Journal of Machine Learning Research 13 (2012) 1469-1587 Submitted 8/11; Published 5/12

Activized Learning: Transforming Passive to Active with Improved Label Complexity*

Steve Hanneke SHANNEKE@STAT.CMU.EDU
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213 USA

Editor: Sanjoy Dasgupta

Abstract

We study the theoretical advantages of active learning over passive learning. Specifically, we prove that, in noise-free classifier learning for VC classes, any passive learning algorithm can be transformed into an active learning algorithm with asymptotically strictly superior label complexity for all nontrivial target functions and distributions. We further provide a general characterization of the magnitudes of these improvements in terms of a novel generalization of the disagreement coefficient. We also extend these results to active learning in the presence of label noise, and find that even under broad classes of noise distributions, we can typically guarantee strict improvements over the known results for passive learning.

Keywords: active learning, selective sampling, sequential design, statistical learning theory, PAC learning, sample complexity

1. Introduction and Background

The recent rapid growth in data sources has spawned an equally rapid expansion in the number of potential applications of machine learning methodologies to extract useful concepts from these data. However, in many cases, the bottleneck in the application process is the need to obtain accurate annotation of the raw data according to the target concept to be learned. For instance, in webpage classification, it is straightforward to rapidly collect a large number of webpages, but training an accurate classifier typically requires a human expert to examine and label a number of these webpages, which may require significant time and effort. For this reason, it is natural to look for ways to reduce the total number of labeled examples required to train an accurate classifier. In the traditional machine learning protocol, here referred to as *passive learning*, the examples labeled by the expert are sampled independently at random, and the emphasis is on designing learning algorithms that make the most effective use of the number of these labeled examples available. However, it is possible to go beyond such methods by altering the protocol itself, allowing the learning algorithm to sequentially *select* the examples to be labeled, based on its observations of the labels of previously-selected examples; this interactive protocol is referred to as *active learning*. The objective in designing this selection mechanism is to focus the expert's efforts toward labeling only the most informative data for the learning process, thus eliminating some degree of redundancy in the information content of the labeled examples.

$*$. Some of these (and related) results previously appeared in the author's doctoral dissertation (Hanneke, 2009b).

It is now well-established that active learning can sometimes provide significant practical and theoretical advantages over passive learning, in terms of the number of labels required to obtain a given accuracy. However, our current understanding of active learning in general is still quite limited in several respects. First, since we are lacking a complete understanding of the potential capabilities of active learning, we are not yet sure to what standards we should aspire for active learning algorithms to meet, and in particular this challenges our ability to characterize how a “good” active learning algorithm should behave. Second, since we have yet to identify a complete set of general principles for the design of effective active learning algorithms, in many cases the most effective known active learning algorithms have problem-specific designs (e.g., designed specifically for linear separators, or decision trees, etc., under specific assumptions on the data distribution), and it is not clear what components of their design can be abstracted and transferred to the design of active learning algorithms for different learning problems (e.g., with different types of classifiers, or different data distributions). Finally, we have yet to fully understand the scope of the relative benefits of active learning over passive learning, and in particular the conditions under which such improvements are achievable, as well as a general characterization of the potential magnitudes of these improvements. In the present work, we take steps toward closing this gap in our understanding of the capabilities, general principles, and advantages of active learning.

Additionally, this work has a second theme, motivated by practical concerns. To date, the machine learning community has invested decades of research into constructing solid, reliable, and well-behaved *passive* learning algorithms, and into understanding their theoretical properties. We might hope that an equivalent amount of effort is *not* required in order to discover and understand effective active learning algorithms. In particular, rather than starting from scratch in the design and analysis of active learning algorithms, it seems desirable to leverage this vast knowledge of passive learning, to whatever extent possible. For instance, it may be possible to design active learning algorithms that *inherit* certain desirable behaviors or properties of a given passive learning algorithm. In this way, we can use a given passive learning algorithm as a *reference point*, and the objective is to design an active learning algorithm with performance guarantees strictly superior to those of the passive algorithm. Thus, if the passive learning algorithm has proven effective in a variety of common learning problems, then the active learning algorithm should be even better for those *same* learning problems. This approach also has the advantage of immediately supplying us with a collection of theoretical guarantees on the performance of the active learning algorithm: namely, improved forms of all known guarantees on the performance of the given passive learning algorithm.

Due to its obvious practical advantages, this general line of informal thinking dominates the existing literature on empirically-tested heuristic approaches to active learning, as most of the published heuristic active learning algorithms make use of a passive learning algorithm as a subroutine (e.g., SVM, logistic regression, k-NN, etc.), constructing sets of labeled examples and feeding them into the passive learning algorithm at various times during the execution of the active learning algorithm (see the references in Section 7). Below, we take a more rigorous look at this general strategy. We develop a reduction-style framework for studying this approach to the design of active learning algorithms relative to a given passive learning algorithm. We then proceed to develop and analyze a variety of such methods, to realize this approach in a very general sense.

Specifically, we explore the following fundamental questions.

- Is there a general procedure that, given any passive learning algorithm, transforms it into an active learning algorithm requiring significantly fewer labels to achieve a given accuracy?
- If so, how large is the reduction in the number of labels required by the resulting active learning algorithm, compared to the number of labels required by the original passive algorithm?
- What are sufficient conditions for an *exponential* reduction in the number of labels required?
- To what extent can these methods be made robust to imperfect or noisy labels?

In the process of exploring these questions, we find that for many interesting learning problems, the techniques in the existing literature are not capable of realizing the full potential of active learning. Thus, exploring this topic in generality requires us to develop novel insights and entirely new techniques for the design of active learning algorithms. We also develop corresponding natural complexity quantities to characterize the performance of such algorithms. Several of the results we establish here are more general than any related results in the existing literature, and in many cases the algorithms we develop use significantly fewer labels than any previously published methods.

1.1 Background

The term *active learning* refers to a family of supervised learning protocols, characterized by the ability of the learning algorithm to pose queries to a teacher, who has access to the target concept to be learned. In practice, the teacher and queries may take a variety of forms: a human expert, in which case the queries may be questions or annotation tasks; nature, in which case the queries may be scientific experiments; a computer simulation, in which case the queries may be particular parameter values or initial conditions for the simulator; or a host of other possibilities. In our present context, we will specifically discuss a protocol known as *pool-based* active learning, a type of sequential design based on a collection of unlabeled examples; this seems to be the most common form of active learning in practical use today (e.g., Settles, 2010; Baldridge and Palmer, 2009; Gangadharaiah, Brown, and Carbonell, 2009; Hoi, Jin, Zhu, and Lyu, 2006; Luo, Kramer, Goldgof, Hall, Samson, Remsen, and Hopkins, 2005; Roy and McCallum, 2001; Tong and Koller, 2001; McCallum and Nigam, 1998). We will not discuss alternative models of active learning, such as *online* (Dekel, Gentile, and Sridharan, 2010) or *exact* (Hegedüs, 1995). In the pool-based active learning setting, the learning algorithm is supplied with a large collection of unlabeled examples (the *pool*), and is allowed to select any example from the pool to request that it be labeled. After observing the label of this example, the algorithm can then select another unlabeled example from the pool to request that it be labeled. This continues sequentially for a number of rounds until some halting condition is satisfied, at which time the algorithm returns a function intended to approximately mimic and generalize the observed labeling behavior. This setting contrasts with *passive learning*, where the learning algorithm is supplied with a collection of *labeled* examples without any interaction.

Supposing the labels received agree with some true target concept, the objective is to use this returned function to approximate the true target concept on future (previously unobserved) data points. The hope is that, by carefully selecting which examples should be labeled, the algorithm can achieve improved accuracy while using fewer labels compared to passive learning. The motivation for this setting is simple. For many modern machine learning problems, unlabeled examples are inexpensive and available in abundance, while annotation is time-consuming or expensive. For instance, this is the case in the aforementioned webpage classification problem, where the pool would be the set of all webpages, and labeling a webpage requires a human expert to examine the website content. Settles (2010) surveys a variety of other applications for which active learning is presently being used. To simplify the discussion, in this work we focus specifically on *binary classification*, in which there are only two possible labels. The results generalize naturally to multiclass classification as well.

As the above description indicates, when studying the advantages of active learning, we are primarily interested in the number of label requests sufficient to achieve a given accuracy, a quantity referred to as the *label complexity* (Definition 1 below). Although active learning has been an active topic in the machine learning literature for many years now, our *theoretical* understanding of this topic was largely lacking until very recently. However, within the past few years, there has been an explosion of progress. These advances can be grouped into two categories: namely, the *realizable case* and the *agnostic case*.

1.1.1 THE REALIZABLE CASE

In the realizable case, we are interested in a particularly strict scenario, where the true label of any example is *determined* by a function of the features (covariates), and where that function has a specific known form (e.g., linear separator, decision tree, union of intervals, etc.); the set of classifiers having this known form is referred to as the *concept space*. The natural formalization of the realizable case is very much analogous to the well-known PAC model for passive learning (Valiant, 1984). In the realizable case, there are obvious examples of learning problems where active learning can provide a significant advantage compared to passive learning; for instance, in the problem of learning *threshold* classifiers on the real line (Example 1 below), a kind of *binary search* strategy for selecting which examples to request labels for naturally leads to *exponential* improvements in label complexity compared to learning from random labeled examples (passive learning). As such, there is a natural attraction to determine how general this phenomenon is. This leads us to think about general-purpose learning strategies (i.e., which can be instantiated for more than merely threshold classifiers on the real line), which exhibit this binary search behavior in various special cases.

The first such general-purpose strategy to emerge in the literature was a particularly elegant strategy proposed by Cohn, Atlas, and Ladner (1994), typically referred to as CAL after its discoverers (Meta-Algorithm 2 below). The strategy behind CAL is the following. The algorithm examines each example in the unlabeled pool in sequence, and if there are two classifiers in the concept space consistent with all previously-observed labels, but which disagree on the label of this next example, then the algorithm requests that label, and otherwise it does not. For this reason, below we refer to the general family of algorithms inspired by CAL as *disagreement-based* methods. Disagreement-based methods are sometimes referred to as "mellow" active learning, since in some sense this is the *least* we can expect from a reasonable active learning algorithm; it never requests the label of an example whose label it can *infer* from information already available, but otherwise makes no attempt to seek out particularly informative examples to request the labels of. That is, the notion of *informativeness* implicit in disagreement-based methods is a *binary* one, so that an example is either informative or not informative, but there is no further ranking of the informativeness of examples. The disagreement-based strategy is quite general, and obviously leads to algorithms that are at least *reasonable*, but Cohn, Atlas, and Ladner (1994) did not study the label complexity achieved by their strategy in any generality.

In a Bayesian variant of the realizable setting, Freund, Seung, Shamir, and Tishby (1997) studied an algorithm known as *query by committee* (QBC), which in some sense represents a Bayesian variant of CAL. However, QBC *does* distinguish between different levels of informativeness beyond simple disagreement, based on the *amount* of disagreement on a random unlabeled example. They were able to analyze the label complexity achieved by QBC in terms of a type of information gain, and found that when the information gain is lower bounded by a positive constant, the algorithm achieves a label complexity exponentially smaller than the known results for passive learning. In particular, this is the case for the threshold learning problem, and also for the problem of learning higher-dimensional (nearly balanced) linear separators when the data satisfy a certain (uniform) distribution. Below, we will not discuss this analysis further, since it is for a slightly different (Bayesian) setting. However, the results below in our present setting do have interesting implications for the Bayesian setting as well, as discussed in the recent work of Yang, Hanneke, and Carbonell (2011).

