beygelzimer, and Langford, 2006). However, Hanneke (2007b) found $\theta_P(\beta+\varepsilon)$ can sometimes be as large as $\frac{1}{\beta+\varepsilon}$, particularly for the "searching in the dark" scenarios discussed above.

Subsequently, Dasgupta, Hsu, and Monteleoni (2007) refined the dependence on $\theta_P(\beta+\varepsilon)$ in this bound (analyzing a different disagreement-based algorithm), replacing $\theta_P(\beta+\varepsilon)^2$ with $\theta_P(\beta+\varepsilon)$. Specifically, they proposed a variant of the A^2 disagreement-based active learning algorithm. Unlike A^2, their algorithm was expressed as a *reduction to empirical risk minimization*, thereby offering some practical advantages. In particular, the algorithm does not explicitly represent the set V, but rather maintains it *implicitly* by a *constraint* on the differences of empirical error rates of concepts whose disagreements are worth querying. More explicitly, they maintain data sets Q_i, L_i, where Q_i are the queries so far (up to round i) and L_i are the agreed-upon labels of all unqueried examples so far. On round i, they consider the concepts h^1, h^0 of minimal $\hat{\mathrm{er}}_{Q_{i-1}}(h)$ subject to $\hat{\mathrm{er}}_{L_{i-1}}(h) = 0$ and $h^1(X_i) = 1$, $h^0(X_i) = 0$. If $\hat{\mathrm{er}}_{Q_{i-1}}(h^1)$ and $\hat{\mathrm{er}}_{Q_{i-1}}(h^0)$ are of similar sizes, they query for Y_i and add it to Q_{i-1} to get Q_i (letting $L_i = L_{i-1}$), and otherwise they take an *inferred* label $\hat{y}_i = \mathrm{argmin}_y \hat{\mathrm{er}}_{Q_{i-1}}(h^y)$ and add it to L_{i-1} to get L_i (letting $Q_i = Q_{i-1}$). Note that this is quite similar to the idea of maintaining a set V of concepts h having $\hat{\mathrm{er}}_{Q_{i-1}}(h)$ close to $\min_{h' \in \mathbb{C}} \hat{\mathrm{er}}_{Q_{i-1}}(h')$, and querying X_i iff $X_i \in \mathrm{DIS}(V)$ (the only difference from this being the L_i constraints, which only serve to decrease the set V).[9] Thus, this algorithm can also be thought of as a disagreement-based active learner. The specific quantification of the "similar sizes" criterion for the difference of empirical error rates come from uniform Bernstein-style concentration inequalities (related to the uniform Bernstein inequality stated in Lemma 7 of Appendix D below). This guarantees that $h^\star$ always satisfies the condition, and thus the algorithm will query whenever some $h \in \mathbb{C}$ with $\mathrm{er}_P(h)$ not much larger than $h^\star$ has $h(X_i) \neq h^\star(X_i)$.

Dasgupta, Hsu, and Monteleoni (2007) analyzed the query complexity of this algorithm, showing that it guarantees $\mathrm{er}_P(\hat{h}) \leq \mathrm{er}_P(h^\star) + \varepsilon$ after a number of queries $\tilde{O}\left(\theta_P(\beta+\varepsilon)\mathsf{d}\left(\frac{\beta^2}{\varepsilon^2}+1\right)\right)$. Compared to the original analysis of Hanneke (2007b), this improves the bound in its dependence on $\theta_P(\beta+\varepsilon)$, reducing from quadratic $\theta_P(\beta+\varepsilon)^2$ to linear $\theta_P(\beta+\varepsilon)$. Again, the conclusion is that the algorithm's query complexity matches the lower bound of Kääriäinen (2006); Beygelzimer, Dasgupta, and Langford (2009) up to logs whenever $\theta_P(\beta+\varepsilon) = \tilde{O}(1)$.

The above techniques, and corresponding analysis in terms of the disagreement coefficient, seeded a vast literature, with many variations on the technique, analysis, and complexity measures, and many examples of scenarios $(\mathbb{C}, P)$ for which $\theta_P(\beta+\varepsilon)$ can be favorably bounded. This branch of the literature is collectively referred to as *disagreement-based active learning* (see e.g., the works of Hanneke, 2009b, 2011, 2012, 2014, 2016b; Balcan, Hanneke, and Vaughan, 2010; Hsu, 2010; El-Yaniv and Wiener, 2012; Friedman, 2009; Mahalanabis, 2011; Koltchinskii, 2010; Wang, 2011; Beygelzimer, Dasgupta, and Langford, 2009; Beygelzimer, Hsu, Langford, and Zhang, 2010; Raginsky and Rakhlin, 2011; Ailon, Begleiter, and Ezra, 2014; Huang, Agarwal, Hsu, Langford, and Schapire, 2015; Wiener, Hanneke, and El-Yaniv, 2015; Hanneke and Yang, 2010, 2015, 2019; Yan, Chaudhuri, and Javidi, 2018, 2019; Gelbhart and El-Yaniv, 2019; Cortes, DeSalvo, Gentile, Mohri, and Zhang, 2019a; Cortes, DeSalvo, Gentile, Mohri, and Zhang, 2019b,c, 2020; DeSalvo, Gentile, and Thune, 2021; Shayestehmanesh, 2020; Puchkin and Zhivotovskiy, 2022). A detailed summary of this line of work is presented in the survey of Hanneke (2014).

In the context of distribution-free analysis, Hanneke and Yang (2015) showed that $\sup_P \theta_P(\beta+\varepsilon) = \mathfrak{s} \wedge \frac{1}{\beta+\varepsilon}$, where $\mathfrak{s}$ is the *star number* of $\mathbb{C}$ (Definition 2), and where the sup is over realizable distributions P (so that, in particular, they satisfy the condition $\mathrm{er}_P(h^\star) \leq \beta$). Thus, in terms of their implications for the distribution-free query complexity $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C})$, these P-dependent analyses of disagreement-based active learning simplify to a bound of the form $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}) = \tilde{O}\left(\left(\mathfrak{s} \wedge \frac{1}{\beta+\varepsilon}\right)\mathsf{d}\left(\frac{\beta^2}{\varepsilon^2}+1\right)\right)$. In particular, such bounds are capable of providing improvements in

[9] Subsequent works of Beygelzimer, Hsu, Langford, and Zhang (2010); Hsu (2010) established similar query complexity bounds for a simpler variant which does not enforce these additional L_i constraints.

distribution-free query complexity over the sample complexity of passive learning $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$ if and only if $\mathfrak{s} < \infty$ (which, as discussed above, is a rather strong restriction). This contrasts with Theorems 1, 3, which provide improvements for *all* concept classes $\mathbb{C}$, regardless of whether $\mathfrak{s}$ is finite or infinite. The role of $\mathfrak{s}$ in Theorem 3 is merely in refining the lower-order term in the special case that $\mathfrak{s} < \infty$.

We note that, while the AVID Agnostic algorithm (Figure 1) itself should not be regarded as a disagreement-based active learner (as its primary advantage over passive learning is *not* based on the restriction of queries to $\mathrm{DIS}(V)$), elements of disagreement-based learning have been incorporated into it for the purpose of the refined *lower-order* term in the upper bound in Theorem 3. Specifically, the choice to query examples in $S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$ in Step 2 (and similarly Step 9) restricts to queries in $D_{k-1} = \mathrm{DIS}(V_{k-1})$. This restriction is directly responsible for the lower-order term being of the form $\tilde{O}\big(\big(\mathfrak{s} \wedge \frac{1}{\varepsilon}\big)\,\mathsf{d}\big)$ rather than $\tilde{O}\big(\frac{\mathsf{d}}{\varepsilon}\big)$ as in Theorem 1. On the other hand, for the purpose of the lead term, this incorporation of disagreement-based queries is unnecessary, and indeed Theorem 1 remains valid *without* this aspect of the algorithm: that is, in Steps 2 and 9, if we simply query all of $S_k^1 \setminus \Delta_{i_k}$, the algorithm still achieves the query complexity bound stated in Theorem 1 with its lower-order term $\tilde{O}\big(\frac{\mathsf{d}}{\varepsilon}\big)$.

The argument leading to the refined lower-order term in Theorem 3 makes use of reasoning directly rooted in the analysis of disagreement-based methods via the disagreement coefficient (Lemma 22), and indeed we present P-dependent refinements of this lower-order term directly expressed in terms of $\theta_P(\beta+\varepsilon)$ in Appendix F.2. In particular, we show (Corollary 28) the lower-order term $\tilde{O}\big(\big(\mathfrak{s} \wedge \frac{1}{\varepsilon}\big)\,\mathsf{d}\big)$ can be replaced by $\tilde{O}\big(\theta_P(\beta+\varepsilon)^2\mathsf{d}\big)$, yielding an overall P-dependent query complexity bound $O\Big(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\big(\frac{1}{\delta}\big)\right)\Big) + \tilde{O}\big(\theta_P(\beta+\varepsilon)^2\mathsf{d}\big)$. we further argue, in Appendix F.1, that it is *not* possible (by any algorithm) to reduce this lower-order term to $\tilde{O}(\theta_P(\beta+\varepsilon)\mathsf{d})$ or even $\tilde{O}(\theta_P(0)\mathsf{d})$, though we also establish intermediate forms of the term, such as $\tilde{O}\Big(\theta_P(\beta+\varepsilon)\mathsf{d}\left(\frac{\beta+\varepsilon}{\varepsilon}\right)\Big)$.

Subregion-based (Margin-based) Active Learning: Shortly after the work of Balcan, Beygelzimer, and Langford (2006), which included an analysis of homogeneous linear classifiers under a uniform distribution, Balcan, Broder, and Zhang (2007) proposed a refinement of disagreement-based active learning specific to linear classifiers. Rather than querying every example in the region of disagreement $\mathrm{DIS}(V)$, they identified a *subregion* $R \subseteq \mathrm{DIS}(V)$ which suffices for the purpose of estimating differences of error rates $\mathrm{er}_P(f) - \mathrm{er}_P(g)$ among $f, g \in V$. The key idea is to choose R so that any $f, g \in V$ has $P_X(\{f \neq g\} \setminus R)$ small, so that R captures *most* of the disagreements between concepts $f, g \in V$ that are far apart. In their case, since they were specifically focusing on homogeneous linear classifiers (i.e., concepts $h_w(x) = \mathbb{1}[\langle w, x\rangle \geq 0]$ on $\mathcal{X} = \mathbb{R}^{\mathsf{d}}$) under P_X uniform in an origin-centered ball, they could describe this region R as a *slab* around the boundary of a current hypothesis $h_{\hat{w}}$: that is, $R = \big\{x \in \mathbb{R}^{\mathsf{d}} : |\langle \hat{w}, x\rangle| \leq b\big\}$ for an appropriate width b (which decreases over time as the algorithm progresses). In other words, the algorithm queries examples X_i with *low margin* under the current hypothesis $\hat{w}$. As such, this technique is referred to as *margin-based active learning*. They analyzed this technique for the realizable case and under a specialized noise condition (Tsybakov noise), and found it provides advantages over disagreement-based learning: in the realizable case, improving the query complexity from $\mathsf{d}^{3/2} \cdot \mathrm{polylog}\big(\frac{1}{\varepsilon\delta}\big)$ to $\mathsf{d} \cdot \mathrm{polylog}\big(\frac{1}{\varepsilon\delta}\big)$ (matching the query complexities achieved by earlier works Freund, Seung, Shamir, and Tishby, 1997; Dasgupta, Kalai, and Monteleoni, 2005; Dasgupta, 2005), while allowing for some robustness to non-realizable distributions P (albeit not fully agnostic). The technique was later extended in various ways, including studying adaptivity to certain noise parameters (Wang and Singh, 2016) and generalizing beyond the uniform distribution, to general isotropic log-concave or s-concave distributions (Balcan and Long, 2013; Balcan and Zhang, 2017).

This idea was extended to general concept classes $\mathbb{C}$ and distributions P, including the agnostic setting, in the work of Zhang and Chaudhuri (2014). Again the idea is to identify a region $R \subseteq \mathrm{DIS}(V)$ for which concepts $f, g \in V$ have only small disagreements outside R: $P_X(\{f \neq g\} \setminus R) \leq \eta$, for a small η. Rather than an explicit region R (as in margin-based active learning), they simply choose a subset of the unlabeled examples via a linear program, which they show (in the analysis) can be related to an optimal choice of such a region. We discuss this technique in detail in Appendix F.3 (where we compose this idea with the AVID principle to produce a refinement of the AVID Agnostic algorithm).

The implication of this refinement of disagreement-based learning is a P-dependent query complexity bound, stated in terms of a subregion-based refinement of the disagreement coefficient, defined as follows (adopting some simplifications from Hanneke, 2016b). As above, define the r-ball $\mathrm{B}_{P_X}(h^\star, r) = \{h \in \mathbb{C} : P_X(x : h(x) \neq h^\star(x)) \leq r\}$ for $r > 0$. Also, for $\eta \geq 0$, define

$$\Phi_{P_X}(\mathrm{B}_{P_X}(h^\star, r), \eta) := \inf\left\{ P_X(R) : \sup_{h \in \mathrm{B}_{P_X}(h^\star, r)} P_X(\{h \neq f\} \setminus R) \leq \eta, R \subseteq \mathcal{X}, f : \mathcal{X} \to \{0,1\} \right\},$$

where R and f are restricted to be measurable. Finally, for $\varepsilon \geq 0$, define the subregion disagreement coefficient (Definition 31) as

$$\varphi_P(\varepsilon, \eta) := \sup_{r > \eta + \varepsilon} \frac{\Phi_{P_X}(\mathrm{B}_{P_X}(h^\star, r), (r - \eta)/c)}{r} \vee 1,$$

for an appropriate universal constant $c > 1$. Their technique provides a P-dependent query complexity bound of the form $\tilde{O}\left(\varphi_P(\varepsilon, 2\beta)\mathsf{d}\left(\frac{\beta^2}{\varepsilon^2} + 1\right)\right)$. In particular, it follows from the definitions that $\varphi_P(\varepsilon, 2\beta) \leq \theta_P(2\beta + \varepsilon)$ (see Appendix F.3), with some examples where the gap is large. Thus, this represents a refinement of the query complexity bounds for disagreement-based active learning discussed above.

As a primary example where $\varphi_P(\varepsilon, 2\beta) \ll \theta_P(\beta + \varepsilon)$, consider again the scenario of homogeneous linear classifiers on $\mathbb{R}^\mathsf{d}$ under P_X an isotropic log-concave distribution (as considered in the margin-based active learning works of Balcan, Broder, and Zhang, 2007; Balcan and Long, 2013 discussed above). In this scenario, Zhang and Chaudhuri (2014) show that $\varphi_P(\varepsilon, 2\beta) = O\left(\log\left(\frac{\beta}{\varepsilon}\right)\right)$ (based on concentration arguments from Balcan and Long, 2013). Thus, in this scenario, the query complexity bound of Zhang and Chaudhuri (2014) is $\tilde{O}\left(\mathsf{d}\left(\frac{\beta^2}{\varepsilon^2} + 1\right)\right)$. In contrast, Hanneke (2007b) showed $\theta_P(\beta + \varepsilon) = \Omega\left(\sqrt{\mathsf{d}} \wedge \frac{1}{\beta + \varepsilon}\right)$ for P_X the uniform distribution on an origin-centered sphere (a special case of isotropic log-concave), so that the query complexity bounds for disagreement-based active learning are roughly $\mathsf{d}^{3/2}\left(\frac{\beta^2}{\varepsilon^2} + 1\right)$, hence are suboptimal by a factor $\sqrt{\mathsf{d}}$.

That said, in the context of distribution-free analysis, it is unclear whether there are advantages from this subregion technique. Specifically, Hanneke (2016b) showed that $\sup_P \varphi_P(\varepsilon, 0) = \mathfrak{s} \wedge \frac{1}{\varepsilon}$ (where the sup is restricted to realizable distributions P), which matches the worst-case value of $\theta_P(\varepsilon)$ (established by Hanneke and Yang, 2015). In (51) of Appendix F.3, we further extend this to $\varphi_P(\varepsilon, \eta)$ (using the fact that $\varphi_P(\varepsilon, \eta) \geq \varphi_P(\eta + \varepsilon, 0)$), establishing that (for $\varepsilon, \eta \geq 0$ with $\eta + \varepsilon \leq 1$)

$$\sup_P \varphi_P(\varepsilon, \eta) = \mathfrak{s} \wedge \frac{1}{\eta + \varepsilon},$$

where again the sup is restricted realizable distributions P. Thus, the implication of the P-dependent query complexity bound of Zhang and Chaudhuri (2014) for bounding the distribution-free query complexity $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C})$ is merely to recover the same query complexity bound $\tilde{O}\left(\left(\mathfrak{s} \wedge \frac{1}{\beta + \varepsilon}\right)\mathsf{d}\left(\frac{\beta^2}{\varepsilon^2} + 1\right)\right)$ already known to hold for disagreement-based active learning. In particular, this means that the above query complexity bound of Zhang and Chaudhuri (2014) is capable of providing improvements in the distribution-free query complexity of active learning, compared to the sample complexity $\mathcal{M}_p(\varepsilon, \delta; \beta, \mathbb{C})$ of passive learning, *if and only if* $\mathfrak{s} < \infty$ (again, a rather strong restriction). Again, this contrasts with Theorems 1, 3, which provide improvements for *all* concept classes $\mathbb{C}$, regardless of $\mathfrak{s}$, with $\mathfrak{s}$ merely influencing refinements in the lower-order term in Theorem 3.

In Appendix F.3, we give a refinement of the AVID Agnostic algorithm, which adopts this subregion technique (in combination with the AVID principle). We show this Subregion-AVID Agnostic algorithm achieves a P-dependent refinement of the lower-order term compared to the original AVID Agnostic algorithm. For instance, one implication of this refinement is replacing the term $\tilde{O}\left(\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right)\mathsf{d}\right)$ in Theorem 3 with $\tilde{O}\left(\varphi_P(\varepsilon, 5\beta)^2\mathsf{d}\right)$, yielding a P-dependent query complexity bound $O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\varphi_P(\varepsilon, 5\beta)^2\mathsf{d}\right)$. It follows from an example in Appendix F.1 that the above quadratic dependence $\varphi_P(\varepsilon, 5\beta)^2$ cannot be reduced to $\varphi_P(\varepsilon, 5\beta)$ (or even $\varphi_P(0, 0)$) without introducing additional factors, though we also establish intermediate forms of the term, such as $\tilde{O}\left(\varphi_P(\varepsilon, 5\beta)\mathsf{d}\left(\frac{\beta + \varepsilon}{\varepsilon}\right)\right)$.

A.3 Background of the Techniques

Having surveyed much of the related work on agnostic active learning above, we conclude our discussion of related work by discussing previous works in the learning theory literature containing ideas related to our main technique (the AVID principle).

Background of the AVID Principle: Arguably the main innovation involved in this work is the decomposition of the space $\mathcal{X}$ into regions $\mathcal{X} \setminus \Delta_{i_k}$ and Δ_{i_k}, and augmenting the predictor $\hat{h}_k$ to be a (shallow) *decision list* of concepts from $\mathbb{C}$. One key inspiration for the main idea underlying the technique is rooted in the works of Bousquet and Zhivotovskiy (2021); Puchkin and Zhivotovskiy (2022) on prediction with an *abstention* option (evaluated with the *Chow loss*). Interestingly, this continues a long precedent of finding useful connections and cross-inspirations between active learning and prediction with abstentions (Mitchell, 1979; El-Yaniv and Wiener, 2010, 2012; Zhang and Chaudhuri, 2014, e.g.,). Specifically, Bousquet and Zhivotovskiy (2021); Puchkin and Zhivotovskiy (2022) consider methods exhibiting a kind of *transition time*, in which they determine that, for some $f, g \in \mathbb{C}$, *abstaining* in a the pairwise disagreement region $\{x : f(x) \neq g(x)\}$, and predicting with f in its complement, comes out to have smaller Chow loss than the overall loss of the best $h \in \mathbb{C}$. Some reasoning very much analogous to this (and directly inspired by it) can be found in one of the base cases of the arguments in the present paper (namely, concerning the "early stopping" case in the algorithm), in which we find that in the case of early stopping (Step 4), we can find $f, g \in V_{k-1}$ and $h_1, h_2 \in \mathbb{C}$, such that predicting with h_1 in $\{f \neq g\} \setminus \Delta_{i_k}$ (rather than abstaining), with f in $\{f = g\} \setminus \Delta_{i_k}$, and with h_2 in Δ_{i_k}, produces a *smaller* overall error rate in compared to the best concept $h^\star \in \mathbb{C}$. Of course, the algorithm and analysis here contain many additional pieces on top of this, but it is interesting that this connection to learning with abstentions still remains present at the core (though it is noteworthy that this connection is qualitatively different from the usual one, in that here we are not replacing abstentions with queries, but rather that a part of the analysis inspires part of our analysis). We remark that this analysis of learning with abstentions by Bousquet and Zhivotovskiy (2021); Puchkin and Zhivotovskiy (2022) was also inspirational for an active learning method in the work of Zhu and Nowak (2022) (though the aim in that work is different from the present work, and the setting is generally not comparable to ours).

At a high level, we can view the technique as also analogous to an idea of Hanneke and Yang (2015) developed for the *benign noise* model: namely, the restriction of the agnostic setting to the case the *Bayes classifier* $h^\star_{\mathrm{Bayes}}(x) \mapsto \mathbb{1}[P(Y = 1|X = x) \geq 1/2]$ is in the concept class $\mathbb{C}$. Hanneke and Yang (2015) prove a query complexity bound for this special case which matches Theorem 3 (and indeed, refines the lower-order term's $\mathfrak{sd}$ dependence to simply $\mathfrak{s}$). In that context, since the $h^\star_{\mathrm{Bayes}} \in \mathbb{C}$, the only source of non-realizability is in the *noisiness* of the conditional label distribution $Y|X$. Thus, if an active learner could *repeatedly query* a given X_t to receive *multiple* conditionally independent samples of Y_t given X_t, it could use the majority vote of these samples to effectively *de-noise* the label of X_t, thereby identifying $h^\star_{\mathrm{Bayes}}(X_t)$. This strategy only fails if $P(Y = 1|X = X_t)$ is very close to $\frac{1}{2}$, in which case this de-noising would require too many queries to be worthwhile, particularly since such noisy examples have very little effect on the *excess* error rate $\mathrm{er}_P(\hat{h}) - \mathrm{er}_P(h^\star_{\mathrm{Bayes}})$. As such, if the active learner cannot identify the optimal label within some number of queries, it should *abandon* the example X_t and move on. Of course, in the model of active learning studied in this work, and in the work of Hanneke and Yang (2015), an active learner cannot actually obtain multiple conditionally independent copies of the label Y_t. However, by appropriate discretization of the space $\mathcal{X}$ based on the structure of the concept class $\mathbb{C}$, Hanneke and Yang (2015) are able to *approximate* this idealized behavior. The resulting algorithm effectively *adapts* to the noisiness of the labels of examples X_t within the equivalence classes induced by this discretization, allocating more queries to the noisier (high-label-variance) regions (and abandoning the regions it finds to be too noisy). In that sense, the high-level idea behind the AVID principle is similar in nature. The goal is to isolate the regions where learning is more challenging, due to higher variance in error difference estimation, and allocate disproportionately more queries to these regions. Of course, in the agnostic case, this is made much more challenging, since the source of non-realizabilityy is not merely label noise, but also model misspecification (i.e., $h^\star_{\mathrm{Bayes}} \notin \mathbb{C}$) so that de-noising the examples may sometimes have little benefit (e.g., it is even possible to have $\beta > 0$ while $P(Y = 1|X) \in \{0, 1\}$). As such, the AVID principle necessarily makes greater use of the structure of the concept class to isolate such regions of high variance in error difference estimation.

It is worth mentioning that other works on active learning have also considered decomposing the space $\mathcal{X}$ into subregions and learning separately in each region (e.g., Cortes, DeSalvo, Gentile, Mohri, and Zhang, 2019a; Cortes, DeSalvo, Gentile, Mohri, and Zhang, 2019b, 2020). However, we note that these works retain the above issue of having a query complexity of the form $c(\beta)\mathsf{d}\frac{\beta^2}{\varepsilon^2}$ for a complexity measure $c(\beta)$ (as discussed in Section 2) such that, in the worst case over distributions P (respecting the β constraint) the results become ultimately no smaller than the sample complexity of passive learning.

The idea of decomposing a predictor into a decision list based on pairwise disagreement regions has an even closer parallel in the recent work of Hanneke, Larsen, and Zhivotovskiy (2024b), which removes a log factor from the lead term in the (first-order) sample complexity of *passive* learning, thereby obtaining an optimal lead term of $\Theta\left(\frac{\beta}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)\right)$. The overall approach in that work is in many ways similar to the technique in the present work, though with some important differences in the actual algorithms. In particular, since the interest in that work is merely removing a factor $\log\left(\frac{1}{\beta}\right)$, it essentially suffices for the algorithm to reduce the best-in-class *error rate* in a region $\mathcal{X}\setminus\Delta$ down to $\frac{\beta}{\log(1/\beta)}$ (for $P_X(\Delta)=O(\beta)$), so that a uniform Bernstein inequality for the error rate of ERM implies the desired result in that region $\mathcal{X}\setminus\Delta$, and a uniform convergence analysis of ERM under the conditional distribution given Δ implies the desired result in the region Δ. In contrast, our interest in the present work is a factor of β in the lead term, with a lower-order term of size $\tilde{O}\left(\frac{\mathsf{d}}{\varepsilon}\right)$, and to achieve this our algorithm aims to reduce (below ε) the *diameter* of a set V_k of surviving concepts, in a region $\mathcal{X}\setminus\Delta$ (with $P_X(\Delta)=O(\beta)$). We achieve this via uniform estimation of error differences, using an appropriate number of samples from these two regions, while precisely controlling the schedule of decreases of this diameter in the algorithm (in part by increasing the Δ region as needed to maintain this schedule of diameter decreases). Nevertheless, the essential inspiration and strategy behind these two algorithms are notably related, perhaps indicating that the AVID principle might in fact be a widely useful idea.

B Additional Definitions and Notation

We provide additional definitions and notation required for the formal analysis. A fundamental quantity in statistical learning theory is the *VC dimension* (Vapnik and Chervonenkis, 1971), which plays an important role in characterizing the optimal query complexity (and optimal sample complexity of passive learning). It is defined as follows.

Definition 4. *For any concept class $\mathbb{C}$, the* VC dimension *of $\mathbb{C}$, denoted by* $\mathrm{VC}(\mathbb{C})$*, is defined as the supremum $n\in\mathbb{N}\cup\{0\}$ for which there exists a sequence $\{x_1,\ldots,x_n\}\in\mathcal{X}^n$ such that $\{(h(x_1),\ldots,h(x_n)):h\in\mathbb{C}\}=\{0,1\}^n$ (i.e., all 2^n classifications are realizable by $\mathbb{C}$).*

For brevity, in all results, proofs, and discussion below (where $\mathbb{C}$ is clear from the context), we will simply denote by $\mathsf{d}:=\mathrm{VC}(\mathbb{C})$. In all statements below, we suppose $\mathsf{d}<\infty$ (see Appendix G). Also note that, by our assumption that $|\mathbb{C}|\geq 3$ (see footnote 1), we always have $\mathsf{d}\geq 1$.

Additional Notation and Conventions: For any distribution P on $\mathcal{X}\times\{0,1\}$, denote by P_X the marginal distribution on $\mathcal{X}$. Throughout, we refer to any sequence $S\in(\mathcal{X}\times\{0,1\})^*$ as a *data set*. For any $x\in\mathbb{R}$, it will be convenient to define $\log(x)=\ln(\max\{x,e\})$, and for $x>0$ we define $\log(x/0)=x/0=\infty$ and $0\log(x/0)=0$. For $a,b\in\mathbb{R}\cup\{\infty\}$, we use $a\wedge b$ or $\min\{a,b\}$ to denote the minimum of a and b, and $a\vee b$ or $\max\{a,b\}$ to denote the maximum of a and b. We will make use of standard big-O notation (O, Ω, Θ effectively hide universal constant factors, while $\tilde{O}$, $\tilde{\Theta}$ effectively hide log factors) to simplify theorem statements. The precise constant and log factors will always be made explicit in the formal proofs. We also adopt a convention regarding conditional probabilities: all claims involving conditional probabilities given a random variable should be interpreted as holding almost surely (i.e., for a *version* of the conditional probability), such as when claiming that an event holds with conditional probability at least $1-\delta$ given a random variable X. We also continue the notational conventions introduced in Section 4, such as $\mathrm{ER}(h)$, $\mathrm{DIS}(\mathbb{C}')$, $\{f\neq g\}$, overloading set notation to treat $A\subseteq\mathcal{X}$ as notationally interchangeable with its labeled extension $A\times\{0,1\}$, extending notation for set-intersection to allow intersections with sequences, and defining empirical estimates $\hat{P}_S(A)=|S\cap A|/|S|$. See Section 4 for details of these conventions.

Measurability: We remark that, formally speaking, an active learning algorithm can be defined simply as a measurable function $\mathbb{A} : (\mathcal{X} \times \{0,1\})^m \times \mathcal{X} \to \{0,1\}$: that is, taking as input an i.i.d. data set $S = \{(X_i, Y_i)\}_{i \leq m}$ and an independent test point X and evaluating to a prediction $\mathbb{A}(S, X) \in \{0,1\}$. In this view, the number of *queries* is merely bookkeeping, keeping track of the dependences of this function on the labels Y_i. For simplicity of presentation, we have adopted the common colloquialism of referring to the function $\hat{h}$ *returned* by $\mathbb{A}(S)$, which in this view simply refers to the function $\mathbb{A}(S, \cdot)$, so that $\mathrm{er}_P(\hat{h})$ is simply the conditional expectation $\mathbb{E}[\mathbb{1}[\mathbb{A}(S, X) \neq Y]|S]$. The measurability of the algorithms $\mathbb{A}$ defined in this work follows from measurability of the individual operations involved in their execution[10] under the standard measure-theoretic assumptions on $(\mathcal{X}, \mathbb{C})$ specified in footnote 1. To simplify the presentation, we do not explicitly discuss this in the proofs.

C The Query Complexity of the AVID Agnostic Algorithm

This section presents a detailed version of Theorem 3, bounding the query complexity of the AVID Agnostic algorithm. Recall the definition of the algorithm and notation from Section 4. Before stating the theorem, we first discuss a few additional technical aspects of the algorithm omitted from the high-level description in Section 4, starting with an explicit specification of the quantities involved. Let c_0, c_1 be universal constants, defined by Lemmas 7 and 8 of Appendix D. We define[11] $C = \frac{11}{10}$, $C'' = \left(\frac{200C^3}{8-5C^3}\right)^2$, and $C' = \frac{\sqrt{C''}}{16}$. For a given $\varepsilon, \delta \in (0,1)$ (arguments to $\mathbb{A}_{\mathrm{avid}}$), as in Section 4 we let $N = \left\lceil \log_C\left(\frac{2}{\varepsilon}\right)\right\rceil$, and for $k \in \mathbb{N}$, let $\varepsilon_k = C^{1-k}$, and we then define $m_k := \left\lceil \frac{300C''c_0}{\varepsilon_k}\left(\mathsf{d}\log\left(\frac{C''c_0}{\varepsilon_k}\right) + \log\left(\frac{1}{\delta}\right)\right)\right\rceil$. The algorithm adaptively allocates data subsets S_k^1, S_k^2, $S_{k,i}^3$, S_k^4 during its execution, as described in Section 4.1. Recall that S_k^1, $S_{k,i}^3$, and S_k^4 are all of size m_k (for any k, i for which they exist). The data subset S_k^2 is of size m'_k, formally defined as follows. For the value i_k and the set Δ_{i_k} as defined in the algorithm at the time that S_k^2 is allocated (either in Step 2 for some value of k, or in Step 9, in which case let $k = N+1$), letting $\hat{p}_k := 2\hat{P}_{S_k^4}(\Delta_{i_k})$, define $m'_k := \left\lceil \frac{C''c_1^2\hat{p}_k}{\varepsilon_k^2}\left(\mathsf{d} + \log\left(\frac{4(3+N-k)^2}{\delta}\right)\right)\right\rceil$.

We remark that, for simplicity of presentation, we have described the algorithm without explicitly discussing what happens if the algorithm *runs out* of unlabeled examples while allocating examples to subsets S_k^2, $S_{k,i}^3$. In this event, the algorithm can simply halt and return an arbitrary predictor $\hat{h}$, as the analysis will account for this event in the δ failure probability. To avoid excessive clutter, we do not explicitly mention this case in the description of the algorithm or allocation of data subsets used therein (i.e., we explicitly discuss this only in the analysis, and indeed only in the final part of the proof; see the discussion at the start of Appendix E).

The following theorem provides a bound on the query complexity achieved by $\mathbb{A}_{\mathrm{avid}}$ along with a bound on the unlabeled data set size sufficient to achieve it. This result represents a detailed version of the upper bound in Theorem 3 of Section 4 (in particular, Theorems 1 and 3 are immediate implications of this result). The constant factors in the big-O will be made explicit in the formal proof. The proof is given in Appendix E.

Theorem 5 (Query Complexity of AVID Agnostic)**.** *For any concept class* $\mathbb{C}$ *with* $\mathrm{VC}(\mathbb{C}) < \infty$*, letting* $\mathsf{d} = \mathrm{VC}(\mathbb{C})$*, for every distribution* P *on* $\mathcal{X} \times \{0,1\}$*, letting* $\beta = \inf_{h \in \mathbb{C}} \mathrm{er}_P(h)$*, for any* $\varepsilon, \delta \in$

[10] The only part requiring some care in this regard is the definition of $\hat{h}_k$ in (3), where formally we require that, given V_{k-1}, Δ_{i_k}, m'_k, the function $(S_k^1, S_k^2, x) \mapsto \hat{h}_k(x)$ should be a measurable function; such a measurable function can be shown to exist assuming $\mathbb{C}$ (and therefore V_{k-1}) satisfies the conditions of footnote 1, following straightforwardly from arguments of (Dudley, 1999).

[11] For simplicity of presentation, the constant C plays two major roles in the algorithm. First, it controls the schedule of *diameter* guarantees ε_k in the algorithm. Second, it controls certain constant factors in uniform concentration guarantees employed in the proof (Lemma 10). If we were to separate these roles, into C_1 and C_2, respectively, the two values exhibit a trade-off. In particular, we can admit a schedule $\varepsilon_k = C_1^{1-k}$ for *any* choice of $1 < C_1 < 2$ by an appropriately large choice of C'' (diverging as $C_1 \to 2$) and corresponding $C_2 > 1$ sufficiently close to 1. The source of this 2 limitation is the multiplicative factor in Lemma 20 which, in the limit as $C'' \to \infty$ and $C_2 \to 1$, becomes $\frac{2}{2-C_1}$. We also remark that we have defined constants that enable the cleanest presentation of the algorithm and analysis. We leave the issue of optimizing the constants to minimize the query complexity for future work.