The first general analysis of the label complexity of active learning in the (non-Bayesian) realizable case came in the breakthrough work of Dasgupta (2005). In that work, Dasgupta proposed a quantity, called the *splitting index*, to characterize the label complexities achievable by active learning. The splitting index analysis is noteworthy for several reasons. First, one can show it provides nearly tight bounds on the *minimax* label complexity for a given concept space and data distribution. In particular, the analysis matches the exponential improvements known to be possible for threshold classifiers, as well as generalizations to higher-dimensional homogeneous linear separators under near-uniform distributions (as first established by Dasgupta, Kalai, and Monteleoni, 2005, 2009). Second, it provides a novel notion of *informativeness* of an example, beyond the simple binary notion of informativeness employed in disagreement-based methods. Specifically, it describes the informativeness of an example in terms of the number of *pairs* of well-separated classifiers for which at least one out of each pair will be contradicted, supposing the least-favorable label. Finally, unlike any other existing work on active learning (present work included), it provides an elegant description of the *trade-off* between the number of label requests and the number of unlabeled examples needed by the learning algorithm. Another interesting byproduct of Dasgupta's work is a better understanding of the *nature* of the improvements achievable by active learning in the general case. In particular, his work clearly illustrates the need to study the label complexity as a quantity that varies depending on the particular target concept and data distribution. We will see this issue arise in many of the examples below.

Coming from a slightly different perspective, Hanneke (2007a) later analyzed the label complexity of active learning in terms of an extension of the *teaching dimension* (Goldman and Kearns, 1995). Related quantities were previously used by Hegedüs (1995) and Hellerstein, Pillaipakkamnatt, Raghavan, and Wilkins (1996) to tightly characterize the number of membership queries sufficient for *Exact* learning; Hanneke (2007a) provided a natural generalization to the *PAC* learning setting. At this time, it is not clear how this quantity relates to the splitting index. From a practical perspective, in some instances it may be easier to calculate (see the work of Nowak, 2008 for a discussion related to this), though in other cases the opposite seems true.

The next progress toward understanding the label complexity of active learning came in the work of Hanneke (2007b), who introduced a quantity called the *disagreement coefficient* (Definition 9 below), accompanied by a technique for analyzing disagreement-based active learning algorithms. In particular, implicit in that work, and made explicit in the later work of Hanneke (2011), was the first general characterization of the label complexities achieved by the original CAL strategy for active learning in the realizable case, stated in terms of the disagreement coefficient. The results of the present work are direct descendants of that 2007 paper, and we will discuss the disagreement coefficient, and results based on it, in substantial detail below. Disagreement-based active learners such as CAL are known to be sometimes suboptimal relative to the splitting index analysis, and therefore the disagreement coefficient analysis sometimes results in larger label complexity bounds than the splitting index analysis. However, in many cases the label complexity bounds based on the disagreement coefficient are surprisingly good considering the simplicity of the methods. Furthermore, as we will see below, the disagreement coefficient has the practical benefit of often being fairly straightforward to calculate for a variety of learning problems, particularly when there is a natural geometric interpretation of the classifiers and the data distribution is relatively smooth. As we discuss below, it can also be used to bound the label complexity of active learning in noisy settings. For these reasons (simplicity of algorithms, ease of calculation, and applicability beyond the realizable case), subsequent work on the label complexity of active learning has tended to favor the disagreement-based approach, making use of the disagreement coefficient to bound the label complexity (Dasgupta, Hsu, and Monteleoni, 2007; Friedman, 2009; Beygelzimer, Dasgupta, and Langford, 2009; Wang, 2009; Balcan, Hanneke, and Vaughan, 2010; Hanneke, 2011; Koltchinskii, 2010; Beygelzimer, Hsu, Langford, and Zhang, 2010; Mahalanabis, 2011; Wang, 2011). A significant part of the present paper focuses on extending and generalizing the disagreement coefficient analysis, while still maintaining the relative ease of calculation that makes the disagreement coefficient so useful.

In addition to many positive results, Dasgupta (2005) also pointed out several negative results, even for very simple and natural learning problems. In particular, for many problems, the minimax label complexity of active learning will be no better than that of passive learning. In fact, Balcan, Hanneke, and Vaughan (2010) later showed that, for a certain type of active learning algorithm—namely, *self-verifying* algorithms, which themselves adaptively determine how many label requests they need to achieve a given accuracy—there are even particular target concepts and data distributions for which *no* active learning algorithm of that type can outperform passive learning. Since all of the above label complexity analyses (splitting index, teaching dimension, disagreement coefficient) apply to certain respective self-verifying learning algorithms, these negative results are also reflected in all of the existing general label complexity analyses.

While at first these negative results may seem discouraging, Balcan, Hanneke, and Vaughan (2010) noted that if we do not require the algorithm to be self-verifying, instead simply measuring the number of label requests the algorithm needs to *find* a good classifier, rather than the number needed to both find a good classifier *and verify* that it is indeed good, then these negative results vanish. In fact, (shockingly) they were able to show that for any concept space with finite VC dimension, and any fixed data distribution, for any given passive learning algorithm there is an active learning algorithm with asymptotically superior label complexity for *every* nontrivial target concept! A positive result of this generality and strength is certainly an exciting advance in our understanding of the advantages of active learning. But perhaps equally exciting are the unresolved questions raised by that work, as there are potential opportunities to strengthen, generalize, simplify, and elaborate on this result. First, note that the above statement allows the active learning algorithm to be specialized to the particular distribution according to which the (unlabeled) data are sampled, and indeed the active learning method used by Balcan, Hanneke, and Vaughan (2010) in their proof has a rather strong direct dependence on the data distribution (which cannot be removed by simply replacing some calculations with data-dependent estimators). One interesting question is whether an alternative approach might avoid this direct distribution-dependence in the algorithm, so that the claim can be strengthened to say that the active algorithm is superior to the passive algorithm for all nontrivial target concepts *and data distributions*. This question is interesting both theoretically, in order to obtain the strongest possible theorem on the advantages of active learning, as well as practically, since direct access to the distribution from which the data are sampled is typically *not* available in practical learning scenarios. A second question left open by Balcan, Hanneke, and Vaughan (2010) regards the *magnitude* of the gap between the active and passive label complexities. Specifically, although they did find particularly nasty learning problems where the label complexity of active learning will be close to that of passive learning (though always better), they hypothesized that for most natural learning problems, the improvements over passive learning should typically be *exponentially large* (as is the case for threshold classifiers); they gave many examples to illustrate this point, but left open the problem of characterizing general sufficient conditions for these exponential improvements to be achievable, even when they are not achievable by self-verifying algorithms. Another question left unresolved by Balcan, Hanneke, and Vaughan (2010) is whether this type of general improvement guarantee might be realized by a computationally *efficient* active learning algorithm. Finally, they left open the question of whether such general results might be further generalized to settings that involve noisy labels. The present work picks up where Balcan, Hanneke, and Vaughan (2010) left off in several respects, making progress on each of the above questions, in some cases completely resolving the question.

1.1.2 The Agnostic Case

In addition to the above advances in our understanding of active learning in the realizable case, there has also been wonderful progress in making these methods robust to imperfect teachers, feature space underspecification, and model misspecification. This general topic goes by the name *agnostic active learning*, from its roots in the agnostic PAC model (Kearns, Schapire, and Sellie, 1994). In contrast to the realizable case, in the *agnostic case*, there is not necessarily a perfect classifier of a known form, and indeed there may even be *label noise* so that there is no perfect classifier of *any* form. Rather, we have a given set of classifiers (e.g., linear separators, or depth-limited decision trees, etc.), and the objective is to identify a classifier whose accuracy is not much worse than the best classifier of that type. Agnostic learning is strictly more general, and often more difficult, than realizable learning; this is true for both passive learning and active learning. However, for a given agnostic learning problem, we might still hope that active learning can achieve a given accuracy using fewer labels than required for passive learning.

The general topic of agnostic active learning got its first taste of real progress from Balcan, Beygelzimer, and Langford (2006a, 2009) with the publication of the A^2 (agnostic active) algorithm. This method is a noise-robust disagreement-based algorithm, which can be applied with essentially arbitrary types of classifiers under arbitrary noise distributions. It is interesting both for its effectiveness and (as with CAL) its elegance. The original work of Balcan, Beygelzimer, and Langford (2006a, 2009) showed that, in some special cases (thresholds, and homogeneous linear separators under a uniform distribution), the A^2 algorithm does achieve improved label complexities compared to the known results for passive learning.

Using a different type of general active learning strategy, Hanneke (2007a) found that the *teaching dimension* analysis (discussed above for the realizable case) can be extended beyond the realizable case, arriving at general bounds on the label complexity under arbitrary noise distributions.

These bounds improve over the known results for passive learning in many cases. However, the algorithm requires direct access to a certain quantity that depends on the noise distribution (namely, the noise rate, defined in Section 6 below), which would not be available in many real-world learning problems.

Later, Hanneke (2007b) established a general characterization of the label complexities achieved by A^2, expressed in terms of the disagreement coefficient. The result holds for arbitrary types of classifiers (of finite VC dimension) and arbitrary noise distributions, and represents the natural generalization of the aforementioned realizable-case analysis of CAL. In many cases, this result shows improvements over the known results for passive learning. Furthermore, because of the simplicity of the disagreement coefficient, the bound can be calculated for a variety of natural learning problems.

Soon after this, Dasgupta, Hsu, and Monteleoni (2007) proposed a new active learning strategy, which is also effective in the agnostic setting. Like A^2, the new algorithm is a noise-robust disagreement-based method. The work of Dasgupta, Hsu, and Monteleoni (2007) is significant for at least two reasons. First, they were able to establish a general label complexity bound for this method based on the disagreement coefficient. The bound is similar in form to the previous label complexity bound for A^2 by Hanneke (2007b), but improves the dependence of the bound on the disagreement coefficient. Second, the proposed method of Dasgupta, Hsu, and Monteleoni (2007) set a new standard for computational and aesthetic simplicity in agnostic active learning algorithms. This work has since been followed by related methods of Beygelzimer, Dasgupta, and Langford (2009) and Beygelzimer, Hsu, Langford, and Zhang (2010). In particular, Beygelzimer, Dasgupta, and Langford (2009) develop a method capable of learning under an essentially arbitrary loss function; they also show label complexity bounds similar to those of Dasgupta, Hsu, and Monteleoni (2007), but applicable to a larger class of loss functions, and stated in terms of a generalization of the disagreement coefficient for arbitrary loss functions.

While the above results are encouraging, the guarantees reflected in these label complexity bounds essentially take the form of (at best) constant factor improvements; specifically, in some cases the bounds improve the dependence on the noise rate factor (defined in Section 6 below), compared to the known results for passive learning. In fact, Kääriäinen (2006) showed that any label complexity bound depending on the noise distribution only via the noise rate cannot do better than this type of constant-factor improvement. This raised the question of whether, with a more detailed description of the noise distribution, one can show improvements in the *asymptotic form* of the label complexity compared to passive learning. Toward this end, Castro and Nowak (2008) studied a certain refined description of the noise conditions, related to the margin conditions of Mammen and Tsybakov (1999), which are well-studied in the passive learning literature. Specifically, they found that in some special cases, under certain restrictions on the noise distribution, the asymptotic form of the label complexity *can* be improved compared to passive learning, and in some cases the improvements can even be *exponential* in magnitude; to achieve this, they developed algorithms specifically tailored to the types of classifiers they studied (threshold classifiers and boundary fragment classes). Balcan, Broder, and Zhang (2007) later extended this result to general homogeneous linear separators under a uniform distribution. Following this, Hanneke (2009a, 2011) generalized these results, showing that both of the published general agnostic active learning algorithms (Balcan, Beygelzimer, and Langford, 2009; Dasgupta, Hsu, and Monteleoni, 2007) can also achieve these types of improvements in the asymptotic form of the label complexity; he further proved general bounds on the label complexities of these methods, again based on the disagreement coefficient, which apply to arbitrary types of classifiers, and which reflect these types of improvements (under conditions on the disagreement coefficient). Wang (2009) later bounded the label complexity of A^2 under somewhat different noise conditions, in particular identifying weaker noise conditions sufficient for these improvements to be exponential in magnitude (again, under conditions on the disagreement coefficient). Koltchinskii (2010) has recently improved on some of Hanneke's results, refining certain logarithmic factors and simplifying the proofs, using a slightly different algorithm based on similar principles. Though the present work discusses only classes of finite VC dimension, most of the above references also contain results for various types of nonparametric classes with infinite VC dimension.

At present, all of the published bounds on the label complexity of agnostic active learning also apply to *self-verifying* algorithms. As mentioned, in the realizable case, it is typically possible to achieve significantly better label complexities if we do not require the active learning algorithm to be self-verifying, since the verification of learning may be more difficult than the learning itself (Balcan, Hanneke, and Vaughan, 2010). We might wonder whether this is also true in the agnostic case, and whether agnostic active learning algorithms that are not self-verifying might possibly achieve significantly better label complexities than the existing label complexity bounds described above. We investigate this in depth below.

1.2 Summary of Contributions

In the present work, we build on and extend the above results in a variety of ways, resolving a number of open problems. The main contributions of this work can be summarized as follows.

- We formally define a notion of a universal activizer, a meta-algorithm that transforms any passive learning algorithm into an active learning algorithm with asymptotically strictly superior label complexities for all nontrivial distributions and target concepts in the concept space.
- We analyze the existing strategy of disagreement-based active learning from this perspective, precisely characterizing the conditions under which this strategy can lead to a universal activizer for VC classes in the realizable case.
- We propose a new type of active learning algorithm, based on shatterable sets, and construct universal activizers for all VC classes in the realizable case based on this idea; in particular, this overcomes the issue of distribution-dependence in the existing results mentioned above.
- We present a novel generalization of the disagreement coefficient, along with a new asymptotic bound on the label complexities achievable by active learning in the realizable case; this new bound is often significantly smaller than the existing results in the published literature.
- We state new concise sufficient conditions for exponential improvements over passive learning to be achievable in the realizable case, including a significant weakening of known conditions in the published literature.
- We present a new general-purpose active learning algorithm for the agnostic case, based on the aforementioned idea involving shatterable sets.
- We prove a new asymptotic bound on the label complexities achievable by active learning in the presence of label noise (the agnostic case), often significantly smaller than any previously published results.

- We formulate a general conjecture on the theoretical advantages of active learning over passive learning in the presence of arbitrary types of label noise.

1.3 Outline of the Paper

The paper is organized as follows. In Section 2, we introduce the basic notation used throughout, formally define the learning protocol, and formally define the label complexity. We also define the notion of an *activizer*, which is a procedure that transforms a passive learning algorithm into an active learning algorithm with asymptotically superior label complexity. In Section 3, we review the established technique of *disagreement-based* active learning, and prove a new result precisely characterizing the scenarios in which disagreement-based active learning can be used to construct an activizer. In particular, we find that in many scenarios, disagreement-based active learning is not powerful enough to provide the desired improvements. In Section 4, we move beyond disagreement-based active learning, developing a new type of active learning algorithm based on *shatterable* sets of points. We apply this technique to construct a simple 3-stage procedure, which we then prove is a universal activizer for any concept space of finite VC dimension. In Section 5, we begin by reviewing the known results for bounding the label complexity of disagreement-based active learning in terms of the disagreement coefficient; we then develop a somewhat more involved procedure, again based on shatterable sets, which takes full advantage of the sequential nature of active learning. In addition to being an activizer, we show that this procedure often achieves dramatically superior label complexities than achievable by passive learning. In particular, we define a novel generalization of the disagreement coefficient, and use it to bound the label complexity of this procedure. This also provides us with concise sufficient conditions for obtaining exponential improvements over passive learning. Continuing in Section 6, we extend our framework to allow for label noise (the agnostic case), and discuss the possibility of extending the results from previous sections to these noisy learning problems. We first review the known results for noise-robust disagreement-based active learning, and characterizations of its label complexity in terms of the disagreement coefficient and Mammen-Tsybakov noise parameters. We then proceed to develop a new type of noise-robust active learning algorithm, again based on shatterable sets, and prove bounds on its label complexity in terms of our aforementioned generalization of the disagreement coefficient. Additionally, we present a general conjecture concerning the existence of activizers for certain passive learning algorithms in the agnostic case. We conclude in Section 7 with a host of enticing open problems for future investigation.

2. Definitions and Notation

For most of the paper, we consider the following formal setting. There is a measurable space $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$, where $\mathcal{X}$ is called the *instance space*; for simplicity, we suppose this is a standard Borel space (Srivastava, 1998) (e.g., $\mathbb{R}^m$ under the usual Borel σ-algebra), though most of the results generalize. A *classifier* is any measurable function $h : \mathcal{X} \to \{-1, +1\}$. There is a set $\mathbb{C}$ of classifiers called the *concept space*. In the *realizable case*, the learning problem is characterized as follows. There is a probability measure $\mathcal{P}$ on $\mathcal{X}$, and a sequence $\mathcal{Z}_X = \{X_1, X_2, \ldots\}$ of independent $\mathcal{X}$-valued random variables, each with distribution $\mathcal{P}$. We refer to these random variables as the sequence of *unlabeled examples*; although in practice, this sequence would typically be large but finite, to simplify the discussion and focus strictly on counting labels, we will suppose this sequence is inexhaustible. There is additionally a special element $f \in \mathbb{C}$, called the *target function*, and we denote by

$Y_i = f(X_i)$; we further denote by $\mathcal{Z} = \{(X_1,Y_1),(X_2,Y_2),\ldots\}$ the sequence of *labeled examples*, and for $m \in \mathbb{N}$ we denote by $\mathcal{Z}_m = \{(X_1,Y_1),(X_2,Y_2),\ldots,(X_m,Y_m)\}$ the finite subsequence consisting of the first m elements of $\mathcal{Z}$. For any classifier h, we define the *error rate* $\mathrm{er}(h) = \mathcal{P}(x : h(x) \neq f(x))$. Informally, the learning objective in the realizable case is to identify some h with small $\mathrm{er}(h)$ using elements from $\mathcal{Z}$, without direct access to f.

An *active learning algorithm* $\mathcal{A}$ is permitted direct access to the $\mathcal{Z}_X$ sequence (the unlabeled examples), but to gain access to the Y_i values it must request them one at a time, in a sequential manner. Specifically, given access to the $\mathcal{Z}_X$ values, the algorithm selects any index $i \in \mathbb{N}$, requests to observe the Y_i value, then having observed the value of Y_i, selects another index i', observes the value of $Y_{i'}$, etc. The algorithm is given as input an integer n, called the *label budget*, and is permitted to observe at most n labels total before eventually halting and returning a classifier $\hat{h}_n = \mathcal{A}(n)$; that is, by definition, an active learning algorithm never attempts to access more than the given budget n number of labels. We will then study the values of n sufficient to guarantee $\mathbb{E}[\mathrm{er}(\hat{h}_n)] \leq \varepsilon$, for any given value $\varepsilon \in (0,1)$. We refer to this as the *label complexity*. We will be particularly interested in the asymptotic dependence on ε in the label complexity, as $\varepsilon \to 0$. Formally, we have the following definition.

Definition 1 *An active learning algorithm $\mathcal{A}$ achieves label complexity $\Lambda(\cdot,\cdot,\cdot)$ if, for every target function f, distribution $\mathcal{P}$, $\varepsilon \in (0,1)$, and integer $n \geq \Lambda(\varepsilon, f, \mathcal{P})$, we have $\mathbb{E}[\mathrm{er}(\mathcal{A}(n))] \leq \varepsilon$.*

This definition of label complexity is similar to one originally studied by Balcan, Hanneke, and Vaughan (2010). It has a few features worth noting. First, the label complexity has an explicit dependence on the target function f and distribution $\mathcal{P}$. As noted by Dasgupta (2005), we need this dependence if we are to fully understand the range of label complexities achievable by active learning; we further illustrate this issue in the examples below. The second feature to note is that the label complexity, as defined here, is simply a sufficient budget size to achieve the specified accuracy. That is, here we are asking only how many label requests are required for the algorithm to achieve a given accuracy (in expectation). However, as noted by Balcan, Hanneke, and Vaughan (2010), this number might not be sufficiently large to *detect* that the algorithm has indeed achieved the required accuracy based only on the observed data. That is, because the number of labeled examples used in active learning can be quite small, we come across the problem that the number of labels needed to *learn* a concept might be significantly smaller than the number of labels needed to *verify* that we have successfully learned the concept. As such, this notion of label complexity is most useful in the *design* of effective learning algorithms, rather than for predicting the number of labels an algorithm should request in any particular application. Specifically, to design effective active learning algorithms, we should generally desire small label complexity values, so that (in the extreme case) if some algorithm $\mathcal{A}$ has smaller label complexity values than some other algorithm $\mathcal{A}'$ for *all* target functions and distributions, then (all other factors being equal) we should clearly prefer algorithm $\mathcal{A}$ over algorithm $\mathcal{A}'$; this is true regardless of whether we have a means to *detect* (verify) how large the improvements offered by algorithm $\mathcal{A}$ over algorithm $\mathcal{A}'$ are for any particular application. Thus, in our present context, performance guarantees in terms of this notion of label complexity play a role analogous to concepts such as *universal consistency* or *admissibility*, which are also generally useful in guiding the design of effective algorithms, but are not intended to be informative in the context of any particular application. See the work of Balcan, Hanneke, and Vaughan (2010) for a discussion of this issue, as it relates to a definition of label complexity similar to that above, as well as other notions of label complexity from the active learning literature (some of which include a verification requirement).

We will be interested in the performance of active learning algorithms, relative to the performance of a given *passive learning algorithm*. In this context, a passive learning algorithm $\mathcal{A}$ takes as input a finite sequence of labeled examples $\mathcal{L} \in \bigcup_n (\mathcal{X} \times \{-1,+1\})^n$, and returns a classifier $\hat{h} = \mathcal{A}(\mathcal{L})$. We allow both active and passive learning algorithms to be randomized: that is, to have independent internal randomness, in addition to the given random data. We define the label complexity for a passive learning algorithm as follows.