$(0,1)$, *if the algorithm* $\mathbb{A}_{\text{avid}}$ *is executed with parameters* (ε,δ)*, with any number* $m \geq M(\varepsilon,\delta;\beta)$ *of i.i.d.-*P *examples, for a value* $M(\varepsilon,\delta;\beta)$ *(defined in Lemma 24) satisfying*

$$M(\varepsilon,\delta;\beta) = O\left(\frac{\beta+\varepsilon}{\varepsilon^2}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)\right) = \tilde{O}\left(\frac{\beta\mathsf{d}}{\varepsilon^2}+\frac{\mathsf{d}}{\varepsilon}\right),$$

then with probability at least $1-\delta$*, the returned predictor* $\hat{h}$ *satisfies* $\operatorname{er}_P(\hat{h}) \leq \inf_{h\in\mathbb{C}}\operatorname{er}_P(h)+\varepsilon$ *and the algorithm makes a number of queries at most* $Q(\varepsilon,\delta;\beta)$ *(defined in Lemma 23) satisfying*

$$\begin{aligned} Q(\varepsilon,\delta;\beta) &= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)+\min\left\{\mathfrak{s}\log\left(\frac{1}{\varepsilon}\right),\frac{1}{\varepsilon}\right\}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)\right) \\ &= \tilde{O}\left(\frac{\beta^2\mathsf{d}}{\varepsilon^2}+\left(\mathfrak{s}\wedge\frac{1}{\varepsilon}\right)\mathsf{d}\right). \end{aligned}$$

Remark on adaptivity to β: We emphasize that the algorithm *does not need to know* β in its execution (i.e., it *adaptively* achieves the above query complexity bound *for all* β). A more subtle point worth noting is that we can also *run* the algorithm without *ourselves* knowing β (to choose m), since the guarantee on query complexity holds for *any* unlabeled sample size $m \geq M(\varepsilon,\delta;\beta)$. For instance, if we run the algorithm with a β-independent number of unlabeled examples $m = \tilde{\Theta}\left(\frac{1}{\varepsilon^2}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)\right)$, the query complexity bound $Q(\varepsilon,\delta;\beta)$ would remain valid as stated in Theorem 5. Additionally, in the proof (see Lemma 24), we show that, in a sense, even the *unlabeled* sample complexity $M(\varepsilon,\delta;\beta)$ is achieved *adaptively*, since the algorithm (with no knowledge of β) only actually uses the first (at most) $M(\varepsilon,\delta;\beta)$ unlabeled examples in the sequence. This is itself an interesting feature. In particular, if we consider an alternative setting where, rather than getting the unlabeled data altogether at the start, the algorithm can adaptively *sample* new unlabeled examples $X_i \sim P_X$ one-at-a-time during execution (i.e., it has access to an *unlabeled example oracle*, which it can use to construct the data subsets S_k^1, S_k^2, $S_{k,i}^3$, S_k^4, during execution), the analysis establishes that the algorithm will succeed while adaptively sampling at most $M(\varepsilon,\delta;\beta)$ unlabeled examples (and querying at most $Q(\varepsilon,\delta;\beta)$ of them), all *without* knowing β (or anything else about P).

D Concentration Inequalities

This section presents a number of useful concentration inequalities, essential to the analysis. We begin with the classic *multiplicative Chernoff bound* (Chernoff, 1952; Bernstein, 1924). We will find the following particular form to be useful; since this is slightly different from the more-typical statements of Chernoff bounds, we include a brief explanation of how this result is derived from the more-standard exponential form.

Lemma 6 (Multiplicative Chernoff bound)**.** *Fix any* $p \in [0,1]$ *and* $n \in \mathbb{N}$*, and let* $B_1,\ldots,B_n$ *be i.i.d.* Bernoulli(p) *random variables. Let* $\bar{B} := \frac{1}{n}\sum_{i=1}^n B_i$*. For any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, the following both hold:*

$$p \leq \max\left\{2\bar{B}, \frac{8}{n}\ln\left(\frac{2}{\delta}\right)\right\},$$
$$\bar{B} \leq \max\left\{2p, \frac{6}{n}\ln\left(\frac{2}{\delta}\right)\right\}.$$

Proof. We include a brief explanation, based on more well-known exponential forms of the Chernoff bound: namely, $\mathbb{P}(\bar{B} < (1/2)p) \leq e^{-np/8}$ and $\mathbb{P}(\bar{B} > 2p) \leq e^{-np/3}$ (see e.g., Zhang, 2023).

For the first inequality in the lemma, we note that it trivially holds if $p < \frac{8}{n}\ln\left(\frac{2}{\delta}\right)$, and otherwise, if $p \geq \frac{8}{n}\ln\left(\frac{2}{\delta}\right)$, then by the above exponential tail bound, we have $\mathbb{P}(\bar{B} < (1/2)p) \leq e^{-np/8} \leq \frac{\delta}{2}$. For the second claimed inequality, note that it trivially holds if $\frac{6}{n}\ln\left(\frac{2}{\delta}\right) \geq 1$, so let us focus on the case $\frac{6}{n}\ln\left(\frac{2}{\delta}\right) < 1$. Note that for $p' \in [0,1]$ and $B_1',\ldots,B_n'$ i.i.d. Bernoulli(p'), and $\bar{B}' = \frac{1}{n}\sum_{i=1}^n B_i'$, for any $x \in \mathbb{R}$ the value of $\mathbb{P}(\bar{B}' > x)$ is non-decreasing in p'. Thus, letting $p' = \max\left\{p, \frac{3}{n}\ln\left(\frac{2}{\delta}\right)\right\} \geq p$, this monotonicity (together with the second exponential tail bound above) implies $\mathbb{P}(\bar{B} > 2p') \leq \mathbb{P}(\bar{B}' > 2p') \leq e^{-np'/3} \leq \frac{\delta}{2}$. The lemma then follows by the union

bound, so that both of these inequalities hold simultaneously with probability at least $1-\delta$. ■

We will also rely heavily on *uniform* concentration inequalities. Toward stating these, we first introduce additional useful notation.

VC dimension of collections of sets: As is standard in the literature, we overload the definition of *VC dimension* (Vapnik and Chervonenkis, 1971) to also allow for collections of sets. Formally, for any non-empty set $\mathcal{Z}$ and any non-empty $\mathcal{A} \subseteq 2^{\mathcal{Z}}$ (i.e., a collection of subsets of $\mathcal{Z}$), the VC dimension of $\mathcal{A}$, denoted by $\mathrm{VC}(\mathcal{A})$, is the supremum $n \in \mathbb{N} \cup \{0\}$ for which there exists $Z \subseteq \mathcal{Z}$ with $|Z| = n$ such that $\{Z \cap A : A \in \mathcal{A}\} = 2^Z$ (i.e., it is possible to pick out any subset of Z by intersection with an appropriate $A \in \mathcal{A}$). Equivalently, $\mathrm{VC}(\mathcal{A})$ is the VC dimension (Definition 4) of the *indicator functions* $\{\mathbb{1}_A : A \in \mathcal{A}\}$.

Uniform concentration term: For any non-empty set $\mathcal{Z}$ and any non-empty $\mathcal{A} \subseteq 2^{\mathcal{Z}}$, for any $n \in \mathbb{N}$ and $\delta \in (0,1)$, define (for a universal constant c_0 defined by Lemma 7 below)

$$\varepsilon(n, \delta; \mathcal{A}) := \frac{c_0}{n}\left(\mathrm{VC}(\mathcal{A}) \log\left(\frac{n}{\mathrm{VC}(\mathcal{A})}\right) + \log\left(\frac{1}{\delta}\right)\right). \tag{4}$$

The following result represents a uniform variant of the classic *Bernstein inequality* (or Bennett inequality) (Bernstein, 1924; Bennett, 1962). It can be derived from results proven by Vapnik and Chervonenkis (1974) (see Hanneke and Kpotufe, 2022 for an explicit derivation, via a layered application of Massart's lemma and Bousquet's inequality). We additionally include implications providing a uniform variant of multiplicative Chernoff bounds, which are easily derived from the stated uniform Bernstein inequality (taking $B = \emptyset$).

Lemma 7 (Uniform Bernstein and multiplicative Chernoff bounds)**.** *There is a finite universal constant $c_0 > 1$ for which the following holds. Fix any $n \in \mathbb{N}$, $\delta \in (0,1)$, any non-empty set $\mathcal{Z}$, and any set $\mathcal{A} \subseteq 2^{\mathcal{Z}}$ with $\mathrm{VC}(\mathcal{A}) < \infty$.*[12] *Define $\varepsilon(n, \delta; \mathcal{A})$ as in (4). Fix any distribution P on $\mathcal{Z}$ and let $Z = \{Z_1, \ldots, Z_n\} \sim P^n$ (i.i.d. P random variables). For any measurable set $A \subseteq \mathcal{Z}$, define its* empirical probability $\hat{P}_Z(A) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}[Z_i \in A]$. *With probability at least $1-\delta$, every $A, B \in \mathcal{A} \cup \{\emptyset\}$ satisfy the following (where $A \oplus B := (A \setminus B) \cup (B \setminus A)$ denotes the symmetric difference)*

$$\left|(\hat{P}_Z(A) - \hat{P}_Z(B)) - (P(A) - P(B))\right| \\ \leq \sqrt{\min\{P(A \oplus B), \hat{P}_Z(A \oplus B)\}\varepsilon(n, \delta; \mathcal{A})} + \varepsilon(n, \delta; \mathcal{A}).$$

Moreover, for any $\varepsilon > 0$ and $\alpha \in (0,1)$ satisfying $\varepsilon(n, \delta; \mathcal{A}) \leq \frac{\alpha^2}{4}\varepsilon$, the above inequality immediately yields the following implications: $\forall A \in \mathcal{A}$,

$$\hat{P}_Z(A) \geq \varepsilon \implies P(A) > (1-\alpha)\varepsilon, \qquad \text{or equivalently, } P(A) \leq (1-\alpha)\varepsilon \implies \hat{P}_Z(A) < \varepsilon$$
$$P(A) \geq \varepsilon \implies \hat{P}_Z(A) > (1-\alpha)\varepsilon, \qquad \text{or equivalently, } \hat{P}_Z(A) \leq (1-\alpha)\varepsilon \implies P(A) < \varepsilon.$$

We also make use of a uniform concentration inequality which refines the classic uniform convergence bound $\sqrt{\frac{1}{n}\left(\mathrm{VC}(\mathcal{A}) + \log\left(\frac{1}{\delta}\right)\right)}$ of Talagrand (1994) in the case that $\bigcup \mathcal{A}$ has small measure under P. The lemma is well-known in the literature, and follows immediately from expectation bounds based on chaining involving an *envelope* function (e.g., Theorem 2.14.1 of van der Vaart and Wellner, 1996) together with Bousquet's inequality (Bousquet, 2002) to achieve high probability. For completeness, we provide a brief direct proof, by simply applying the uniform convergence bound of Talagrand (1994) to the samples from the conditional distribution given a set $D \supseteq \bigcup \mathcal{A}$.

Lemma 8. *There is a finite universal constant $c_1 \geq 1$ for which the following holds. Let $\mathcal{A}$ be as in Lemma 7, and suppose $D \subseteq \mathcal{Z}$ is a measurable set such that $\forall A \in \mathcal{A}$, $A \subseteq D$. Then for the same quantities as Lemma 7, if $P(D) \geq \frac{9}{n}\ln\left(\frac{4}{\delta}\right)$, then with probability at least $1-\delta$, $\forall A \in \mathcal{A}$*

$$\left|\hat{P}_Z(A) - P(A)\right| \leq c_1\sqrt{\frac{P(D)}{n}\left(\mathrm{VC}(\mathcal{A}) + \log\left(\frac{1}{\delta}\right)\right)}.$$

[12] We suppose standard mild measure-theoretic restrictions on $\mathcal{A}$ and the σ-algebra of $\mathcal{Z}$, from empirical process theory: namely, the image-admissible Suslin condition (Dudley, 1999).

Proof. Note that the samples in $Z \cap D$ are conditionally i.i.d. $P(\cdot|D)$ given $|Z \cap D|$. For each $A \in \mathcal{A}$, denote by $\hat{P}_Z(A|D) := \hat{P}_{Z\cap D}(A)$ (or 0 if $|Z \cap D| = 0$). Applying the uniform convergence bound of Talagrand (1994) to the samples in $Z \cap D$ under the conditional distribution given $|Z \cap D|$, together with the law of total probability, yields that, with probability at least $1 - \frac{\delta}{2}$, $\forall A \in \mathcal{A}$,

$$\left|\hat{P}_Z(A|D) - P(A|D)\right| \le c_1' \sqrt{\frac{1}{|Z \cap D|}\left(\mathrm{VC}(\mathcal{A}) + \log\left(\frac{2}{\delta}\right)\right)}, \tag{5}$$

for a finite universal constant $c_1' \ge 1$. Moreover, by Bernstein's inequality (see Theorem 2.10 of Boucheron, Lugosi, and Massart, 2013), with probability at least $1 - \frac{\delta}{2}$,

$$\left|\hat{P}_Z(D) - P(D)\right| \le \sqrt{\frac{2P(D)}{n}\ln\left(\frac{4}{\delta}\right)} + \frac{1}{n}\ln\left(\frac{4}{\delta}\right) \le 2\sqrt{\frac{P(D)}{n}\ln\left(\frac{4}{\delta}\right)}, \tag{6}$$

where the last inequality is due to the assumption that $P(D) \ge \frac{9}{n}\ln\left(\frac{4}{\delta}\right)$. By the union bound, these two events occur simultaneously with probability at least $1 - \delta$. Suppose this occurs. In particular, by the assumption that $P(D) \ge \frac{9}{n}\ln\left(\frac{4}{\delta}\right)$, (6) further implies

$$\frac{1}{n}|Z \cap D| = \hat{P}_Z(D) \ge P(D) - 2\sqrt{\frac{P(D)}{n}\ln\left(\frac{4}{\delta}\right)} \ge \frac{1}{3}P(D),$$

so that the right hand side of (5) is at most

$$c_1'\sqrt{\frac{3}{nP(D)}\left(\mathrm{VC}(\mathcal{A}) + \log\left(\frac{2}{\delta}\right)\right)}.$$

Combining this with (5) and (6) implies that $\forall A \in \mathcal{A}$, since $A \subseteq D$,

$$\begin{aligned}
\left|\hat{P}_Z(A) - P(A)\right| &= \left|\hat{P}_Z(D)\hat{P}_Z(A|D) - P(D)P(A|D)\right| \\
&\le P(D)\left|\hat{P}_Z(A|D) - P(A|D)\right| + \hat{P}_Z(A|D)2\sqrt{\frac{P(D)}{n}\ln\left(\frac{4}{\delta}\right)} \\
&\le c_1'\sqrt{\frac{3P(D)}{n}\left(\mathrm{VC}(\mathcal{A}) + \log\left(\frac{2}{\delta}\right)\right)} + 2\sqrt{\frac{P(D)}{n}\ln\left(\frac{4}{\delta}\right)} \\
&\le c_1\sqrt{\frac{P(D)}{n}\left(\mathrm{VC}(\mathcal{A}) + \log\left(\frac{1}{\delta}\right)\right)},
\end{aligned}$$

where $c_1 := c_1'\sqrt{6} + 2\sqrt{\ln(4e)}$ (recalling $\log(x) := \ln(x \vee e)$). ■

E Proof of Theorem 5: Query Complexity of the AVID Agnostic Algorithm

The formal proof of Theorem 5, given at the end of this section, will be built up from a sequence of lemmas, roughly following the outline presented in Section 4.2.

Throughout this section, we fix an arbitrary concept class $\mathbb{C}$ (with $\mathsf{d} := \mathrm{VC}(\mathbb{C}) < \infty$) and distribution P on $\mathcal{X} \times \{0,1\}$, let $\beta = \inf_{h\in\mathbb{C}} \mathrm{er}_P(h)$, fix any $\varepsilon, \delta \in (0,1)$ (where ε, δ are inputs to the AVID algorithm), let $(X_1, Y_1), (X_2, Y_2), \ldots$ be independent P-distributed examples, and we let all values (N, ε_k, m_k, m_k', etc.) be defined as in Appendix C, based on these values ε, δ, and the examples $(X_1, Y_1), (X_2, Y_2), \ldots$. Also let $h^\star \in \mathbb{C}$ denote any concept with

$$\mathrm{er}_P(h^\star) < \inf_{h\in\mathbb{C}} \mathrm{er}_P(h) + \frac{\varepsilon}{10^4}. \tag{7}$$

For full generality, we do not assume there exists a minimizer achieving the infimum on the right hand side; rather, any choice of $h^\star$ satisfying this *near*-minimality property will suffice for our purposes in the analysis below.

To simplify the proof, we will establish the sequence of lemmas under a scenario where the algorithm is executed with an *inexhaustible* source of examples (for the adaptive allocation of data subsets): i.e., an infinite sequence $(X_1, Y_1), (X_2, Y_2), \ldots$ of independent P-distributed examples. However, it will follow from these lemmas that, with high probability, the algorithm only depends on a *finite* prefix $(X_1, Y_1), \ldots, (X_m, Y_m)$, for a sufficiently large $m = M(\varepsilon, \delta; \beta)$ as in Theorem 5 (see Lemma 24). At the end of the section, when combining the lemmas into a formal proof of Theorem 5, we will return to the standard setting where the algorithm has access *only* to such a finite prefix. In that context, the event that the algorithm attempts to access any examples (X_t, Y_t) with $t > m$ will be accounted for as part of the allowed δ-probability failure event, and thus (as mentioned in Appendix C) in such a case the algorithm can simply halt and return an arbitrary predictor $\hat{h}$. As mentioned in the remark following Theorem 5, the fact that the algorithm *adaptively* decides how many unlabeled examples to use is itself an interesting feature, as it means the algorithm can be considered adaptive to β even in its use of unlabeled examples.

Before proceeding with the proof, we first introduce some convenient notation regarding the values of k and i encountered in the algorithm. If the algorithm returns in Step 9, denote by $K := N + 1$, and otherwise, let K be the maximum value of k reached in the '*For*' loop in the algorithm; we argue in Lemma 10 below that the algorithm terminates eventually, with high probability, so that this latter case coincides with the case of returning in Step 4, with K being the value of k on which this occurs. Let $\mathcal{K} := \{1, \ldots, K \wedge N\}$: that is, the set of values of k encountered in the '*For*' loop in the algorithm. Also, for each $k \in \mathcal{K}$, denote by $\mathcal{I}_k$ the values of i encountered by the algorithm on round k; in particular, for $k < K$, $\mathcal{I}_k = \{i_k, \ldots, i_{k+1}\}$. In the case $K = N + 1$, for convenience also denote by $\mathcal{I}_{N+1} := \{i_{N+1}\}$.

We begin with a lemma which motivates our choice of sample size m_k for S_k^1, $S_{k,i}^3$, S_k^4. Recall $m_k := \left\lceil \frac{300C''c_0}{\varepsilon_k} \left(\mathsf{d} \log\left(\frac{C''c_0}{\varepsilon_k}\right) + \log\left(\frac{1}{\delta}\right)\right)\right\rceil$. Also recall our convention (adopted throughout this work) of treating sets $D \subseteq \mathcal{X}$ as notationally interchangeable with their labeled extension $D \times \{0,1\}$, such as in $A \cap D$ or $A \setminus D$ for $A \subseteq \mathcal{X} \times \{0,1\}$.

Lemma 9. *Fix any set $D \subseteq \mathcal{X}$ and define a family of subsets of $\mathcal{X} \times \{0,1\}$:*

$$\mathcal{A} = \Big\{ ((\mathrm{ER}(f) \cap \{f = g\}) \cup (\mathrm{ER}(h) \cap \{f \neq g\})) \setminus D : f, g, h \in \mathbb{C} \Big\}$$
$$\cup \Big\{ (\{f \neq g\} \times \{0,1\}) \setminus D : f, g \in \mathbb{C} \Big\} \cup \Big\{ \mathrm{ER}(h) \setminus D : h \in \mathbb{C} \Big\}.$$

For any $n \in \mathbb{N}$ and $\delta' \in (0,1)$, let $\varepsilon(n, \delta'; \mathcal{A})$ be defined as in (4). For each $k \in \{1, \ldots, N+1\}$, letting $\delta_k := \frac{\delta \varepsilon_{k+3}^2}{72}$, it holds that

$$\varepsilon(m_k, \delta_k; \mathcal{A}) < \frac{\varepsilon_k}{C''}. \tag{8}$$

Proof. We begin by bounding $\mathrm{VC}(\mathcal{A})$, as needed to evaluate $\varepsilon(m_k, \delta_k; \mathcal{A})$. Define the following families of subsets of $\mathcal{X} \times \{0,1\}$:

$$\mathcal{A}_0 := \{\mathrm{ER}(h) : h \in \mathbb{C}\} \cup \{\emptyset, \mathcal{X} \times \{0,1\}\},$$
$$\mathcal{A}_1 := \{\{f \neq g\} \times \{0,1\} : f, g \in \mathbb{C}\} \cup \{\mathcal{X} \times \{0,1\}\},$$
$$\mathcal{A}_2 := \{((A \setminus C) \cup (B \cap C)) \setminus D : A, B \in \mathcal{A}_0, C \in \mathcal{A}_1\}.$$

First note that $\mathcal{A} \subseteq \mathcal{A}_2$. To see this, note that for any $f, g, h \in \mathbb{C}$, taking $A = \mathrm{ER}(f)$, $B = \mathrm{ER}(h)$, $C = \{f \neq g\} \times \{0,1\}$, we have that $((\mathrm{ER}(f) \cap \{f = g\}) \cup (\mathrm{ER}(h) \cap \{f \neq g\})) \setminus D = ((A \setminus C) \cup (B \cap C)) \setminus D \in \mathcal{A}_2$. Similarly, for any $f, g \in \mathbb{C}$, taking $A = \emptyset$, $B = \mathcal{X} \times \{0,1\}$, $C = \{f \neq g\} \times \{0,1\}$ reveals $(\{f \neq g\} \times \{0,1\}) \setminus D = ((A \setminus C) \cup (B \cap C)) \setminus D \in \mathcal{A}_2$. Finally, for $h, f \in \mathbb{C}$, taking $A = \mathrm{ER}(h)$, $B = \emptyset$, $C = \{f \neq f\} \times \{0,1\} = \emptyset$ reveals $\mathrm{ER}(h) \setminus D = ((A \setminus C) \cup (B \cap C)) \setminus D \in \mathcal{A}_2$.

Next we bound $\mathrm{VC}(\mathcal{A}_2)$. It is immediate from the definition that $\mathrm{VC}(\{\mathrm{ER}(h) : h \in \mathbb{C}\}) = \mathsf{d}$. Moreover, this implies $\mathrm{VC}(\mathcal{A}_0) \leq \mathsf{d} + 2$ (Vidyasagar, 2003, Lemma 4.11). Also note that $\mathcal{A}_1 \subseteq \{A \oplus B : A, B \in \mathcal{A}_0\}$, where $A \oplus B := (A \setminus B) \cup (B \setminus A)$ is the symmetric difference: that is, trivially $(\mathcal{X} \times \{0,1\}) \oplus \emptyset = \mathcal{X} \times \{0,1\}$, and for any $f, g \in \mathbb{C}$, $\{f \neq g\} \times \{0,1\} = \mathrm{ER}(f) \oplus \mathrm{ER}(g)$. Thus, any element of $\mathcal{A}_2$ can be expressed as a fixed function of four sets $A, B, A', B' \in \mathcal{A}_0$: namely $(A, B, A', B') \mapsto ((A \setminus (A' \oplus B')) \cup (B \cap (A' \oplus B'))) \setminus D$. Based on this fact, well-known results about the effect of such combinations on the VC dimension imply $\mathrm{VC}(\mathcal{A}_2) = O(\mathrm{VC}(\mathcal{A}_0))$: explicitly, Theorem 4.5 of Vidyasagar (2003) implies $\mathrm{VC}(\mathcal{A}_2) \leq 25\mathrm{VC}(\mathcal{A}_0) \leq 25(\mathsf{d} + 2)$. By the

assumption that $|\mathbb{C}| \geq 3$ (footnote 1) we know $\mathsf{d} \geq 1$, so that $25(\mathsf{d}+2) \leq 75\mathsf{d}$. Altogether, we have $\mathrm{VC}(\mathcal{A}) \leq 75\mathsf{d}$.

With this in mind, we may note that (also using that $\mathsf{d} \geq 1$ and $C'' \geq 9C^3$)

$$\begin{aligned} m_k &\geq \frac{300C''c_0}{\varepsilon_k}\left(\mathsf{d}\log\left(\frac{C''c_0}{\varepsilon_k}\right)+\log\left(\frac{1}{\delta}\right)\right) \\ &\geq \frac{150C''c_0}{\varepsilon_k}\left(\mathsf{d}\log\left(\frac{C''c_0}{\varepsilon_k}\right)+\log\left(\frac{9C^3}{\delta\varepsilon_k}\right)\right) \\ &\geq \frac{2C''c_0}{\varepsilon_k}\left(\mathrm{VC}(\mathcal{A})\log\left(\frac{C''c_0}{\varepsilon_k}\right)+\log\left(\frac{1}{\delta_k}\right)\right). \end{aligned} \tag{9}$$

In particular, if $\mathrm{VC}(\mathcal{A}) \geq 1$, then by Corollary 4.1 of Vidyasagar (2003), (9) implies

$$m_k > \frac{C''c_0}{\varepsilon_k}\left(\mathrm{VC}(\mathcal{A})\log\left(\frac{m_k}{\mathrm{VC}(\mathcal{A})}\right)+\log\left(\frac{1}{\delta_k}\right)\right). \tag{10}$$

Moreover, if $\mathrm{VC}(\mathcal{A}) = 0$, then recalling we define $0\log(1/0) = 0$, (9) trivially implies (10) in this case as well. Thus, regardless of the value of $\mathrm{VC}(\mathcal{A})$, by definition of $\varepsilon(m_k, \delta_k; \mathcal{A})$, the claim in (8) follows from (10). ■

We continue the proof with a lemma conveniently summarizing several uniform concentration bounds which are useful in various places throughout the rest of the proof. In particular, the lemma focuses on concentration inequalities in the $\mathcal{X} \setminus \Delta_{i_k}$ region of focus of the learning algorithm. It will therefore be convenient to explicitly define the portion of the functions in $V_{k-1}^{(4)}$ specific to this region: namely, for every $k \in \{1, \ldots, K\}$, define[13]

$$V_{k-1}^{(3)} := \{f\mathbb{1}_{\{f=g\}} + h\mathbb{1}_{\{f\neq g\}} : f, g \in V_{k-1}, h \in \mathbb{C}\}.$$

Lemma 10. *On an event E_0 of probability at least $1 - \frac{\delta}{4}$, for every $k \in \{1, \ldots, K\}$, it holds that*

$$\begin{aligned} \forall h, h' \in V_{k-1}^{(3)}, \Bigg| &\left(\hat{P}_{S_k^1}(\mathrm{ER}(h) \cap D_{k-1} \setminus \Delta_{i_k}) - \hat{P}_{S_k^1}(\mathrm{ER}(h') \cap D_{k-1} \setminus \Delta_{i_k})\right) \\ &- \left(P(\mathrm{ER}(h) \setminus \Delta_{i_k}) - P(\mathrm{ER}(h') \setminus \Delta_{i_k})\right)\Bigg| \\ &< \sqrt{P_X(\{h \neq h'\} \setminus \Delta_{i_k})\frac{\varepsilon_k}{C''}} + \frac{\varepsilon_k}{C''}, \end{aligned} \tag{11}$$

and for every $k \in \mathcal{K}$ and every $i \in \mathcal{I}_k$, $\forall f, g \in \mathbb{C}$,

$$\hat{P}_{S_{k,i}^3}(\{f \neq g\} \setminus \Delta_i) \geq \varepsilon_{k+2} \implies P_X(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+3} \tag{12}$$

$$\hat{P}_{S_{k,i}^3}(\{f \neq g\} \setminus \Delta_i) \leq \varepsilon_{k+2} \implies P_X(\{f \neq g\} \setminus \Delta_i) < \varepsilon_{k+1}, \tag{13}$$

and moreover, $\max \mathcal{I}_k \leq \frac{1}{\varepsilon_{k+3}}$. In particular, the latter implies the algorithm eventually terminates (in Step 9 if $K = N+1$, or in Step 4 if $K \leq N$).

Proof. Consider any $k \in \{1, \ldots, N+1\}$ having a non-zero probability of $k \leq K$. Let δ_k be as in Lemma 9. Recall that the data set S_k^1 is independent of all data involved in rounds $k' < k$ in the algorithm, whereas the event $k \leq K$ and (in this event) the set Δ_{i_k} are entirely determined by data involved in rounds $k' < k$. Thus, even conditioned on the event that $k \leq K$ and and the set Δ_{i_k}, the data set S_k^1 remains conditionally i.i.d.-P. Therefore, letting $\mathcal{A}_k$ denote the set $\mathcal{A}$ as defined in Lemma 9 with $D = \Delta_{i_k}$, applying the uniform Bernstein inequality (Lemma 7 in Appendix D) with this $\mathcal{A}_k$ under the conditional distribution given the event $k \leq K$ and the set Δ_{i_k} implies that, with conditional probability at least $1 - \delta_k$ given the event $k \leq K$ and the set Δ_{i_k}, it holds that $\forall A, B \in \mathcal{A}_k$,

$$\left|\left(\hat{P}_{S_k^1}(A) - \hat{P}_{S_k^1}(B)\right) - \left(P(A) - P(B)\right)\right| \leq \sqrt{P(A \oplus B)\varepsilon(m_k, \delta_k; \mathcal{A}_k)} + \varepsilon(m_k, \delta_k; \mathcal{A}_k). \tag{14}$$

[13] Since this work focuses on binary classification, $V_{k-1}^{(3)}$ can equivalently be stated as $\{\mathrm{Maj}(f, g, h) : f, g \in V_{k-1}, h \in \mathbb{C}\}$, where $\mathrm{Maj}(f, g, h)(x) = \mathbb{1}[f(x) + g(x) + h(x) \geq 2]$ is the majority vote function. The definition of $V_{k-1}^{(3)}$ above expresses a more-general form, which, as we discuss in Section G, also extends to *multiclass* classification.

By the law of total probability, on an event $E_{0,k}$ of probability at least $1-\delta_k$, if $k \le K$ (and thus Δ_{i_k} and $\mathcal{A}_k$ are defined) then (14) holds $\forall A, B \in \mathcal{A}_k$.

In particular, on the event $E_{0,k}$, supposing $k \le K$, if we consider any $h, h' \in V^{(3)}_{k-1}$, then for the sets $A = \mathrm{ER}(h) \setminus \Delta_{i_k} \in \mathcal{A}_k$ and $B = \mathrm{ER}(h') \setminus \Delta_{i_k} \in \mathcal{A}_k$, we may note that the symmetric difference $A \oplus B = (\mathrm{ER}(h) \oplus \mathrm{ER}(h')) \setminus \Delta_{i_k} = (\{h \ne h'\} \times \{0,1\}) \setminus \Delta_{i_k}$, so that together with (8) of Lemma 9, (14) implies

$$\left|\left(\hat{P}_{S^1_k}(\mathrm{ER}(h) \setminus \Delta_{i_k}) - \hat{P}_{S^1_k}(\mathrm{ER}(h') \setminus \Delta_{i_k})\right) - \left(P(\mathrm{ER}(h) \setminus \Delta_{i_k}) - P(\mathrm{ER}(h') \setminus \Delta_{i_k})\right)\right| < \sqrt{P_X(\{h \ne h'\} \setminus \Delta_{i_k})\frac{\varepsilon_k}{C''}} + \frac{\varepsilon_k}{C''}. \tag{15}$$

To arrive at the claim in (11), we merely note that for any $f, g, f', g' \in V_{k-1}$ and $h, h' \in \mathbb{C}$, letting $\mathrm{DL}(f,g,h) := f\mathbb{1}_{\{f=g\}} + h\mathbb{1}_{\{f\ne g\}}$ and $\mathrm{DL}(f',g',h') := f'\mathbb{1}_{\{f'=g'\}} + h'\mathbb{1}_{\{f'\ne g'\}}$, for any $x \in \mathcal{X} \setminus D_{k-1}$, we have $g(x) = f(x) = f'(x) = g'(x)$, so that $\mathrm{DL}(f,g,h)(x) = f(x) = f'(x) = \mathrm{DL}(f',g',h')(x)$. Thus, any $h, h' \in V^{(3)}_{k-1}$ have $h(x) = h'(x)$ for all $x \notin D_{k-1}$, and therefore

$$\hat{P}_{S^1_k}(\mathrm{ER}(h) \cap D_{k-1} \setminus \Delta_{i_k}) - \hat{P}_{S^1_k}(\mathrm{ER}(h') \cap D_{k-1} \setminus \Delta_{i_k}) = \hat{P}_{S^1_k}(\mathrm{ER}(h) \setminus \Delta_{i_k}) - \hat{P}_{S^1_k}(\mathrm{ER}(h') \setminus \Delta_{i_k}),$$

so that (11) follows from (15). To unify the discussion below, for any $k \in \{1, \ldots, N+1\}$ with probability zero of $k \le K$, also denote by $E_{0,k}$ the event (of probability one) that $k > K$.

Turning now to the claims in (12) and (13), consider any (k,i) having non-zero probability that $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$. Note that, since $S^3_{k,i}$ is a data set of size m_k, allocated from the remaining *unused* unlabeled data upon reaching Step 5 with values (k,i) (noting this can happen at most once in the algorithm), the samples in $S^3_{k,i}$ are conditionally i.i.d.-P given $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$, and moreover, $S^3_{k,i}$ is conditionally independent of Δ_i given the events that $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$. In the event that $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$, let $\mathcal{A}_{k,i}$ denote the set $\mathcal{A}$ as defined in Lemma 9 with $D = \Delta_i$. Recalling again our definition of $C = \frac{11}{10}$ and $C'' \ge 32C^5\left(\frac{C}{C-1}\right)^2$, note that for $\alpha = 1 - \frac{1}{C}$, (8) of Lemma 9 implies $\varepsilon(m_k, \delta_k; \mathcal{A}_{k,i}) < \frac{\varepsilon_k}{C''} < \frac{\alpha^2}{4}\varepsilon_{k+2} < \frac{\alpha^2}{4}\varepsilon_{k+1}$. Therefore, applying Lemma 7 of Appendix D under the conditional distribution given the events that $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$ and the set Δ_i, we have that with conditional probability at least $1-\delta_k$, $\forall f, g \in \mathbb{C}$, the set $(\{f \ne g\} \setminus \Delta_i) \times \{0,1\} \in \mathcal{A}_{k,i}$ satisfies

$$\hat{P}_{S^3_{k,i}}(\{f \ne g\} \setminus \Delta_i) \ge \varepsilon_{k+2} \implies P_X(\{f \ne g\} \setminus \Delta_i) > (1-\alpha)\varepsilon_{k+2} = \varepsilon_{k+3}$$

and

$$\hat{P}_{S^3_{k,i}}(\{f \ne g\} \setminus \Delta_i) \le \varepsilon_{k+2} = (1-\alpha)\varepsilon_{k+1} \implies P_X(\{f \ne g\} \setminus \Delta_i) < \varepsilon_{k+1}.$$

By the law of total probability, there is an event $E_{0,k,i}$ of probability at least $1-\delta_k$, on which, if $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$, then the above inequalities hold $\forall f, g \in \mathbb{C}$. To unify cases, for any (k,i) with $k \le N$ and $i \le 1/\varepsilon_{k+3}$ having probability zero of satisfying $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$, also define $E_{0,k,i}$ as the event (of probability one) that either $k \notin \mathcal{K}$ or $i \notin \mathcal{I}_k$.

We have thus established (11) for all $k \le K$ and (12 - 13) for all $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$ with $i \le 1/\varepsilon_{k+3}$, on the event $E_0 := \left(\bigcap_{k \le N+1} E_{0,k}\right) \cap \bigcap_{k \le N} \bigcap_{i \le 1/\varepsilon_{k+3}} E_{0,k,i}$. By the union bound, E_0 fails with probability at most

$$\sum_{k=1}^{N+1}\left(\delta_k + \sum_{i \le 1/\varepsilon_{k+3}} \delta_k\right) \le \sum_{k=1}^{N+1}\left(1 + \frac{1}{\varepsilon_{k+3}}\right)\delta_k \le \sum_{k=1}^{N+1}\frac{2\delta_k}{\varepsilon_{k+3}} = \sum_{k=1}^{N+1}\frac{\delta}{36}\varepsilon_{k+3} < \frac{\delta}{4},$$

where the equality follows from our definition of $\delta_k = \frac{\delta\varepsilon^2_{k+3}}{72}$ (from Lemma 9) and the last inequality follows from our choice of $C = \frac{11}{10}$.

Finally, we argue that, on the event E_0, for any $k \in \mathcal{K}$, the maximum value of $i \in \mathcal{I}_k$ satisfies $i \le 1/\varepsilon_{k+3}$. We argue this by induction. Specifically, we will argue that, for any $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$, $P_X(\mathcal{X} \setminus \Delta_i) \le 1 - i\varepsilon_{k+3}$. For the purpose of induction, suppose that for some $k \in \mathcal{K}$, we have $P_X(\mathcal{X} \setminus \Delta_{i_k}) \le 1 - i_k\varepsilon_{k+3}$ (which is trivially satisfied for $k = 1$, since $i_k = 0$, which can therefore serve as a base case for induction). Taking this i_k as a base case for a further nested

induction on $i \in \mathcal{I}_k$ (noting that i_k is the minimum element of $\mathcal{I}_k$), suppose that for some $i \in \mathcal{I}_k$ we have $P_X(\mathcal{X} \setminus \Delta_i) \leq 1 - i\varepsilon_{k+3}$. Since probabilities are non-negative, this necessarily implies $i \leq 1/\varepsilon_{k+3}$. Then note that, if i is not the maximal element of $\mathcal{I}_k$, the algorithm augments Δ_i in Step 7, so that $\Delta_{i+1} = \Delta_i \cup \{f \neq g\}$ for (f, g) defined in Step 6. By the criterion in Step 5, we further know that $\hat{P}_{S^3_{k,i}}(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+2}$. Since $i \leq 1/\varepsilon_{k+3}$, the event E_0 implies (12) holds, which therefore implies $P_X(\Delta_{i+1} \setminus \Delta_i) = P_X(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+3}$, so that $P_X(\mathcal{X} \setminus \Delta_{i+1}) = P_X(\mathcal{X} \setminus \Delta_i) - P_X(\Delta_{i+1} \setminus \Delta_i) < 1 - (i+1)\varepsilon_{k+3}$, thus extending the inductive hypothesis. By the principle of induction, this establishes that $P_X(\mathcal{X} \setminus \Delta_i) \leq 1 - i\varepsilon_{k+3}$ for every $i \in \mathcal{I}_k$. In particular, returning to the induction on k, in the event that this k is not the maximal element of $\mathcal{K}$, we have $i_{k+1} \in \mathcal{I}_k$, so that $P_X(\mathcal{X} \setminus \Delta_{i_{k+1}}) \leq 1 - i_{k+1}\varepsilon_{k+3} \leq 1 - i_{k+1}\varepsilon_{(k+1)+3}$, which therefore extends the inductive hypothesis for k. By the principle of induction, we have thus established that every $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$ satisfy $P_X(\mathcal{X} \setminus \Delta_i) \leq 1 - i\varepsilon_{k+3}$. In particular, since probabilities are non-negative, this immediately implies any such (k, i) satisfy $i \leq 1/\varepsilon_{k+3}$, as claimed. Thus, on the event E_0, we have established all of the claimed inequalities: (11) for all $k \in \{1, \ldots, K\}$, and (12 - 13) for all $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$, which further satisfy $\max \mathcal{I}_k \leq \frac{1}{\varepsilon_{k+3}}$. ■

The following is an obvious implication of Lemma 10, which will be useful to state explicitly for later reference.

Lemma 11. *On the event E_0, for every $k \in \mathcal{K}$ and $i \in \mathcal{I}_k$, if the algorithm reaches Step 6 with these values (k, i), then for f, g as defined there,*

$$P_X(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+3}.$$

Moreover, on the event E_0, every $k \in \{1, \ldots, K\}$ with $\Delta_{i_k} \neq \emptyset$ satisfies $P_X(\Delta_{i_k}) > \varepsilon_{k+2}$.

Proof. By the condition in Step 5, if the algorithm reaches Step 6 then $\hat{P}_{S^3_{k,i}}(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+2}$. By (12) of Lemma 10, on the event E_0, this implies $P_X(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+3}$.

Turning now to the second claim, suppose again that E_0 occurs, and first note that this claim is trivially satisfied if $\Delta_{i_K} = \emptyset$. To address the remaining case, suppose $\Delta_{i_K} \neq \emptyset$, and consider the minimum value $k' \in \{1, \ldots, K\}$ for which $\Delta_{i_{k'}} \neq \emptyset$. By definition we have $\Delta_{i_1} = \Delta_0 = \emptyset$, which implies we must have $k' \geq 2$. By minimality of k', we also know that the algorithm reaches Step 6 at least once during round $k = k' - 1$ of the '*For*' loop, in particular with $i = i_k$. Thus, letting (f, g) be as defined in Step 6 for these values $(k, i) = (k' - 1, i_{k'-1})$, by the first claim in the lemma, we have $P_X(\Delta_{i_{k'}}) \geq P_X(f \neq g) = P_X(\{f \neq g\} \setminus \Delta_{i_{k'-1}}) > \varepsilon_{k'+2}$. Thus, since $\Delta_{i_{k''}}$ is non-decreasing in k'', and minimality of k' implies all k'' with $\Delta_{i_{k''}} \neq \emptyset$ have $k'' \geq k'$, we conclude that every $k'' \in \{1, \ldots, K\}$ with $\Delta_{i_{k''}} \neq \emptyset$ satisfies $P_X(\Delta_{i_{k''}}) \geq P_X(\Delta_{i_{k'}}) > \varepsilon_{k'+2} \geq \varepsilon_{k''+2}$. ■

Next we state a bound on the diameters of V_{k-1} and $V^{(3)}_{k-1}$, useful for Lemmas 15, 20, and 22.

Lemma 12. *On the event E_0, for every $k \in \{1, \ldots, K\}$,*

$$\sup_{f,g \in V_{k-1}} P_X(\{f \neq g\} \setminus \Delta_{i_k}) \leq \varepsilon_k \tag{16}$$

$$\text{and} \quad \sup_{f,g \in V^{(3)}_{k-1}} P_X(\{f \neq g\} \setminus \Delta_{i_k}) \leq 3\varepsilon_k. \tag{17}$$

Proof. Throughout this proof, we suppose the event E_0 holds. The inequality (16) is trivially satisfied for $k = 1$, recalling that $V_0 = \mathbb{C}$ and $\varepsilon_1 = C^0 = 1$. For the remaining case, fix any $k' \in \{2, \ldots, K\}$ and consider the round $k = k' - 1$ in the '*For*' loop (noting that, by definition of K, we have $k = k' - 1 \in \mathcal{K}$ regardless of whether $K = N + 1$ or $K \leq N$). Since $k + 1 = k' \leq K$, we know the algorithm reaches Step 8 in round k (i.e., it does not terminate early in Step 4 during round k). In particular, this means the condition in Step 5 fails for the value $i = i_{k+1} = \max \mathcal{I}_k$: that is, $\max_{f,g \in V_k} \hat{P}_{S^3_{k,i}}(\{f \neq g\} \setminus \Delta_{i_{k+1}}) \leq \varepsilon_{k+2}$. By (13) of Lemma 10, this implies

$$\sup_{f,g \in V_{k'-1}} P_X(\{f \neq g\} \setminus \Delta_{i_{k'}}) = \sup_{f,g \in V_k} P_X(\{f \neq g\} \setminus \Delta_{i_{k+1}}) < \varepsilon_{k+1} = \varepsilon_{k'}.$$

This completes the proof of (16) for every $k \in \{1, \ldots, K\}$.

To show (17), let $k \in \{1, \ldots, K\}$, and for any $f, g \in V_{k-1}$ and $h \in \mathbb{C}$, denote by $\mathrm{DL}(f, g, h) := f\mathbb{1}_{\{f=g\}} + h\mathbb{1}_{\{f \neq g\}} \in V_{k-1}^{(3)}$. Note that for any $f, g, f', g' \in V_{k-1}$, $h, h' \in \mathbb{C}$, and $x \in \mathcal{X}$, if $g(x) = f(x) = f'(x) = g'(x)$, then $\mathrm{DL}(f, g, h)(x) = \mathrm{DL}(f', g', h')(x)$. Therefore,

$$P_X(\{\mathrm{DL}(f, g, h) \neq \mathrm{DL}(f', g', h')\} \setminus \Delta_{i_k}) \leq P_X((\{f \neq g\} \cup \{f' \neq g'\} \cup \{f \neq f'\}) \setminus \Delta_{i_k})$$
$$\leq P_X(\{f \neq g\} \setminus \Delta_{i_k}) + P_X(\{f' \neq g'\} \setminus \Delta_{i_k}) + P_X(\{f \neq f'\} \setminus \Delta_{i_k}) \leq 3\varepsilon_k,$$

where the last inequality is by (16). This completes the proof of the lemma. ■

The following Lemmas 13 and 14 concern concentration of empirical errors in the set $S_k^2 \cap \Delta_{i_k}$, which will be useful in establishing guarantees on the quality of $\hat{h}_k$ (in Lemma 15) and of the functions in V_k (in Lemmas 16 and 17) below. We first need to argue that the $\hat{p}_k$ quantities approximate $P_X(\Delta_{i_k})$, which leads to the data sets S_k^2 being of appropriate size for concentration of empirical error rates.

Lemma 13. *There is an event E_1 of probability at least $1 - \frac{\delta}{4}$ such that, on $E_0 \cap E_1$, $\forall k \in \{1, \ldots, K\}$, the quantity $\hat{p}_k := 2\hat{P}_{S_k^4}(\Delta_{i_k})$ (as defined above) satisfies*

$$P_X(\Delta_{i_k}) \leq \hat{p}_k \leq 4P_X(\Delta_{i_k}). \tag{18}$$

Proof. Consider any $k \in \{1, \ldots, N+1\}$ having non-zero probability that $k \leq K$. Note that the execution of the algorithm does not depend on S_k^4 at any time prior to Step 2 of round k (or Step 9 if $k = N+1$), supposing this step is even reached in the algorithm (i.e., $k \leq K$). Thus, since the event that $k \leq K$ and the set Δ_{i_k} are both completely determined by events occurring prior to this first time the examples in S_k^4 are used by the algorithm, we have that S_k^4 is independent of these. Thus, conditioned on the event that $k \leq K$ and on the random variable Δ_{i_k}, we have that for the sequence of m_k examples (X_t, Y_t) comprising S_k^4, the corresponding sequence of indicator random variables $\mathbb{1}[X_t \in \Delta_{i_k}]$ are conditionally independent $\mathrm{Bernoulli}(P_X(\Delta_{i_k}))$ random variables. Therefore, applying a multiplicative Chernoff bound (Lemma 6 of Appendix D) under the conditional distribution given the event $k \leq K$ and the random variable Δ_{i_k}, together with the law of total probability, we have that on an event $E_{1,k}$ of probability at least $1 - \frac{\delta\varepsilon_k}{44}$, if $k \leq K$, then

$$P_X(\Delta_{i_k}) \leq \max\left\{2\hat{P}_{S_k^4}(\Delta_{i_k}), \frac{8}{m_k}\ln\left(\frac{88}{\delta\varepsilon_k}\right)\right\}, \tag{19}$$

$$\hat{P}_{S_k^4}(\Delta_{i_k}) \leq \max\left\{2P_X(\Delta_{i_k}), \frac{6}{m_k}\ln\left(\frac{88}{\delta\varepsilon_k}\right)\right\}. \tag{20}$$

For simplicity, for any $k \in \{1, \ldots, N+1\}$ having probability zero of $k \leq K$, simply define $E_{1,k}$ as the event of probability one that $k > K$, so that the above claim also holds (vacuously) for such values k. Define an event $E_1 = \bigcap_{k=1}^{N+1} E_{1,k}$, and note that, by the union bound, E_1 occurs with probability at least $1 - \sum_{k=1}^{N+1} \frac{\delta\varepsilon_k}{44} \geq 1 - \frac{\delta}{4}$.

We now argue these inequalities further imply the simpler inequalities stated in (18), on the additional event E_0. Suppose the event $E_0 \cap E_1$ holds, and let $k \in \{1, \ldots, K\}$. If $\Delta_{i_k} = \emptyset$, (18) trivially holds since $\hat{p}_k = 0 = P_X(\Delta_{i_k})$. To address the remaining case, suppose $\Delta_{i_k} \neq \emptyset$. By the final claim in Lemma 11, we have $P_X(\Delta_{i_k}) > \varepsilon_{k+2}$. Also note that, by definition of m_k (and recalling $|\mathbb{C}| \geq 2$, which implies $\mathsf{d} \geq 1$), we have $\frac{8}{m_k}\ln\left(\frac{88}{\delta\varepsilon_k}\right) < \varepsilon_{k+2}$. In particular, these imply $2P_X(\Delta_{i_k}) > 2\varepsilon_{k+2} > \frac{6}{m_k}\ln\left(\frac{88}{\delta\varepsilon_k}\right)$, so that the right hand side of (20) equals $2P_X(\Delta_{i_k})$ and hence $\hat{p}_k \leq 4P_X(\Delta_{i_k})$. Moreover, since $\frac{8}{m_k}\ln\left(\frac{88}{\delta\varepsilon_k}\right) < \varepsilon_{k+2} < P_X(\Delta_{i_k})$, the "max" on the right hand side of (19) cannot be achieved by the second term (as it is smaller than the quantity on the left hand side), so it must be achieved by the first term. Therefore, $P_X(\Delta_{i_k}) \leq 2\hat{P}_{S_k^4}(\Delta_{i_k}) = \hat{p}_k$. ■

Using Lemma 13 to bound the size of the data set S_k^2 (which is based on $\hat{p}_k$), we are now ready to establish a concentration inequality for the error rates in the Δ_{i_k} region in the following lemma.

Lemma 14. *There is an event E_2 of probability at least $1 - \frac{\delta}{4}$ such that, on the event $E_0 \cap E_1 \cap E_2$, $\forall k \in \{1, \ldots, K\}$,*

$$\sup_{h \in \mathbb{C}} \left|\hat{P}_{S_k^2}(\mathrm{ER}(h) \cap \Delta_{i_k}) - P(\mathrm{ER}(h) \cap \Delta_{i_k})\right| \leq \frac{\varepsilon_k}{\sqrt{C''}}. \tag{21}$$

Proof. Consider any $k \in \{1, \dots, N+1\}$ having non-zero probability of $k \leq K$. Supposing $k \leq K$ occurs, define a collection of sets $\mathcal{A}'_k := \{\mathrm{ER}(h) \cap \Delta_{i_k} : h \in \mathbb{C}\}$. Note that $\mathrm{VC}(\mathcal{A}'_k) \leq \mathsf{d}$ (which is immediate from the definition of VC dimension). We aim to apply Lemma 8 of Appendix D, a refinement of the uniform convergence bound of Talagrand (1994), which accounts for an *envelope* set $D \supseteq \bigcup \mathcal{A}'_k$; specifically, we instantiate the various sets and variables in Lemma 8 to be $\mathcal{Z} = \mathcal{X} \times \{0,1\}$, $n = m'_k$, $\mathcal{A} = \mathcal{A}'_k$, envelope set $D = \Delta_{i_k}$, data set $Z = S^2_k$, and confidence parameter $\delta/(4(3+N-k)^2)$, and we apply the lemma under the conditional distribution given the event $k \leq K$ and given the random variables Δ_{i_k} and m'_k. Since the event that $k \leq K$, and the random variables Δ_{i_k} and m'_k, are all completely determined by examples allocated to data sets *before* allocating examples to the data set S^2_k, we may note that the m'_k examples comprising S^2_k are conditionally independent P-distributed random variables given the event that $k \leq K$ and given the random variables Δ_{i_k} and m'_k. Thus, applying Lemma 8 of Appendix D under the conditional distribution given the event that $k \leq K$ and given the random variables Δ_{i_k} and m'_k, together with the law of total probability, we have that on an event $E_{2,k}$ of probability at least $1 - \frac{\delta}{4(3+N-k)^2}$, if $k \leq K$ and $P_X(\Delta_{i_k}) \geq \frac{9}{m'_k} \ln\left(\frac{16(3+N-k)^2}{\delta}\right)$, then

$$\sup_{h \in \mathbb{C}} \left| \hat{P}_{S^2_k}(\mathrm{ER}(h) \cap \Delta_{i_k}) - P(\mathrm{ER}(h) \cap \Delta_{i_k}) \right| \leq c_1 \sqrt{\frac{P_X(\Delta_{i_k})}{m'_k} \left(\mathsf{d} + \log\left(\frac{4(3+N-k)^2}{\delta}\right)\right)}. \quad (22)$$

For simplicity, for any $k \in \{1, \dots, N+1\}$ having zero probability of $k \leq K$, let $E_{2,k}$ denote the event of probability one that $k > K$. Finally, define $E_2 = \bigcap_{k=1}^{N+1} E_{2,k}$, and note that, by the union bound, E_2 holds with probability at least $1 - \sum_{k=1}^{N+1} \frac{\delta}{4(3+N-k)^2} \geq 1 - \frac{\delta}{4} \sum_{j=2}^{\infty} \frac{1}{j^2} \geq 1 - \frac{\delta}{4}$.

Now suppose the event $E_0 \cap E_1 \cap E_2$ occurs, and consider any $k \in \{1, \dots, K\}$. If $\Delta_{i_k} = \emptyset$ then (21) holds trivially since the left hand side of (21) is then zero. To address the remaining case, suppose $\Delta_{i_k} \neq \emptyset$. By the final claim in Lemma 11, we have $P_X(\Delta_{i_k}) > \varepsilon_{k+2}$. Moreover, by Lemma 13 we have $\hat{p}_k \geq P_X(\Delta_{i_k})$. Recalling $m'_k := \left\lceil \frac{C'' c_1^2 \hat{p}_k}{\varepsilon_k^2} \left(\mathsf{d} + \log\left(\frac{4(3+N-k)^2}{\delta}\right)\right)\right\rceil$, these imply

$$m'_k > \frac{C'' c_1^2}{C^4 \varepsilon_{k+2}} \ln\left(\frac{4(3+N-k)^2}{\delta}\right) > \frac{9}{\varepsilon_{k+2}} \ln\left(\frac{16(3+N-k)^2}{\delta}\right),$$

where the last inequality follows from $c_1 \geq 1$ and $C'' \geq 18C^4$. Thus, $\frac{9}{m'_k} \ln\left(\frac{16(3+N-k)^2}{\delta}\right) < \varepsilon_{k+2} < P_X(\Delta_{i_k})$. By the definition of E_2, it follows that (22) holds. Moreover, since $\hat{p}_k \geq P_X(\Delta_{i_k})$, we have that $m'_k \geq \frac{C'' c_1^2 P_X(\Delta_{i_k})}{\varepsilon_k^2} \left(\mathsf{d} + \log\left(\frac{4(3+N-k)^2}{\delta}\right)\right)$, so that the right hand side of (22) is at most $\frac{\varepsilon_k}{\sqrt{C''}}$, thus establishing (21). ■

Combining the concentration inequality from Lemma 14 with (11) of Lemma 10 together with (17) of Lemma 12 yields a concentration inequality for the differences $\hat{\mathrm{er}}^{1,2}_k(h) - \hat{\mathrm{er}}^{1,2}_k(h')$ among $h, h' \in V_{k-1}$, recalling the definition from (1):

$$\hat{\mathrm{er}}^{1,2}_k(h) := \hat{P}_{S^1_k}(\mathrm{ER}(h) \cap D_{k-1} \setminus \Delta_{i_k}) + \hat{P}_{S^2_k}(\mathrm{ER}(h) \cap \Delta_{i_k}).$$

In fact, the implication is stronger than this, admitting functions $h, h' \in V^{(4)}_{k-1}$. In particular, for any $k \in \{1, \dots, K\}$, note that $V^{(4)}_{k-1}$ in (2) can equivalently be defined as

$$V^{(4)}_{k-1} = \left\{ h_1 \mathbb{1}_{\mathcal{X} \setminus \Delta_{i_k}} + h_2 \mathbb{1}_{\Delta_{i_k}} : h_1 \in V^{(3)}_{k-1}, h_2 \in \mathbb{C} \right\}.$$

The following lemma provides a concentration inequality for $\hat{\mathrm{er}}^{1,2}_k(h) - \hat{\mathrm{er}}^{1,2}_k(h')$ among functions $h, h' \in V^{(4)}_{k-1}$.

Lemma 15. *On the event $E_0 \cap E_1 \cap E_2$, for every $k \in \{1, \dots, K\}$ we have*

$$\sup_{h, h' \in V^{(4)}_{k-1}} \left| \left(\hat{\mathrm{er}}^{1,2}_k(h) - \hat{\mathrm{er}}^{1,2}_k(h')\right) - \left(\mathrm{er}_P(h) - \mathrm{er}_P(h')\right) \right| \leq \frac{\varepsilon_k}{4C'}, \quad (23)$$

recalling that $C' := \frac{\sqrt{C''}}{16}$. Moreover, (23) implies

$$\mathrm{er}_P\left(\hat{h}_k\right) \leq \inf_{h \in V^{(4)}_{k-1}} \mathrm{er}_P(h) + \frac{\varepsilon_k}{4C'}. \quad (24)$$

Proof. Suppose the event $E_0 \cap E_1 \cap E_2$ holds and consider any $k \in \{1, \dots, K\}$. Note that for any $h = h_1 \mathbb{1}_{\mathcal{X} \setminus \Delta_{i_k}} + h_2 \mathbb{1}_{\Delta_{i_k}} \in V_{k-1}^{(4)}$ we have $\mathrm{er}_P(h) = P(\mathrm{ER}(h_1) \setminus \Delta_{i_k}) + P(\mathrm{ER}(h_2) \cap \Delta_{i_k})$ and $\hat{\mathrm{er}}_k^{1,2}(h) = \hat{P}_{S_k^1}(\mathrm{ER}(h_1) \cap D_{k-1} \setminus \Delta_{i_k}) + \hat{P}_{S_k^2}(\mathrm{ER}(h_2) \cap \Delta_{i_k})$.

Consider any $h_1, h_1' \in V_{k-1}^{(3)}$ and any $h_2, h_2' \in \mathbb{C}$, and let $h = h_1 \mathbb{1}_{\mathcal{X} \setminus \Delta_{i_k}} + h_2 \mathbb{1}_{\Delta_{i_k}}$ and $h' = h_1' \mathbb{1}_{\mathcal{X} \setminus \Delta_{i_k}} + h_2' \mathbb{1}_{\Delta_{i_k}}$. By Lemma 14, we have $\forall h_2'' \in \{h_2, h_2'\}$,

$$\left| \hat{P}_{S_k^2}(\mathrm{ER}(h_2'') \cap \Delta_{i_k}) - P(\mathrm{ER}(h_2'') \cap \Delta_{i_k}) \right| \leq \frac{\varepsilon_k}{\sqrt{C''}}.$$

Additionally, since (17) of Lemma 12 implies $P_X(\{h_1 \neq h_1'\} \setminus \Delta_{i_k}) \leq 3\varepsilon_k$, the inequality (11) of Lemma 10 implies

$$\left| \hat{P}_{S_k^1}(\mathrm{ER}(h_1) \cap D_{k-1} \setminus \Delta_{i_k}) - \hat{P}_{S_k^1}(\mathrm{ER}(h_1') \cap D_{k-1} \setminus \Delta_{i_k}) - (P(\mathrm{ER}(h_1) \setminus \Delta_{i_k}) - P(\mathrm{ER}(h_1') \setminus \Delta_{i_k})) \right|$$
$$< \sqrt{\frac{3\varepsilon_k^2}{C''}} + \frac{\varepsilon_k}{C''} \leq \frac{2\varepsilon_k}{\sqrt{C''}}, \tag{25}$$

where the last inequality follows from $C'' \geq 14$. Combining these with the triangle inequality (namely, $|((\hat{a} + \hat{b}) - (\hat{a}' + \hat{b}')) - ((a + b) - (a' + b'))| \leq |(\hat{a} - \hat{a}') - (a - a')| + |\hat{b} - b| + |\hat{b}' - b'|$) yields that

$$\left| \left(\hat{\mathrm{er}}_k^{1,2}(h) - \hat{\mathrm{er}}_k^{1,2}(h')\right) - \left(\mathrm{er}_P(h) - \mathrm{er}_P(h')\right) \right| < \frac{4\varepsilon_k}{\sqrt{C''}} = \frac{\varepsilon_k}{4C'}.$$

To see that (23) implies (24), note that, by the definition of $\hat{h}_k$ in (3), $h = \hat{h}_k$ has minimal $\hat{\mathrm{er}}_k^{1,2}(h)$ among all $h \in V_{k-1}^{(4)}$: that is, $\forall h \in V_{k-1}^{(4)}$, $\hat{\mathrm{er}}_k^{1,2}\left(\hat{h}_k\right) - \hat{\mathrm{er}}_k^{1,2}\left(h\right) \leq 0$. Together with (23), this implies that $\forall h \in V_{k-1}^{(4)}$, $\mathrm{er}_P\left(\hat{h}_k\right) - \mathrm{er}_P\left(h\right) \leq \hat{\mathrm{er}}_k^{1,2}\left(\hat{h}_k\right) - \hat{\mathrm{er}}_k^{1,2}\left(h\right) + \frac{\varepsilon_k}{4C'} \leq \frac{\varepsilon_k}{4C'}$. ■

In particular, Lemma 15 immediately implies the following lemma concerning the quality of the functions h in V_k.

Lemma 16. *On the event $E_0 \cap E_1 \cap E_2$, $\forall k \in \mathcal{K}$, $\forall h \in V_{k-1}$, the following implications hold:*

$$h \in V_k \implies \mathrm{er}_P\left(h\right) \leq \mathrm{er}_P\left(\hat{h}_k\right) + \frac{5\varepsilon_k}{4C'}, \tag{26}$$

$$\mathrm{er}_P\left(h\right) \leq \mathrm{er}_P\left(\hat{h}_k\right) + \frac{3\varepsilon_k}{4C'} \implies h \in V_k. \tag{27}$$

Proof. Suppose the event $E_0 \cap E_1 \cap E_2$ occurs and consider any $k \in \mathcal{K}$ and $h \in V_{k-1}$. In particular, note that we also have $h \in V_{k-1}^{(4)}$, since letting $f = g = h$, we have $h = f\mathbb{1}_{\{f=g\}} + h\mathbb{1}_{\{f \neq g\}} \in V_{k-1}^{(3)}$, and thus $h = h\mathbb{1}_{\mathcal{X} \setminus \Delta_{i_k}} + h\mathbb{1}_{\Delta_{i_k}} \in V_{k-1}^{(4)}$.

If $h \in V_k$, then by definition of V_k in Step 3 we have $\hat{\mathrm{er}}_k^{1,2}\left(h\right) - \hat{\mathrm{er}}_k^{1,2}\left(\hat{h}_k\right) \leq \frac{\varepsilon_k}{C'}$. Together with (23) of Lemma 15, this implies $\mathrm{er}_P\left(h\right) - \mathrm{er}_P\left(\hat{h}_k\right) \leq \frac{\varepsilon_k}{C'} + \frac{\varepsilon_k}{4C'} = \frac{5\varepsilon_k}{4C'}$, which establishes (26).

On the other hand, if $\mathrm{er}_P\left(h\right) - \mathrm{er}_P\left(\hat{h}_k\right) \leq \frac{3\varepsilon_k}{4C'}$, then (23) of Lemma 15 implies $\hat{\mathrm{er}}_k^{1,2}\left(h\right) - \hat{\mathrm{er}}_k^{1,2}\left(\hat{h}_k\right) \leq \frac{3\varepsilon_k}{4C'} + \frac{\varepsilon_k}{4C'} = \frac{\varepsilon_k}{C'}$. Thus, any such h is retained in V_k, which establishes (27). ■

The main implication of Lemma 16 pertains to the early stopping case in Step 4, which we turn to next. Recall $h^\star$ denotes an (arbitrary) concept in $\mathbb{C}$ with $\mathrm{er}_P(h^\star) < \inf_{h \in \mathbb{C}} \mathrm{er}_P(h) + \frac{\varepsilon}{10^4}$. In the following lemma, in addition to arguing that the predictor $\hat{h}$ returned in Step 4 has low excess error rate compared to $h^\star$ (in fact, *negative*), this lemma also reveals a second major role of this early stopping case: it ensures that on all rounds k in which the algorithm does *not* terminate in Step 4, we retain $h^\star \in V_k$.

Lemma 17. *On the event $E_0 \cap E_1 \cap E_2$, the following implications hold for every $k \in \mathcal{K}$:*

- *If $\mathbb{A}_{\mathrm{avid}}$ does* not *return in Step 4 on round k, then $h^\star \in V_k$.*
- *If $\mathbb{A}_{\mathrm{avid}}$ returns in Step 4 on round k, then $\mathrm{er}_P\left(\hat{h}_k\right) < \mathrm{er}_P\left(h^\star\right)$.*

Proof. Suppose the event $E_0 \cap E_1 \cap E_2$ occurs. We will prove the first claim by induction on k. As a base case, we trivially have $h^\star \in \mathbb{C} = V_0$. Now, for the purpose of induction, let $k \in \mathcal{K}$ be such that $h^\star \in V_{k-1}$. Also note (as discussed in the proof of Lemma 16) that this also implies $h^\star \in V_{k-1}^{(4)}$. If $h^\star \notin V_k$, then by (27) of Lemma 16, we have $\mathrm{er}_P\big(h^\star\big) - \mathrm{er}_P\big(\hat{h}_k\big) > \frac{3\varepsilon_k}{4C'}$. In particular, this implies that if $V_k \neq \emptyset$, then together with (23) of Lemma 15, we have

$$\min_{h \in V_k} \hat{\mathrm{er}}_k^{1,2}\big(h\big) - \hat{\mathrm{er}}_k^{1,2}\big(\hat{h}_k\big) \geq \min_{h \in V_{k-1}} \hat{\mathrm{er}}_k^{1,2}\big(h\big) - \hat{\mathrm{er}}_k^{1,2}\big(\hat{h}_k\big) \geq \inf_{h \in V_{k-1}} \mathrm{er}_P\big(h\big) - \mathrm{er}_P\big(\hat{h}_k\big) - \frac{\varepsilon_k}{4C'}$$
$$> \mathrm{er}_P\big(h^\star\big) - \mathrm{er}_P\big(\hat{h}_k\big) - \frac{\varepsilon}{10^4} - \frac{\varepsilon_k}{4C'} > \frac{3\varepsilon_k}{4C'} - \frac{\varepsilon}{10^4} - \frac{\varepsilon_k}{4C'} > \frac{\varepsilon_k}{4C'},$$

where the last inequality follows from $\frac{\varepsilon}{10^4} < \frac{2\varepsilon_N}{10^4} \leq \frac{\varepsilon_k}{4C'}$. Thus, either $V_k = \emptyset$ or $\min_{h \in V_k} \hat{\mathrm{er}}_k^{1,2}\big(h\big) - \hat{\mathrm{er}}_k^{1,2}\big(\hat{h}_k\big) > \frac{\varepsilon_k}{4C'}$, so that either way the algorithm will return in Step 4 in this case. Therefore, if the algorithm does *not* return in Step 4 on round k, it must be that $h^\star \in V_k$. This completes the proof of the first claim, by the principle of induction.

Finally, we turn to the second claim. Suppose, for some $k \in \mathcal{K}$, the algorithm returns in Step 4 on round k. In particular, either $k = 1$, in which case $h^\star \in \mathbb{C} = V_{k-1}$, or $k > 1$, in which case (since the algorithm did not return in Step 4 on round $k-1$) the first claim in the lemma implies $h^\star \in V_{k-1}$. Again note that this also implies $h^\star \in V_{k-1}^{(4)}$. If $h^\star \notin V_k$, then (27) of Lemma 16 implies $\mathrm{er}_P\big(h^\star\big) > \mathrm{er}_P\big(\hat{h}_k\big) + \frac{3\varepsilon_k}{4C'}$. Otherwise, if $h^\star \in V_k$, the condition in Step 4 implies $\hat{\mathrm{er}}_k^{1,2}\big(h^\star\big) - \hat{\mathrm{er}}_k^{1,2}\big(\hat{h}_k\big) > \frac{\varepsilon_k}{4C'}$. Together with (23) of Lemma 15, this implies

$$\mathrm{er}_P\big(h^\star\big) - \mathrm{er}_P\big(\hat{h}_k\big) \geq \hat{\mathrm{er}}_k^{1,2}\big(h^\star\big) - \hat{\mathrm{er}}_k^{1,2}\big(\hat{h}_k\big) - \frac{\varepsilon_k}{4C'} > 0.$$

Thus, in either case, we have $\mathrm{er}_P\big(h^\star\big) > \mathrm{er}_P\big(\hat{h}_k\big)$, which establishes the second claim. ■

Lemmas 15, 16, and 17 together have a particularly nice implication, which, although not strictly needed for the proof of Theorem 5, is worth noting (and will be useful in Appendix F). Specifically, we have the following corollary.