Definition 2 *A passive learning algorithm $\mathcal{A}$ achieves label complexity $\Lambda(\cdot,\cdot,\cdot)$ if, for every target function f, distribution $\mathcal{P}$, $\varepsilon \in (0,1)$, and integer $n \geq \Lambda(\varepsilon, f, \mathcal{P})$, we have $\mathbb{E}\left[\mathrm{er}\left(\mathcal{A}\left(\mathcal{Z}_n\right)\right)\right] \leq \varepsilon$.*

Although technically some algorithms may be able to achieve a desired accuracy without any observations, to make the general results easier to state (namely, those in Section 5), unless otherwise stated we suppose label complexities (both passive and active) take strictly positive values, among $\mathbb{N} \cup \{\infty\}$; note that label complexities (both passive and active) can be infinite, indicating that the corresponding algorithm might not achieve expected error rate ε for *any* $n \in \mathbb{N}$. Both the passive and active label complexities are defined as a number of labels sufficient to guarantee the *expected* error rate is at most ε. It is also common in the literature to discuss the number of label requests sufficient to guarantee the error rate is at most ε with *high probability* $1-\delta$ (e.g., Balcan, Hanneke, and Vaughan, 2010). In the present work, we formulate our results in terms of the expected error rate because it simplifies the discussion of asymptotics, in that we need only study the behavior of the label complexity as the single argument ε approaches 0, rather than the more complicated behavior of a function of ε and δ as both ε and δ approach 0 at various relative rates. However, we note that analogous results for these high-probability guarantees on the error rate can be extracted from the proofs below without much difficulty, and in several places we explicitly state results of this form.

Below we employ the standard notation from asymptotic analysis, including $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, $\omega(\cdot)$, $\Theta(\cdot)$, $\ll$, and $\gg$. In all contexts below not otherwise specified, the asymptotics are always considered as $\varepsilon \to 0$ when considering a function of ε, and as $n \to \infty$ when considering a function of n; also, in any expression of the form "$x \to 0$," we always mean the limit *from above* (i.e., $x \downarrow 0$). For instance, when considering nonnegative functions of ε, $\lambda_a(\varepsilon)$ and $\lambda_p(\varepsilon)$, the above notations are defined as follows. We say $\lambda_a(\varepsilon) = o(\lambda_p(\varepsilon))$ when $\lim\limits_{\varepsilon\to 0} \frac{\lambda_a(\varepsilon)}{\lambda_p(\varepsilon)} = 0$, and this is equivalent to writing $\lambda_p(\varepsilon) = \omega(\lambda_a(\varepsilon))$, $\lambda_a(\varepsilon) \ll \lambda_p(\varepsilon)$, or $\lambda_p(\varepsilon) \gg \lambda_a(\varepsilon)$. We say $\lambda_a(\varepsilon) = O(\lambda_p(\varepsilon))$ when $\limsup\limits_{\varepsilon\to 0} \frac{\lambda_a(\varepsilon)}{\lambda_p(\varepsilon)} < \infty$, which can equivalently be expressed as $\lambda_p(\varepsilon) = \Omega(\lambda_a(\varepsilon))$. Finally, we write $\lambda_a(\varepsilon) = \Theta(\lambda_p(\varepsilon))$ to mean that both $\lambda_a(\varepsilon) = O(\lambda_p(\varepsilon))$ and $\lambda_a(\varepsilon) = \Omega(\lambda_p(\varepsilon))$ are satisfied. We also use the standard notation for the limit of a sequence of sets, such as $\lim\limits_{r\to 0} A_r$, defined by the property $\mathbb{1}_{\lim\limits_{r\to 0} A_r} = \lim\limits_{r\to 0} \mathbb{1}_{A_r}$ (if the latter exists), where $\mathbb{1}_A$ is the indicator function for the set A.

Define the class of functions $\mathrm{Polylog}(1/\varepsilon)$ as those $g : (0,1) \to [0,\infty)$ such that, for some $k \in [0,\infty)$, $g(\varepsilon) = O(\log^k(1/\varepsilon))$. For a label complexity Λ, also define the set $\mathrm{Nontrivial}(\Lambda)$ as the collection of all pairs $(f,\mathcal{P})$ of a classifier and a distribution such that, $\forall \varepsilon > 0, \Lambda(\varepsilon, f, \mathcal{P}) < \infty$, and $\forall g \in \mathrm{Polylog}(1/\varepsilon), \Lambda(\varepsilon, f, \mathcal{P}) = \omega(g(\varepsilon))$.

In this context, an *active meta-algorithm* is a procedure $\mathcal{A}_a$ taking as input a passive algorithm $\mathcal{A}_p$ and a label budget n, such that for any passive algorithm $\mathcal{A}_p$, $\mathcal{A}_a(\mathcal{A}_p, \cdot)$ is an active learning algorithm. We define an *activizer* for a given passive algorithm as follows.

Definition 3 *We say an active meta-algorithm $\mathcal{A}_a$* activizes *a passive algorithm $\mathcal{A}_p$ for a concept space $\mathbb{C}$ if the following holds. For any label complexity Λ_p achieved by $\mathcal{A}_p$, the active learning algorithm $\mathcal{A}_a(\mathcal{A}_p, \cdot)$ achieves a label complexity Λ_a such that, for every $f \in \mathbb{C}$ and every distribution $\mathcal{P}$ on $\mathcal{X}$ with $(f, \mathcal{P}) \in \mathrm{Nontrivial}(\Lambda_p)$, there exists a constant $c \in [1, \infty)$ such that*

$$\Lambda_a(c\varepsilon, f, \mathcal{P}) = o\left(\Lambda_p(\varepsilon, f, \mathcal{P})\right).$$

In this case, $\mathcal{A}_a$ is called an activizer *for $\mathcal{A}_p$ with respect to $\mathbb{C}$, and the active learning algorithm $\mathcal{A}_a(\mathcal{A}_p, \cdot)$ is called the $\mathcal{A}_a$-*activized *$\mathcal{A}_p$.*

We also refer to any active meta-algorithm $\mathcal{A}_a$ that activizes *every* passive algorithm $\mathcal{A}_p$ for $\mathbb{C}$ as a *universal activizer* for $\mathbb{C}$. One of the main contributions of this work is establishing that such universal activizers do exist for any VC class $\mathbb{C}$.

A bit of explanation is in order regarding Definition 3. We might interpret it as follows: an *activizer* for $\mathcal{A}_p$ strongly improves (in a little-o sense) the label complexity for all *nontrivial* target functions and distributions. Here, we seek a meta-algorithm that, when given $\mathcal{A}_p$ as input, results in an active learning algorithm with strictly superior label complexities. However, there is a sense in which some distributions $\mathcal{P}$ or target functions f are *trivial* relative to $\mathcal{A}_p$. For instance, perhaps $\mathcal{A}_p$ has a *default* classifier that it is naturally biased toward (e.g., with minimal $\mathcal{P}(x : h(x) = +1)$, as in the Closure algorithm of Helmbold, Sloan, and Warmuth, 1990), so that when this default classifier is the target function, $\mathcal{A}_p$ achieves a constant label complexity. In these trivial scenarios, we cannot hope to *improve* over the behavior of the passive algorithm, but instead can only hope to *compete* with it. The *sense* in which we wish to compete may be a subject of some controversy, but the implication of Definition 3 is that the label complexity of the activized algorithm should be strictly better than every nontrivial upper bound on the label complexity of the passive algorithm. For instance, if $\Lambda_p(\varepsilon, f, \mathcal{P}) \in \mathrm{Polylog}(1/\varepsilon)$, then we are guaranteed $\Lambda_a(\varepsilon, f, \mathcal{P}) \in \mathrm{Polylog}(1/\varepsilon)$ as well, but if $\Lambda_p(\varepsilon, f, \mathcal{P}) = O(1)$, we are still only guaranteed $\Lambda_a(\varepsilon, f, \mathcal{P}) \in \mathrm{Polylog}(1/\varepsilon)$. This serves the purpose of defining a framework that can be studied without requiring too much obsession over small additive terms in trivial scenarios, thus focusing the analyst's efforts toward nontrivial scenarios where $\mathcal{A}_p$ has relatively *large* label complexity, which are precisely the scenarios for which active learning is truly needed. In our proofs, we find that in fact $\mathrm{Polylog}(1/\varepsilon)$ can be replaced with $O(\log(1/\varepsilon))$, giving a slightly broader definition of "nontrivial," for which all of the results below still hold. Section 7 discusses open problems regarding this issue of trivial problems.

The definition of $\mathrm{Nontrivial}(\cdot)$ also only requires the activized algorithm to be effective in scenarios where the passive learning algorithm has *reasonable* behavior (i.e., finite label complexities); this is only intended to keep with the reduction-based style of the framework, and in fact this restriction can easily be lifted using a trick from Balcan, Hanneke, and Vaughan (2010) (aggregating the activized algorithm with another algorithm that is always reasonable).

Finally, we also allow a constant factor c loss in the ε argument to Λ_a. We allow this to be an arbitrary constant, again in the interest of allowing the analyst to focus only on the most significant aspects of the problem; for most reasonable passive learning algorithms, we typically expect $\Lambda_p(\varepsilon, f, \mathcal{P}) = \mathrm{Poly}(1/\varepsilon)$, in which case c can be set to 1 by adjusting the leading constant factors of

Λ_a. A careful inspection of our proofs reveals that c can always be set arbitrarily close to 1 without affecting the theorems below (and in fact, we can even get $c = (1+o(1))$, a function of ε).

Throughout this work, we will adopt the usual notation for probabilities, such as $\mathbb{P}(\mathrm{er}(\hat{h}) > \varepsilon)$, and as usual we interpret this as measuring the corresponding event in the (implicit) underlying probability space. In particular, we make the usual implicit assumption that all sets involved in the analysis are measurable; where this assumption does not hold, we may turn to outer probabilities, though we will not make further mention of these technical details. We will also use the notation $P^k(\cdot)$ to represent k-dimensional product measures; for instance, for a measurable set $A \subseteq \mathcal{X}^k$, $\mathcal{P}^k(A) = \mathbb{P}((X'_1,\ldots,X'_k) \in A)$, for independent $\mathcal{P}$-distributed random variables $X'_1,\ldots,X'_k$. Additionally, to simplify notation, we will adopt the convention that $\mathcal{X}^0 = \{\varnothing\}$ and $\mathcal{P}^0(\mathcal{X}^0) = 1$. Throughout, we will denote by $\mathbb{1}_A(z)$ the indicator function for a set A, which has the value 1 when $z \in A$ and 0 otherwise; additionally, at times it will be more convenient to use the bipolar indicator function, defined as $\mathbb{1}^{\pm}_A(z) = 2\mathbb{1}_A(z) - 1$.