Corollary 18. *On the event $E_0 \cap E_1 \cap E_2$, $\forall k \in \mathcal{K}$,*

$$V_k \subseteq \left\{h \in \mathbb{C} : \mathrm{er}_P(h) - \mathrm{er}_P(h^\star) \leq \frac{3\varepsilon_k}{2C'}\right\}.$$

Proof. Suppose the event $E_0 \cap E_1 \cap E_2$ occurs and consider any $k \in \mathcal{K}$. Since $k - 1 < K$, Lemma 17 implies $h^\star \in V_{k-1}$ in the case $k \geq 2$, while the case $k = 1$ has $h^\star \in V_0$ by definition of $V_0 = \mathbb{C}$. Together with (24) of Lemma 15, this implies $\mathrm{er}_P\big(\hat{h}_k\big) \leq \mathrm{er}_P\big(h^\star\big) + \frac{\varepsilon_k}{4C'}$. Combined with (26) of Lemma 16, we have that every $h \in V_k$ satisfies $\mathrm{er}_P\big(h\big) \leq \mathrm{er}_P\big(\hat{h}_k\big) + \frac{5\varepsilon_k}{4C'} \leq \mathrm{er}_P\big(h^\star\big) + \frac{3\varepsilon_k}{2C'}$. ■

At this point, we may note that Lemmas 15 and 17 together completely address the error guarantee for the $\hat{h}$ returned by $\mathbb{A}_{\mathrm{avid}}$, on the event $E_0 \cap E_1 \cap E_2$. as summarized in the following lemma.

Lemma 19. *On the event $E_0 \cap E_1 \cap E_2$, $\mathbb{A}_{\mathrm{avid}}$ eventually terminates, and the function $\hat{h}$ it returns satisfies $\mathrm{er}_P\big(\hat{h}\big) \leq \inf_{h \in \mathbb{C}} \mathrm{er}_P\big(h\big) + \varepsilon$.*

Proof. Suppose the event $E_0 \cap E_1 \cap E_2$ occurs. If the algorithm terminates in Step 4 in some round $k \in \mathcal{K}$, by definition we have $\hat{h} = \hat{h}_k$, and thus Lemma 17 implies $\mathrm{er}_P\big(\hat{h}\big) < \mathrm{er}_P\big(h^\star\big) < \inf_{h \in \mathbb{C}} \mathrm{er}_P\big(h\big) + \frac{\varepsilon}{10^4}$. On the other hand, if the algorithm does not return in Step 4 on any round $k \in \mathcal{K}$, then we have $K = N + 1$ (recalling Lemma 10 implies the algorithm eventually terminates), so that by definition $\hat{h} = \hat{h}_{N+1}$. Since in this case Lemma 17 implies $h^\star \in V_N$ (and hence $h^\star \in V_N^{(4)}$), (24) of Lemma 15 implies $\mathrm{er}_P\big(\hat{h}\big) = \mathrm{er}_P\big(\hat{h}_{N+1}\big) \leq \mathrm{er}_P\big(h^\star\big) + \frac{\varepsilon_{N+1}}{4C'} < \inf_{h \in \mathbb{C}} \mathrm{er}_P\big(h\big) + \frac{\varepsilon}{10^4} + \frac{\varepsilon}{8C'} < \inf_{h \in \mathbb{C}} \mathrm{er}_P\big(h\big) + \varepsilon$. ■

With the analysis of error guarantees complete, we turn now to establishing the bound $Q(\varepsilon, \delta; \beta)$ on the number of queries, as claimed in Theorem 5. This will be comprised of two main parts. First, we

argue that the set Δ_{i_k} never grows too large: specifically, recalling $\beta := \inf_{h\in\mathbb{C}} \mathrm{er}_P(h)$, Lemma 20 will establish that $P_X(\Delta_{i_k}) = O(\beta)$, which in turn allows us to bound the number of queries in $S_k^2 \cap \Delta_{i_k}$ on each round (in the proof of Lemma 23). Second, in the proof of Lemma 22, we bound $P_X(D_{k-1} \setminus \Delta_{i_k}) \le \mathfrak{s}\varepsilon_k$, by reasoning in terms of the disagreement coefficient (Hanneke, 2007b), relating the latter to the star number via a result of Hanneke and Yang (2015). This in turn allows us to bound the number of queries in $S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$ on each round (in the proof of Lemma 23).

We begin with the first of these parts, stated in the following lemma. We remark that this lemma plays a special role in constraining the allowed values of the constant C, as the argument breaks down if C is taken too large. On the other hand, the proof also reveals that it is possible to decrease the factor "5" in this lemma to *any* value $c > 2$ by taking $C > 1$ appropriately close to 1 and by an appropriately large choice of the constant C'' (and hence C'). See footnote 11 for further discussion.

Lemma 20. *On the event $E_0 \cap E_1 \cap E_2$, for all $k \in \{1,\dots,K\}$ and $i \in \mathcal{I}_k$,*

$$P_X(\Delta_i) \le 5 \inf_{h\in\mathbb{C}} P(\mathrm{ER}(h) \cap \Delta_i) \le 5\beta.$$

Proof. We will argue that, on $E_0 \cap E_1 \cap E_2$, for any $h_0 \in \mathbb{C}$, each region $\Delta_{i+1} \setminus \Delta_i$ (defined in Step 7) satisfies

$$P(\mathrm{ER}(h_0) \cap \Delta_{i+1} \setminus \Delta_i) > \left(1 - \frac{C^3}{2} - \frac{5C^3}{4C'}\right) P_X(\Delta_{i+1} \setminus \Delta_i) = \frac{1}{5} P_X(\Delta_{i+1} \setminus \Delta_i), \tag{28}$$

so that each addition to Δ_i "chops off" a piece of $\mathrm{ER}(h_0)$ of measure proportional to the increase in measure $P_X(\Delta_{i+1}) - P_X(\Delta_i) = P_X(\Delta_{i+1} \setminus \Delta_i)$. The claim in the lemma then follows immediately from (28), since it holds trivially for $i = 0$ (recalling $\Delta_0 = \emptyset$), and if any $k \in \{1,\dots,K\}$ and $i \in \mathcal{I}_k$ has $i \ge 1$, then applying (28) inductively yields

$$P(\mathrm{ER}(h_0) \cap \Delta_i) = \sum_{j=0}^{i-1} P(\mathrm{ER}(h_0) \cap \Delta_{j+1} \setminus \Delta_j) > \sum_{j=0}^{i-1} \frac{1}{5} P_X(\Delta_{j+1} \setminus \Delta_j) = \frac{1}{5} P_X(\Delta_i).$$

Taking the infimum over all $h_0 \in \mathbb{C}$ then implies the lemma.

We proceed now with the formal proof of (28). Suppose the event $E_0 \cap E_1 \cap E_2$ occurs, and for the purpose of analyzing the *increases* $\Delta_{i+1} \setminus \Delta_i$ of the Δ_i set (which only occur in Step 7), consider any $k \in \mathcal{K}$ and any $i \in \mathcal{I}_k$ with $i < \max \mathcal{I}_k$ (equivalently, the algorithm reaches Step 7 with this (k,i)). Let (f,g) be as defined in Step 6 for this (k,i) so that $\Delta_{i+1} \setminus \Delta_i = \{f \ne g\} \setminus \Delta_i$.

Note that $\{f \ne g\} \times \{0,1\} = (\mathrm{ER}(f) \cap \{f \ne g\}) \cup (\mathrm{ER}(g) \cap \{f \ne g\})$, so that (lower-bounding 'max' by 'average')

$$\begin{aligned}
&\max_{f' \in \{f,g\}} P(\mathrm{ER}(f') \cap \{f \ne g\} \setminus \Delta_{i_k}) \\
&\ge \frac{1}{2} P(\mathrm{ER}(f) \cap \{f \ne g\} \setminus \Delta_{i_k}) + \frac{1}{2} P(\mathrm{ER}(g) \cap \{f \ne g\} \setminus \Delta_{i_k}) \ge \frac{1}{2} P_X(\{f \ne g\} \setminus \Delta_{i_k}),
\end{aligned}$$

where in fact the last inequality holds with equality (since, for $\{0,1\}$ labels, $(\mathrm{ER}(f) \cap \{f \ne g\})$ and $(\mathrm{ER}(g) \cap \{f \ne g\})$ are disjoint). Thus, $\exists f' \in \{f,g\}$ with

$$P(\mathrm{ER}(f') \cap \{f \ne g\} \setminus \Delta_{i_k}) \ge \frac{1}{2} P_X(\{f \ne g\} \setminus \Delta_{i_k}).$$

Let $h' = f' \mathbb{1}_{\{f=g\} \setminus \Delta_{i_k}} + h_0 \mathbb{1}_{\{f \ne g\} \setminus \Delta_{i_k}} + f' \mathbb{1}_{\Delta_{i_k}}$ and note that $h' \in V_{k-1}^{(4)}$. Also recall that $\hat{\mathrm{er}}_k^{1,2}(\hat{h}_k) = \min_{h \in V_{k-1}^{(4)}} \hat{\mathrm{er}}_k^{1,2}(h)$, and hence $\hat{\mathrm{er}}_k^{1,2}(h') \ge \hat{\mathrm{er}}_k^{1,2}(\hat{h}_k)$. Since we also have $f' \in V_{k-1}^{(4)}$ (as discussed in the proof of Lemma 16), Lemma 15 implies

$$\mathrm{er}_P(f') - \mathrm{er}_P(h') \le \hat{\mathrm{er}}_k^{1,2}(f') - \hat{\mathrm{er}}_k^{1,2}(h') + \frac{\varepsilon_k}{4C'} \le \hat{\mathrm{er}}_k^{1,2}(f') - \hat{\mathrm{er}}_k^{1,2}(\hat{h}_k) + \frac{\varepsilon_k}{4C'} \le \frac{5\varepsilon_k}{4C'}, \tag{29}$$

where the last inequality is due to $f' \in \{f,g\} \subseteq V_k$, recalling the definition of V_k in Step 3.

Moreover, by definition of f' and h', we have

$$\begin{aligned}
\mathrm{er}_P(f') - \mathrm{er}_P(h') &= P(\mathrm{ER}(f') \cap \{f \ne g\} \setminus \Delta_{i_k}) - P(\mathrm{ER}(h_0) \cap \{f \ne g\} \setminus \Delta_{i_k}) \\
&\ge \frac{1}{2} P_X(\{f \ne g\} \setminus \Delta_{i_k}) - P(\mathrm{ER}(h_0) \cap \{f \ne g\} \setminus \Delta_{i_k}).
\end{aligned}$$

Equivalently: $P(\mathrm{ER}(h_0) \cap \{f \neq g\} \setminus \Delta_{i_k}) \geq \frac{1}{2} P_X(\{f \neq g\} \setminus \Delta_{i_k}) - \big(\mathrm{er}_P(f') - \mathrm{er}_P(h')\big)$. Combining this with (29), we conclude that

$$P(\mathrm{ER}(h_0) \cap \{f \neq g\} \setminus \Delta_{i_k}) \geq \frac{1}{2} P_X(\{f \neq g\} \setminus \Delta_{i_k}) - \frac{5\varepsilon_k}{4C'}. \tag{30}$$

Also note that, since $\Delta_i \supseteq \Delta_{i_k}$, we have

$$\Delta_{i+1} \setminus \Delta_i \;=\; \{f \neq g\} \setminus \Delta_i \;=\; (\{f \neq g\} \setminus \Delta_{i_k}) \setminus (\{f \neq g\} \cap \Delta_i \setminus \Delta_{i_k}),$$

so that

$$P(\mathrm{ER}(h_0) \cap \Delta_{i+1} \setminus \Delta_i) \geq P(\mathrm{ER}(h_0) \cap \{f \neq g\} \setminus \Delta_{i_k}) - P_X(\{f \neq g\} \cap \Delta_i \setminus \Delta_{i_k}). \tag{31}$$

Moreover, again since $\Delta_i \supseteq \Delta_{i_k}$, we have

$$P_X(\{f \neq g\} \cap \Delta_i \setminus \Delta_{i_k}) = P_X(\{f \neq g\} \setminus \Delta_{i_k}) - P_X(\{f \neq g\} \setminus \Delta_i). \tag{32}$$

Combining (31), (32), and (30) yields that

$$\begin{aligned} &P(\mathrm{ER}(h_0) \cap \Delta_{i+1} \setminus \Delta_i) \\ &\geq P(\mathrm{ER}(h_0) \cap \{f \neq g\} \setminus \Delta_{i_k}) - P_X(\{f \neq g\} \setminus \Delta_{i_k}) + P_X(\{f \neq g\} \setminus \Delta_i) \\ &\geq P_X(\{f \neq g\} \setminus \Delta_i) - \frac{1}{2} P_X(\{f \neq g\} \setminus \Delta_{i_k}) - \frac{5\varepsilon_k}{4C'}. \end{aligned} \tag{33}$$

Lemma 12 and the fact that $f, g \in V_k \subseteq V_{k-1}$ imply $P_X(\{f \neq g\} \setminus \Delta_{i_k}) \leq \varepsilon_k$, and Lemma 11 implies $P_X(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+3}$. Together, we have

$$P_X(\{f \neq g\} \setminus \Delta_{i_k}) < C^3 P_X(\{f \neq g\} \setminus \Delta_i).$$

Additionally, again since $P_X(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+3}$, we have that $\frac{5\varepsilon_k}{4C'} < \frac{5C^3}{4C'} P_X(\{f \neq g\} \setminus \Delta_i)$. Combining these inequalities with (33), and recalling $\Delta_{i+1} \setminus \Delta_i = \{f \neq g\} \setminus \Delta_i$, yields that

$$P(\mathrm{ER}(h_0) \cap \Delta_{i+1} \setminus \Delta_i) > \left(1 - \frac{C^3}{2} - \frac{5C^3}{4C'}\right) P_X(\Delta_{i+1} \setminus \Delta_i).$$

Recalling that $C' = \frac{\sqrt{C''}}{16} = \frac{25C^3}{16 - 10C^3}$, we have $\frac{C^3}{2} + \frac{5C^3}{4C'} = \frac{4}{5}$, so that the right hand side above equals $\frac{1}{5} P_X(\Delta_{i+1} \setminus \Delta_i)$, which establishes (28). ■

Next we turn to the second part of the argument outlined above: bounding the number of queries in the sets $S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$. We begin by stating a known fact, due to Hanneke and Yang (2015, Theorem 10): namely, that the *disagreement coefficient* (Hanneke, 2007b) is upper bounded by the *star number* (Hanneke and Yang, 2015) (indeed, Theorem 10 of Hanneke and Yang, 2015 shows the relation is even *sharp* in the worst case over h and distributions P'_X).

Lemma 21 (Hanneke and Yang, 2015)**.** *For any measurable $h : \mathcal{X} \to \{0,1\}$, any distribution P'_X on $\mathcal{X}$, and any $r > 0$, defining the r-*ball centered at h *as $\mathrm{B}_{P'_X}(h, r) := \{h' \in \mathbb{C} : P'_X(h' \neq h) \leq r\}$, it holds that*

$$P'_X\big(\mathrm{DIS}\big(\mathrm{B}_{P'_X}(h, r)\big)\big) \leq \mathfrak{s} r.$$

Toward bounding the number of queries in the sets $S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$ in the algorithm, the following lemma establishes a bound on $P_X(D_{k-1} \setminus \Delta_{i_k})$ by a straightforward application of Lemma 21 to the *conditional* probabilities $P_X(D_{k-1} | \mathcal{X} \setminus \Delta_{i_k})$, in combination with a diameter bound supplied by (16) of Lemma 12.

Lemma 22. *On the event $E_0 \cap E_1 \cap E_2$, for every $k \in \{1, \ldots, K\}$, $P_X(D_{k-1} \setminus \Delta_{i_k}) \leq \mathfrak{s}\varepsilon_k$.*

Proof. Suppose the event $E_0 \cap E_1 \cap E_2$ holds and consider any $k \in \{1, \ldots, K\}$. If $P_X(\mathcal{X} \setminus \Delta_{i_k}) = 0$, we trivially have that $P_X(D_{k-1} \setminus \Delta_{i_k}) = 0 \leq \mathfrak{s}\varepsilon_k$. To address the remaining case, suppose $P_X(\mathcal{X} \setminus \Delta_{i_k}) > 0$, and denote by $P_k := P_X(\cdot | \mathcal{X} \setminus \Delta_{i_k})$. By (16) of Lemma 12 we have

$$\sup_{f,g \in V_{k-1}} P_k(f \neq g) \leq \frac{\varepsilon_k}{P_X(\mathcal{X} \setminus \Delta_{i_k})}.$$

In particular, since $k-1<K$, Lemma 17 implies $h^\star \in V_{k-1}$ in the case $k \geq 2$, while the case $k=1$ has $h^\star \in V_0$ by definition of $V_0 = \mathbb{C}$. Thus, the above inequality implies

$$V_{k-1} \subseteq \mathrm{B}_{P_k}\left(h^\star, \frac{\varepsilon_k}{P_X(\mathcal{X} \setminus \Delta_{i_k})}\right). \tag{34}$$

Together with Lemma 21, this implies

$$P_k(D_{k-1}) = P_k(\mathrm{DIS}(V_{k-1})) \leq P_k\left(\mathrm{DIS}\left(\mathrm{B}_{P_k}\left(h^\star, \frac{\varepsilon_k}{P_X(\mathcal{X} \setminus \Delta_{i_k})}\right)\right)\right) \leq \mathfrak{s}\frac{\varepsilon_k}{P_X(\mathcal{X} \setminus \Delta_{i_k})}. \tag{35}$$

We therefore have that

$$P_X(D_{k-1} \setminus \Delta_{i_k}) = P_k(D_{k-1})P_X(\mathcal{X} \setminus \Delta_{i_k}) \leq \mathfrak{s}\varepsilon_k.$$

■

We are now ready to state a lemma bounding the total number of queries in the algorithm, by a combination of Lemmas 13, 20, and 22 together with a multiplicative Chernoff bound argument. For convenience, this lemma also supplies an upper bound on the sizes m'_k of the data sets S^2_k, which will be of further use when establishing the bound $M(\varepsilon, \delta; \beta)$ on the total number of unlabeled examples sufficient for the execution of the algorithm (Lemma 24 below). Specifically, in the following lemma, for any $k \in \{1, \ldots, N+1\}$, denote by

$$\overline{m}'_k := \frac{25C''c_1^2\beta}{\varepsilon_k^2}\left(\mathsf{d} + \log\left(\frac{4(3+N-k)^2}{\delta}\right)\right),$$
$$\text{and} \quad M_2 := \frac{700C''c_1^2\beta}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{4e^4}{\delta}\right)\right).$$

Lemma 23. *There is an event E_3 of probability at least $1-\frac{\delta}{4}$, such that on $\bigcap_{j=0}^3 E_j$, $\forall k \in \{1, \ldots, K\}$, the following claims hold: $m'_k \leq \overline{m}'_k$,*

$$\left|S^1_k \cap D_{k-1} \setminus \Delta_{i_k}\right| \leq 3\mathfrak{s}\varepsilon_k m_k, \tag{36}$$
$$\left|S^2_k \cap \Delta_{i_k}\right| \leq 2P_X(\Delta_{i_k})m'_k \leq 10\beta\overline{m}'_k. \tag{37}$$

Moreover, we have $\sum_{k=1}^{N+1}\overline{m}'_k \leq M_2$, and the total number of queries by $\mathbb{A}_{\mathrm{avid}}$ is at most $Q(\varepsilon, \delta; \beta)$, where

$$Q(\varepsilon, \delta; \beta) := 10\beta M_2 + \min\{M_1, (3/2)\mathfrak{s}\varepsilon(N+1)m_{N+1}\}$$
$$= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right) + \min\left\{\mathfrak{s}\log\left(\frac{1}{\varepsilon}\right), \frac{1}{\varepsilon}\right\}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right). \tag{38}$$

Proof. By Lemma 13, on $E_0 \cap E_1 \cap E_2$, $\forall k \in \{1, \ldots, K\}$, $P_X(\Delta_{i_k}) \leq \hat{p}_k \leq 4P_X(\Delta_{i_k})$. Recall the definition of $m'_k := \left\lceil \frac{C''c_1^2\hat{p}_k}{\varepsilon_k^2}\left(\mathsf{d} + \log\left(\frac{4(3+N-k)^2}{\delta}\right)\right)\right\rceil$. Thus, if $\Delta_{i_k} = \emptyset$, we have $\hat{p}_k = 0$, hence $m'_k = 0$, and the implication $m'_k \leq \overline{m}'_k$ trivially follows. Otherwise, on $E_0 \cap E_1 \cap E_2$, if $\Delta_{i_k} \neq \emptyset$, the final claim in Lemma 11 implies $P_X(\Delta_{i_k}) > \varepsilon_{k+2}$, so that together m'_k is at most

$$\left\lceil \frac{C''c_1^2 4P_X(\Delta_{i_k})}{\varepsilon_k^2}\left(\mathsf{d} + \log\left(\frac{4(3+N-k)^2}{\delta}\right)\right)\right\rceil \leq \frac{5C''c_1^2P_X(\Delta_{i_k})}{\varepsilon_k^2}\left(\mathsf{d} + \log\left(\frac{4(3+N-k)^2}{\delta}\right)\right).$$

Since Lemma 20 implies $P_X(\Delta_{i_k}) \leq 5\beta$ on $E_0 \cap E_1 \cap E_2$, we conclude that $m'_k \leq \overline{m}'_k$.

We next turn to establishing (36). Consider any $k \in \{1, \ldots, N+1\}$ having non-zero probability that $k \leq K$. Given that $k \leq K$, note that V_{k-1} and Δ_{i_k} have no dependence on S^1_k, so that the samples in S^1_k are conditionally i.i.d.-P given the event that $k \leq K$ and given the random variables V_{k-1} and Δ_{i_k}. Therefore, applying a multiplicative Chernoff bound (Lemma 6 of Appendix D) under the conditional distribution given the event $k \leq K$ and the random variables V_{k-1} and Δ_{i_k}, with conditional probability at least $1 - \frac{\delta}{8k(k+1)}$,

$$\left|S^1_k \cap D_{k-1} \setminus \Delta_{i_k}\right| \leq 2m_k P_X(D_{k-1} \setminus \Delta_{i_k}) + 6\ln\left(\frac{16k(k+1)}{\delta}\right). \tag{39}$$

In particular, by the law of total probability, this implies that for every $k \in \{1, \ldots, N+1\}$, with probability at least $1 - \frac{\delta}{8k(k+1)}$, if $k \leq K$ then (39) holds. Letting E_3' denote the event that (39) holds for every $k \in \{1, \ldots, K\}$, by the union bound, E_3' holds with probability at least $1 - \frac{\delta}{8}$. Combining (39) with Lemma 22, we have that on the event $E_0 \cap E_1 \cap E_2 \cap E_3'$, $\forall k \in \{1, \ldots, K\}$,

$$\left|S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}\right| \leq 2\mathfrak{s}\varepsilon_k m_k + 6\ln\left(\frac{16k(k+1)}{\delta}\right) \leq 3\mathfrak{s}\varepsilon_k m_k,$$

where the rightmost inequality follows from recalling $m_k := \left\lceil \frac{300C''c_0}{\varepsilon_k}\left(\mathsf{d}\log\left(\frac{C''c_0}{\varepsilon_k}\right) + \log\left(\frac{1}{\delta}\right)\right)\right\rceil$, which satisfies $\varepsilon_k m_k \geq 6\ln\left(\frac{16k(k+1)}{\delta}\right)$. Thus, we have established (36).

We argue the left inequality in (37) similarly. Consider any $k \in \{1, \ldots, N+1\}$ having non-zero probability of $k \leq K$. Given $k \leq K$, note that Δ_{i_k} has no dependence on S_k^2 or m_k', so that the m_k' samples in S_k^2 are conditionally i.i.d.-P given the event $k \leq K$ and given the random variables Δ_{i_k} and m_k'. Therefore, applying a multiplicative Chernoff bound (Lemma 6 of Appendix D) under the conditional distribution given the event $k \leq K$ and the random variables Δ_{i_k} and m_k', with conditional probability at least $1 - \frac{\delta}{8(3+N-k)^2}$,

$$\left|S_k^2 \cap \Delta_{i_k}\right| \leq \max\left\{2P_X(\Delta_{i_k})m_k', 6\ln\left(\frac{16(3+N-k)^2}{\delta}\right)\right\}. \tag{40}$$

By the law of total probability, we have that for every $k \in \{1, \ldots, N+1\}$, with probability at least $1 - \frac{\delta}{8(3+N-k)^2}$, if $k \leq K$ then (40) holds. Letting E_3'' denote the event that (40) holds for every $k \in \{1, \ldots, K\}$, by the union bound, E_3'' holds with probability at least $1 - \sum_{k=1}^{N+1} \frac{\delta}{8(3+N-k)^2} \geq 1 - \frac{\delta}{8}$. Let $E_3 = E_3' \cap E_3''$, and note that, by the union bound, E_3 holds with probability at least $1 - \frac{\delta}{4}$. For the remainder of the proof, let us suppose the event $\bigcap_{j=0}^{3} E_j$ occurs.

To arrive at the simpler claimed inequalities in (37), we follow a similar argument to the final part of the proof of Lemma 14. Explicitly, we first note that for any $k \in \{1, \ldots, K\}$, if $\Delta_{i_k} = \emptyset$, we trivially have $|S_k^2 \cap \Delta_{i_k}| = 0 = 2P_X(\Delta_{i_k})m_k' \leq 10\beta\overline{m}_k'$. On the other hand, if $\Delta_{i_k} \neq \emptyset$, the final claim in Lemma 11 implies $P_X(\Delta_{i_k}) > \varepsilon_{k+2}$, and combined with Lemma 13 this further implies $\hat{p}_k \geq P_X(\Delta_{i_k}) > \varepsilon_{k+2}$. Therefore, in this case,

$$2P_X(\Delta_{i_k})m_k' > \frac{2C''c_1^2}{C^4}\left(\mathsf{d} + \ln\left(\frac{4(3+N-k)^2}{\delta}\right)\right) \geq 6\ln\left(\frac{16(3+N-k)^2}{\delta}\right),$$

where the rightmost inequality follows from $c_1 \geq 1$ and $C'' \geq 6C^4$. Thus, the left inequality in (37) follows from (40). The right inequality in (37) follows immediately from the fact (established above) that $m_k' \leq \overline{m}_k'$, together with the fact (from Lemma 20) that $P_X(\Delta_{i_k}) \leq 5\beta$.

The remaining claims in the lemma follow from reasoning about convergence of the relevant series. Specifically, recalling that $\varepsilon_k = C^{1-k}$, $N = \left\lceil \log_C\left(\frac{2}{\varepsilon}\right)\right\rceil$, and $C = \frac{11}{10}$, we note that $\sum_{k=1}^{N+1} \frac{1}{\varepsilon_k^2} = \frac{1}{C^2-1}\left(C^{2(N+1)} - 1\right) \leq \frac{28}{\varepsilon^2}$ and

$$\sum_{k=1}^{N+1} \frac{1}{\varepsilon_k^2}\ln(3+N-k) = C^{2N}\sum_{j=0}^{N} C^{-2j}\ln(2+j) \leq C^{2N}\cdot 10 \leq \frac{49}{\varepsilon^2}.$$

Recalling $\overline{m}_k' = \frac{25C''c_1^2\beta}{\varepsilon_k^2}\left(\mathsf{d} + 2\ln(3+N-k) + \ln\left(\frac{4}{\delta}\right)\right)$, we have

$$\sum_{k=1}^{N+1} \overline{m}_k' \leq \frac{25C''c_1^2\beta}{\varepsilon^2}\left(28\mathsf{d} + 2\cdot 49 + 28\ln\left(\frac{4}{\delta}\right)\right) \leq M_2. \tag{41}$$

To obtain the query bound $Q(\varepsilon, \delta; \beta)$ in (38), note that the total number of queries is precisely

$$\left(\sum_{k=1}^{K}\left|S_k^2 \cap \Delta_{i_k}\right|\right) + \left(\sum_{k=1}^{K}\left|S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}\right|\right). \tag{42}$$

By (37), the first term in (42) is upper bounded by $10\beta \cdot \sum_{k=1}^{N+1} \overline{m}'_k$, and (41) implies this is at most $10\beta M_2$. The second term in (42) is trivially upper bounded by $M_1 := \sum_{k=1}^{N+1} m_k$. Moreover, noting that $\varepsilon_k m_k$ is increasing in k, (36) implies the second term in (42) is also upper bounded by $3\mathfrak{s}\varepsilon_{N+1} \cdot m_{N+1} \cdot (N+1) \le (3/2)\mathfrak{s}\varepsilon(N+1)m_{N+1}$. Together with the definition of $Q(\varepsilon, \delta; \beta)$ from (38), we have that the total number of queries (42) is at most $Q(\varepsilon, \delta; \beta)$.

The bound on the asymptotic form of $Q(\varepsilon, \delta; \beta)$ in (38) follows immediately from the definitions. Specifically, by definition of M_2, we have $10\beta M_2 = O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right)$. Moreover, since $\varepsilon_{N+1} \ge \frac{\varepsilon}{2C}$, we have $(3/2)\mathfrak{s}\varepsilon(N+1)m_{N+1} = O\left(\mathfrak{s}\log\left(\frac{1}{\varepsilon}\right)\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$, while (since each $k \le N+1$ has $\varepsilon_k \ge \frac{\varepsilon}{2C}$), $M_1 = \sum_{k=1}^{N+1} m_k \le \sum_{k=1}^{N+1} \frac{301C''c_0}{\varepsilon_k}\left(\mathsf{d}\log\left(\frac{2CC''c_0}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right) = O\left(\frac{1}{\varepsilon}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$ by evaluating the geometric series. ■

As a final step before composing these lemmas into a proof of Theorem 5, we state an explicit bound on the number of unlabeled examples used by the algorithm. Much of this analysis is already implied by the above lemmas: namely, by definition, the number of examples allocated to data sets S_k^1 and S_k^4 is precisely $2M_1 = 2\sum_{k=1}^{N+1} m_k$, and Lemma 23 implies the number of examples allocated to data sets S_k^2 is at most M_2. What remains is to bound the number of examples allocated to the data sets $S_{k,i}^3$, which hinges on bounding the number of iterations of the '*While*' loop for each k. We have already noted, in Lemma 10, that $\max \mathcal{I}_k \le \frac{1}{\varepsilon_{k+3}}$ on the event E_0, which already suffices to establish a coarse bound $\tilde{O}\left(\frac{\mathsf{d}}{\varepsilon^2}\right)$. However, we will need a slight refinement to obtain the claimed upper bound, which will follow from a combination of Lemmas 11 and 20.

Lemma 24. *On the event $\bigcap_{j=0}^{3} E_j$, the total number of examples allocated to data sets S_k^1, S_k^4 ($k \le N+1$), S_k^2 ($k \le K$), and $S_{k,i}^3$ ($k \in \mathcal{K}$, $i \in \mathcal{I}_k$) is at most*

$$M(\varepsilon, \delta; \beta) := 3M_1 + M_2 + \frac{100C^4\beta m_N}{\varepsilon} = O\left(\frac{\beta+\varepsilon}{\varepsilon^2}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right).$$

Proof. Suppose the event $\bigcap_{j=0}^{3} E_j$ occurs. By definition, the number of examples allocated to data sets S_k^1 and S_k^4 is m_k each, for $k \in \{1, \dots, N+1\}$, so that the total number of such examples is $\sum_{k=1}^{N+1} 2m_k = 2M_1$. Also, by the first claim in Lemma 23, the number m'_k of examples allocated to each S_k^2 data set (for $k \in \{1, \dots, K\}$) satisfies $m'_k \le \overline{m}'_k$. Moreover, Lemma 23 also establishes that $\sum_{k=1}^{N+1} \overline{m}'_k \le M_2$. Together, we have that the total number of examples allocated to data sets S_k^2 is $\sum_{k=1}^{K} m'_k \le M_2$. Thus, to complete the proof of Lemma 24, it suffices to bound the total number of examples allocated to data sets $S_{k,i}^3$ ($k \in \mathcal{K}$, $i \in \mathcal{I}_k$).

Toward this end, recall that for each $k \in \mathcal{K}$, each $S_{k,i}^3$ is of size m_k, and is allocated if and when the algorithm reaches Step 5 with values (k, i). Thus, if $k = K$ (which, by the final claim in Lemma 10, occurs only if the algorithm returns in Step 4 in round k), then *no* examples are allocated to any $S_{k,i}^3$ sets in round k, whereas if $k < K$, then the number of $S_{k,i}^3$ data sets allocated during round k is precisely the number of distinct values of i encountered in round k: that is, $|\mathcal{I}_k|$. Moreover, note that since each time through the '*While*' loop increments i, each $k \in \mathcal{K}$ with $k < K$ has $|\mathcal{I}_k| = i_{k+1} - i_k + 1$. It follows that the total number of examples allocated to data sets $S_{k,i}^3$ in the algorithm is precisely $\sum_{k \in \mathcal{K}: k < K} m_k(i_{k+1} - i_k + 1)$.

Next we upper bound $i_{k+1} - i_k$ for each $k \in \mathcal{K}$ with $k < K$. Specifically, for any such k, note that $\Delta_{i_{k+1}} \setminus \Delta_{i_k} = \bigcup_{i=i_k}^{i_{k+1}-1} (\Delta_{i+1} \setminus \Delta_i)$, and by definition the sets $\Delta_{i+1} \setminus \Delta_i$ are disjoint over i. Moreover, by Lemma 11 and the definition of Δ_{i+1} in Step 7, any $i \in \{i_k, \dots, i_{k+1} - 1\}$ has $P_X(\Delta_{i+1} \setminus \Delta_i) > \varepsilon_{k+3}$. Therefore,

$$P_X(\Delta_{i_{k+1}} \setminus \Delta_{i_k}) = \sum_{i=i_k}^{i_{k+1}-1} P_X(\Delta_{i+1} \setminus \Delta_i) \ge (i_{k+1} - i_k)\varepsilon_{k+3}.$$

On the other hand, by Lemma 20, $P_X(\Delta_{i_{k+1}} \setminus \Delta_{i_k}) \le P_X(\Delta_{i_{k+1}}) \le 5\beta$. Combining these inequalities, we conclude that $(i_{k+1} - i_k) \le \frac{5C^3\beta}{\varepsilon_k}$. Combined with the facts that $m_k \le m_N$ and

$\sum_{k\in\mathcal{K}:k<K} m_k \le M_1$, altogether we have

$$\sum_{k\in\mathcal{K}:k<K} m_k(i_{k+1}-i_k+1) \le M_1 + m_N \sum_{k=1}^{N} \frac{5C^3\beta}{\varepsilon_k} \le M_1 + \frac{100C^4\beta m_N}{\varepsilon},$$

where the last inequality follows by evaluating the geometric series and recalling $\varepsilon_N \ge \frac{\varepsilon}{2C}$. This completes the proof that the total number of examples allocated to data sets S_k^1, $S_{k,i}^3$, S_k^2, S_k^4 is at most $M(\varepsilon,\delta;\beta)$. The claimed asymptotic form of $M(\varepsilon,\delta;\beta)$ follows immediately from the definitions of the quantities involved: namely, by definition, $3M_1 = \Theta\big(\frac{1}{\varepsilon}\left(\mathsf{d}\log\big(\frac{1}{\varepsilon}\big)+\log\big(\frac{1}{\delta}\big)\right)\big)$, $M_2 = \Theta\Big(\frac{\beta}{\varepsilon^2}\left(\mathsf{d}+\log\big(\frac{1}{\delta}\big)\right)\Big)$, and (since $\varepsilon_N \ge \frac{\varepsilon}{2C}$) $\frac{100C^4\beta m_N}{\varepsilon} = \Theta\Big(\frac{\beta}{\varepsilon^2}\left(\mathsf{d}\log\big(\frac{1}{\varepsilon}\big)+\log\big(\frac{1}{\delta}\big)\right)\Big)$. ■

We are now ready to combine the above lemmas into a complete proof of Theorem 5.