We will require a few additional definitions for the discussion below. For any classifier $h : \mathcal{X} \to \{-1,+1\}$ and finite sequence of labeled examples $\mathcal{L} \in \bigcup_m (\mathcal{X} \times \{-1,+1\})^m$, define the *empirical error rate* $\mathrm{er}_{\mathcal{L}}(h) = |\mathcal{L}|^{-1} \sum_{(x,y)\in\mathcal{L}} \mathbb{1}_{\{-y\}}(h(x))$; for completeness, define $\mathrm{er}_{\emptyset}(h) = 0$. Also, for $\mathcal{L} = \mathcal{Z}_m$, the first m labeled examples in the data sequence, abbreviate this as $\mathrm{er}_m(h) = \mathrm{er}_{\mathcal{Z}_m}(h)$. For any probability measure P on $\mathcal{X}$, set of classifiers $\mathcal{H}$, classifier h, and $r > 0$, define $\mathrm{B}_{\mathcal{H},P}(h,r) = \{g \in \mathcal{H} : P(x : h(x) \neq g(x)) \leq r\}$; when $P = \mathcal{P}$, the distribution of the unlabeled examples, and $\mathcal{P}$ is clear from the context, we abbreviate this as $\mathrm{B}_{\mathcal{H}}(h,r) = \mathrm{B}_{\mathcal{H},\mathcal{P}}(h,r)$; furthermore, when $P = \mathcal{P}$ and $\mathcal{H} = \mathbb{C}$, the concept space, and both $\mathcal{P}$ and $\mathbb{C}$ are clear from the context, we abbreviate this as $\mathrm{B}(h,r) = \mathrm{B}_{\mathbb{C},\mathcal{P}}(h,r)$. Also, for any set of classifiers $\mathcal{H}$, and any sequence of labeled examples $\mathcal{L} \in \bigcup_m (\mathcal{X} \times \{-1,+1\})^m$, define $\mathcal{H}[\mathcal{L}] = \{h \in \mathcal{H} : \mathrm{er}_{\mathcal{L}}(h) = 0\}$; for any $(x,y) \in \mathcal{X} \times \{-1,+1\}$, abbreviate $\mathcal{H}[(x,y)] = \mathcal{H}[\{(x,y)\}] = \{h \in \mathcal{H} : h(x) = y\}$.

We also adopt the usual definition of "shattering" used in learning theory (e.g., Vapnik, 1998). Specifically, for any set of classifiers $\mathcal{H}$, $k \in \mathbb{N}$, and $S = (x_1,\ldots,x_k) \in \mathcal{X}^k$, we say $\mathcal{H}$ *shatters* S if, $\forall (y_1,\ldots,y_k) \in \{-1,+1\}^k$, $\exists h \in \mathcal{H}$ such that $\forall i \in \{1,\ldots,k\}$, $h(x_i) = y_i$; equivalently, $\mathcal{H}$ shatters S if $\exists \{h_1,\ldots,h_{2^k}\} \subseteq \mathcal{H}$ such that for each $i,j \in \{1,\ldots,2^k\}$ with $i \neq j$, $\exists \ell \in \{1,\ldots,k\}$ with $h_i(x_\ell) \neq h_j(x_\ell)$. To simplify notation, we will also say that $\mathcal{H}$ shatters $\varnothing$ if and only if $\mathcal{H} \neq \{\}$. As usual, we define the *VC dimension* of $\mathbb{C}$, denoted d, as the largest integer k such that $\exists S \in \mathcal{X}^k$ shattered by $\mathbb{C}$ (Vapnik and Chervonenkis, 1971; Vapnik, 1998). To focus on nontrivial problems, we will only consider concept spaces $\mathbb{C}$ with $d > 0$ in the results below. Generally, any such concept space $\mathbb{C}$ with $d < \infty$ is called a *VC class*.

2.1 Motivating Examples

Throughout this paper, we will repeatedly refer to a few canonical examples. Although themselves quite toy-like, they represent the boiled-down essence of some important distinctions between various types of learning problems. In some sense, the process of grappling with the fundamental distinctions raised by these types of examples has been a driving force behind much of the recent progress in understanding the label complexity of active learning.

The first example is perhaps the most classic, and is clearly the first that comes to mind when considering the potential for active learning to provide strong improvements over passive learning.

Example 1 *In the problem of learning* threshold *classifiers, we consider* $\mathcal{X} = [0,1]$ *and* $\mathbb{C} = \{h_z(x) = \mathbb{1}^{\pm}_{[z,1]}(x) : z \in (0,1)\}$.

There is a simple universal activizer for threshold classifiers, based on a kind of binary search. Specifically, suppose $n \in \mathbb{N}$ and that $\mathcal{A}_p$ is any given passive learning algorithm. Consider the points in $\{X_1, X_2, \ldots, X_m\}$, for $m = 2^{n-1}$, and sort them in increasing order: $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$. Also initialize $\ell = 0$ and $u = m+1$, and define $X_{(0)} = 0$ and $X_{(m+1)} = 1$. Now request the label of $X_{(i)}$ for $i = \lfloor (\ell + u)/2 \rfloor$ (i.e., the median point between ℓ and u); if the label is -1, let $\ell = i$, and otherwise let $u = i$; repeat this (requesting this median point, then updating ℓ or u accordingly) until we have $u = \ell + 1$. Finally, let $\hat{z} = X_{(u)}$, construct the labeled sequence $\mathcal{L} = \{(X_1, h_{\hat{z}}(X_1)), \ldots, (X_m, h_{\hat{z}}(X_m))\}$, and return the classifier $\hat{h} = \mathcal{A}_p(\mathcal{L})$.

Since each label request at least halves the set of integers between ℓ and u, the total number of label requests is at most $\log_2(m) + 1 = n$. Supposing $f \in \mathbb{C}$ is the target function, this procedure maintains the invariant that $f(X_{(\ell)}) = -1$ and $f(X_{(u)}) = +1$. Thus, once we reach $u = \ell + 1$, since f is a threshold, it must be some h_z with $z \in (\ell, u]$; therefore every $X_{(j)}$ with $j \leq \ell$ has $f(X_{(j)}) = -1$, and likewise every $X_{(j)}$ with $j \geq u$ has $f(X_{(j)}) = +1$; in particular, this means $\mathcal{L}$ equals $\mathcal{Z}_m$, the *true* labeled sequence. But this means $\hat{h} = \mathcal{A}_p(\mathcal{Z}_m)$. Since $n = \log_2(m) + 1$, this active learning algorithm will achieve an equivalent error rate to what $\mathcal{A}_p$ achieves with m labeled examples, but using only $\log_2(m) + 1$ label requests. In particular, this implies that if $\mathcal{A}_p$ achieves label complexity Λ_p, then this active learning algorithm achieves label complexity Λ_a such that $\Lambda_a(\varepsilon, f, \mathcal{P}) \leq \log_2 \Lambda_p(\varepsilon, f, \mathcal{P}) + 2$; as long as $1 \ll \Lambda_p(\varepsilon, f, \mathcal{P}) < \infty$, this is $o(\Lambda_p(\varepsilon, f, \mathcal{P}))$, so that this procedure activizes $\mathcal{A}_p$ for $\mathbb{C}$.

The second example we consider is almost equally simple (only increasing the VC dimension from 1 to 2), but is far more subtle in terms of how we must approach its analysis in active learning.

Example 2 *In the problem of learning* interval *classifiers, we consider* $\mathcal{X} = [0,1]$ *and* $\mathbb{C} = \{h_{[a,b]}(x) = \mathbb{1}^{\pm}_{[a,b]}(x) : 0 < a \leq b < 1\}$.

For the intervals problem, we can also construct a universal activizer, though slightly more complicated. Specifically, suppose again that $n \in \mathbb{N}$ and that $\mathcal{A}_p$ is any given passive learning algorithm. We first request the labels $\{Y_1, Y_2, \ldots, Y_{\lceil n/2 \rceil}\}$ of the first $\lceil n/2 \rceil$ examples in the sequence. If every one of these labels is -1, then we immediately return the all-negative constant classifier $\hat{h}(x) = -1$. Otherwise, consider the points $\{X_1, X_2, \ldots, X_m\}$, for $m = \max\left\{2^{\lfloor n/4 \rfloor - 1}, n\right\}$, and sort them in increasing order $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$. For some value $i \in \{1, \ldots, \lceil n/2 \rceil\}$ with $Y_i = +1$, let j_+ denote the corresponding index j such that $X_{(j)} = X_i$. Also initialize $\ell_1 = 0$, $u_1 = \ell_2 = j_+$, and $u_2 = m+1$, and define $X_{(0)} = 0$ and $X_{(m+1)} = 1$. Now if $\ell_1 + 1 < u_1$, request the label of $X_{(i)}$ for $i = \lfloor (\ell_1 + u_1)/2 \rfloor$ (the median point between ℓ_1 and u_1); if the label is -1, let $\ell_1 = i$, and otherwise let $u_1 = i$; repeat this (requesting this median point, then updating ℓ_1 or u_1 accordingly) until we have $u_1 = \ell_1 + 1$. Now if $\ell_2 + 1 < u_2$, request the label of $X_{(i)}$ for $i = \lfloor (\ell_2 + u_2)/2 \rfloor$ (the median point between ℓ_2 and u_2); if the label is -1, let $u_2 = i$, and otherwise let $\ell_2 = i$; repeat this (requesting this median point, then updating u_2 or ℓ_2 accordingly) until we have $u_2 = \ell_2 + 1$. Finally, let $\hat{a} = u_1$ and $\hat{b} = \ell_2$, construct the labeled sequence $\mathcal{L} = \left\{\left(X_1, h_{[\hat{a},\hat{b}]}(X_1)\right), \ldots, \left(X_m, h_{[\hat{a},\hat{b}]}(X_m)\right)\right\}$, and return the classifier $\hat{h} = \mathcal{A}_p(\mathcal{L})$.

Since each label request in the second phase halves the set of values between either ℓ_1 and u_1 or ℓ_2 and u_2, the total number of label requests is at most $\min\{m, \lceil n/2 \rceil + 2\log_2(m) + 2\} \leq n$. Suppose $f \in \mathbb{C}$ is the target function, and let $w(f) = \mathcal{P}(x : f(x) = +1)$. If $w(f) = 0$, then with probability 1 the algorithm will return the constant classifier $\hat{h}(x) = -1$, which has $\operatorname{er}(\hat{h}) = 0$ in this case. Otherwise, if $w(f) > 0$, then for any $n \geq \frac{2}{w(f)} \ln \frac{1}{\varepsilon}$, with probability at least $1 - \varepsilon$, there exists

$i \in \{1, \ldots, \lceil n/2 \rceil\}$ with $Y_i = +1$. Let H_+ denote the event that such an i exists. Supposing this is the case, the algorithm will make it into the second phase. In this case, the procedure maintains the invariant that $f(X_{(\ell_1)}) = -1$, $f(X_{(u_1)}) = f(X_{(\ell_2)}) = +1$, and $f(X_{(u_2)}) = -1$, where $\ell_1 < u_1 \le \ell_2 < u_2$. Thus, once we have $u_1 = \ell_1 + 1$ and $u_2 = \ell_2 + 1$, since f is an interval, it must be some $h_{[a,b]}$ with $a \in (\ell_1, u_1]$ and $b \in [\ell_2, u_1)$; therefore, every $X_{(j)}$ with $j \le \ell_1$ or $j \ge u_2$ has $f(X_{(j)}) = -1$, and likewise every $X_{(j)}$ with $u_1 \le j \le \ell_2$ has $f(X_{(j)}) = +1$; in particular, this means $\mathcal{L}$ equals $\mathcal{Z}_m$, the *true* labeled sequence. But this means $\hat{h} = \mathcal{A}_p(\mathcal{Z}_m)$. Supposing $\mathcal{A}_p$ achieves label complexity Λ_p, and that $n \ge \max\left\{8 + 4\log_2 \Lambda_p(\varepsilon, f, \mathcal{P}), \frac{2}{w(f)} \ln \frac{1}{\varepsilon}\right\}$, then $m \ge 2^{\lfloor n/4 \rfloor - 1} \ge \Lambda_p(\varepsilon, f, \mathcal{P})$ and $\mathbb{E}\left[\mathrm{er}(\hat{h})\right] \le \mathbb{E}\left[\mathrm{er}(\hat{h})\mathbb{1}_{H_+}\right] + (1 - \mathbb{P}(H_+)) \le \mathbb{E}\left[\mathrm{er}(\mathcal{A}_p(\mathcal{Z}_m))\right] + \varepsilon \le 2\varepsilon$. In particular, this means this active learning algorithm achieves label complexity Λ_a such that, for any $f \in \mathbb{C}$ with $w(f) = 0$, $\Lambda_a(2\varepsilon, f, \mathcal{P}) = 0$, and for any $f \in \mathbb{C}$ with $w(f) > 0$, $\Lambda_a(2\varepsilon, f, \mathcal{P}) \le \max\left\{8 + 4\log_2 \Lambda_p(\varepsilon, f, \mathcal{P}), \frac{2}{w(f)} \ln \frac{1}{\varepsilon}\right\}$. If $(f, \mathcal{P}) \in \mathrm{Nontrivial}(\Lambda_p)$, then $\frac{2}{w(f)} \ln \frac{1}{\varepsilon} = o(\Lambda_p(\varepsilon, f, \mathcal{P}))$ and $8 + 4\log_2 \Lambda_p(\varepsilon, f, \mathcal{P}) = o(\Lambda_p(\varepsilon, f, \mathcal{P}))$, so that $\Lambda_a(2\varepsilon, f, \mathcal{P}) = o(\Lambda_p(\varepsilon, f, \mathcal{P}))$. Therefore, this procedure activizes $\mathcal{A}_p$ for $\mathbb{C}$.

This example also brings to light some interesting phenomena in the analysis of the label complexity of active learning. Note that unlike the thresholds example, we have a much stronger dependence on the target function in these label complexity bounds, via the $w(f)$ quantity. This issue is fundamental to the problem, and cannot be avoided. In particular, when $\mathcal{P}([0,x])$ is continuous, this is the very issue that makes the *minimax* label complexity for this problem (i.e., $\min_{\Lambda_a} \max_{f \in \mathbb{C}} \Lambda_a(\varepsilon, f, \mathcal{P})$) *no better* than passive learning (Dasgupta, 2005). Thus, this problem emphasizes the need for any informative label complexity analysis of active learning to explicitly describe the dependence of the label complexity on the target function, as advocated by Dasgupta (2005). This example also highlights the *unverifiability* phenomenon explored by Balcan, Hanneke, and Vaughan (2010), since in the case of $w(f) = 0$, the error rate of the returned classifier is *zero*, but (for nondegenerate $\mathcal{P}$) there is no way for the algorithm to verify this fact based only on the finite number of labels it observes. In fact, Balcan, Hanneke, and Vaughan (2010) have shown that under continuous $\mathcal{P}$, for any $f \in \mathbb{C}$ with $w(f) = 0$, the number of labels required to both *find* a classifier of small error rate *and verify* that the error rate is small based only on observable quantities is essentially *no better* than for passive learning.

These issues are present to a small degree in the intervals example, but were easily handled in a very natural way. The target-dependence shows up only in an initial phase of waiting for a positive example, and the always-negative classifiers were handled by setting a *default* return value. However, we can amplify these issues so that they show up in more subtle and involved ways. Specifically, consider the following example, studied by Balcan, Hanneke, and Vaughan (2010).

Example 3 *In the problem of learning* unions of i intervals*, we consider* $\mathcal{X} = [0,1]$ *and* $\mathbb{C} = \left\{h_{\mathbf{z}}(x) = \mathbb{1}^{\pm}_{\bigcup_{j=1}^{i}[z_{2j-1}, z_{2j}]}(x) : 0 < z_1 \le z_2 \le \ldots \le z_{2i} < 1\right\}$.

The challenge of this problem is that, because sometimes $z_j = z_{j+1}$ for some j values, we do not know how many intervals are required to minimally represent the target function: only that it is at most i. This issue will be made clearer below. We can essentially think of any effective strategy here as having two components: one component that searches (perhaps randomly) with the purpose of identifying at least one example from each decision region, and another component that refines our estimates of the end-points of the regions the first component identifies. Later, we will go through the behavior of a universal activizer for this problem in detail.

3. Disagreement-Based Active Learning

At present, perhaps the best-understood active learning algorithms are those choosing their label requests based on disagreement among a set of remaining candidate classifiers. The canonical algorithm of this type, a version of which we discuss below in Section 5.1, was proposed by Cohn, Atlas, and Ladner (1994). Specifically, for any set $\mathcal{H}$ of classifiers, define the *region of disagreement*:

$$\operatorname{DIS}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h_1, h_2 \in \mathcal{H} \text{ s.t. } h_1(x) \neq h_2(x)\}.$$

The basic idea of disagreement-based algorithms is that, at any given time in the algorithm, there is a subset $V \subseteq \mathbb{C}$ of remaining candidates, called the *version space*, which is guaranteed to contain the target f. When deciding whether to request a particular label Y_i, the algorithm simply checks whether $X_i \in \operatorname{DIS}(V)$: if so, the algorithm requests Y_i, and otherwise it does not. This general strategy is reasonable, since for any $X_i \notin \operatorname{DIS}(V)$, the label agreed upon by V must be $f(X_i)$, so that we would get no information by requesting Y_i; that is, for $X_i \notin \operatorname{DIS}(V)$, we can accurately *infer* Y_i based on information already available. This type of algorithm has recently received substantial attention, not only for its obvious elegance and simplicity, but also because (as we discuss in Section 6) there are natural ways to extend the technique to the general problem of learning with label noise and model misspecification (the *agnostic* setting). The details of disagreement-based algorithms can vary in how they update the set V and how frequently they do so, but it turns out almost all disagreement-based algorithms share many of the same fundamental properties, which we describe below.

3.1 A Basic Disagreement-Based Active Learning Algorithm

In Section 5.1, we discuss several known results on the label complexities achievable by these types of active learning algorithms. However, for now let us examine a very basic algorithm of this type. The following is intended to be a simple representative of the family of disagreement-based active learning algorithms. It has been stripped down to the bare essentials of what makes such algorithms work. As a result, although the gap between its label complexity and that achieved by passive learning is not necessarily as large as those achieved by the more sophisticated disagreement-based active learning algorithms of Section 5.1, it has the property that whenever those more sophisticated methods have label complexities asymptotically superior to those achieved by passive learning, that guarantee will also be true for this simpler method, and vice versa. The algorithm operates in only 2 phases. In the first, it uses one batch of label requests to reduce the version space V to a subset of $\mathbb{C}$; in the second, it uses another batch of label requests, this time only requesting labels for points in $\operatorname{DIS}(V)$. Thus, we have isolated precisely that aspect of disagreement-based active learning that involves improvements due to only requesting the labels of examples in the region of disagreement. The procedure is formally defined as follows, in terms of an estimator $\hat{P}_n(\operatorname{DIS}(V))$ specified below.

Meta-Algorithm 0 Input: passive algorithm $\mathcal{A}_p$, label budget n Output: classifier $\hat{h}$
0. Request the first $\lfloor n/2 \rfloor$ labels $\{Y_1, \ldots, Y_{\lfloor n/2 \rfloor}\}$, and let $t \leftarrow \lfloor n/2 \rfloor$ 1. Let $V = \{h \in \mathbb{C} : \mathrm{er}_{\lfloor n/2 \rfloor}(h) = 0\}$ 2. Let $\hat{\Delta} \leftarrow \hat{P}_n(\mathrm{DIS}(V))$ 3. Let $\mathcal{L} \leftarrow \{\}$ 4. For $m = \lfloor n/2 \rfloor + 1, \ldots \lfloor n/2 \rfloor + \lfloor n/(4\hat{\Delta}) \rfloor$ 5. If $X_m \in \mathrm{DIS}(V)$ and $t < n$, request the label Y_m of X_m, and let $\hat{y} \leftarrow Y_m$ and $t \leftarrow t+1$ 6. Else let $\hat{y} \leftarrow h(X_m)$ for an arbitrary $h \in V$ 7. Let $\mathcal{L} \leftarrow \mathcal{L} \cup \{(X_m, \hat{y})\}$ 8. Return $\mathcal{A}_p(\mathcal{L})$

Meta-Algorithm 0 depends on a data-dependent estimator $\hat{P}_n(\mathrm{DIS}(V))$ of $\mathcal{P}(\mathrm{DIS}(V))$, which we can define in a variety of ways using only *unlabeled* examples. In particular, for the theorems below, we will take the following definition for $\hat{P}_n(\mathrm{DIS}(V))$, designed to be a confidence upper bound on $\mathcal{P}(\mathrm{DIS}(V))$. Let $\mathcal{U}_n = \{X_{n^2+1}, \ldots, X_{2n^2}\}$. Then define

$$\hat{P}_n(\mathrm{DIS}(V)) = \max\left\{\frac{2}{n^2}\sum_{x \in \mathcal{U}_n} \mathbb{1}_{\mathrm{DIS}(V)}(x), \frac{4}{n}\right\}. \tag{1}$$

Meta-Algorithm 0 is divided into two stages: one stage where we focus on reducing V, and a second stage where we construct the sample $\mathcal{L}$ for the passive algorithm. This might intuitively seem somewhat wasteful, as one might wish to use the requested labels from the first stage to augment those in the second stage when constructing $\mathcal{L}$, thus feeding all of the observed labels into the passive algorithm $\mathcal{A}_p$. Indeed, this can improve the label complexity in some cases (albeit only by a constant factor); however, in order to get the *general* property of being an activizer for *all* passive algorithms $\mathcal{A}_p$, we construct the sample $\mathcal{L}$ so that the conditional distribution of the $\mathcal{X}$ components in $\mathcal{L}$ given $|\mathcal{L}|$ is $\mathcal{P}^{|\mathcal{L}|}$, so that it is (conditionally) an i.i.d. sample, which is essential to our analysis. The choice of the number of (unlabeled) examples to process in the second stage guarantees (by a Chernoff bound) that the "$t < n$" constraint in Step 5 is redundant; this is a trick we will employ in several of the methods below. As explained above, because $f \in V$, this implies that every $(x,y) \in \mathcal{L}$ has $y = f(x)$.