Proof of Theorem 5. By the union bound, the event $\bigcap_{j=0}^{3} E_j$ has probability at least $1-\delta$. By Lemma 24, on $\bigcap_{j=0}^{3} E_j$, $\mathbb{A}_{\text{avid}}$ uses at most $M(\varepsilon,\delta;\beta)$ (as defined in the lemma) of the examples in the sequence; in particular, this means that if we were to run the algorithm with a finite sequence $(X_1,Y_1),\dots,(X_m,Y_m)$, for any $m \ge M(\varepsilon,\delta;\beta)$, then on the event $\bigcap_{j=0}^{3} E_j$, the behavior of the algorithm (e.g., queries, returned $\hat{h}$) is identical to the *idealized* setting the above lemmas were established under (where there is an unlimited supply of examples), and hence the claims in the above lemmas remain valid. Thus, for any sample size $m \ge M(\varepsilon,\delta;\beta)$, on the event $\bigcap_{j=0}^{3} E_j$, by Lemma 19 we have $\mathrm{er}_P(\hat{h}) \le \inf_{h\in\mathbb{C}} \mathrm{er}_P(h) + \varepsilon$, and by Lemma 23 the total number of queries is at most $Q(\varepsilon,\delta;\beta)$ as defined therein. ■

A remark on intersecting with D_{k-1} in Δ_{i_k}: We remark that Theorem 5 remains valid if we restrict either (or both) h_1, h_2 to be in V_{k-1} in the definition (2) of $V_{k-1}^{(4)}$. The entire proof remains valid (applying the same change to $V_{k-1}^{(3)}$), with the only exception being the first inequality in Lemma 20, which should then replace $\mathbb{C}$ by V_{k-1}. This change is of no consequence to the second inequality in the lemma since the proof of Lemma 17 in fact implies that $h^\star \in V_{k-1}$ holds simultaneously (on $E_0\cap E_1\cap E_2$) for *all* functions $h^\star \in \mathbb{C}$ satisfying (7) (and hence $\inf_{h\in V_{k-1}} \mathrm{er}_P(h) = \beta$). Moreover, with this restriction to require $h_2 \in V_{k-1}$ in $V_{k-1}^{(4)}$, we can extend the intersection with D_{k-1} to the Δ_{i_k} region: that is, instead of querying all of $S_k^2 \cap \Delta_{i_k}$ (in Step 2) or $S_{N+1}^2 \cap \Delta_{i_{N+1}}$ (in Step 9), we can instead merely query the subset $S_k^2 \cap D_{k-1} \cap \Delta_{i_k}$ (in Step 2) or $S_{N+1}^2 \cap D_N \cap \Delta_{i_{N+1}}$ (in Step 9). With this change, we must then also modify the definition of $\hat{\mathrm{er}}_k^{1,2}(h)$ in (1) to $\hat{\mathrm{er}}_k^{1,2}(h) := \hat{P}_{S_k^1}(\mathrm{ER}(h)\cap D_{k-1}\setminus\Delta_{i_k}) + \hat{P}_{S_k^2}(\mathrm{ER}(h)\cap D_{k-1}\cap\Delta_{i_k})$. The argument in Lemma 15 extends to this modified definition of $\hat{\mathrm{er}}_k^{1,2}(h)$, since (as in the proof of (11) in Lemma 10) we are only interested in error *differences*, which, for $h,h'\in V_{k-1}$, satisfy $\hat{P}_{S_k^2}(\mathrm{ER}(h)\cap D_{k-1}\cap\Delta_{i_k}) - \hat{P}_{S_k^2}(\mathrm{ER}(h')\cap D_{k-1}\cap\Delta_{i_k}) = \hat{P}_{S_k^2}(\mathrm{ER}(h)\cap\Delta_{i_k}) - \hat{P}_{S_k^2}(\mathrm{ER}(h')\cap\Delta_{i_k})$. Indeed, with additional modifications to the proof, we can then even slightly refine the query complexity analysis, since if we replace the sets $\mathrm{ER}(h)\cap\Delta_{i_k}$ with $\mathrm{ER}(h)\cap D_{k-1}\cap\Delta_{i_k}$ in Lemma 14, the *envelope* set in the application of Lemma 8 in the proof of Lemma 14 can be chosen as $D_{k-1}\cap\Delta_{i_k}$, so that we can refine the definition of $\hat{p}_k$ to $2\hat{P}_{S_k^4}(D_{k-1}\cap\Delta_{i_k}) + O(\varepsilon_k)$. However, since these changes concern only the leading term $\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\big(\frac{1}{\delta}\big)\right)$ in Theorem 5, which is already optimal (perfectly matching the lower bounds of Kääriäinen, 2006; Beygelzimer, Dasgupta, and Langford, 2009), they are completely inconsequential to the theorem. We have therefore stated the algorithm without these modifications, for simplicity. However, this modified variant would be interesting in the context of P-dependent analysis, where it can lead to refinements to the leading term in the upper bound under certain favorable distributions. We leave the investigation of such refinements as an interesting direction for future work (focusing our P-dependent analysis in Appendix F on refining the *lower-order* term).

F Distribution-Dependent Analysis

In addition to analysis based on the star number (Hanneke and Yang, 2015), the active learning literature includes a variety of *distribution-dependent* complexity measures which have been used to analyze the query complexity in various contexts (see Appendix A). In this section, we will add to this line of work a distribution-dependent analysis of $\mathbb{A}_{\mathrm{avid}}$ which replaces the star number $\mathfrak{s}$ in Theorem 3 by a (never-larger) distribution-dependent quantity (Theorem 27), which can be further upper-bounded in terms of a simpler and more-familiar quantity: namely, a quadratic θ^2 dependence in the *disagreement coefficient* (Hanneke, 2007b; Definition 25 below). We also show (in Appendix F.1) that it is *not possible* (by *any* algorithm) to obtain a lower-order term which replaces the star number in Theorem 3 with the disagreement coefficient θ itself, so that the aforementioned θ^2 quadratic dependence generally cannot be reduced to linear (without introducing other factors). We will also present (in Appendix F.3) a slight refinement of $\mathbb{A}_{\mathrm{avid}}$, which replaces the region of disagreement D_{k-1} by a carefully-chosen *subregion*, following the technique of Zhang and Chaudhuri (2014); Balcan, Broder, and Zhang (2007), which yields a corresponding refinement of the distribution-dependent query complexity bound. For instance, in the case of learning homogeneous linear classifiers under a uniform (or isotropic log-concave) distribution, this recovers a known query complexity bound $\tilde{O}\left(\mathsf{d}\frac{\beta^2}{\varepsilon^2} + \mathsf{d}\right)$ (and indeed, improves log factors in the lead term compared to prior works).

The Disagreement Coefficient: In the context of *agnostic* active learning, the most commonly-used P-dependent complexity measure is the *disagreement coefficient*, introduced by Hanneke (2007b), defined as follows.

Definition 25. *For any concept class $\mathbb{C}$ and distribution P_X on $\mathcal{X}$, for any measurable function $f : \mathcal{X} \to \{0,1\}$, for any $\varepsilon \geq 0$, the* disagreement coefficient*, denoted by $\theta_{P_X,f}(\varepsilon)$, is defined as*

$$\theta_{P_X,f}(\varepsilon) := \sup_{r>\varepsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}_{P_X}(f,r)))}{r} \vee 1,$$

*where $\mathrm{B}_{P_X}(f,r) := \{h \in \mathbb{C} : P_X(h \neq f) \leq r\}$ denotes the r-*ball centered at *f, and $\mathrm{DIS}(\mathbb{C}') := \{x \in \mathcal{X} : \exists h, h' \in \mathbb{C}', h(x) \neq h'(x)\}$ denotes the* region of disagreement *(as in Section 4). For any distribution P on $\mathcal{X} \times \{0,1\}$ and $\varepsilon > 0$, for $h^\star$ as in (7),[14] define $\theta_P(\varepsilon) := \theta_{P_X,h^\star}(\varepsilon)$.*

There are many works establishing bounds on the disagreement coefficient for commonly-studied classes $\mathbb{C}$ under various restrictions on the distribution P (see Hanneke, 2014, for a detailed summary). As discussed in Appendix A, the disagreement coefficient commonly appears in analyses of the query complexity of *disagreement-based* active learning methods (e.g., Hanneke, 2007b, 2009b, 2011, 2014; Dasgupta, Hsu, and Monteleoni, 2007). Since the lower-order term in Theorem 3 arises from the analysis of queries in the region of disagreement $\mathrm{DIS}(V_{k-1})$ of V_{k-1}, one might naturally wonder whether we can replace $\mathfrak{s}$ with $\theta_P(\varepsilon)$ in the upper bound in Theorem 3. Hanneke and Yang (2015) have shown that $\sup_P \theta_P(\varepsilon) = \mathfrak{s} \wedge \frac{1}{\varepsilon}$ for $\varepsilon \in (0,1]$, which implies that if we could replace $\mathfrak{s}$ by $\theta_P(\varepsilon)$ it would indeed represent a distribution-dependent refinement of the upper bound in Theorem 3. However, it turns out this is *not possible* (by *any* algorithm) for some classes $\mathbb{C}$, as we demonstrate by an example in Appendix F.1. Following this, in Appendix F.2, we find that it is possible to achieve a lower-order term $\tilde{O}(\mathsf{d}\theta_P(\beta+\varepsilon)^2)$, and indeed this is achieved by $\mathbb{A}_{\mathrm{avid}}$. This quadratic dependence unfortunately means the upper bound is sometimes loose (i.e., sometimes larger than that in Theorem 3). However, as an intermediate step, we also establish a query complexity bound (Theorem 27) expressed in terms of a modified disagreement coefficient which is *never larger* than the $\mathfrak{s}$-dependent query complexity bound in Theorem 3 (though which is more difficult to evaluate due to a more-involved definition).

[14] When $h^\star$ is not uniquely defined, in principle we can define $\theta_P(\varepsilon)$ as the *infimum* value among all choices of such $h^\star$. It is also possible to define $h^\star$ as an ε-independent fixed function, even when $\mathrm{er}_P(h)$ does not have a minimizer in $\mathbb{C}$, by choosing it as an element of the $L_1(P_X)$-closure of $\mathbb{C}$ having $\mathrm{er}_P(h^\star) = \inf_{h\in\mathbb{C}} \mathrm{er}_P(h)$: see (Hanneke, 2012) for a proof that such an $h^\star$ always exists when $\mathrm{VC}(\mathbb{C}) < \infty$. In particular, with such an $h^\star$, the limiting value $\theta_P(0) := \theta_{P_X,h^\star}(0)$ is also well-defined.

F.1 Impossibility of Replacing $\mathfrak{s}$ with $\theta_P(\varepsilon)$

In this section, we present an example demonstrating that *no algorithm* can achieve a P-dependent query complexity bound which replaces $\mathfrak{s}$ by $\theta_P(\varepsilon)$ in Theorem 3.

An Example: Consider the following concept class (see Hanneke, 2007b, for a related construction). Let $\mathcal{X} = \mathbb{Z}$ (the integers) and define a concept class

$$\mathbb{C}_{\text{ts}} := \{\mathbb{1}_{\{-t\}\cup[t,\infty)} : t \in \mathbb{N}\}.$$

In other words, each $h \in \mathbb{C}_{\text{ts}}$ defines a *threshold* classifier on the *positive* integers and a *singleton* classifier on the *negative* integers, and the position $-t$ of the singleton point mirrors the position t of the threshold boundary point.

Fix any $\varepsilon, \beta \in (0, 1/3)$, denote by $n = \frac{1-2\beta}{2\varepsilon}$, and for simplicity suppose $n \in \mathbb{N}$. Define a marginal distribution P_X on $\mathcal{X}$ as follows: $\forall x \in \{1, \ldots, n\}$, $P_X(\{x\}) = \frac{2\beta}{n}$ and $P_X(\{-x\}) = 2\varepsilon$. Note that this completely specifies P_X. Now define a family of probability distributions $P_1, \ldots, P_n$: for each $t \in \{1, \ldots, n\}$, P_t has marginal distribution P_X on $\mathcal{X}$ and conditional distribution $\forall x \in \mathcal{X}$

$$P_t(Y = 1|X = x) = \begin{cases} 1, & \text{if } x = -t \\ 0, & \text{if } x \in \{-1, \ldots, -n\} \setminus \{-t\} \\ \frac{1}{2}, & \text{otherwise} \end{cases}.$$

In particular, note that each P_t satisfies $\inf_{h\in\mathbb{C}_{\text{ts}}} \text{er}_{P_t}(h) = \beta$ and $h^\star$ is uniquely equal $\mathbb{1}_{\{-t\}\cup[t,\infty)}$.[15]

Upper-bounding $\theta_{P_t}(\varepsilon)$: For any P_t, we will argue $\theta_{P_t}(\varepsilon) \le \theta_{P_X,h^\star}(0) = O(\frac{1}{\beta})$. For any $r > 0$, if $h \in \text{B}_{P_X}(h^\star, r)$, then letting $k_h := |\{h \ne h^\star\} \cap \{1, \ldots, n\}|$, since each $x \in \{1, \ldots, n\}$ has $P_X(\{x\}) = \frac{2\beta}{n}$, we must have $k_h \le k_r := \left\lfloor \frac{rn}{2\beta} \right\rfloor$. Since h and $h^\star$ both implement threshold functions in this region $\{1, \ldots, n\}$, the k_h elements in $\{h \ne h^\star\} \cap \{1, \ldots, n\}$ are a contiguous segment: either $\{t, \ldots, t + k_h - 1\}$ or $\{t - k_h, \ldots, t - 1\}$. In either case, we have $\{h \ne h^\star\} \cap \{1, \ldots, n\} \subseteq \{t - k_r, \ldots, t + k_r - 1\} \cap \{1, \ldots, n\}$. Moreover, since $h = \mathbb{1}_{\{-t'\}\cup[t',\infty)}$ for some $t' \in \mathbb{N}$, this further implies $\{h \ne h^\star\} \cap \{-1, \ldots, -n\} \subseteq \{-(t - k_r), \ldots, -(t + k_r)\} \cap \{-1, \ldots, -n\}$. Since $\text{DIS}(\text{B}_{P_X}(h^\star, r))$ is just the union of these $\{h \ne h^\star\}$ regions among all $h \in \text{B}_{P_X}(h^\star, r)$, we have

$$\begin{aligned} &\text{DIS}(\text{B}_{P_X}(h^\star, r)) \cap \{-n, \ldots, -1, 1, \ldots, n\} \\ &\subseteq (\{-(t - k_r), \ldots, -(t + k_r)\} \cap \{-1, \ldots, -n\}) \cup (\{t - k_r, \ldots, t + k_r - 1\} \cap \{1, \ldots, n\}) \\ &\implies P_X(\text{DIS}(\text{B}_{P_X}(h^\star, r))) \le (2k_r + 1)2\varepsilon + 2k_r\frac{2\beta}{n} \le \frac{r(1 - 2\beta)}{\beta} + 2\varepsilon + 2r. \end{aligned}$$

We also note that any $r < 2\varepsilon$ has $\text{B}_{P_X}(h^\star, r) = \{h^\star\}$. Altogether,

$$\theta_{P_t}(\varepsilon) \le \theta_{P_X,h^\star}(0) \le \sup_{r\ge 2\varepsilon} \frac{r((1 - 2\beta)/\beta) + 2\varepsilon + 2r}{r} = \frac{1 - 2\beta}{\beta} + 3 = O\left(\frac{1}{\beta}\right).$$

Lower-bounding the query complexity: On the other hand, we will argue that the query complexity is $\Omega(\frac{1}{\varepsilon})$ under the assumption $P \in \{P_1, \ldots, P_n\}$. Note that every $h : \mathcal{X} \to \{0, 1\}$ has the *same* value of $P_t(\text{ER}(h) \setminus \{-1, \ldots, -n\}) = \beta$. Together with the definition of P_t in $\{-1, \ldots, -n\}$, this implies any $h : \mathcal{X} \to \{0, 1\}$ with $\{h \ne h^\star\} \cap \{-1, \ldots, -n\} \ne \emptyset$ has $\text{er}_{P_t}(h) - \text{er}_{P_t}(h^\star) \ge 2\varepsilon$. Thus, the problem of learning, to an excess error ε (under the assumption that $P \in \{P_t : t \in \{1, \ldots, n\}\}$) is equivalent to the problem of *identifying* the value $t \in \{1, \ldots, n\}$ for which the distribution $P = P_t$: that is, if an algorithm returns $\hat{h}$ with $\text{er}_{P_t}(\hat{h}) - \text{er}_{P_t}(h^\star) \le \varepsilon$, the unique $x \in \{1, \ldots, n\}$ for which $\hat{h}(-x) = 1$ satisfies $x = t$. Moreover, in the active learning problem defined by these distributions, for every $x \notin \{-1, \ldots, -n\}$, the conditional distribution $P_t(Y = 1|X = x)$ of responses to queries for examples at x is *invariant* to t, so that such queries reveal no information about which

[15] Indeed, these distributions P_t even satisfy the stronger *benign noise* property: i.e., $\inf_{h\in\mathbb{C}_{\text{ts}}} \text{er}_{P_t}(h) =$ Bayes risk (a setting studied by Hanneke, 2009b; Hanneke and Yang, 2015). Thus, the argument in this section further implies the impossibility of using $\theta_P(\varepsilon)$ in the lower-order term under benign noise.

t has $P = P_t$, and hence without loss of generality we can restrict to active learning algorithms that do not query outside $\{-1, \ldots, -n\}$. The unlabeled examples also reveal no such information, since all P_t have the same marginal distribution P_X. Altogether, the active learning problem for this set of distributions is information-theoretically no easier (in terms of query complexity) than the problem of actively identifying a *singleton* classifier on $\{-1, \ldots, -n\}$ in the realizable case under marginal $\mathrm{Uniform}(\{-1, \ldots, -n\})$.[16] It is well known that the minimax query complexity of this latter problem (with confidence parameter $\delta = 1/3$) is $\Omega(n) = \Omega(\frac{1}{\varepsilon})$ (Dasgupta, 2004, 2005; Hanneke, 2014; Hanneke and Yang, 2015), which therefore serves as a lower bound on the minimax query complexity for $P \in \{P_t : t \in \{1, \ldots, n\}\}$: that is, for every active learning algorithm, there exists $P \in \{P_t : t \in \{1, \ldots, n\}\}$ for which, with probability at least δ, it either makes $\Omega(\frac{1}{\varepsilon})$ queries or returns $\hat{h}$ with $\mathrm{er}_P(\hat{h}) > \inf_{h \in \mathbb{C}_{\mathrm{ts}}} \mathrm{er}_P(h) + \varepsilon$.

Conclusion that $\theta_{P_t}(\varepsilon)$ is not achievable in the lower-order term: From the above arguments, we can conclude that for the class $\mathbb{C}_{\mathrm{ts}}$, it is *not possible* to replace $\mathfrak{s}$ by $\theta_P(\varepsilon)$ (or indeed $\theta_P(0)$) in the upper bound of Theorem 3 to obtain a P-dependent refinement of the upper bound. Formally, for *every* active learning algorithm guaranteeing that, for every P with $\inf_{h \in \mathbb{C}_{\mathrm{ts}}} \mathrm{er}_P(h) \leq \beta$, with probability at least $2/3$ (i.e., $\delta = 1/3$), it returns $\hat{h}$ with $\mathrm{er}_P(\hat{h}) \leq \inf_{h \in \mathbb{C}_{\mathrm{ts}}} \mathrm{er}_P(h) + \varepsilon$, there exists a distribution P satisfying this for which $\theta_P(\varepsilon) \leq \frac{1-2\beta}{\beta} + 3 = O(\frac{1}{\beta})$, yet with probability at least $1/3$, the algorithm makes a number of queries $\Omega(\frac{1}{\varepsilon})$. We have argued this conclusion for any choices of $\varepsilon, \beta \in (0, 1/3)$ (with $n \in \mathbb{N}$ for simplicity). In particular, for $\varepsilon \ll \beta \ll \sqrt{\varepsilon}$ (e.g., $\beta \approx \varepsilon^{2/3}$), such a distribution P has $\frac{\beta^2}{\varepsilon^2} + \theta_P(\varepsilon) = \frac{\beta^2}{\varepsilon^2} + O(\frac{1}{\beta}) \ll \frac{1}{\varepsilon}$, so that replacing $\mathfrak{s}$ with $\theta_P(\varepsilon)$ in Theorem 3 *cannot* yield a valid query complexity bound (holding for all P) for *any* active learning algorithm. Indeed, we have established that this conclusion also holds for $\theta_P(0)$ (as defined in footnote 14).

We will see in Corollary 28 of Appendix F.2 that $\mathbb{A}_{\mathrm{avid}}$ *does* achieve an upper-bound $\tilde{O}\big(\mathsf{d}\theta_P(\beta + \varepsilon)^2\big)$ on the lower-order term: a *quadratic* dependence on the disagreement coefficient. This conclusion is compatible with the above scenario, since $\frac{\beta^2}{\varepsilon^2} + \frac{1}{\beta^2} = \Omega\big(\frac{1}{\varepsilon}\big)$ for the full range of β, ε.

F.2 Replacing $\mathfrak{s}$ with $\theta_P(\beta + \varepsilon)^2$

Appendix F.1 implies the disagreement coefficient $\theta_P(\varepsilon)$, as defined in Definition 25, cannot be used as a P-dependent substitute for the star number $\mathfrak{s}$ in Theorem 3 (at least, not with a *linear* dependence). In this section, we will argue that the AVID Agnostic algorithm $\mathbb{A}_{\mathrm{avid}}$ *does* achieve a P-dependent lower-order term which is at most *quadratic* in the disagreement coefficient: namely, $\tilde{O}(\mathsf{d}\theta_P(\beta + \varepsilon)^2)$. We will argue this by first establishing a P-dependent refinement of Theorem 5 based on a modified disagreement coefficient (Definition 26) which is *never larger* than the star number. While this quantity itself is often more-difficult to calculate, compared to the original disagreement coefficient $\theta_P(\varepsilon)$, fortunately it is always upper bounded by $O\Big(\frac{\beta^2}{\varepsilon^2} + \theta_P(\beta + \varepsilon)^2\Big)$. In particular, this means that for any P with $\theta_P(0) < \infty$, the asymptotic dependence on ε, δ in the lower-order term in Theorem 3 can be reduced to $\mathrm{polylog}\big(\frac{1}{\varepsilon\delta}\big)$.

Specifically, the modified disagreement coefficient we consider can be expressed as the value $\theta_{P_\Delta, h^\star}(\varepsilon)$ produced under a *restriction* of P_X to a subregion $\mathcal{X} \setminus \Delta$ of size at least $1 - O(\beta)$. Toward stating the definition, we first extend Definition 25 to allow for general measures μ: that is, for any measure μ on $\mathcal{X}$ and measurable $f : \mathcal{X} \to \{0, 1\}$, define $\mathrm{B}_\mu(f, r) := \{h \in \mathbb{C} : \mu(h \neq f) \leq r\}$,

[16] Formally, for any active learning algorithm $\mathbb{A}$, under distributions $P \in \{P_t : t \in \{1, \ldots, n\}\}$, we can convert $\mathbb{A}$ into an active learner $\mathbb{A}'$ for realizable-case singletons under $\mathrm{Uniform}(\{-1, \ldots, -n\})$ with at most the query complexity of $\mathbb{A}$ under such distributions P. Specifically, given any number m of i.i.d. unlabeled examples $X_1, \ldots, X_m \sim \mathrm{Uniform}(\{-1, \ldots, -n\})$, define independent random variables (also independent of $X_1, \ldots, X_m$) $B_1, \ldots, B_m \sim \mathrm{Bernoulli}(2\beta)$, $X'_1, \ldots, X'_m \sim \mathrm{Uniform}(\{1, \ldots, n\})$, and $Y'_1, \ldots, Y'_m \sim \mathrm{Bernoulli}(\frac{1}{2})$. For each $i \leq m$, let $X''_i = X_i$ if $B_i = 0$ and $X''_i = X'_i$ if $B_i = 1$. Then $\mathbb{A}'$ runs $\mathbb{A}$ with unlabeled data $X''_1, \ldots, X''_m$; whenever $\mathbb{A}$ queries an X''_i with $B_i = 0$, $\mathbb{A}'$ queries for the label Y_i of X_i and gives this as a response to the query, and whenever $\mathbb{A}$ queries an X''_i with $B_i = 1$, $\mathbb{A}'$ gives Y'_i as a response to the query. Note that the corresponding data sequence and responses observed by $\mathbb{A}$ are indeed identical to running $\mathbb{A}$ under $P = P_t$, where $-t$ is the singleton location for the realizable-case singleton problem $P_t(\cdot|\{-1, \ldots, -n\})$. Thus, the query complexity of $\mathbb{A}'$ identifying the t for the realizable-case singletons distribution $P_t(\cdot|\{-1, \ldots, -n\})$ is at most that of $\mathbb{A}$ identifying this t when $P = P_t$.

and for $\varepsilon \geq 0$ define $\theta_{\mu,f}(\varepsilon) := \sup_{r>\varepsilon} \frac{\mu(\mathrm{DIS}(\mathrm{B}_\mu(f,r)))}{r} \vee 1$. We then consider the following definition: a *region-excluded disagreement coefficient*.

Definition 26. *For any distribution P on $\mathcal{X} \times \{0,1\}$ and any measurable $\Delta \subseteq \mathcal{X}$, define a measure $A \mapsto P_\Delta(A) := P_X(A \setminus \Delta)$. For any $\varepsilon, \tau \geq 0$, for $h^\star \in \mathbb{C}$ as in (7) (under P),[17] define*

$$\theta_P(\varepsilon;\tau) := \sup_{\Delta \subseteq \mathcal{X}: P_X(\Delta) \leq \tau} \theta_{P_\Delta, h^\star}(\varepsilon).$$

We can equivalently define $\theta_P(\varepsilon;\tau)$ as the disagreement coefficient under a worst-case *conditional* distribution $P_X(\cdot|\mathcal{X} \setminus \Delta)$: that is,

$$\theta_P(\varepsilon;\tau) = \sup_{\Delta \subseteq \mathcal{X}: P_X(\Delta) \leq \tau} \theta_{P_X(\cdot|\mathcal{X}\setminus\Delta), h^\star}(\varepsilon / P_X(\mathcal{X} \setminus \Delta)), \tag{43}$$

where we define $\theta_{P_X(\cdot|\mathcal{X}\setminus\Delta), h^\star}(\varepsilon / P_X(\mathcal{X} \setminus \Delta)) = 1$ in the case $P_X(\mathcal{X} \setminus \Delta) = 0$ (which coincides with the value $\theta_{P_\Delta, h^\star}(\varepsilon)$ for such Δ).

We may note that $\theta_P(\varepsilon;\tau)$ indeed provides a *refinement* of the star number, in that it is *never larger*. Specifically, since Hanneke and Yang (2015) have shown

$$\sup_{P_X} \sup_{h \in \mathbb{C}} \theta_{P_X,h}(\varepsilon) = \mathfrak{s} \wedge \frac{1}{\varepsilon} \tag{44}$$

for every $\varepsilon \in (0,1]$, the expression in (43) of $\theta_P(\varepsilon;\tau)$ as the disagreement coefficient under *conditional* distributions immediately implies

$$\theta_P(\varepsilon;\tau) \leq \mathfrak{s} \wedge \frac{1}{\varepsilon}. \tag{45}$$

Thus, replacing $\mathfrak{s}$ in Theorem 3 by $\theta_P(\varepsilon;\tau)$ would indeed yield a (never-larger) P-dependent refinement.

We give examples below (Appendix F.2.1) of calculating and upper-bounding $\theta_P(\varepsilon;\tau)$ under various scenarios $(\mathbb{C}, P)$. We remark that, due to the supremum over regions Δ, the quantity $\theta_P(\varepsilon;\tau)$ is often *much* more involved to calculate or bound compared to the original disagreement coefficient $\theta_P(\varepsilon)$ in Definition 25. We might therefore think of $\theta_P(\varepsilon;\tau)$ as a kind of *intermediate* complexity measure, which is useful in that it provides a P-dependent refinement of $\mathfrak{s}$, while also admitting general upper bounds which are more accessible than directly calculating $\theta_P(\varepsilon;\tau)$. Concretely, there are at least *weak* relations between $\theta_P(\varepsilon;\tau)$ and the more-familiar disagreement coefficient from Definition 25: namely, $\theta_P(\varepsilon) \leq \theta_P(\varepsilon;\tau)$ and

$$\begin{aligned}\theta_P(\varepsilon;\tau) &\leq \sup_{r>\varepsilon} \frac{P_X(\mathrm{DIS}(\mathrm{B}_{P_X}(h^\star, \tau + r)))}{r} \vee 1 \\ &\leq \theta_P(\tau+\varepsilon)\left(\frac{\tau+\varepsilon}{\varepsilon}\right) \leq \theta_P(\tau+\varepsilon)^2 + \left(\frac{\tau+\varepsilon}{\varepsilon}\right)^2.\end{aligned} \tag{46}$$

These upper bounds on $\theta_P(\varepsilon;\tau)$ are noteworthy since $\theta_P(\tau+\varepsilon)$ is typically significantly easier to calculate compared to directly calculating $\theta_P(\varepsilon;\tau)$ (and there are already many works deriving bounds on $\theta_P(\tau+\varepsilon)$ for various scenarios; see Hanneke, 2014).

The quantity $\theta_P(\varepsilon;\tau)$ is particularly well-suited for the analysis of $\mathbb{A}_{\mathrm{avid}}$, since the algorithm explicitly maintains low diameter of V_k under a region-excluded measure $A \mapsto P_X(A \setminus \Delta_{i_k})$. Specifically, Lemma 12 implies $V_{k-1} \subseteq \mathrm{B}_{P_{\Delta_{i_k}}}(h^\star, \varepsilon_k)$, while Lemma 20 implies $P_X(\Delta_{i_k}) \leq 5\beta$, so that $P_X(D_{k-1} \setminus \Delta_{i_k}) \leq \theta_P(\varepsilon_k; 5\beta)\varepsilon_k$, and hence the number of queries in $S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$ is $O(\theta_P(\varepsilon_k;5\beta)\varepsilon_k m_k) = \tilde{O}(\theta_P(\varepsilon;5\beta)\mathsf{d})$. Formally, this leads to the following result, which simply replaces $\mathfrak{s}$ with $\theta_P(\varepsilon;5\beta)$ in the lower-order term compared to Theorem 5. Due to (45), the query complexity bound in this result is *never larger* than that of Theorem 5 (and below we discuss scenarios where it is strictly smaller). We remark that, based on the comment preceding Lemma 20, the factor "5" in $\theta_P(\varepsilon;5\beta)$ in this theorem can be reduced to any value $c > 2$ by appropriately adjusting the constants C, C'' in the algorithm.

[17] The remarks concerning the choice of $h^\star$ in footnote 14 also apply here, noting that the lemmas concerning $h^\star$ in Appendix E actually apply simultaneously to all functions $h^\star \in \mathbb{C}$ satisfying (7).

Theorem 27 (Distribution-dependent Query Complexity of AVID Agnostic)**.** *For any concept class $\mathbb{C}$ with $\mathrm{VC}(\mathbb{C}) < \infty$, letting $\mathsf{d} = \mathrm{VC}(\mathbb{C})$, for every distribution P on $\mathcal{X} \times \{0,1\}$, letting $\beta = \inf_{h\in\mathbb{C}} \mathrm{er}_P(h)$, for any $\varepsilon, \delta \in (0,1)$, if the algorithm $\mathbb{A}_{\mathrm{avid}}$ is executed with parameters (ε, δ), with any number $m \geq M(\varepsilon,\delta;\beta)$ of i.i.d.-P examples (for $M(\varepsilon,\delta;\beta)$ as in Theorem 5, defined in Lemma 24), then with probability at least $1-\delta$, the returned predictor $\hat{h}$ satisfies $\mathrm{er}_P(\hat{h}) \leq \inf_{h\in\mathbb{C}} \mathrm{er}_P(h) + \varepsilon$ and the algorithm makes a number of queries at most $Q(\varepsilon,\delta;P)$ satisfying*

$$\begin{aligned} Q(\varepsilon,\delta;P) &= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right) + \min\left\{\theta_P(\varepsilon;5\beta)\log\left(\frac{1}{\varepsilon}\right), \frac{1}{\varepsilon}\right\}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)\right) \\ &= \tilde{O}\left(\mathsf{d}\frac{\beta^2}{\varepsilon^2} + \mathsf{d}\theta_P(\varepsilon;5\beta)\right). \end{aligned}$$

Proof. The result follows identically to Theorem 5, with only one minor change: replacing $\mathfrak{s}$ with $2C\theta_P(\varepsilon;5\beta)$ in Lemma 22. Note that this one change will suffice, since every subsequent appearance of $\mathfrak{s}$ in the proof is due to its appearance in Lemma 22, and hence changing $\mathfrak{s}$ to $2C\theta_P(\varepsilon;5\beta)$ in this lemma allows us to make the same change in every subsequent appearance of $\mathfrak{s}$ in the proof.

To see why Lemma 22 remains valid with this change, first note that its proof establishes that, on the event $E_0 \cap E_1 \cap E_2$, in the non-trivial case of $P_X(\mathcal{X} \setminus \Delta_{i_k}) \neq 0$, (35) holds. Rather than relaxing the third expression in (35) using the star number, we can instead relax it using $\theta_P(\varepsilon;5\beta)$: that is, for P_k as defined in that context, (35) implies

$$P_X(D_{k-1}\setminus\Delta_{i_k}) = P_k(D_{k-1})P_X(\mathcal{X}\setminus\Delta_{i_k}) \leq P_k\left(\mathrm{DIS}\left(\mathrm{B}_{P_k}\left(h^\star, \frac{\varepsilon_k}{P_X(\mathcal{X}\setminus\Delta_{i_k})}\right)\right)\right) P_X(\mathcal{X}\setminus\Delta_{i_k})$$
$$= P_{\Delta_{i_k}}\left(\mathrm{DIS}\left(\mathrm{B}_{P_{\Delta_{i_k}}}(h^\star,\varepsilon_k)\right)\right) \leq \theta_P(\varepsilon/(2C);5\beta)\varepsilon_k,$$

where the last inequality follows from Definition 26, the fact that $\varepsilon_k > \frac{\varepsilon}{2C}$, and the fact (from Lemma 20) that $P_X(\Delta_{i_k}) \leq 5\beta$. We then note that (as in Corollary 7.2 of Hanneke, 2014) for any $\Delta \subseteq \mathcal{X}$,

$$\theta_{P_\Delta,h^\star}(\varepsilon/(2C)) = \sup_{r>\varepsilon} \frac{P_\Delta(\mathrm{DIS}(\mathrm{B}_{P_\Delta}(h^\star, r/(2C))))}{r/(2C)} \vee 1 \leq 2C \sup_{r>\varepsilon} \frac{P_\Delta(\mathrm{DIS}(\mathrm{B}_{P_\Delta}(h^\star,r)))}{r} \vee 1,$$

and therefore $\theta_P(\varepsilon/(2C);5\beta) \leq 2C\theta_P(\varepsilon;5\beta)$. Altogether, we have that Lemma 22 remains valid while replacing $\mathfrak{s}$ with $2C\theta_P(\varepsilon;5\beta)$. ■

We emphasize that $\mathbb{A}_{\mathrm{avid}}$ *does not need to know* the value $\theta_P(\varepsilon;5\beta)$ (or anything else about P) to achieve this query complexity: that is, it is *adaptive* to the value of $\theta_P(\varepsilon;5\beta)$.