To give some basic intuition for how this algorithm behaves, consider the example of learning threshold classifiers (Example 1); to simplify the explanation, for now we ignore the fact that $\hat{P}_n$ is only an estimate, as well as the "$t < n$" constraint in Step 5 (both of which will be addressed in the general analysis below). In this case, suppose the target function is $f = h_z$. Let $a = \max\{X_i : X_i < z, 1 \le i \le \lfloor n/2 \rfloor\}$ and $b = \min\{X_i : X_i \ge z, 1 \le i \le \lfloor n/2 \rfloor\}$. Then $V = \{h_{z'} : a < z' \le b\}$ and $\mathrm{DIS}(V) = (a,b)$, so that the second phase of the algorithm only requests labels for a number of points in the region (a,b). With probability $1-\varepsilon$, the probability mass in this region is at most $O(\log(1/\varepsilon)/n)$, so that $|\mathcal{L}| \ge \ell_{n,\varepsilon} = \Omega(n^2/\log(1/\varepsilon))$; also, since the labels in $\mathcal{L}$ are all correct, and the X_m values in $\mathcal{L}$ are conditionally iid (with distribution $\mathcal{P}$) given $|\mathcal{L}|$, we see that the conditional distribution of $\mathcal{L}$ given $|\mathcal{L}| = \ell$ is the same as the (unconditional) distribution of $\mathcal{Z}_\ell$. In particular, if $\mathcal{A}_p$ achieves label complexity Λ_p, and $\hat{h}_n$ is the classifier returned by Meta-Algorithm 0 applied to

$\mathcal{A}_p$, then for any $n = \Omega\left(\sqrt{\Lambda_p(\varepsilon, f, \mathcal{P})\log(1/\varepsilon)}\right)$ chosen so that $\ell_{n,\varepsilon} \geq \Lambda_p(\varepsilon, f, \mathcal{P})$, we have

$$\mathbb{E}\left[\operatorname{er}\left(\hat{h}_n\right)\right] \leq \varepsilon + \sup_{\ell \geq \ell_{n,\varepsilon}} \mathbb{E}\left[\operatorname{er}\left(\mathcal{A}_p(\mathcal{Z}_\ell)\right)\right] \leq \varepsilon + \sup_{\ell \geq \Lambda_p(\varepsilon, f, \mathcal{P})} \mathbb{E}\left[\operatorname{er}\left(\mathcal{A}_p(\mathcal{Z}_\ell)\right)\right] \leq 2\varepsilon.$$

This indicates the active learning algorithm achieves label complexity Λ_a with $\Lambda_a(2\varepsilon, f, \mathcal{P}) = O\left(\sqrt{\Lambda_p(\varepsilon, f, \mathcal{P})\log(1/\varepsilon)}\right)$. In particular, if $\infty > \Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon))$, then $\Lambda_a(2\varepsilon, f, \mathcal{P}) = o(\Lambda_p(\varepsilon, f, \mathcal{P}))$. Therefore, Meta-Algorithm 0 is a universal activizer for the space of threshold classifiers.

In contrast, consider the problem of learning interval classifiers (Example 2). In this case, suppose the target function f has $\mathcal{P}(x : f(x) = +1) = 0$, and that $\mathcal{P}$ is uniform in $[0,1]$. Since (with probability one) every $Y_i = -1$, we have $V = \{h_{[a,b]} : \{X_1, \ldots, X_{\lfloor n/2 \rfloor}\} \cap [a,b] = \emptyset\}$. But this contains classifiers $h_{[a,a]}$ for every $a \in (0,1) \setminus \{X_1, \ldots, X_{\lfloor n/2 \rfloor}\}$, so that $\operatorname{DIS}(V) = (0,1) \setminus \{X_1, \ldots, X_{\lfloor n/2 \rfloor}\}$. Thus, $\mathcal{P}(\operatorname{DIS}(V)) = 1$, and $|\mathcal{L}| = O(n)$; that is, $\mathcal{A}_p$ gets run with no more labeled examples than simple passive learning would use. This indicates we should not expect Meta-Algorithm 0 to be a universal activizer for interval classifiers. Below, we formalize this by constructing a passive learning algorithm $\mathcal{A}_p$ that Meta-Algorithm 0 does not activize in scenarios of this type.

3.2 The Limiting Region of Disagreement

In this subsection, we generalize the examples from the previous subsection. Specifically, we prove that the performance of Meta-Algorithm 0 is intimately tied to a particular limiting set, referred to as the *disagreement core*. A similar definition was given by Balcan, Hanneke, and Vaughan (2010) (there referred to as the *boundary*, for reasons that will become clear below); it is also related to certain quantities in the work of Hanneke (2007b, 2011) described below in Section 5.1.

Definition 4 *Define the* disagreement core *of a classifier f with respect to a set of classifiers $\mathcal{H}$ and probability measure P as*

$$\partial_{\mathcal{H},P} f = \lim_{r \to 0} \operatorname{DIS}\left(\mathrm{B}_{\mathcal{H},P}(f, r)\right).$$

When $P = \mathcal{P}$, the data distribution on $\mathcal{X}$, and $\mathcal{P}$ is clear from the context, we abbreviate this as $\partial_{\mathcal{H}} f = \partial_{\mathcal{H},\mathcal{P}} f$; if additionally $\mathcal{H} = \mathbb{C}$, the full concept space, which is clear from the context, we further abbreviate this as $\partial f = \partial_{\mathbb{C}} f = \partial_{\mathbb{C},\mathcal{P}} f$.

As we will see, disagreement-based algorithms often tend to focus their label requests around the disagreement core of the target function. As such, the concept of the disagreement core will be essential in much of our discussion below. We therefore go through a few examples to build intuition about this concept and its properties. Perhaps the simplest example to start with is $\mathbb{C}$ as the class of *threshold* classifiers (Example 1), under $\mathcal{P}$ uniform on $[0,1]$. For any $h_z \in \mathbb{C}$ and sufficiently small $r > 0$, $\mathrm{B}(f, r) = \{h_{z'} : |z' - z| \leq r\}$, and $\operatorname{DIS}(\mathrm{B}(f, r)) = [z - r, z + r)$. Therefore, $\partial h_z = \lim_{r \to 0} \operatorname{DIS}(\mathrm{B}(h_z, r)) = \lim_{r \to 0} [z - r, z + r) = \{z\}$. Thus, in this case, the disagreement core of h_z with respect to $\mathbb{C}$ and $\mathcal{P}$ is precisely the decision boundary of the classifier. As a slightly more involved example, consider again the example of *interval* classifiers (Example 2), again under $\mathcal{P}$ uniform on $[0,1]$. Now for any $h_{[a,b]} \in \mathbb{C}$ with $b - a > 0$, for any sufficiently small $r > 0$, $\mathrm{B}(h_{[a,b]}, r) = \{h_{[a',b']} : |a - a'| + |b - b'| \leq r\}$, and $\operatorname{DIS}(\mathrm{B}(h_{[a,b]}, r)) = [a - r, a + r) \cup (b - r, b + r]$. Therefore, $\partial h_{[a,b]} = \lim_{r \to 0} \operatorname{DIS}(\mathrm{B}(h_{[a,b]}, r)) = \lim_{r \to 0} [a - r, a + r) \cup (b - r, b + r] = \{a, b\}$. Thus, in this case as well, the disagreement core of $h_{[a,b]}$ with respect to $\mathbb{C}$ and $\mathcal{P}$ is again the decision boundary of the classifier.

As the above two examples illustrate, ∂f often corresponds to the decision boundary of f in some geometric interpretation of $\mathcal{X}$ and f. Indeed, under fairly general conditions on $\mathbb{C}$ and $\mathcal{P}$, the disagreement core of f does correspond to (a subset of) the set of points dividing the two label regions of f; for instance, Friedman (2009) derives sufficient conditions, under which this is the case. In these cases, the behavior of disagreement-based active learning algorithms can often be interpreted in the intuitive terms of seeking label requests near the decision boundary of the target function, to refine an estimate of that boundary. However, in some more subtle scenarios this is no longer the case, for interesting reasons. To illustrate this, let us continue the example of interval classifiers from above, but now consider $h_{[a,a]}$ (i.e., $h_{[a,b]}$ with $a = b$). This time, for any $r \in (0,1)$ we have $\mathrm{B}(h_{[a,a]}, r) = \{h_{[a',b']} \in \mathbb{C} : b' - a' \leq r\}$, and $\mathrm{DIS}(\mathrm{B}(h_{[a,a]}, r)) = (0,1)$. Therefore, $\partial h_{[a,a]} = \lim_{r \to 0} \mathrm{DIS}(\mathrm{B}(h_{[a,a]}, r)) = \lim_{r \to 0} (0,1) = (0,1)$.

This example shows that in some cases, the disagreement core does not correspond to the decision boundary of the classifier, and indeed has $\mathcal{P}(\partial f) > 0$. Intuitively, as in the above example, this typically happens when the decision surface of the classifier is in some sense *simpler* than it could be. For instance, consider the space $\mathbb{C}$ of *unions of two intervals* (Example 3 with $i = 2$) under uniform $\mathcal{P}$. The classifiers $f \in \mathbb{C}$ with $\mathcal{P}(\partial f) > 0$ are precisely those representable (up to probability zero differences) as a single interval. The others (with $0 < z_1 < z_2 < z_3 < z_4 < 1$) have $\partial h_{\mathbf{z}} = \{z_1, z_2, z_3, z_4\}$. In these examples, the $f \in \mathbb{C}$ with $\mathcal{P}(\partial f) > 0$ are not only simpler than other nearby classifiers in $\mathbb{C}$, but they are also in some sense *degenerate* relative to the rest of $\mathbb{C}$; however, it turns out this is not always the case, as there exist scenarios $(\mathbb{C}, \mathcal{P})$, even with $d = 2$, and even with *countable* $\mathbb{C}$, for which *every* $f \in \mathbb{C}$ has $\mathcal{P}(\partial f) > 0$; in these cases, every classifier is in some important sense *simpler* than some other subset of nearby classifiers in $\mathbb{C}$.

In Section 3.3, we show that the label complexity of disagreement-based active learning is intimately tied to the disagreement core. In particular, scenarios where $\mathcal{P}(\partial f) > 0$, such as those mentioned above, lead to the conclusion that disagreement-based methods are sometimes insufficient for activized learning. This motivates the design of more sophisticated methods in Section 4, which overcome this deficiency, along with a corresponding refinement of the definition of "disagreement core " in Section 5.2 that eliminates the above issue with "simple" classifiers.

3.3 Necessary and Sufficient Conditions for Disagreement-Based Activized Learning

In the specific case of Meta-Algorithm 0, for large n we may intuitively expect it to focus its second batch of label requests in and around the disagreement core of the target function. Thus, whenever $\mathcal{P}(\partial f) = 0$, we should expect the label requests to be quite focused, and therefore the algorithm should achieve smaller label complexity compared to passive learning. On the other hand, if $\mathcal{P}(\partial f) > 0$, then the label requests will *not* become focused beyond a constant fraction of the space, so that the improvements achieved by Meta-Algorithm 0 over passive learning should be, at best, a constant factor. This intuition is formalized in the following general theorem, the proof of which is included in Appendix A.

Theorem 5 *For any VC class $\mathbb{C}$, Meta-Algorithm 0 is a universal activizer for $\mathbb{C}$ if and only if every $f \in \mathbb{C}$ and distribution $\mathcal{P}$ has $\mathcal{P}(\partial_{\mathbb{C},\mathcal{P}} f) = 0$.*

While the formal proof is given in Appendix A, the general idea is simple. As we always have $f \in V$, any $\hat{y}$ inferred in Step 6 must equal $f(x)$, so that all of the labels in $\mathcal{L}$ are correct. Also, as n grows large, classic results on passive learning imply the diameter of the set V will become small, shrinking to zero as $n \to \infty$ (Vapnik and Chervonenkis, 1971; Vapnik, 1982; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989). Therefore, as $n \to \infty$, $\mathrm{DIS}(V)$ should converge to a subset of ∂f, so that in the case $\mathcal{P}(\partial f) = 0$, we have $\hat{\Delta} \to 0$; thus $|\mathcal{L}| \gg n$, which implies an asymptotic strict improvement in label complexity over the passive algorithm $\mathcal{A}_p$ that $\mathcal{L}$ is fed into in Step 8. On the other hand, since ∂f is defined by classifiers arbitrarily close to f, it is unlikely that any finite sample of correctly labeled examples can contradict enough classifiers to make $\mathrm{DIS}(V)$ significantly smaller than ∂f, so that we always have $\mathcal{P}(\mathrm{DIS}(V)) \geq \mathcal{P}(\partial f)$. Therefore, if $\mathcal{P}(\partial f) > 0$, then $\hat{\Delta}$ converges to some nonzero constant, so that $|\mathcal{L}| = O(n)$, representing only a constant factor improvement in label complexity. In fact, as is implied from this sketch (and is proven in Appendix A), the targets f and distributions $\mathcal{P}$ for which Meta-Algorithm 0 achieves asymptotic strict improvements for all passive learning algorithms (for which f and $\mathcal{P}$ are nontrivial) are precisely those (and only those) for which $\mathcal{P}(\partial_{\mathbb{C},\mathcal{P}} f) = 0$.