Together with (46), the above result further implies a (sometimes loose) relaxation, in which the lower-order term has a *quadratic* dependence on $\theta_P(\beta+\varepsilon)$, as formally stated in the following corollary (compare this with Appendix F.1, which showed it is impossible to generally reduce this $\theta_P(\beta+\varepsilon)^2$ term to a linear term $\theta_P(\beta+\varepsilon)$ or even $\theta_P(0)$).

Corollary 28. *The query complexity bound $Q(\varepsilon,\delta;P)$ in Theorem 27 (achieved by $\mathbb{A}_{\mathrm{avid}}$) satisfies*

$$\begin{aligned} Q(\varepsilon,\delta;P) &= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\min\left\{\mathsf{d}\theta_P(\beta+\varepsilon)\left(\frac{\beta+\varepsilon}{\varepsilon}\right), \frac{\mathsf{d}}{\varepsilon}\right\}\right) \\ &= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\min\left\{\mathsf{d}\theta_P(\beta+\varepsilon)^2, \frac{\mathsf{d}}{\varepsilon}\right\}\right). \end{aligned}$$

Proof. Due to the first two inequalities in (46), and $\theta_P(5\beta+\varepsilon) \leq \theta_P(\beta+\varepsilon)$, the second term in the expression of $Q(\varepsilon,\delta;P)$ in Theorem 27 is at most

$$O\left(\min\left\{\theta_P(\beta+\varepsilon)\log\left(\frac{1}{\varepsilon}\right)\left(\frac{\beta+\varepsilon}{\varepsilon}\right), \frac{1}{\varepsilon}\right\}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)\right). \tag{47}$$

Relaxing $\mathsf{d}\log\left(\frac{1}{\varepsilon}\right)+\log\left(\frac{1}{\delta}\right) \leq \log\left(\frac{1}{\varepsilon}\right)\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)$ and noting that

$$\theta_P(\beta+\varepsilon)\log^2\left(\frac{1}{\varepsilon}\right)\left(\frac{\beta+\varepsilon}{\varepsilon}\right) \leq \theta_P(\beta+\varepsilon)^2\log^4\left(\frac{1}{\varepsilon}\right) + \left(\frac{\beta+\varepsilon}{\varepsilon}\right)^2,$$

and $\left(\frac{\beta+\varepsilon}{\varepsilon}\right)^2 \leq 4\frac{\beta^2}{\varepsilon^2}+4$, the quantity (47) is at most

$$O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)+\min\left\{\theta_P(\beta+\varepsilon)^2\log^4\left(\frac{1}{\varepsilon}\right),\frac{1}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)\right\}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)\right).$$

Adding this to the first term in the expression of $Q(\varepsilon,\delta;P)$, the result follows. ■

In particular, Corollary 28 implies that, whenever $\theta_P(0)<\infty$, the dependence on β,ε in the query complexity bound in Theorem 27 is of order $\frac{\beta^2}{\varepsilon^2}+\mathrm{polylog}\left(\frac{1}{\varepsilon}\right)$. For instance, see (Hanneke, 2014, Chapter 7) for some general conditions on $(\mathbb{C},P)$ under which this occurs. We remark that the *first* bound in Corollary 28 is at least never larger than the upper bound in Theorem 1, since we always have $\theta_P(\beta+\varepsilon)\left(\frac{\beta+\varepsilon}{\varepsilon}\right)\leq\frac{1}{\varepsilon}$; however, we note that this is *not* the case for the *second* upper bound in Corollary 28. Beyond these basic observations, there exist scenarios $(\mathbb{C},P)$ where both upper bounds in Corollary 28 are *loose* compared to Theorem 27, to such an extent that they are sometimes even larger than the $\mathfrak{s}$-dependent bound in Theorem 5 (see Example 6 below). It is for this reason that we have chosen to express Theorem 27 in terms of the more-complicated quantity $\theta_P(\varepsilon;\tau)$, to provide a starting point for P-dependent analysis that is at least never worse than Theorem 5.

F.2.1 Examples

We next present some examples illustrating the values of the lower-order terms in Theorem 27 and Corollary 28 by bounding the quantities $\theta_P(\varepsilon;\tau)$ and $\theta_P(\beta+\varepsilon)^2$. Specifically, Example 1 achieves this via the relation to $\mathfrak{s}$, Example 2 provides a simple scenario with $\mathfrak{s}=\infty$ where it is possible to directly bound $\theta_P(\varepsilon;\tau)$, Example 3 expresses a bound on $\theta_P(\beta+\varepsilon)$ which is known in the literature, but when combined with Corollary 28 provides an improved P-dependent query complexity bound compared to previous works. Example 5 revisits the example from Appendix F.1 to illustrate that $\theta_P(\varepsilon;5\beta)$ provides a valid lower-order term for this example. In Appendix F.4, we will present additional examples of P-dependent query complexity bounds, for some classes with $\mathrm{VC}(\mathbb{C})=\infty$, via P_X-dependent *covering numbers*.

Example 1 (Thresholds)**.** Due to (45), any $\mathbb{C}$ with finite star number $\mathfrak{s}$ admits a bounded $\theta_P(\varepsilon;\tau)$. A simple example of this is *threshold* classifiers: namely, $\mathcal{X}=\mathbb{R}$ and $\mathbb{C}=\{\mathbb{1}_{[t,\infty)}:t\in\mathbb{R}\}$. This class has $\mathfrak{s}=2$ (Hanneke and Yang, 2015), and hence $\theta_P(\varepsilon;\tau)\leq 2$ for any P.

Example 2 (Linear classifiers under 1-sparse distributions)**.** To illustrate a simple example where $\theta_P(\varepsilon;\tau)=O(1)$ while $\mathfrak{s}=\infty$, consider the class $\mathbb{C}$ of linear classifiers in $\mathcal{X}=\mathbb{R}^p$, $p\geq 2$: that is, $\mathbb{C}=\{x\mapsto\mathbb{1}_{\langle w,x\rangle+b\geq 0}:w\in\mathbb{R}^p,b\in\mathbb{R}\}$. This class has $\mathfrak{s}=\infty$ (Hanneke and Yang, 2015). However, if we consider P_X as a distribution supported entirely on *one axis* (e.g., $\mathrm{Uniform}([0,1]\times\{0\}^{p-1})$), then it is a simple exercise to show that $\theta_P(\varepsilon;\tau)\leq 2$: the concepts in $\mathrm{B}_{P_\Delta}(h^\star,r)$ are those that disagree with $h^\star$ on at most r measure (under P_Δ) either to the left or right of where the $h^\star$ separator intersects the axis, so that $\mathrm{DIS}(\mathrm{B}_{P_\Delta}(h^\star,r))$ is simply the union of these two (at most) r-measure regions, hence has P_Δ measure at most $2r$.

While the above examples merely recover known results, the following example derives a previously-unknown P-dependent query complexity bound, which significantly improves over the best previously-known bound for this scenario.

Example 3 (Rectangles)**.** Consider the case $\mathcal{X}=\mathbb{R}^p$, $p\geq 1$, and $\mathbb{C}=\{\mathbb{1}_{[a_1,b_1]\times\cdots\times[a_p,b_p]}:a_1\leq b_1,\ldots,a_p\leq b_p\}$: the class of *axis-aligned rectangles* (Mitchell, 1979). This class is known to have $\mathfrak{s}=\infty$ (Hanneke and Yang, 2015). Consider $P_X=\mathrm{Uniform}([0,1]^p)$ (the example trivially extends to any product distribution P_X with marginals on each axis having continuous CDFs) and any P with well-defined $h^\star\in\mathrm{argmin}_{h\in\mathbb{C}}\mathrm{er}_P(h)$ satisfying $P_X(\{x:h^\star(x)=1\})=:\lambda>0$. The optimal first-order query complexity under these conditions is not yet precisely known. However, Wiener, Hanneke, and El-Yaniv (2015) have shown that $\theta_P(\beta+\varepsilon)=O\left(\frac{\mathsf{d}}{\lambda}\log(\mathsf{d})\wedge\frac{1}{\beta+\varepsilon}\right)$ for this scenario, and based on this, the best known query complexity upper bound is of the form $\tilde{O}\left(\min\left\{\frac{\mathsf{d}^2}{\lambda}\frac{(\beta+\varepsilon)^2}{\varepsilon^2},\mathsf{d}\frac{\beta+\varepsilon}{\varepsilon^2}\right\}\right)$. We can derive a bound which improves over this, as follows. We first recall that Theorem 1 provides a query complexity bound $\tilde{O}\left(\mathsf{d}\frac{\beta^2}{\varepsilon^2}+\frac{\mathsf{d}}{\varepsilon}\right)$, which *already* improves over the query complexity bound

of Wiener, Hanneke, and El-Yaniv (2015) in all regimes with $\varepsilon \ll \beta \ll 1$ (for every λ). However, we can further refine the lower-order term by introducing a dependence on λ. Specifically, the first bound in Corollary 28 provides a query complexity bound $\tilde{O}\left(\mathsf{d}\frac{\beta^2}{\varepsilon^2} + \frac{\mathsf{d}^2(\beta+\varepsilon)}{\lambda\varepsilon}\right)$, which offers a refinement over Theorem 1 whenever $\lambda \gg \mathsf{d}(\beta+\varepsilon)$. Moreover, the second bound in Corollary 28 provides a query complexity bound $\tilde{O}\left(\mathsf{d}\frac{\beta^2}{\varepsilon^2} + \frac{\mathsf{d}^3}{\lambda^2}\right)$. In particular, for $\lambda = \Theta(1)$ and $\beta = \tilde{O}(\sqrt{\varepsilon})$, this yields a query complexity bound $\mathrm{poly}(\mathsf{d})\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$, which was only available in the bound of Wiener, Hanneke, and El-Yaniv (2015) in the more-restrictive regime $\beta = \tilde{O}(\varepsilon)$. We leave open the question of identifying the *optimal* query complexity for this scenario. In particular, one concrete technical question toward that end would be to determine whether, for $\lambda > 2\tau$, $\theta_P(\varepsilon;\tau) = \tilde{O}\left(\frac{\mathsf{d}}{\lambda}\right)$.

Example 4 (Linear Classifiers)**.** Consider the commonly-studied concept class of *linear classifiers*, defined as: $\mathcal{X} = \mathbb{R}^{\mathsf{d}-1}$ ($\mathsf{d} \geq 3$) and $\mathbb{C} = \{h_{w,b} : w \in \mathbb{R}^{\mathsf{d}-1}, b \in \mathbb{R}\}$, where $h_{w,b}(x) = \mathbb{1}[\langle w, x\rangle + b \geq 0]$. This is perhaps the most well-studied concept class in the active learning literature. Its VC dimension satisfies $\mathrm{VC}(\mathbb{C}) = \mathsf{d}$ (Vapnik and Chervonenkis, 1974), and while its star number satisfies $\mathfrak{s} = \infty$ (Hanneke and Yang, 2015), the disagreement coefficient has been shown to be bounded or sublinear under various distributional conditions (Hanneke, 2007b, 2014; Balcan, Hanneke, and Vaughan, 2010; Friedman, 2009; Mahalanabis, 2011; Wiener, Hanneke, and El-Yaniv, 2015). These results compose directly with Corollary 28 to yield previously-unknown bounds on the query complexity under these same conditions. For instance, if P_X is a mixture of a finite number t of multivariate Gaussian distributions with full-rank diagonal covariance matrices, then Wiener, Hanneke, and El-Yaniv (2015) provide a bound $\theta_P(r) \leq c_{\mathsf{d},t}\log^{\mathsf{d}-2}\left(\frac{1}{r}\right)$ for a (d,t)-dependent constant $c_{\mathsf{d},t}$. Plugging into Corollary 28 (or rather, the explicit bounds in the proof thereof) yields a novel query complexity bound of order $\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right) + c_{\mathsf{d},t}^2\log^{2(\mathsf{d}-2)}\left(\frac{1}{\beta+\varepsilon}\right)\log^4\left(\frac{1}{\varepsilon}\right)\left(\mathsf{d}+\log\left(\frac{1}{\delta}\right)\right)$. More generally, if P_X admits a density with respect to the Lebesgue measure on $\mathbb{R}^{\mathsf{d}-1}$, then (taking $h^\star$ as in footnote 14) Hanneke (2014) argues that $\theta_P(r) = o\left(\frac{1}{r}\right)$ (where the specific form of this function $\theta_P(r)$ varies depending on P). Recalling that (as $\varepsilon \to 0$) the lower-order term becomes relevant only in the regime $\beta \ll \sqrt{\varepsilon}$, combining this with Corollary 28 yields a query complexity bound which often provides refinements over Theorem 1. In particular, under sufficient regularity conditions on P_X (see the proof of Hanneke, 2014) to ensure this $o\left(\frac{1}{r}\right)$ function further satisfies $\theta_P(r)\log^2\left(\frac{1}{r}\right) = o\left(\frac{1}{r}\right)$, the resulting asymptotic dependence on (ε,β) is of the form $\frac{\beta^2}{\varepsilon^2} + o\left(\frac{1}{\varepsilon}\right)$. Moreover, if additionally the density of P_X is bounded and has finite-diameter support, and if the hyperplane boundary corresponding to $h^\star$ passes through a continuity point of this density in its support, then Hanneke (2014) argues $\theta_P(r) = O(1)$, so that Corollary 28 yields a query complexity bound with asymptotic dependence on (ε,β) of the form $\frac{\beta^2}{\varepsilon^2} + \log^4\left(\frac{1}{\varepsilon}\right)$. Moreover, under the further restrictions (density bounded away from 0, compactness of the support), Friedman (2009); Mahalanabis (2011) argue $\theta_P(r)$ is asymptotically bounded by $O(\mathsf{d})$ (for the precise statement, see the original works, or discussion thereof by Hanneke, 2014).

Example 5 (Coupled thresholds and singletons)**.** Let us revisit the example from Appendix F.1, for which we argued that $\theta_P(\varepsilon)$ cannot itself be used to replace the star number in Theorem 3 (for *any* algorithm). We will here explain how the region-excluded disagreement coefficient $\theta_P(\varepsilon;5\beta)$ explicitly corrects for the issue with $\theta_P(\varepsilon)$ in this example. Specifically, consider again $\mathcal{X} = \mathbb{Z}$, and $\mathbb{C} = \mathbb{C}_{\mathrm{ts}} := \{\mathbb{1}_{\{-t\}\cup[t,\infty)} : t \in \mathbb{N}\}$, the class of *coupled* thresholds and singletons. Let $\varepsilon,\beta \in (0,1/3)$, let $n = \frac{1-2\beta}{2\varepsilon}$ (and assume $n \in \mathbb{N}$), and consider again the distributions P_t, $t \in \{1,\ldots,n\}$, as defined in Appendix F.1: that is, all P_t have marginal P_X on $\mathcal{X}$, where for $x \in \{1,\ldots,n\}$, $P_X(\{x\}) = \frac{2\beta}{n}$, $P_X(\{-x\}) = 2\varepsilon$, and for $x \in \{-1,\ldots,-n\}$, $P_t(Y = 1|X = x) = \mathbb{1}[x = -t]$, while every $x \notin \{-1,\ldots,-n\}$ has $P_t(Y = 1|X = x) = \frac{1}{2}$. Note that, for $\Delta = [0,\infty)$, we have $P_X(\Delta) = 2\beta$. Moreover, for $h^\star$ as defined under P_t, we have $\mathrm{B}_{P_\Delta}(h^\star,4\varepsilon) = \mathbb{C}_{\mathrm{ts}}$ (since only the disagreements on the *singleton* part are measured by P_Δ). This implies $\mathrm{DIS}(\mathrm{B}_{P_\Delta}(h^\star,4\varepsilon)) = \mathbb{Z}\setminus\{0\}$, so that $P_\Delta(\mathrm{DIS}(\mathrm{B}_{P_\Delta}(h^\star,4\varepsilon))) = 1-2\beta$. Therefore, $\theta_P(\varepsilon;5\beta) \geq \theta_P(\varepsilon;2\beta) \geq \frac{P_\Delta(\mathrm{DIS}(\mathrm{B}_{P_\Delta}(h^\star,4\varepsilon)))}{4\varepsilon} = \frac{1-2\beta}{4\varepsilon}$. Since we always have $\theta_P(\varepsilon;5\beta) \leq \frac{1}{\varepsilon}$, we conclude that $\theta_P(\varepsilon;5\beta) = \Theta\left(\frac{1}{\varepsilon}\right)$. As argued in Appendix F.1, the minimax optimal query complexity (constraining to $P \in \{P_t : t\{1,\ldots,n\}\}$) is $\Omega\left(\frac{1}{\varepsilon}\right)$, so that, unlike $\theta_P(\varepsilon)$, the quantity $\theta_P(\varepsilon;5\beta)$ is an appropriate replacement for $\mathfrak{s}$ in Theorem 3. Of course, Theorem 27 shows this replacement is *always* valid, so the point here is merely to illustrate how the exclusion of the Δ region in the definition of

$\theta_P(\varepsilon; 5\beta)$ is precisely the right type of correction, compared to $\theta_P(\varepsilon)$, for this example, as it explicitly removes the issue underlying the failure of $\theta_P(\varepsilon)$: namely, the fact that the *threshold* portion of the concepts $\mathbb{1}_{\{-t'\}\cup[t',\infty)}$ is irrelevant to the learning problem inherent in the P_t distributions. It is also worth noting that the *first* upper bound $\theta_P(\varepsilon; 5\beta) \leq \theta_P(5\beta+\varepsilon)\left(\frac{5\beta+\varepsilon}{\varepsilon}\right)$ from (46) also yields a value $\Theta\left(\frac{1}{\varepsilon}\right)$ (since this upper bound is *never* larger than $\frac{1}{\varepsilon}$). However, the *second* upper bound $\theta_P(5\beta+\varepsilon)^2 + \left(\frac{5\beta+\varepsilon}{\varepsilon}\right)^2$ can be significantly looser for this example, in most regimes of ε, β (namely, $\beta \neq \Theta(\sqrt{\varepsilon})$).

F.2.2 The Error Disagreement Coefficient

It is also possible to derive Corollary 28 via another intermediate P-dependent variant of the disagreement coefficient: namely, the *error disagreement coefficient*, defined as follows.

Definition 29. *For any probability distribution P on $\mathcal{X} \times \{0,1\}$, for any $\varepsilon \geq 0$, define*

$$\theta_P^{\mathrm{er}}(\varepsilon) := \sup_{r>\varepsilon} \frac{P_X(\mathrm{DIS}(\mathbb{C}_P(r)))}{r} \vee 1,$$

*where $\mathbb{C}_P(r) := \{h \in \mathbb{C} : \mathrm{er}_P(h) - \inf_{h' \in \mathbb{C}} \mathrm{er}_P(h') \leq r\}$ is known as the r-*minimal set.

Similarly to $\theta_P(\varepsilon; \tau)$, the quantity $\theta_P^{\mathrm{er}}(\varepsilon)$ has direct relations to the original disagreement coefficient from Definition 25. Specifically, for $h^\star$ as in (7),[18] since $\mathrm{B}_{P_X}(h^\star, r/2) \subseteq \mathbb{C}_P(r) \subseteq \mathrm{B}_{P_X}(h^\star, 2(\beta+r))$ for any $r > \varepsilon$, we immediately have

$$\frac{1}{2}\theta_P(\varepsilon/2) \leq \theta_P^{\mathrm{er}}(\varepsilon) \leq 2\theta_P(2(\beta+\varepsilon))\left(\frac{\beta+\varepsilon}{\varepsilon}\right) \leq \theta_P(2(\beta+\varepsilon))^2 + 4\left(\frac{\beta+\varepsilon}{\varepsilon}\right)^2. \tag{48}$$

By definition, we always have $\theta_P^{\mathrm{er}}(\varepsilon) \leq \frac{1}{\varepsilon}$. However, unlike $\theta_P(\varepsilon; \tau)$ in (45), the quantity $\theta_P^{\mathrm{er}}(\varepsilon)$ is *not* always upper-bounded by the star number $\mathfrak{s}$ (see Example 6 below), so that we need be careful when replacing $\mathfrak{s}$ by $\theta_P^{\mathrm{er}}(\varepsilon)$ in Theorem 3.

It is also worth noting that $\theta_P^{\mathrm{er}}(\varepsilon)$ is often not as easy to use for studying specific scenarios, compared to $\theta_P(\varepsilon)$, due to the dependence on the conditional distribution $Y|X$ (whereas $\theta_P(\varepsilon)$ depends only on P_X and $h^\star$). Nonetheless, below we will state a query complexity bound in terms of $\theta_P^{\mathrm{er}}(\varepsilon)$ (Theorem 30) which is sometimes smaller than that in Theorem 27 (as we illustrate in examples below), and moreover (together with (48)) provides another route to proving the query complexity bound in Corollary 28.

The quantity $\theta_P^{\mathrm{er}}(\varepsilon)$ essentially arises naturally in many existing analyses of disagreement-based active learning (e.g., Hanneke, 2009b, 2011, 2014; Koltchinskii, 2010; Foster, Rakhlin, Simchi-Levi, and Xu, 2021), wherein certain algorithms are shown to makes queries in a subset of $\mathrm{DIS}(\mathbb{C}_P(\varepsilon'))$ for an appropriate $\varepsilon' \geq \varepsilon$ (decreasing as the algorithm runs). In those contexts, it is traditional to upper bound $P_X(\mathrm{DIS}(\mathbb{C}_P(\varepsilon')))$ by $\theta_P(r(\varepsilon'))r(\varepsilon')$, where $r(\varepsilon') \geq \sup_{h \in \mathbb{C}_P(\varepsilon')} P_X(h \neq h^\star)$: for instance, $r(\varepsilon') = 2(\beta+\varepsilon')$ suffices in the agnostic setting. However, one can alternatively upper bound $P_X(\mathrm{DIS}(\mathbb{C}_P(\varepsilon')))$ by $\theta_P^{\mathrm{er}}(\varepsilon')\varepsilon'$. Such arguments are also valid in the context of $\mathbb{A}_{\mathrm{avid}}$, since Corollary 18 implies $P_X(D_{k-1}) \leq P_X(\mathrm{DIS}(\mathbb{C}_P(\varepsilon_k))) \leq \theta_P^{\mathrm{er}}(\varepsilon_k)\varepsilon_k$, so that the number of queries in $S_k^1 \cap D_{k-1}$ in round k is of order $\theta_P^{\mathrm{er}}(\varepsilon_k)\varepsilon_k m_k = \tilde{O}(\theta_P^{\mathrm{er}}(\varepsilon_k)\mathsf{d})$, which will lead to a lower-order term $\tilde{O}(\theta_P^{\mathrm{er}}(\varepsilon)\mathsf{d})$. Together with reasoning similar to the proof of Theorem 27, this implies the following.

Theorem 30. *Under the same conditions as Theorem 27, with probability at least $1-\delta$ the predictor $\hat{h}$ returned by $\mathbb{A}_{\mathrm{avid}}$ satisfies $\mathrm{er}_P(\hat{h}) \leq \inf_{h \in \mathbb{C}} \mathrm{er}_P(h) + \varepsilon$ and the algorithm makes a number of queries at most $Q(\varepsilon, \delta; P)$ satisfying*

$$\begin{aligned} Q(\varepsilon, \delta; P) &= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right) + \min\left\{\theta_P^{\mathrm{er}}(\varepsilon)\log\left(\frac{1}{\varepsilon}\right), \frac{1}{\varepsilon}\right\}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right) \\ &= \tilde{O}\left(\mathsf{d}\frac{\beta^2}{\varepsilon^2} + \mathsf{d}\theta_P^{\mathrm{er}}(\varepsilon)\right). \end{aligned}$$

[18] The relation sharpens to $\theta_P(\varepsilon) \leq \theta_P^{\mathrm{er}}(\varepsilon) \leq \theta_P(2\beta+\varepsilon)\left(\frac{2\beta+\varepsilon}{\varepsilon}\right)$ if we take $h^\star \in \mathrm{argmin}_{h \in \mathbb{C}} \mathrm{er}_P(h)$, supposing this exists (or otherwise, taking $h^\star$ as discussed in footnote 14).

Since $\theta^{\mathrm{er}}_P(\varepsilon) \le \frac{1}{\varepsilon}$, the query complexity bound in Theorem 30 is never larger than that in Theorem 1. However, unlike Theorem 27, since the quantity $\theta^{\mathrm{er}}_P(\varepsilon)$ is *not* always upper-bounded by the star number $\mathfrak{s}$, the query complexity bound in Theorem 30 is sometimes *larger* than that in Theorem 3 (see Example 6 below). That said, the quantities $\theta^{\mathrm{er}}_P(\varepsilon)$ and $\theta_P(\varepsilon; 5\beta)$ are generally incomparable (see Examples 6 and 7), so that either bound may be useful depending on the scenario being studied. Moreover, in light of (48), Theorem 30 is also useful for providing another route to establishing Corollary 28, which is therefore an immediate corollary of *either* Theorem 27 or Theorem 30.

As mentioned, depending on $(\mathbb{C}, P)$, the quantitative difference between $\theta^{\mathrm{er}}_P(\varepsilon)$ and $\theta_P(\varepsilon; 5\beta)$ can be better or worse. We illustrate this in the following two examples.

Example 6 ($\theta^{\mathrm{er}}_P(\varepsilon) \gg \mathfrak{s} \ge \theta_P(\varepsilon; 5\beta)$)**.** As mentioned, for some scenarios, $\theta^{\mathrm{er}}_P(\varepsilon)$ can be quite large, even larger than the *star number* $\mathfrak{s}$, so that the bound in Theorem 30 becomes even worse than the P-independent bound in Theorem 3 (in contrast to Theorem 27, which is never worse than Theorem 3). For instance, consider a *singletons* class: $\mathcal{X} = \{1, \dots, \frac{2}{\beta}\}$, $\mathbb{C} = \{\mathbb{1}_{\{t\}} : t \in \mathcal{X}\}$, where $\beta \in (0, 1/2)$ satisfies $\frac{2}{\beta} \in \mathbb{N}$ for simplicity. Let $P_X = \mathrm{Uniform}(\mathcal{X})$, $P(Y = 1|X = x) = \frac{\beta}{2(1-\beta)}$ for all $x \in \mathcal{X}$. Note that every $h \in \mathbb{C}$ has $\mathrm{er}_P(h) = \beta$. Moreover, $\mathfrak{s} = |\mathcal{X}| - 1 = \frac{2-\beta}{\beta}$. However, consider $0 < \varepsilon \ll \beta$. Since $\mathbb{C}_P(\varepsilon) = \mathbb{C}$ and $\mathrm{DIS}(\mathbb{C}) = \mathcal{X}$, we have $\theta^{\mathrm{er}}_P(\varepsilon) = \frac{1}{\varepsilon} = \Theta\left(\mathfrak{s}\frac{\beta}{\varepsilon}\right)$. For instance, for $\beta = \varepsilon^{2/3}$, the bound in Theorem 3 is (ignoring logs) of order $\frac{\beta^2}{\varepsilon^2} + \frac{1}{\beta} = \frac{2}{\varepsilon^{2/3}} \ll \frac{1}{\varepsilon} = \theta^{\mathrm{er}}_P(\varepsilon)$. In light of (48), this example also witnesses a scenario where both of the query complexity bounds in Corollary 28 are worse than the $\mathfrak{s}$-dependent bound in Theorem 3; more directly, in this example, we have $\theta_P(\beta + \varepsilon)\left(\frac{\beta+\varepsilon}{\varepsilon}\right) = \frac{1}{\varepsilon} \gg \mathfrak{s}$. In contrast, for any $r < \frac{\beta}{2}$ and $\Delta \subset \mathcal{X}$, we have $\mathrm{DIS}(\mathrm{B}_{P_\Delta}(h^\star, r)) \in \{\emptyset, \Delta\}$ (depending whether the x with $h^\star(x) = 1$ is in Δ or not), so that $P_\Delta(\mathrm{DIS}(\mathrm{B}_{P_\Delta}(h^\star, r))) = 0$; this immediately implies $\theta_P(\varepsilon; \tau) \le \frac{2}{\beta}$ (indeed, by careful reasoning, we can observe that any $\tau \ge \frac{\beta}{2}$ has $\theta_P(\varepsilon; \tau) = \frac{2}{\beta} - 1 = \mathfrak{s}$). More generally, by (45), the bound in Theorem 27 is never worse than that in Theorem 3.

On the other hand, there are scenarios $(\mathbb{C}, P)$ where the opposite occurs, so that in general neither quantity $\theta_P(\varepsilon; 5\beta)$ nor $\theta^{\mathrm{er}}_P(\varepsilon)$ dominates the other. This is illustrated in the following example.

Example 7 ($\theta_P(\varepsilon; 5\beta) \gg \theta^{\mathrm{er}}_P(\varepsilon)$)**.** Consider again the class from Appendix F.1 (and Example 5): that is, $\mathcal{X} = \mathbb{Z}$ and $\mathbb{C} = \mathbb{C}_{\mathrm{ts}} := \{\mathbb{1}_{\{-t\}\cup[t,\infty)} : t \in \mathbb{N}\}$. Let $\varepsilon, \beta \in (0, 1/3)$ with $\varepsilon \ll \beta$, and define P with marginal P_X on $\mathcal{X}$ as defined in Appendix F.1: that is, $n = \frac{1-2\beta}{2\varepsilon}$, and for $x \in \{1, \dots, n\}$, $P_X(\{x\}) = \frac{2\beta}{n}$, $P_X(\{-x\}) = 2\varepsilon$. However, rather than the distributions P_t described there, consider the family P'_t, $t \in \{1, \dots, n\}$, with $P'_t(Y = 1|X = x) = \beta + (1 - 2\beta)\mathbb{1}_{\{-t\}\cup[t,\infty)}(x)$ for every $x \in \mathcal{X}$. These distributions represent a scenario with *uniform classification noise*. For $P = P'_t$ for any $t \in \{1, \dots, n\}$, letting $h^\star = \mathbb{1}_{\{-t\}\cup[t,\infty)}$, it is easy to see that $\mathrm{er}_P(h^\star) = \inf_{h\in\mathbb{C}} \mathrm{er}_P(h) = \beta$. Moreover, $\theta^{\mathrm{er}}_P(\varepsilon) = \theta_P(\varepsilon/(1-2\beta)) \le \theta_P(0) \le \frac{1-2\beta}{\beta} + 3$, where the last inequality was established in Appendix F.1. In contrast, we argued in Example 5 that $\theta_P(\varepsilon; 5\beta) \ge \frac{1-2\beta}{4\varepsilon} \gg \frac{1-2\beta}{\beta} + 3 \ge \theta^{\mathrm{er}}_P(\varepsilon)$. Thus, in this scenario, $\theta_P(\varepsilon; 5\beta)$ is larger than $\theta^{\mathrm{er}}_P(\varepsilon)$.

A natural question is whether the gaps between $\theta_P(\varepsilon; 5\beta)$ and $\theta^{\mathrm{er}}_P(\varepsilon)$ in Examples 6 and 7 can also arise in cases where the smaller of the two corresponding query complexity bounds (either Theorem 27 or 30) is actually nearly-*sharp* (in a minimax analysis over a family of distributions). This is straightforward to obtain, by defining a family of possible distributions P each obtained as a uniform *mixture* of one of the above two scenarios and the simple 2-point construction of Kääriäinen (2006) giving rise to the $\frac{\beta^2}{\varepsilon^2}$ lower bound. For brevity, we omit the details of this.

The fact that Theorem 27 at least provides a starting point for P-dependent analysis which is never worse than the P-independent bound in Theorem 3 is a desirable feature. In contrast, the bound in Theorem 30 is sometimes better and sometimes worse than that in Theorem 3, so that one should be careful when using Theorem 30. Nonetheless, as illustrated in Example 7, there are at least some scenarios where $\theta^{\mathrm{er}}_P(\varepsilon)$ may be useful for describing favorable scenarios (particularly concerning the $Y|X$ conditional distribution).

F.3 Querying in Subregions of the Region of Disagreement

In the active learning literature, one technique for going beyond disagreement-based queries is to query examples in a carefully selected *subregion* $R \subseteq \mathrm{DIS}(V)$ of the region of disagreement of the set V of surviving concepts. This idea originates in the work of Balcan, Broder, and Zhang (2007) on *margin-based* active learning of homogeneous linear classifiers under certain marginals P_X in realizable and Tsybakov-noise scenarios, and was extended to a technique for general concept classes and the agnostic case by Zhang and Chaudhuri (2014) (see Appendix A for further discussion of the history). For instance, the most well-known case of this technique providing improvements over disagreement-based queries (see Example 8 below) is homogeneous linear classifiers under a uniform distribution on a sphere (alternatively, any isotropic log-concave distribution), where the query complexity of this technique is $\tilde{O}\left(\mathsf{d}\frac{(\beta+\varepsilon)^2}{\varepsilon^2}\right)$ (Zhang and Chaudhuri, 2014) (minimax optimal up to log factors), compared to disagreement-based active learning for which the best known bound is $\tilde{O}\left(\mathsf{d}^{3/2} \cdot \frac{(\beta+\varepsilon)^2}{\varepsilon^2}\right)$ (Dasgupta, Hsu, and Monteleoni, 2007).

In this section, we show this technique is also compatible with the AVID principle, and propose a refinement of the AVID Agnostic algorithm which replaces $D_{k-1} = \mathrm{DIS}(V_{k-1})$ in Steps 2 and 9 with a well-chosen *subregion* $R_{k-1} \subseteq D_{k-1}$. We argue that this change does not affect the validity of Theorem 5, and admits refined P-dependent query complexity bounds compared to those presented in Appendix F.2. In particular, this shows that the AVID principle can recover the optimal query complexity of homogeneous linear classifiers under the uniform distribution, and generally any isotropic log-concave distribution (indeed, with improved log factors compared to prior works).

The basic argument (building from the original ideas of Balcan, Broder, and Zhang, 2007, and Zhang and Chaudhuri, 2014) is that, rather than querying all examples in $S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$ in Step 2 of $\mathbb{A}_{\mathrm{avid}}$, the algorithm identifies a *subset* of these examples $Q_k \subseteq S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$ which suffices for the purpose of updating V_k in Step 3. Specifically, we aim to identify a subset $Q_k \subseteq S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}$ for which, for any $h, h' \in V_{k-1}$,

$$\left|\hat{P}_{S_k^1}(\{h \neq h'\} \cap Q_k) - \hat{P}_{S_k^1}(\{h \neq h'\} \cap D_{k-1} \setminus \Delta_{i_k})\right| \leq \frac{\varepsilon_k}{3C''}.$$

In other words, most of the significant disagreements in $\mathcal{X} \setminus \Delta_{i_k}$ among concepts in V_{k-1} are captured in the Q_k set. In particular, this retains the guarantees of Lemmas 15 and 16 with only minor adjustments to the constants in the bounds (accounting for the potential $\frac{\varepsilon_k}{8C'}$ probability disagreements that are lost).

Formally, consider the algorithm $\mathbb{A}_{\mathrm{avid}}^{\mathrm{sub}}$ stated in Figure 2, where the Q_k data subset is defined below. The values C, C', C'', N, m_k and data subsets S_k^1, S_k^4 are all as defined in $\mathbb{A}_{\mathrm{avid}}$. The data subsets S_k^2, $S_{k,i}^3$ are defined analogously to $\mathbb{A}_{\mathrm{avid}}$ except allocated in the corresponding steps of $\mathbb{A}_{\mathrm{avid}}^{\mathrm{sub}}$: that is, if and when the algorithm reaches Step 2 with a value k, or reaches Step 9 (in which case let $k = N+1$), then for the value i_k and the set Δ_{i_k} as defined at that time in the algorithm, letting $\hat{p}_k := 2\hat{P}_{S_k^4}(\Delta_{i_k})$, the algorithm allocates to S_k^2 the next $m_k' := \left\lceil \frac{C'' c_1^2 \hat{p}_k}{\varepsilon_k^2}\left(\mathsf{d} + \log\left(\frac{4(3+N-k)^2}{\delta}\right)\right)\right\rceil$ consecutive examples not previously allocated to any data subset, and likewise, if and when the algorithm reaches Step 5 with values (k, i), it allocates to $S_{k,i}^3$ the next m_k consecutive examples which have not yet been allocated to any data subset.

We define the Q_k data subset via a technique analogous to the work of Zhang and Chaudhuri (2014), specified via a discrete *linear program* with a finite number of constraints imposed by the set of realizable classifications of S_k^1. Let $t_k = \sum_{k'=1}^{k-1} m_{k'}$, and recall $S_k^1 := \{(X_{t_k+1}, Y_{t_k+1}), \ldots, (X_{t_k+m_k}, Y_{t_k+m_k})\}$. The algorithm inductively constructs sets $V_k \subseteq \mathbb{C}$ in Step 3 (analogous to the V_k sets in $\mathbb{A}_{\mathrm{avid}}$). For any given $k \in \{1, \ldots, N+1\}$, denote by

$$V_{k-1}(S_k^1) := \{(h(X_{t_k+1}), \ldots, h(X_{t_k+m_k})) : h \in V_{k-1}\} \subseteq \{0,1\}^{m_k},$$

the set of V_{k-1}-realizable classifications of S_k^1. For the set Δ_{i_k} (which is defined inductively based on previous rounds of the algorithm, analogously to the set Δ_{i_k} in $\mathbb{A}_{\mathrm{avid}}$), consider the following integer linear program with binary variables[19] $\zeta_{1,0}, \zeta_{1,1}, q_1, \ldots, \zeta_{m_k,0}, \zeta_{m_k,1}, q_{m_k}$.

[19]For simplicity, in this work, we present a technique based on an *integer* linear program, to arrive at a deterministic querying strategy. It is straightforward to extend the result to allow for non-integer solutions

Algorithm $\mathbb{A}^{\text{sub}}_{\text{avid}}$
Input: Error parameter ε, Confidence parameter δ, Unlabeled data $X_1, \ldots, X_m$
Output: Classifier $\hat{h}$
0. Initialize $i = i_1 = 0$, $\Delta_0 = \emptyset$, $V_0 = \mathbb{C}$
1. For $k = 1, \ldots, N$
2. Query all examples in Q_k and $S^2_k \cap \Delta_{i_k}$
3. $V_k \leftarrow \left\{h \in V_{k-1} : \hat{\text{er}}^{1,2}_k(h) \leq \hat{\text{er}}^{1,2}_k(\hat{h}_k) + \frac{\varepsilon_k}{C'}\right\}$
4. If $V_k = \emptyset$ or $\hat{\text{er}}^{1,2}_k(\hat{h}_k) < \min_{h \in V_k} \hat{\text{er}}^{1,2}_k(h) - \frac{\varepsilon_k}{4C'}$, Then Return $\hat{h} := \hat{h}_k$
5. While $\max_{f,g \in V_k} \hat{P}_{S^3_{k,i}}(\{f \neq g\} \setminus \Delta_i) > \varepsilon_{k+2}$
6. $(f, g) \leftarrow \text{argmax}_{(f',g') \in V_k^2} \hat{P}_{S^3_{k,i}}(\{f' \neq g'\} \setminus \Delta_i)$
7. $\Delta_{i+1} \leftarrow \Delta_i \cup \{f \neq g\}$, and update $i \leftarrow i + 1$
8. $i_{k+1} \leftarrow i$
9. Query all examples in Q_{N+1} and $S^2_{N+1} \cap \Delta_{i_{N+1}}$ and Return $\hat{h} := \hat{h}_{N+1}$

Figure 2: The Subregion-AVID Agnostic algorithm.

LP_k:

$$\begin{aligned}
&\text{minimize} && \sum_{t=1}^{m_k} q_t \\
&\text{subject to} && \forall (y_1, \ldots, y_{m_k}) \in V_{k-1}(S^1_k),\ \frac{1}{m_k} \sum_{t=1}^{m_k} \zeta_{t,1-y_t} \mathbb{1}[X_{t_k+t} \notin \Delta_{i_k}] \leq \frac{\varepsilon_k}{6C''} \\
& && \forall t \in \{1, \ldots, m_k\},\ \zeta_{t,0} + \zeta_{t,1} + q_t = 1 \\
& && \zeta_{1,0}, \zeta_{1,1}, q_1, \ldots, \zeta_{m_k,0}, \zeta_{m_k,1}, q_{m_k} \in \{0, 1\}
\end{aligned}$$

In particular, note that the solution only depends on the *unlabeled* examples $X_{t_k+1}, \ldots, X_{t_k+m_k}$, and thus the algorithm may use the solution of this optimization problem when determining an appropriate set Q_k of queries in Steps 2 and 9. Denote by $q^k_1, \ldots, q^k_{m_k}$ the respective values of the $q_1, \ldots, q_{m_k}$ variables at the solution found by LP_k. Then define Q_k as a subsequence of S^1_k:

$$Q_k := \{(X_{t_k+t}, Y_{t_k+t}) : 1 \leq t \leq m_k, q^k_t = 1\}.$$

Let us generalize the definition of $\hat{P}_{S^1_k}$ to involve intersections with the subsequence Q_k: for any set $A \subseteq \mathcal{X} \times \{0, 1\}$, $\hat{P}_{S^1_k}(A \cap Q_k) := \frac{1}{m_k} \sum_{t=1}^{m_k} q^k_t \cdot \mathbb{1}[(X_{t_k+t}, Y_{t_k+t}) \in A]$, and $\hat{P}_{S^1_k}(\text{ER}(f) \setminus Q_k) := \frac{1}{m_k} \sum_{t=1}^{m_k} (1 - q^k_t) \cdot \mathbb{1}[(X_{t_k+t}, Y_{t_k+t}) \in A]$. As usual, we also overload this notation for $A \subseteq \mathcal{X}$, such as sets $\{f \neq g\}$, interpreting such sets A as synonymous with their labeled extension $A \times \{0, 1\}$.

The algorithm also relies on the following modifications to the definition of $\hat{\text{er}}^{1,2}_k$:

$$\forall h, \hat{\text{er}}^{1,2}_k(h) := \hat{P}_{S^1_k}(\text{ER}(h) \cap Q_k) + \hat{P}_{S^2_k}(\text{ER}(h) \cap \Delta_{i_k}). \tag{49}$$

The definitions of $V^{(4)}_{k-1}$ and $\hat{h}_k$ are then defined as in (2) and (3) based on the set V_{k-1} defined in $\mathbb{A}^{\text{sub}}_{\text{avid}}$ and the modified definition of $\hat{\text{er}}^{1,2}_k$ in (49). This completes the specification of the $\mathbb{A}^{\text{sub}}_{\text{avid}}$ algorithm.

We state a query complexity guarantee for this algorithm, phrased in terms of a variant of a *subregion disagreement coefficient*. As in Appendix F.2, we first present the known definition from the literature, which serves both as a starting point for the modified version and as a more-accessible quantity useful for upper-bounding the new quantity. Specifically, the following definition (a refinement of the disagreement coefficient from Definition 25) was proposed by Zhang and Chaudhuri (2014) (see also Hanneke, 2016b).[20]

$\zeta_{t,0}, \zeta_{t,1} \in [0, 1]$ to the LP, resulting in a *randomized* querying strategy (see Zhang and Chaudhuri, 2014). This makes no significant difference to the query complexity bound (see Hanneke, 2016b, for a related discussion), but may be more attractive from a computational perspective.

[20]The variant stated here is phrased slightly differently, to simplify the definition. In particular, $\varphi_P(\varepsilon, 0)$ is equivalent to a quantity $\varphi^{01}_c(\varepsilon)$ studied by Hanneke (2016b), which is only slightly different than the original

Definition 31. *For any measure μ on $\mathcal{X}$, any $V \subseteq \mathbb{C}$, and any $\eta \geq 0$, define*

$$\Phi_\mu(V, \eta) := \inf\left\{ \mu(R) : \sup_{g \in V} \mu(\{g \neq f\} \setminus R) \leq \eta, \text{ measurable } R \subseteq \mathcal{X} \text{ and } f : \mathcal{X} \to \{0,1\} \right\}.$$

For any distribution P_X on $\mathcal{X}$ and any measurable $h : \mathcal{X} \to \{0,1\}$, for any $\varepsilon, \alpha \geq 0$, define

$$\varphi_{P_X,h}(\varepsilon, \alpha) := \sup_{r > \alpha + \varepsilon} \frac{\Phi_{P_X}(\mathrm{B}_{P_X}(h, r), (r - \alpha)/(36CC''))}{r} \vee 1.$$

In particular, for any distribution P on $\mathcal{X} \times \{0,1\}$, letting $h^\star$ be as in (7) (or see footnote 14), define $\varphi_P(\varepsilon, \alpha) := \varphi_{P_X,h^\star}(\varepsilon, \alpha)$.

The quantity $\Phi_{P_X}(V, \eta)$ identifies the smallest $P_X(R)$ among regions $R \subseteq \mathcal{X}$ for which functions $g \in V$ do not disagree much outside the region R (i.e., they have at most η disagreement with a fixed function f on $\mathcal{X} \setminus R$). In particular, we can upper bound $\Phi_{P_X}(V, \eta)$ by taking $R = \mathrm{DIS}(V)$ and any $f \in V$, which satisfies $\sup_{g \in V} P_X(\{g \neq f\} \setminus R) = 0 \leq \eta$, so that $\Phi_{P_X}(V, \eta) \leq P_X(\mathrm{DIS}(V))$. It immediately follows that the quantity $\varphi_{P_X,h}(\varepsilon, \alpha)$ is never larger than the disagreement coefficient:

$$\varphi_{P_X,h}(\varepsilon, \alpha) \leq \theta_{P_X,h}(\alpha + \varepsilon). \tag{50}$$

Indeed, there are several known examples of scenarios $(\mathbb{C}, P_X, h^\star)$ where $\varphi_P(\varepsilon, \alpha)$ is substantially smaller than $\theta_P(\alpha + \varepsilon)$ (Zhang and Chaudhuri, 2014). One example (discussed formally in Example 8 below) is the class $\mathbb{C}$ of *homogeneous linear classifiers* in $\mathbb{R}^{\mathsf{d}}$ under P_X a uniform distribution on an origin-centered sphere, where $\theta_P(0) = \Theta\left(\sqrt{\mathsf{d}}\right)$ and $\varphi_P(\varepsilon, \alpha) = O\left(\log\left(\frac{\alpha+\varepsilon}{\varepsilon}\right)\right)$ (Hanneke, 2007b; Balcan, Broder, and Zhang, 2007; Zhang and Chaudhuri, 2014).

By (50) and (44), we also always have $\varphi_P(\varepsilon, \alpha) \leq \mathfrak{s} \wedge \frac{1}{\alpha+\varepsilon}$ for any $\varepsilon, \alpha \geq 0$ with $\alpha + \varepsilon \leq 1$. Indeed, as with $\theta_P(\varepsilon)$ in (44), this inequality turns out to be sharp in the worst case. Specifically, Hanneke (2016b) has shown that $\sup_{P_X} \sup_{h \in \mathbb{C}} \varphi_{P_X,h}(\varepsilon, 0) = \mathfrak{s} \wedge \frac{1}{\varepsilon}$ for $\varepsilon \in (0, 1]$. Additionally, by definition we have $\varphi_{P_X,h}(\varepsilon, \alpha) \geq \varphi_{P_X,h}(\alpha + \varepsilon, 0)$. Since (50) implies $\varphi_{P_X,h}(\varepsilon, \alpha) \leq \theta_{P_X,h}(\alpha + \varepsilon)$, it immediately follows from combining this result of Hanneke (2016b) with (44) that for any $\varepsilon, \alpha \geq 0$ with $\alpha + \varepsilon \leq 1$,

$$\sup_{P_X} \sup_{h \in \mathbb{C}} \varphi_{P_X,h}(\varepsilon, \alpha) = \mathfrak{s} \wedge \frac{1}{\alpha + \varepsilon}. \tag{51}$$

Due to (50) and the example in Appendix F.1, we know it is not possible to replace $\mathfrak{s}$ with $\varphi_P(\varepsilon, \alpha)$ in Theorem 3 for any $\alpha \geq 0$ (for any algorithm). However, similarly to the modification $\theta_P(\varepsilon; \tau)$ of $\theta_P(\varepsilon)$ presented in Appendix F.2, we can modify Definition 31 appropriately to provide a quantity suitable for developing a query complexity bound for $\mathbb{A}^{\mathrm{sub}}_{\mathrm{avid}}$. Specifically, as in Definition 26, let us first generalize the definition of $\varphi_{P_X,h}(\varepsilon, \alpha)$ to general *measures* μ: that is, for any measure μ on $\mathcal{X}$ and measurable function $h : \mathcal{X} \to \{0,1\}$, for any $\varepsilon, \alpha \geq 0$, define $\varphi_{\mu,h}(\varepsilon, \alpha) := \sup_{r > \alpha + \varepsilon} \frac{\Phi_\mu(\mathrm{B}_\mu(h,r), (r-\alpha)/(36CC''))}{r} \vee 1$. Then consider the following definition, representing a *region-excluded subregion disagreement coefficient*.

Definition 32. *For any distribution P on $\mathcal{X} \times \{0,1\}$ and any measurable $\Delta \subseteq \mathcal{X}$, define a measure $A \mapsto P_\Delta(A) := P_X(A \setminus \Delta)$. For any $\varepsilon > 0$ and $\alpha, \tau \geq 0$, for $h^\star \in \mathbb{C}$ as in (7) (or see footnotes 14, 17), define*

$$\varphi_P(\varepsilon, \alpha; \tau) := \sup_{\Delta \subseteq \mathcal{X} : P_X(\Delta) \leq \tau} \varphi_{P_\Delta,h^\star}(\varepsilon, \alpha).$$

As was the case for $\theta_P(\varepsilon; \tau)$, we can equivalently define $\varphi_P(\varepsilon, \alpha; \tau)$ as the subregion disagreement coefficient under a worst-case *conditional* distribution $P_X(\cdot | \mathcal{X} \setminus \Delta)$: that is,

$$\varphi_P(\varepsilon, \alpha; \tau) = \sup_{\Delta \subseteq \mathcal{X} : P_X(\Delta) \leq \tau} \varphi_{P_X(\cdot|\mathcal{X}\setminus\Delta),h^\star}(\varepsilon / P_X(\mathcal{X} \setminus \Delta), \alpha / P_X(\mathcal{X} \setminus \Delta)), \tag{52}$$

where we define $\varphi_{P_X(\cdot|\mathcal{X}\setminus\Delta),h^\star}(\varepsilon / P_X(\mathcal{X} \setminus \Delta), \alpha / P_X(\mathcal{X} \setminus \Delta)) = 1$ in the case $P_X(\mathcal{X} \setminus \Delta) = 0$ (which coincides with the value $\varphi_{P_\Delta,h^\star}(\varepsilon, \alpha)$ for such Δ). In particular, combining this equivalent definition with (51) yields that

$$\varphi_P(\varepsilon, \alpha; \tau) \leq \mathfrak{s} \wedge \frac{1}{\alpha + \varepsilon} \vee 1. \tag{53}$$

quantity studied by Zhang and Chaudhuri (2014) in that it considers *binary* functions rather than fractional values in $[0, 1]$. Hanneke (2016b) has shown this change to binary values makes little quantitative difference compared to the quantity of Zhang and Chaudhuri (2014).

Thus, replacing $\mathfrak{s}$ in Theorem 3 by $\varphi_P(\varepsilon, \alpha; \tau)$ would yield a (never-larger) P-dependent refinement.

As with $\theta_P(\varepsilon; \tau)$, the quantity $\varphi_P(\varepsilon, \alpha; \tau)$ itself may often be challenging to calculate. Fortunately, again as with $\theta_P(\varepsilon; \tau)$, it can be upper-bounded by expressions that are more-easily calculated (though at the expense of some slack, so that they are no longer upper-bounded by $\mathfrak{s}$). We might therefore think of $\varphi_P(\varepsilon, \alpha; \tau)$ as an intermediate complexity measure (analogous to $\theta_P(\varepsilon; \tau)$), which is useful in providing a starting point for a P-dependent refinement of $\mathfrak{s}$ which is never larger than $\mathfrak{s}$, and which admits general upper bounds which are more accessible than directly calculating $\varphi_P(\varepsilon, \alpha; \tau)$. Specifically, it follows immediately from Definition 32 that we always have a lower bound $\varphi_P(\varepsilon, \alpha) \le \varphi_P(\varepsilon, \alpha; \tau)$, and an upper bound

$$\begin{aligned}\varphi_P(\varepsilon, \alpha; \tau) &\le \sup_{r > \alpha+\varepsilon} \frac{\Phi_{P_X}(\mathrm{B}_{P_X}(h^\star, \tau + r), (r-\alpha)/(36CC''))}{r} \vee 1 \\ &\le \varphi_P(\varepsilon, \alpha+\tau)\left(\frac{\alpha+\tau+\varepsilon}{\alpha+\varepsilon}\right) \le \varphi_P(\varepsilon, \alpha+\tau)^2 + \left(\frac{\alpha+\tau+\varepsilon}{\alpha+\varepsilon}\right)^2. \end{aligned} \tag{54}$$

Making use of the quantity $\varphi_P(\varepsilon, \alpha; \tau)$ to analyze $\mathbb{A}^{\mathrm{sub}}_{\mathrm{avid}}$ analogously to the analysis of $\mathbb{A}_{\mathrm{avid}}$ based on $\theta_P(\varepsilon; \tau)$ in Theorem 27, we arrive at the following theorem.

Theorem 33 (Distribution-dependent Query Complexity of Subregion AVID Agnostic)**.** *For any concept class $\mathbb{C}$ with $\mathrm{VC}(\mathbb{C}) < \infty$, letting $\mathsf{d} = \mathrm{VC}(\mathbb{C})$, for every distribution P on $\mathcal{X} \times \{0,1\}$, letting $\beta = \inf_{h \in \mathbb{C}} \mathrm{er}_P(h)$, for any $\varepsilon, \delta \in (0,1)$, if the algorithm $\mathbb{A}^{\mathrm{sub}}_{\mathrm{avid}}$ is executed with parameters (ε, δ), with any number $m \ge M(\varepsilon, \delta; \beta)$ of i.i.d.-P examples (for $M(\varepsilon, \delta; \beta)$ as in Theorem 5, defined in Lemma 24), then with probability at least $1-\delta$, the returned predictor $\hat{h}$ satisfies $\mathrm{er}_P(\hat{h}) \le \inf_{h \in \mathbb{C}} \mathrm{er}_P(h) + \varepsilon$ and the algorithm makes a number of queries at most $Q(\varepsilon, \delta; P)$ satisfying*

$$\begin{aligned} Q(\varepsilon, \delta; P) &= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right) + \min\left\{\varphi_P(\varepsilon, 0; 5\beta)\log\left(\frac{1}{\varepsilon}\right), \frac{1}{\varepsilon}\right\}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right) \\ &= \tilde{O}\left(\mathsf{d}\frac{\beta^2}{\varepsilon^2} + \mathsf{d}\varphi_P(\varepsilon, 0; 5\beta)\right). \end{aligned}$$

Proof Sketch. The proof of this theorem follows nearly identically to the proof of Theorem 5. We will merely highlight the changes compared to the original proof. Specifically, throughout the proof, we first replace all definitions from $\mathbb{A}_{\mathrm{avid}}$ with the corresponding definitions from $\mathbb{A}^{\mathrm{sub}}_{\mathrm{avid}}$ (e.g., Δ_i, i_k, $\hat{h}_k$, V_k, $D_{k-1} = \mathrm{DIS}(V_{k-1})$, S^2_k, $S^3_{k,i}$, K, $\mathcal{K}$, etc.) so that all definitions in the proof refer to the respective quantities in the $\mathbb{A}^{\mathrm{sub}}_{\mathrm{avid}}$ algorithm. Since the only definitional change in $\mathbb{A}^{\mathrm{sub}}_{\mathrm{avid}}$ compared to $\mathbb{A}_{\mathrm{avid}}$ is in the use of Q_k rather than $S^1_k \cap D_{k-1} \setminus \Delta_{i_k}$, to provide the ε error guarantee it will suffice to argue that the inequality (11) of Lemma 10 remains valid (only slightly larger) with this change: namely, on the event E_0,

$$\begin{aligned} \forall h, h' \in V^{(3)}_{k-1}, \Big| &\left(\hat{P}_{S^1_k}(\mathrm{ER}(h) \cap Q_k) - \hat{P}_{S^1_k}(\mathrm{ER}(h') \cap Q_k)\right) \\ &\qquad - \left(P(\mathrm{ER}(h) \setminus \Delta_{i_k}) - P(\mathrm{ER}(h') \setminus \Delta_{i_k})\right)\Big| \\ &< \sqrt{P_X(\{h \ne h'\} \setminus \Delta_{i_k})\frac{\varepsilon_k}{C''}} + \frac{2\varepsilon_k}{C''}. \end{aligned} \tag{55}$$

Note that this is only larger than the bound in (11) by an additive $\frac{\varepsilon_k}{C''}$ (which we will argue below is inconsequential to the proof). As was true of (11), we argue that (55) in fact follows immediately from (15), as follows. Consider the values $\zeta^k_{1,0}, \zeta^k_{1,1}, q^k_1, \ldots, \zeta^k_{m_k,0}, \zeta^k_{m_k,1}, q^k_{m_k}$ at the solution of the LP_k optimization. Due to the first constraint in LP_k, we know every $f \in V_{k-1}$ has

$$\frac{1}{m_k}\sum_{t=1}^{m_k} \zeta^k_{t,1-f(X_{t_k+t})} \mathbb{1}[X_{t_k+t} \notin \Delta_{i_k}] \le \frac{\varepsilon_k}{6C''}.$$

Moreover, due to the second constraint in LP_k, for every $f, g \in V_{k-1}$, any $X_{t_k+t} \in \{f \ne g\}$ has $\zeta^k_{t,1-f(X_{t_k+t})} + \zeta^k_{t,1-g(X_{t_k+t})} + q^k_t = 1$, so that $q^k_t = 0 \implies \zeta^k_{t,1-f(X_{t_k+t})} + \zeta^k_{t,1-g(X_{t_k+t})} = 1$. Together, we have

$$\hat{P}_{S^1_k}((\{f \ne g\} \setminus \Delta_{i_k}) \setminus Q_k) \le \frac{1}{m_k}\sum_{t=1}^{m_k}\left(\zeta^k_{t,1-f(X_{t_k+t})} + \zeta^k_{t,1-g(X_{t_k+t})}\right)\mathbb{1}[X_{t_k+t} \notin \Delta_{i_k}] \le \frac{\varepsilon_k}{3C''}.$$

Recall that every $h \in V_{k-1}^{(3)}$ is of the form $h = \mathrm{DL}(f', g', h') := f' \mathbb{1}_{\{f'=g'\}} + h' \mathbb{1}_{\{f' \neq g'\}}$ for some $f', g' \in V_{k-1}$ and $h' \in \mathbb{C}$. Consider any two such functions $h, h' \in V_{k-1}^{(3)}$, where $h = \mathrm{DL}(f_1, g_1, h_1)$ and $h' = \mathrm{DL}(f_2, g_2, h_2)$ for $f_1, g_1, f_2, g_2 \in V_{k-1}$ and $h_1, h_2 \in \mathbb{C}$. Note that

$$\{h \neq h'\} \subseteq \{f_1 \neq f_2\} \cup \{g_1 \neq g_2\} \cup \{f_1 \neq g_1\}.$$

Therefore, the union bound implies

$$\begin{aligned}
&\hat{P}_{S_k^1}((\{h \neq h'\} \setminus \Delta_{i_k}) \setminus Q_k) \\
&\leq \hat{P}_{S_k^1}((\{f_1 \neq f_2\} \setminus \Delta_{i_k}) \setminus Q_k) + \hat{P}_{S_k^1}((\{g_1 \neq g_2\} \setminus \Delta_{i_k}) \setminus Q_k) + \hat{P}_{S_k^1}((\{f_1 \neq g_1\} \setminus \Delta_{i_k}) \setminus Q_k) \\
&\leq \frac{\varepsilon_k}{C''}.
\end{aligned}$$

Also note that, due to the indicator $\mathbb{1}[X_{t_k+t} \notin \Delta_{i_k}]$ in the first constraint of LP_k, at the solution to LP_k, every $X_{t_k+t} \notin \Delta_{i_k}$ has $q_t^k = 0$, so that any $h \in V_{k-1}^{(3)}$ has $\mathrm{ER}(h) \cap Q_k = (\mathrm{ER}(h) \setminus \Delta_{i_k}) \cap Q_k$. Altogether, we have that every $h, h' \in V_{k-1}^{(3)}$ satisfy

$$\begin{aligned}
&\left|\left(\hat{P}_{S_k^1}(\mathrm{ER}(h) \cap Q_k) - \hat{P}_{S_k^1}(\mathrm{ER}(h') \cap Q_k)\right) - \left(\hat{P}_{S_k^1}(\mathrm{ER}(h) \setminus \Delta_{i_k}) - \hat{P}_{S_k^1}(\mathrm{ER}(h') \setminus \Delta_{i_k})\right)\right| \\
&= \left|\hat{P}_{S_k^1}((\mathrm{ER}(h') \setminus \Delta_{i_k}) \setminus Q_k) - \hat{P}_{S_k^1}((\mathrm{ER}(h) \setminus \Delta_{i_k}) \setminus Q_k)\right| \\
&\leq \hat{P}_{S_k^1}((\{h \neq h'\} \setminus \Delta_{i_k}) \setminus Q_k) \leq \frac{\varepsilon_k}{C''}. \qquad (56)
\end{aligned}$$

Together with (15) we arrive at the claimed inequality (55).

In the context of the rest of the proof of Theorem 5, the only place (11) is used is in (25) in the proof of Lemma 15. In that context, substituting (55) yields the same conclusion: namely, for $h_1, h_1' \in V_{k-1}^{(3)}$, since (17) of Lemma 12 implies $P_X(\{h_1 \neq h_1'\} \setminus \Delta_{i_k}) \leq 3\varepsilon_k$, (55) implies

$$\begin{aligned}
&\left|\hat{P}_{S_k^1}(\mathrm{ER}(h_1) \cap Q_k) - \hat{P}_{S_k^1}(\mathrm{ER}(h_1') \cap Q_k) - (P(\mathrm{ER}(h_1) \setminus \Delta_{i_k}) - P(\mathrm{ER}(h_1') \setminus \Delta_{i_k}))\right| \\
&\leq \sqrt{\frac{3\varepsilon_k^2}{C''}} + \frac{2\varepsilon_k}{C''} \leq \frac{2\varepsilon_k}{\sqrt{C''}},
\end{aligned}$$

where the last inequality follows from $C'' \geq 100$. Therefore, the conclusion of Lemma 15 remains valid (with the modified definition of $\hat{\mathrm{er}}_k^{1,2}$ from (49)). The rest of the proof of the error bound (Lemma 19), and unlabeled sample size $M(\varepsilon, \delta; \beta)$ (Lemma 24), and size of $P_X(\Delta_{i_k})$ (Lemma 20) follow verbatim from this fact.

It remains only to establish the claimed bound $Q(\varepsilon, \delta; P)$ on the number of queries. In the context of the proof of Theorem 5, this effectively means replacing (36) of Lemma 23 with a bound on $|Q_k|$ based on $\varphi_P(\varepsilon, 0; 5\beta)$, on an event E_3' of probability at least $1 - \frac{\delta}{8}$ (which replaces the event E_3' defined in the proof of Lemma 23).

Toward this end, consider any $k \in \{1, \ldots, N+1\}$ having non-zero probability of $k \leq K$. Given the event that $k \leq K$ and the random variables V_{k-1} and Δ_{i_k}, fix a measurable function $h_k : \mathcal{X} \to \{0, 1\}$ and a measurable set $R_k \subseteq \mathcal{X}$ (dependent on V_{k-1} and Δ_{i_k} but *not* on S_k^1) such that

$$\sup_{g \in V_{k-1}} P_X((\{g \neq h_k\} \setminus R_k) \setminus \Delta_{i_k}) \leq \frac{\varepsilon_k}{18C''} \qquad (57)$$

$$\text{and} \quad P_X(R_k \setminus \Delta_{i_k}) \leq \Phi_{P_{\Delta_{i_k}}}\left(V_{k-1}, \frac{\varepsilon_k}{18C''}\right) + \frac{\varepsilon_k}{2}. \qquad (58)$$

Such a pair (h_k, R_k) is guaranteed to exist by the definition of $\Phi_\mu(\cdot, \cdot)$ in Definition 31.

We aim to argue that the constraints in LP_k are satisfied by taking $\zeta_{t,h_k(X_{t_k+t})} = \mathbb{1}[X_{t_k+t} \notin R_k]$ and $q_t = \mathbb{1}[X_{t_k+t} \in R_k]$, via a uniform multiplicative Chernoff bound (Lemma 7 of Appendix D). Toward this end, define a collection $\tilde{\mathcal{A}}_k$ of subsets of $\mathcal{X}$:

$$\tilde{\mathcal{A}}_k := \{(\{g \neq h_k\} \setminus R_k) \setminus \Delta_{i_k} : g \in V_{k-1}\}.$$

Note that $\mathrm{VC}(\tilde{\mathcal{A}}_k) \leq \mathsf{d}$. Let $\tilde{\delta}_k := \frac{\delta \varepsilon_{k+3}}{144}$. We bound $\varepsilon(m_k, \tilde{\delta}_k; \tilde{\mathcal{A}}_k)$ by reasoning similar to the proof of Lemma 9. Specifically, we have

$$m_k \geq \frac{150C''c_0}{\varepsilon_k}\left(\mathsf{d}\log\left(\frac{C''c_0}{\varepsilon_k}\right) + \log\left(\frac{C''c_0}{\delta\varepsilon_k}\right)\right) > \frac{108C''c_0}{\varepsilon_k}\left(\mathsf{d}\log\left(\frac{54C''c_0}{\varepsilon_k}\right) + \log\left(\frac{1}{\tilde{\delta}_k}\right)\right),$$

where the last inequality is by $c_0 \geq 1$, $C'' > 144C^3$, and $(C'')^{150/108} > 54C''$. By Corollary 4.1 of Vidyasagar (2003), this implies

$$m_k > \frac{54C''c_0}{\varepsilon_k}\left(\mathsf{d}\log\left(\frac{m_k}{\mathsf{d}}\right) + \log\left(\frac{1}{\tilde{\delta}_k}\right)\right),$$

so that

$$\varepsilon\left(m_k, \tilde{\delta}_k; \tilde{\mathcal{A}}_k\right) < \frac{\varepsilon_k}{54C''}. \tag{59}$$

Letting $\alpha = \frac{2}{3}$, we therefore have $\varepsilon\left(m_k, \tilde{\delta}_k; \tilde{\mathcal{A}}_k\right) < \frac{\alpha^2}{4}\frac{\varepsilon_k}{6C''}$. Together with (57) and Lemma 7 of Appendix D, we have that with conditional probability at least $1 - \tilde{\delta}_k$ given the event that $k \leq K$ and the random variables V_{k-1}, R_k, and Δ_{i_k},

$$\sup_{g \in V_{k-1}} \hat{P}_{S_k^1}((\{g \neq h_k\} \setminus R_k) \setminus \Delta_{i_k}) < \frac{\varepsilon_k}{6C''}. \tag{60}$$

By the law of total probability, there is an event $E'_{3,k}$ of probability at least $1 - \tilde{\delta}_k$ such that, on $E'_{3,k}$, if $k \leq K$, then (60) holds. To unify notation, for any $k \in \{1, \ldots, N+1\}$ having probability zero of $k \leq K$, define $E'_{3,k}$ as the event (of probability one) that $k > K$, so that this conclusion also vacuously holds for such values k.

In particular, for any $k \in \{1, \ldots, N+1\}$, suppose the events $E'_{3,k}$ and $k \leq K$ occur. For each $t \in \{1, \ldots, m_k\}$, let $\zeta'_{t,0} = \mathbb{1}[h_k(X_{t_k+t}) = 0]\mathbb{1}[X_{t_k+t} \notin (R_k \setminus \Delta_{i_k})]$, $\zeta'_{t,1} = \mathbb{1}[h_k(X_{t_k+t}) = 1]\mathbb{1}[X_{t_k+t} \notin (R_k \setminus \Delta_{i_k})]$, $q'_t = \mathbb{1}[X_{t_k+t} \in R_k \setminus \Delta_{i_k}]$. Note that these values satisfy the second and third constraints on $\zeta_{t,0}, \zeta_{t,1}, q_t$ in LP_k. Moreover, (60) implies that $\forall g \in V_{k-1}$,

$$\frac{1}{m_k}\sum_{t=1}^{m_k} \zeta'_{t,1-g(X_{t_k+t})}\mathbb{1}[X_{t_k+t} \notin \Delta_{i_k}] = \hat{P}_{S_k^1}((\{g \neq h_k\} \setminus R_k) \setminus \Delta_{i_k}) < \frac{\varepsilon_k}{6C''},$$

so that the first constraint in LP_k is also satisfied by this choice of $\zeta_{t,0}, \zeta_{t,1}, q_t$. Since the values $\zeta^k_{t,0}, \zeta^k_{t,1}, q^k_t$ at the solution of LP_k *minimize* $\sum_{t=1}^{m_k} q_t$ among all choices of $\zeta_{t,0}, \zeta_{t,1}, q_t$ satisfying the constraints, we conclude that the above values of q'_t satisfy

$$|Q_k| = \sum_{t=1}^{m_k} q^k_t \leq \sum_{t=1}^{m_k} q'_t = m_k \hat{P}_{S_k^1}(R_k \setminus \Delta_{i_k}). \tag{61}$$

Next we upper bound the right hand side of (61) via a multiplicative Chernoff bound (Lemma 6 of Appendix D). Consider again any $k \in \{1, \ldots, N+1\}$ having non-zero probability of $k \leq K$. Given the event $k \leq K$ and the random variables R_k and Δ_{k-1}, Lemma 6 of Appendix D implies that, with conditional probability at least $1 - \tilde{\delta}_k$,

$$\hat{P}_{S_k^1}(R_k \setminus \Delta_{i_k}) \leq \max\left\{2P_X(R_k \setminus \Delta_{i_k}), \frac{6}{m_k}\ln\left(\frac{2}{\tilde{\delta}_k}\right)\right\} \leq \max\{2P_X(R_k \setminus \Delta_{i_k}), \varepsilon_k\}, \tag{62}$$

where the last inequality follows from (59) and straightforward reasoning about numerical constant factors. By the law of total probability, there is an event $E''_{3,k}$ of probability at least $1 - \tilde{\delta}_k$ on which, if $k \leq K$, then (62) holds. To unify notation, for any $k \in \{1, \ldots, N+1\}$ having probability zero of $k \leq K$, also define $E''_{3,k}$ as the event (of probability one) that $k > K$, so that this conclusion also vacuously holds for such values k.

Define $E'_3 = \bigcap_{k=1}^{N+1} E'_{3,k} \cap E''_{3,k}$. By the union bound, the event E'_3 fails with probability at most

$$\sum_{k=1}^{N+1} 2\tilde{\delta}_k = \sum_{k=1}^{N+1} \frac{\delta\varepsilon_{k+3}}{72} < \frac{\delta}{8},$$

where the last inequality follows from our choice of $C = \frac{11}{10}$.

Altogether, on the event E'_3, for every $k \in \{1, \ldots, K\}$, (61), (62), and (58) together imply

$$|Q_k| \leq 2m_k\Phi_{P_{\Delta_{i_k}}}\left(V_{k-1}, \frac{\varepsilon_k}{18C''}\right) + m_k\varepsilon_k. \tag{63}$$

It remains to relate the right hand side of (63) to the quantity $\varphi_P(\varepsilon, 0; 5\beta)$. For the remainder of the proof, suppose the event $E_0 \cap E_1 \cap E_2 \cap E_3$ holds (with E_3' in the definition of E_3 from the proof of Lemma 23 replaced by the above definition of E_3'). Consider any $k \in \{1, \ldots, K\}$. Recall that Lemma 17 implies $h^\star \in V_{k-1}$, which together with Lemma 12 implies $V_{k-1} \subseteq \mathrm{B}_{P_{\Delta_{i_k}}}(h^\star, \varepsilon_k)$. Thus, since the definition of $\Phi_\mu(\cdot, \cdot)$ is non-decreasing in its first argument, we have

$$\Phi_{P_{\Delta_{i_k}}}\left(V_{k-1}, \frac{\varepsilon_k}{18C''}\right) \leq \Phi_{P_{\Delta_{i_k}}}\left(\mathrm{B}_{P_{\Delta_{i_k}}}(h^\star, \varepsilon_k), \frac{\varepsilon_k}{18C''}\right).$$

Also recall that Lemma 20 implies $P_X(\Delta_{i_k}) \leq 5\beta$, so that Δ_{i_k} is among the sets Δ considered in the supremum in the definition of $\varphi_P(\varepsilon_k, 0; 5\beta)$. Additionally, note that $\varepsilon_k \geq \varepsilon_{N+1} > \frac{\varepsilon}{2C}$. It follows that

$$\begin{aligned}
&\Phi_{P_{\Delta_{i_k}}}\left(\mathrm{B}_{P_{\Delta_{i_k}}}(h^\star, \varepsilon_k), \frac{\varepsilon_k}{18C''}\right) = \frac{\Phi_{P_{\Delta_{i_k}}}\left(\mathrm{B}_{P_{\Delta_{i_k}}}(h^\star, \varepsilon_k), \frac{\varepsilon_k}{18C''}\right)}{\varepsilon_k}\varepsilon_k \\
&\leq \sup_{r > \varepsilon/2C} \frac{\Phi_{P_{\Delta_{i_k}}}\left(\mathrm{B}_{P_{\Delta_{i_k}}}(h^\star, r), \frac{r}{18C''}\right)}{r}\varepsilon_k = \sup_{r > \varepsilon} \frac{\Phi_{P_{\Delta_{i_k}}}\left(\mathrm{B}_{P_{\Delta_{i_k}}}\left(h^\star, \frac{r}{2C}\right), \frac{r}{36CC''}\right)}{r/(2C)}\varepsilon_k \\
&\leq 2C \sup_{r > \varepsilon} \frac{\Phi_{P_{\Delta_{i_k}}}\left(\mathrm{B}_{P_{\Delta_{i_k}}}(h^\star, r), \frac{r}{36CC''}\right)}{r}\varepsilon_k \leq 2C\varphi_P(\varepsilon, 0; 5\beta)\varepsilon_k.
\end{aligned}$$

Altogether, we have that every $k \in \{1, \ldots, K\}$ satisfies

$$|Q_k| \leq (4C\varphi_P(\varepsilon, 0; 5\beta) + 1)\, m_k\varepsilon_k \leq (4C+1)\varphi_P(\varepsilon, 0; 5\beta) m_k \varepsilon_k.$$

Substituting $|Q_k|$ in place of $|S_k^1 \cap D_{k-1} \setminus \Delta_{i_k}|$ in the proof of Lemma 23, and using the above bound on $|Q_k|$ in place of (36), we arrive at a bound $Q(\varepsilon, \delta; P)$ on the total number of queries

$$\begin{aligned}
&Q(\varepsilon, \delta; P) := 10\beta M_2 + \min\left\{M_1, \frac{4C+1}{2}\varphi_P(\varepsilon, 0; 5\beta)\varepsilon(N+1)m_{N+1}\right\} \\
&= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \min\left\{\varphi_P(\varepsilon, 0; 5\beta)\log\left(\frac{1}{\varepsilon}\right), \frac{1}{\varepsilon}\right\}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right).
\end{aligned}$$

■

As with $\mathbb{A}_{\mathrm{avid}}$, the algorithm $\mathbb{A}_{\mathrm{avid}}^{\mathrm{sub}}$ *does not need to know* the value $\varphi_P(\varepsilon, 0; 5\beta)$ (or anything else about P) to achieve this query complexity: that is, it is *adaptive* to the value $\varphi_P(\varepsilon, 0; 5\beta)$.

Together with (54), Theorem 33 further implies a (sometimes loose) relaxation in terms of $\varphi_P(\varepsilon, 5\beta)$, which is often easier to evaluate for given scenarios $(\mathbb{C}, P)$. This is stated formally in the following corollary. As mentioned above, the example in Appendix F.1 shows that it is not generally possible to reduce the $\varphi_P(\varepsilon, 5\beta)^2$ dependence to a linear $\varphi_P(\varepsilon, 5\beta)$ (or even any $\varphi_P(0, \alpha)$).

Corollary 34. *The query complexity bound $Q(\varepsilon, \delta; P)$ in Theorem 33 (achieved by $\mathbb{A}_{\mathrm{avid}}^{\mathrm{sub}}$) satisfies*

$$\begin{aligned}
Q(\varepsilon, \delta; P) &= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\min\left\{\mathsf{d}\varphi_P(\varepsilon, 5\beta)\left(\frac{\beta+\varepsilon}{\varepsilon}\right), \frac{\mathsf{d}}{\varepsilon}\right\}\right) \\
&= O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\min\left\{\mathsf{d}\varphi_P(\varepsilon, 5\beta)^2, \frac{\mathsf{d}}{\varepsilon}\right\}\right).
\end{aligned}$$

Proof. Due to the first two inequalities in (54), the second term in the expression of $Q(\varepsilon, \delta; P)$ in Theorem 27 is at most

$$O\left(\min\left\{\varphi_P(\varepsilon, 5\beta)\log\left(\frac{1}{\varepsilon}\right)\left(\frac{\beta+\varepsilon}{\varepsilon}\right), \frac{1}{\varepsilon}\right\}\left(\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right). \tag{64}$$

Relaxing $\mathsf{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \leq \log\left(\frac{1}{\varepsilon}\right)\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)$ and noting that

$$\varphi_P(\varepsilon, 5\beta)\log^2\left(\frac{1}{\varepsilon}\right)\left(\frac{\beta+\varepsilon}{\varepsilon}\right) \leq \varphi_P(\varepsilon, 5\beta)^2\log^4\left(\frac{1}{\varepsilon}\right) + \left(\frac{\beta+\varepsilon}{\varepsilon}\right)^2,$$

and $\left(\frac{\beta+\varepsilon}{\varepsilon}\right)^2 \leq 4\frac{\beta^2}{\varepsilon^2} + 4$, the quantity (64) is at most

$$O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right) + \min\left\{\varphi_P(\varepsilon, 5\beta)^2 \log^4\left(\frac{1}{\varepsilon}\right), \frac{1}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)\right\}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right).$$

Adding this to the first term in the expression of $Q(\varepsilon, \delta; P)$, the result follows. ■

An immediate consequence of Corollary 34 is that, whenever $\sup_{\alpha\in[0,5]} \varphi_P(\varepsilon, \alpha) = \operatorname{polylog}\left(\frac{1}{\varepsilon}\right)$, the dependence on β, ε in the query complexity bound in Theorem 33 is of order $\frac{\beta^2}{\varepsilon^2} + \operatorname{polylog}\left(\frac{1}{\varepsilon}\right)$. As was true of Corollary 28, (53) implies the first upper bound in Corollary 34 is never larger than the upper bound in Theorem 1; however, this is not always the case for the second upper bound in Corollary 34. Moreover, unlike Theorem 33, *both* upper bounds in Corollary 34 can sometimes be loose compared to the $\mathfrak{s}$-dependent bound in Theorem 5 (e.g., Example 6 in Appendix F.2.2). For this reason, as with Theorem 27, Theorem 33 is useful despite having a quantity $\varphi_P(\varepsilon, 0; 5\beta)$ that is more challenging to calculate, as it provides a starting point for P-dependent analysis that is at least never worse than Theorem 5.

To illustrate a well-known scenario where the technique presented in this subsection provides improvements over the basic $\mathbb{A}_{\text{avid}}$ algorithm, consider the following example.

Example 8 (Homogeneous linear classifiers, uniform distribution)**.** As an implication of Corollary 34, we find that $\mathbb{A}^{\text{sub}}_{\text{avid}}$ recovers a near-optimal query complexity bound for learning homogeneous linear classifiers under any marginal P_X that is isotropic log-concave. Let $d \geq 2$. For any $x, w \in \mathbb{R}^d$, denote by $h_w(x) = \mathbb{1}[\langle w, x\rangle \geq 0]$. In this scenario, we suppose $\mathcal{X} = \mathbb{R}^d$, $\mathbb{C} = \{h_w : w \in \mathbb{R}^d, \|w\| = 1\}$ (for which $\mathrm{VC}(\mathbb{C}) = d$), and P_X is any isotropic log-concave distribution (Balcan and Long, 2013) (for instance, $P_X = \mathrm{Uniform}(\{x : \|x\| = 1\})$ is one such distribution). In other words, $\mathbb{C}$ is the class of linear classifiers whose hyperplane decision boundary passes through the origin. This scenario has a long history of interest in the active learning literature (see Section A), featuring prominently (with P_X a uniform distribution) in the original A^2 paper of Balcan, Beygelzimer, and Langford (2005, 2006, 2009), which studied the case $\beta \lesssim \varepsilon/\sqrt{d}$ and showed a query complexity bound $\tilde{O}\left(d^2 \log\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)\right)$ in this regime. Later works refined this, via subregion-based techniques. Building on the works of Balcan, Broder, and Zhang (2007); Balcan and Long (2013) (which studied more-restrictive noise models), Zhang and Chaudhuri (2014) obtain a query complexity bound $\tilde{O}\left(d\frac{\beta^2}{\varepsilon^2} + d\right)$. Here we argue this query complexity bound can be recovered from Corollary 34 (indeed, with improvements by log factors in the lead term). Specifically, Zhang and Chaudhuri (2014) show (based on results of Balcan and Long, 2013) that $\varphi_P(\varepsilon, 5\beta) = O\left(\log\left(\frac{\beta}{\varepsilon}\right)\right)$. Plugging into Corollary 34 (rather, the expression obtained in the proof thereof), we obtain a query complexity bound

$$O\left(\frac{\beta^2}{\varepsilon^2}\left(d + \log\left(\frac{1}{\delta}\right)\right) + \log^2\left(\frac{\beta}{\varepsilon}\right)\log^4\left(\frac{1}{\varepsilon}\right)\left(d + \log\left(\frac{1}{\delta}\right)\right)\right).$$

Compared to the result of Zhang and Chaudhuri (2014), this improves the lead term by a factor $\log^2\left(\frac{\beta}{\varepsilon}\right)$ (though at the expense of additional log factors in the lower-order term). We also note that this query complexity bound represents a refinement of what would be obtained from Corollary 28, since even for the special case of P_X uniform on an origin-centered sphere, $\theta_P(\beta+\varepsilon) = \Theta\left(\sqrt{\mathsf{d}} \wedge \frac{1}{\beta+\varepsilon}\right)$ (Hanneke, 2007b).

F.4 Classes with Infinite VC Dimension via Covering Numbers

As one final remark about P-dependent query complexity bounds, we note that it is also possible to derive interesting query complexity improvements over passive learning even for classes with $\mathrm{VC}(\mathbb{C}) = \infty$ under conditions on P commonly studied in the *nonparametric* passive learning literature: namely, bounded *covering numbers*.

Denote by $\mathcal{N}(\varepsilon, \mathbb{C}, \mathrm{L}_1(P_X))$ the minimal size of a proper ε-cover: that is, the size of the smallest $\mathbb{C}' \subseteq \mathbb{C}$ for which $\sup_{h\in\mathbb{C}} \min_{h'\in\mathbb{C}'} P_X(h \neq h') \leq \varepsilon$. Being able to construct such a cover from unlabeled examples requires some additional structure beyond finite covering numbers under P_X (e.g.,

finite expected empirical covering numbers suffices; see e.g., van der Vaart and Wellner, 1996). Let us suppose such conditions are satisfied by $(\mathbb{C}, P_X)$, so that (since an active learner can be assumed to have access to an abundant supply of *unlabeled* examples) we may assume we have access to a valid $(\varepsilon/2)$-proper-cover $\mathbb{C}_{\varepsilon/2}$ under $\mathrm{L}_1(P_X)$ of size $O(\mathcal{N}(\varepsilon/2, \mathbb{C}, \mathrm{L}_1(P_X)))$. Constructing this cover $\mathbb{C}_{\varepsilon/2}$ does not affect the query complexity, since it only requires the use of unlabeled examples.

We can then run $\mathbb{A}_{\text{avid}}$ using $\mathbb{C}_{\varepsilon/2}$ in place of $\mathbb{C}$. Since $\mathrm{VC}(\mathbb{C}_{\varepsilon/2}) = O(\log(\mathcal{N}(\varepsilon/2, \mathbb{C}, \mathrm{L}_1(P_X))))$, and $\inf_{h \in \mathbb{C}_{\varepsilon/2}} \mathrm{er}_P(h) \leq \inf_{h \in \mathbb{C}} \mathrm{er}_P(h) + \frac{\varepsilon}{2} =: \beta + \frac{\varepsilon}{2}$, we thereby obtain from Theorem 1 a P_X-dependent query complexity bound

$$O\left(\frac{\beta^2}{\varepsilon^2} \log(\mathcal{N}(\varepsilon/2, \mathbb{C}, \mathrm{L}_1(P_X))/\delta)\right) + \tilde{O}\left(\frac{1}{\varepsilon} \log(\mathcal{N}(\varepsilon/2, \mathbb{C}, \mathrm{L}_1(P_X)))\right).$$

This result can then be composed with bounds on the covering numbers $\mathcal{N}(\varepsilon/2, \mathbb{C}, \mathrm{L}_1(P_X))$ of various classes $\mathbb{C}$ under various conditions on P_X known from the literature. For instance, this provides an improved query complexity for *boundary fragment* classes (a class defined by smoothness conditions on the decision boundaries of concepts in $\mathbb{C}$) under near-uniform distributions P_X on $[0,1]^{k+1}$ compared to the results established by Wang (2011) (see Wang, 2011; Tsybakov, 2004 for the precise definitions and covering numbers).

G Extensions and Future Directions

We conclude with some extensions and several interesting open questions and future directions.

Extension to Multiclass Classification: We can easily generalize the result to hold for *multiclass classification*: that is, where $\mathcal{Y}$ is a general label space, $\mathbb{C}$ is a family of measurable functions $h : \mathcal{X} \to \mathcal{Y}$, P is a distribution on $\mathcal{X} \times \mathcal{Y}$, and we still define $\mathrm{er}_P(h) := P((x,y) : h(x) \neq y) = P(\mathrm{ER}(h))$. The exact same upper bound extends to this setting if we replace d with $\max\{\mathrm{VC}(\mathcal{A}), \mathsf{d}_G\}$ where $\mathcal{A}$ is as in Lemma 9 (replacing $\{0,1\}$ with $\mathcal{Y}$ there) and d_G denotes the *graph dimension* of $\mathbb{C}$ (Natarajan, 1989). The star number $\mathfrak{s}$ is still defined as in Definition 2 (see Hanneke, 2024). The proof holds with only superficial modifications to rely solely on $\mathrm{VC}(\mathcal{A})$ (for the $\mathcal{X} \setminus \Delta_{i_k}$ concentration) and d_G (for concentration in Δ_{i_k}). We further note that this dimension $\max\{\mathrm{VC}(\mathcal{A}), \mathsf{d}_G\}$ is at most $O(\mathsf{d}_N(\mathbb{C}) \log(|\mathcal{Y}|))$, where $\mathsf{d}_N(\mathbb{C})$ is the *Natarajan dimension* of $\mathbb{C}$ (Natarajan, 1989); this follows by a similar argument as used to bound $\mathrm{VC}(\mathcal{A})$ in the proof of Lemma 9, using a generalization of Sauer's lemma for the multiclass setting proven by Haussler and Long (1995).

For a bounded number of labels $|\mathcal{Y}|$, this again leads to essentially optimal query complexity, as a lower bound $\Omega\left(\frac{\mathsf{d}_N(\mathbb{C})\beta^2}{\varepsilon^2}\right)$ based on the Natarajan dimension $\mathsf{d}_N(\mathbb{C})$ can be shown (similarly to the lower bound for binary classification).

However, for unbounded label spaces ($|\mathcal{Y}| = \infty$) the learnability and optimal sample complexity of passive learning in the realizable case are known to depend on a dimension called the *DS dimension* (Brukhim, Carmon, Dinur, Moran, and Yehudayoff, 2022; Daniely and Shalev-Shwartz, 2014) which is between the Natarajan dimension and graph dimension. This raises an important question: What is the optimal query complexity for multiclass agnostic active learning?

Extension to Stream-based Active Learning: For simplicity, we have defined the learning model as so-called *pool-based* active learning, in that the learning algorithm was given the entire sequence $X_1, \ldots, X_m$ of unlabeled examples as input, and can query any example, in any order. However, it is also common to consider an alternative protocol called *stream-based* active learning (or *selective sampling*): namely, where the active learner observes the unlabeled examples X_t one-at-a-time *in sequence*, and for each, decides whether or not to query, and can never revisit that decision later. In the literature on stream-based active learning, it is common to express the guarantees of the active learning in two parts: (1) a bound on the error guarantee expressed as a function of the number m of unlabeled examples processed, and (2) a bound on the number of queries it makes among the first m examples (e.g., Dasgupta, Hsu, and Monteleoni, 2007).

We note that $\mathbb{A}_{\text{avid}}$ can easily be re-expressed as a stream-based active learner. Specifically, rather than limiting the '*For*' loop in Step 1 to $N = O\left(\log\left(\frac{1}{\varepsilon}\right)\right)$ rounds, we can simply let the algorithm run until it has allocated as many unlabeled examples m as we wish. Rather than allocating all of the

S_k^1, S_k^4 data subsets at the start, we can simply allocate these sets if and when the algorithm reaches the k^{th} iteration of the '*For*' loop, at which point the algorithm collects the next m_k examples to allocate to S_k^1, querying each of these examples X_t iff $X_t \in D_{k-1} \setminus \Delta_{i_k}$. Likewise, it then collects the next m_k examples to allocate to S_k^4 (without making any queries), to calculate the value m'_k (where, in this case, we should suitably replace the value $3 + N - k$ in the log term in m'_k to remove the dependence on ε: for instance, replacing it with $k + 2$ would suffice for the present discussion). It then collects the next m'_k examples to allocate to the data subset S_k^2, querying each of these examples X_t iff $X_t \in \Delta_{i_k}$. It then moves on to execute Steps 3-4. Similarly, upon each time it reaches Step 5, it simply collects the next m_k unlabeled examples to construct $S_{k,i}^3$ (without making any queries), which then enables it to execute Steps 5-7. We can execute this until any number m of unlabeled examples have been processed, and define the predictor at such a time as the $\hat{h}_k$ for the last iteration k for which Step 2 was able to completely execute. If the algorithm ever satisfies the early stopping criterion in Step 4 for some iteration k, we can simply take $\hat{h}_k$ as its final predictor. We can then derive the corresponding excess error bound and query bound from the above analysis of the query complexity and unlabeled sample complexity: namely, with probability at least $1 - \delta$, the predictor $\hat{h}$ produced after m unlabeled examples satisfies

$$\operatorname{er}_P(\hat{h}) - \inf_{h \in \mathbb{C}} \operatorname{er}_P(h) = O\left(\sqrt{\beta\left(\mathsf{d}\log\left(\frac{m}{\mathsf{d}}\right) + \log\left(\frac{1}{\delta}\right)\right)} + \frac{1}{m}\left(\mathsf{d}\log\left(\frac{m}{\mathsf{d}}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

and its number of queries is bounded by

$$O\left(\beta m + \min\left\{\mathfrak{s}\log\left(\frac{m}{\mathsf{d}}\right)\left(\mathsf{d}\log\left(\frac{m}{\mathsf{d}}\right) + \log\left(\frac{1}{\delta}\right)\right), \sqrt{\frac{m}{\beta}\left(\mathsf{d}\log\left(\frac{m}{\mathsf{d}}\right) + \log\left(\frac{1}{\delta}\right)\right)}, m\right\}\right)$$
$$= O(\beta m) + \tilde{O}\left(\min\left\{\mathfrak{s}\mathsf{d}, \sqrt{\frac{m\mathsf{d}}{\beta}}, m\right\}\right).$$

Here the βm term is where the improvements over passive learning provided by the AVID principle are reflected in the query bound (as the above excess error bound is nearly as small as the best achievable excess error guarantees for passive learning with m labeled examples; Vapnik and Chervonenkis, 1974; Devroye and Lugosi, 1995; Hanneke, Larsen, and Zhivotovskiy, 2024b). In particular, the above guarantees compare favorably to previous analyses of stream-based active learning (e.g., Dasgupta, Hsu, and Monteleoni, 2007) in the regime of moderate-size β, where, for the same excess error guarantee, the bounds on the number of queries include a term such as $\tilde{O}(\theta_P(\beta)\beta m)$, which becomes of order $\left(\mathfrak{s} \wedge \frac{1}{\beta}\right)\beta m$ in the worst case over P, and hence is no better than m when $\mathfrak{s} = \infty$. In contrast, in this regime of moderate-size β, where the βm term dominates, we obtain a factor β improvement in the number of queries.

We also remark that the analysis above also supplies an "anytime" guarantee, where the algorithm can simply be executed indefinitely, and the above excess error bound and query bound hold simultaneously for every m (where, again, if the algorithm ever satisfies the condition in Step 4, its predictor should simply be defined as the corresponding $\hat{h}_k$ forevermore, and it need not query any further examples in the sequence).

The Optimal Lower-Order Term: As discussed above, while the leading term in Theorem 3 is exacty optimal (perfectly matching a lower bound), the lower-order term in the upper bound in Theorem 3 presents a small gap (in the dependence on d) compared to the best known lower bound (Hanneke and Yang, 2015). As discussed, some aspects of this gap (concerning $\frac{1}{\varepsilon} + \mathsf{d}$ vs $\frac{\mathsf{d}}{\varepsilon}$) *cannot* be improved if the dependence on $\mathbb{C}$ is only expressed via d and $\mathfrak{s}$: that is, without introducing new complexity measures. We leave open the question of formulating such an always-sharp complexity measure, that is, the question: What is the optimal form of the query complexity $\Theta(\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}))$ for all classes $\mathbb{C}$? However, aside from this gap, there is a gap which might be improvable even in expressions of the bound purely in terms of d and $\mathfrak{s}$: namely, the term $\mathfrak{s}\mathsf{d}$ in the upper bound. I conjecture this can be reduced to simply $\mathfrak{s}$: that is, $\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}) = O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathsf{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\min\left\{\mathfrak{s}, \frac{\mathsf{d}}{\varepsilon}\right\}\right)$ for every concept class $\mathbb{C}$.

Proper Learning: As noted above, the $\mathbb{A}_{\text{avid}}$ algorithm is an *improper* learner, meaning its returned predictor $\hat{h}$ might not be an element of the concept class $\mathbb{C}$ (rather, it is a shallow decision list built from concepts in $\mathbb{C}$). It is an interesting open question to determine whether there exist *proper* active learners achieving the query complexity bound in either Theorem 1 or 3 for every concept class $\mathbb{C}$. It follows from Corollary 18 that, in the return case in Step 9, it would suffice to return $\hat{h}$ equal any element of V_N. Thus, the main challenge in obtaining a proper learner is in the early-stopping case in Step 4. In this return case, we have effectively verified that $\text{er}_P\big(\hat{h}_k\big)$ is *better* than $\text{er}_P\big(h^\star\big)$ (Lemma 17). However, the resolution of the error estimates $\hat{\text{er}}_k^{1,2}$ at this stage might not yet be sufficient to find an $h \in V_{k-1}$ nearly as good. Indeed, for this reason, any such early return case in an active learning algorithm may be problematic for proper learning.

On the other hand, we remark that, for all previous known separations between proper and improper sample complexities, the respective proofs break down if the learner is given access to the marginal distribution P_X or a sufficiently large unlabeled data set (Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020; Hanneke, Larsen, and Zhivotovskiy, 2024b; Daniely and Shalev-Shwartz, 2014; Montasser, Hanneke, and Srebro, 2019; Asilis, Devic, Sharan, and Teng, 2025a; Asilis, Høgsgaard, and Velegkas, 2025b). Since, for the purpose of merely bounding the *query complexity*, we may suppose an active learner has access to a large unlabeled data set, this hints that such improvements might indeed be achievable by proper active learners, or otherwise, a novel technique is needed for establishing such a separation between proper and improper active learning.

Computational Efficiency: The focus of this work has been solely on the information-theoretic query complexity of agnostic active learning, without any computational or resource constraints beyond the number of queries and unlabeled examples. However, computational considerations are of course also important to consider. To actually achieve the agnostic learning guarantee of ε excess error is typically thought to be computationally intractable for many concept classes, without distribution restrictions. Nonetheless, it would be interesting to determine whether, at least at some level, the improvements in the leading term reflected in Theorems 1 and 3 might also be reflected in a computationally efficient method, for some classes $\mathbb{C}$ (e.g., linear classifiers) under some restrictions on the distribution P which enable computational tractability yet for which such query complexity bounds are not captured by prior results (e.g., by $\theta_P(\varepsilon)$).

Beyond this, a classical approach to obtaining computationally efficient algorithms in *practice* is to introduce *convex relaxations* of the various optimization problems involved in a given algorithm. In the literature on passive learning, the theory of error bounds for empirical risk minimization has been extended to allow for convex relaxations of the 0-1 loss, called a *surrogate loss*, while still guaranteeing bounds on the excess error rate under appropriate assumptions on P relating excess surrogate risks to excess error rates (Bartlett, Jordan, and McAuliffe, 2006; Zhang, 2004). Prior work on *disagreement-based* active learning has been found to compose well with this theory of surrogate losses. Specifically, Hanneke and Yang (2019); Hanneke (2014) express disagreement-based active learning algorithms, in which the optimization problems defining the query criterion and the learner's final predictor are relaxed to convex programs expressed in terms of any given surrogate loss. For such algorithms, they derive query complexity bounds (based on the disagreement coefficient $\theta_P(\varepsilon)$) holding under the same conditions studied by the passive learning works (Bartlett, Jordan, and McAuliffe, 2006; Zhang, 2004). It is thus a natural question to determine whether such a theory can be made to work for the algorithmic principles underlying $\mathbb{A}_{\text{avid}}$ (i.e., the AVID principle), leading to an algorithm only requiring computationally tractable convex optimization problems based on a given surrogate loss, and expressing query complexity improvements over passive learning (of the type found in Theorems 1 and 3) under these same conditions on P relating excess surrogate risks to the excess error rates. This approach is made challenging in the context of $\mathbb{A}_{\text{avid}}$, due to its use of improper predictors $\hat{h}_k$, and even more-so due to the maximization in Steps 5 and 6 (whereas convex surrogate losses would typically only allow tractability of *minimization* problems).

As a step toward such a technique, an interesting intermediate question is whether Theorems 1 and 3 can be achieved by an active learning algorithm expressed as a *reduction to an empirical risk minimization (ERM) oracle*: that is, where the access to the concept class $\mathbb{C}$ is restricted to solving optimization problems of the form $\text{argmin}_{h\in\mathbb{C}} \hat{\text{er}}_S(h)$ for data sets S (or possibly a *weighted* ERM). This would be particularly interesting if these data sets S are only constructed from subsets of the labeled examples (X_t, Y_t) queried by the algorithm (perhaps plus one additional example (X_t, y) with an artificial label y, which may be needed when deciding whether to query X_t). Previous works

by Beygelzimer, Hsu, Langford, and Zhang (2010); Hsu (2010) have expressed *disagreement-based* active learning algorithms as reductions to such ERM oracles. It is therefore a natural question to consider whether the AVID principle can also be implemented based only on such oracles (and such an implementation could also be an important step toward enabling the above composition with the theory of surrogate losses).

Unlabeled Sample Complexity: Theorem 5 reveals that, to achieve the stated query complexity bound with $\mathbb{A}_{\text{avid}}$, it suffices to have access to a number of *unlabeled* examples $M(\varepsilon, \delta; \beta) = O\left(\frac{\beta+\varepsilon}{\varepsilon^2}\left(\mathtt{d}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$. In comparison, we can obtain an obvious lower bound on the number of unlabeled examples necessary to achieve any query complexity bound by a lower bound on the sample complexity of fully-supervised *passive* learning (Devroye and Lugosi, 1995): i.e., $\Omega\left(\frac{\beta+\varepsilon}{\varepsilon^2}\left(\mathtt{d} + \log\left(\frac{1}{\delta}\right)\right)\right)$. Thus, the upper bound $M(\varepsilon, \delta; \beta)$ in Theorem 5 can be improved by at most a $\log\left(\frac{1}{\varepsilon}\right)$ factor. This naturally raises the question: Is it possible to achieve a near-optimal query complexity $\Theta(\mathrm{QC}_a(\varepsilon, \delta; \beta, \mathbb{C}))$ with an algorithm which uses a number of unlabeled examples $O\left(\frac{\beta+\varepsilon}{\varepsilon^2}\left(\mathtt{d} + \log\left(\frac{1}{\delta}\right)\right)\right)$? Such a result would then be optimal simultaneously in *both* the number of queries *and* the number of unlabeled examples. To date, this is not even known to be achievable by fully-supervised *passive* learning, the best known upper bound having an additive $\tilde{O}\left(\frac{\mathtt{d}}{\varepsilon}\right)$ term (Hanneke, Larsen, and Zhivotovskiy, 2024b). Thus, for now, a more-approachable question would be whether it is possible to match the query complexity bound in Theorem 3 using a number of unlabeled examples suboptimal only in log factors in the lower-order term, that is: Is there an algorithm achieving a query complexity upper bound $O\left(\frac{\beta^2}{\varepsilon^2}\left(\mathtt{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\left(\mathfrak{s} \wedge \frac{1}{\varepsilon}\right)\mathtt{d}\right)$ which uses a number of unlabeled examples at most $O\left(\frac{\beta}{\varepsilon^2}\left(\mathtt{d} + \log\left(\frac{1}{\delta}\right)\right)\right) + \tilde{O}\left(\frac{\mathtt{d}}{\varepsilon}\right)$? As an intermediate step, it would already be interesting to determine whether this many unlabeled examples suffices to achieve the query complexity bound in Theorem 1.

Tsybakov Noise: Beyond the above directions, there are a number of further extensions of this work that seem ripe for exploration. One natural direction is extending the techniques in this work to the case of *Tsybakov noise* (Mammen and Tsybakov, 1999; Tsybakov, 2004; Massart and Nédélec, 2006). The optimal query complexity under Tsybakov noise was already identified by Hanneke and Yang (2015) (aside from similar gaps to the $\frac{\mathtt{d}}{\varepsilon}$ vs $\frac{1}{\varepsilon} + \mathtt{d}$ issue discussed above, which require introducing a new complexity measure to resolve). However, the algorithmic techniques in the present work are significantly simpler, and moreover, have the potential to dramatically reduce the number of *unlabeled* examples required for learning, compared to the technique of Hanneke and Yang (2015). I conjecture that the AVID principle is capable of yielding near-optimal query complexity guarantees under Tsybakov noise (with a number of unlabeled examples of the same order as the sample complexity of supervised learning, up to log factors); however, obtaining such guarantees may require a more-sophisticated usage of the principle, such as by the creation of multiple different regions Δ, coinciding with different levels of variance of excess error estimates. Indeed, an analogous *tiered* allocation of queries was key to the original analysis of the query complexity under Tsybakov noise by Hanneke and Yang (2015).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper formally proves (appropriate formalizations of) the claims made in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly mentions a number of questions left open by this work (e.g., refining a lower-order term, proper learning, expression as a reduction to ERM, efficient relaxations via surrogate losses).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main body includes formal definitions and a proof outline, and a complete formal proof is provided in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: This is a purely theoretical work, and as such does not include experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This is a purely theoretical work, and as such does not include an experimental component relying on data or code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This is a purely theoretical work, and as such does not include an experimental component.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This is a purely theoretical work, and as such does not include experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This is a purely theoretical work, and as such does not include an experimental component relying on computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and fully conform to the guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a purely theoretical work, and as such is not likely to have broader societal impacts.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This is a purely theoretical work, and as such there is no associated data or model being released.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not make use of such assets. All relevant literature is properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The work in this paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The work in this paper does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not involved in any aspect of this research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.