There are some general conditions under which the zero-probability disagreement cores condition of Theorem 5 will hold. For instance, it is not difficult to show this will always hold when $\mathcal{X}$ is countable. Furthermore, with some effort one can show it will hold for most classes having VC dimension one (e.g., any countable $\mathbb{C}$ with $d = 1$). However, as we have seen, not all spaces $\mathbb{C}$ satisfy this zero-probability disagreement cores property. In particular, for the interval classifiers studied in Section 3.2, we have $\mathcal{P}(\partial h_{[a,a]}) = \mathcal{P}((0,1)) = 1$. Indeed, the aforementioned special cases aside, for *most* nontrivial spaces $\mathbb{C}$, one can construct distributions $\mathcal{P}$ that in some sense make $\mathbb{C}$ mimic the intervals problem, so that we should typically expect disagreement-based methods will *not* be activizers. For detailed discussions of various scenarios where the $\mathcal{P}(\partial_{\mathbb{C},\mathcal{P}} f) = 0$ condition is (or is not) satisfied for various $\mathbb{C}$, $\mathcal{P}$, and f, see the works of Hanneke (2009b), Hanneke (2007b), Hanneke (2011), Balcan, Hanneke, and Vaughan (2010), Friedman (2009), Wang (2009) and Wang (2011).

4. Beyond Disagreement: A Basic Activizer

Since the zero-probability disagreement cores condition of Theorem 5 is not always satisfied, we are left with the question of whether there could be other techniques for active learning, beyond simple disagreement-based methods, which could activize *every* passive learning algorithm for *every* VC class. In this section, we present an entirely new type of active learning algorithm, unlike anything in the existing literature, and we show that indeed it is a universal activizer for any class $\mathbb{C}$ of finite VC dimension.

4.1 A Basic Activizer

As mentioned, the case $\mathcal{P}(\partial f) = 0$ is already handled nicely by disagreement-based methods, since the label requests made in the second stage of Meta-Algorithm 0 will become focused into a small region, and $\mathcal{L}$ therefore grows faster than n. Thus, the primary question we are faced with is what to do when $\mathcal{P}(\partial f) > 0$. Since (loosely speaking) we have $\mathrm{DIS}(V) \to \partial f$ in Meta-Algorithm 0, $\mathcal{P}(\partial f) > 0$ corresponds to scenarios where the label requests of Meta-Algorithm 0 will not become focused beyond a certain extent; specifically, as we show in Appendix B (Lemmas 35 and 36), $\mathcal{P}(\mathrm{DIS}(V) \oplus \partial f) \to 0$ almost surely, where $\oplus$ is the symmetric difference, so that we expect Meta-Algorithm 0 will request labels for at least some constant fraction of the examples in $\mathcal{L}$.

On the one hand, this is definitely a major problem for disagreement-based methods, since it prevents them from improving over passive learning in those cases. On the other hand, if we do not restrict ourselves to disagreement-based methods, we may actually be able to exploit properties of this scenario, so that it works to our *advantage*. In particular, in addition to the fact that $\mathcal{P}(\mathrm{DIS}(V) \oplus \partial_{\mathbb{C}} f) \to 0$, we show in Appendix B (Lemma 35) that $\mathcal{P}(\partial_V f \oplus \partial_{\mathbb{C}} f) = 0$ (almost surely) in Meta-Algorithm 0; this implies that for sufficiently large n, a random point x_1 in $\mathrm{DIS}(V)$ is likely to be in $\partial_V f$. We can exploit this fact by using x_1 to split V into two subsets: $V[(x_1,+1)]$ and $V[(x_1,-1)]$. Now, if $x_1 \in \partial_V f$, then (by definition of the disagreement core) $\inf_{h \in V[(x_1,+1)]} \mathrm{er}(h) = \inf_{h \in V[(x_1,-1)]} \mathrm{er}(h) = 0$. Therefore, for almost every point $x \notin \mathrm{DIS}(V[(x_1,+1)])$, the label agreed upon for x by classifiers in $V[(x_1,+1)]$ should be $f(x)$. Likewise, for almost every point $x \notin \mathrm{DIS}(V[(x_1,-1)])$, the label agreed upon for x by classifiers in $V[(x_1,-1)]$ should be $f(x)$. Thus, we can accurately *infer* the label of any point $x \notin \mathrm{DIS}(V[(x_1,+1)]) \cap \mathrm{DIS}(V[(x_1,-1)])$ (except perhaps a zero-probability subset). With these sets $V[(x_1,+1)]$ and $V[(x_1,-1)]$ in hand, there is no longer a need to request the labels of points for which either of them has agreement about the label, and we can focus our label requests to the region $\mathrm{DIS}(V[(x_1,+1)]) \cap \mathrm{DIS}(V[(x_1,-1)])$, which may be *much smaller* than $\mathrm{DIS}(V)$. Now if $\mathcal{P}(\mathrm{DIS}(V[(x_1,+1)]) \cap \mathrm{DIS}(V[(x_1,-1)])) \to 0$, then the label requests will become focused to a shrinking region, and by the same reasoning as for Theorem 5 we can asymptotically achieve strict improvements over passive learning by a method analogous to Meta-Algorithm 0 (with the above changes).

Already this provides a significant improvement over disagreement-based methods in many cases; indeed, in some cases (such as intervals) this fully addresses the nonzero-probability disagreement core issue in Theorem 5. In other cases (such as unions of two intervals), it does not completely address the issue, since for some targets we do not have $\mathcal{P}(\mathrm{DIS}(V[(x_1,+1)]) \cap \mathrm{DIS}(V[(x_1,-1)])) \to 0$. However, by repeatedly applying this same reasoning, we *can* address the issue in full generality. Specifically, if $\mathcal{P}(\mathrm{DIS}(V[(x_1,+1)]) \cap \mathrm{DIS}(V[(x_1,-1)])) \nrightarrow 0$, then $\mathrm{DIS}(V[(x_1,+1)]) \cap \mathrm{DIS}(V[(x_1,-1)])$ essentially converges to a region $\partial_{\mathbb{C}[(x_1,+1)]} f \cap \partial_{\mathbb{C}[(x_1,-1)]} f$, which has nonzero probability, and is nearly equivalent to $\partial_{V[(x_1,+1)]} f \cap \partial_{V[(x_1,-1)]} f$. Thus, for sufficiently large n, a random x_2 in $\mathrm{DIS}(V[(x_1,+1)]) \cap \mathrm{DIS}(V[(x_1,-1)])$ will likely be in $\partial_{V[(x_1,+1)]} f \cap \partial_{V[(x_1,-1)]} f$. In this case, we can repeat the above argument, this time splitting V into four sets ($V[(x_1,+1)][(x_2,+1)]$, $V[(x_1,+1)][(x_2,-1)]$, $V[(x_1,-1)][(x_2,+1)]$, and $V[(x_1,-1)][(x_2,-1)]$), each with infimum error rate equal zero, so that for a point x in the region of agreement of any of these four sets, the agreed-upon label will (almost surely) be $f(x)$, so that we can infer that label. Thus, we need only request the labels of those points in the *intersection* of all four regions of disagreement. We can further repeat this process as many times as needed, until we get a partition of V with shrinking probability mass in the intersection of the regions of disagreement, which (as above) can then be used to obtain asymptotic improvements over passive learning.

Note that the above argument can be written more concisely in terms of *shattering*. That is, any $x \in \mathrm{DIS}(V)$ is simply an x such that V shatters $\{x\}$; a point $x \in \mathrm{DIS}(V[(x_1,+1)]) \cap \mathrm{DIS}(V[(x_1,-1)])$ is simply one for which V shatters $\{x_1, x\}$, and for any $x \notin \mathrm{DIS}(V[(x_1,+1)]) \cap \mathrm{DIS}(V[(x_1,-1)])$, the label y we infer about x has the property that the set $V[(x,-y)]$ does not shatter $\{x_1\}$. This continues for each repetition of the above idea, with x in the intersection of the four regions of disagreement simply being one for which V shatters $\{x_1, x_2, x\}$, and so on. In particular, this perspective makes it clear that we need only repeat this idea at most d times to get a shrinking intersection region, since no set of $d+1$ points is shatterable. Note that there may be unobservable factors (e.g., the target function) determining the appropriate number of iterations of this idea sufficient to have a shrinking probability of requesting a label, while maintaining the accuracy of inferred labels. To address this, we can simply try all $d+1$ possibilities, and then select one of the resulting $d+1$ classifiers via a kind of tournament of pairwise comparisons. Also, in order to reduce the probability of a mistaken inference due to $x_1 \notin \partial_V f$ (or similarly for later x_i), we can replace each single x_i with multiple samples, and then take a majority vote over whether to infer the label, and which label to infer if we do so; generally, we can think of this as estimating certain probabilities, and below we write these estimators as $\hat{P}_m$, and discuss the details of their implementation later. Combining Meta-Algorithm 0 with the above reasoning motivates a new type of active learning algorithm, referred to as Meta-Algorithm 1 below, and stated as follows.

Meta-Algorithm 1
Input: passive algorithm $\mathcal{A}_p$, label budget n
Output: classifier $\hat{h}$

Request the first $m_n = \lfloor n/3 \rfloor$ labels, $\{Y_1, \ldots, Y_{m_n}\}$, and let $t \leftarrow m_n$
Let $V = \{h \in \mathbb{C} : \mathrm{er}_{m_n}(h) = 0\}$
For $k = 1, 2, \ldots, d+1$
  $\hat{\Delta}^{(k)} \leftarrow \hat{P}_{m_n}\left(x : \hat{P}\left(S \in \mathcal{X}^{k-1} : V \text{ shatters } S \cup \{x\} | V \text{ shatters } S\right) \geq 1/2\right)$
  Let $\mathcal{L}_k \leftarrow \{\}$
  For $m = m_n + 1, \ldots, m_n + \lfloor n/(6 \cdot 2^k \hat{\Delta}^{(k)}) \rfloor$
    If $\hat{P}_m\left(S \in \mathcal{X}^{k-1} : V \text{ shatters } S \cup \{X_m\} | V \text{ shatters } S\right) \geq 1/2$ and $t < \lfloor 2n/3 \rfloor$
      Request the label Y_m of X_m, and let $\hat{y} \leftarrow Y_m$ and $t \leftarrow t+1$
    Else, let $\hat{y} \leftarrow \operatorname*{argmax}_{y \in \{-1,+1\}} \hat{P}_m\left(S \in \mathcal{X}^{k-1} : V[(X_m, -y)] \text{ does not shatter } S | V \text{ shatters } S\right)$
    Let $\mathcal{L}_k \leftarrow \mathcal{L}_k \cup \{(X_m, \hat{y})\}$
Return ActiveSelect($\{\mathcal{A}_p(\mathcal{L}_1), \mathcal{A}_p(\mathcal{L}_2), \ldots, \mathcal{A}_p(\mathcal{L}_{d+1})\}, \lfloor n/3 \rfloor, \{X_{m_n + \max_k |\mathcal{L}_k| + 1}, \ldots\}$)

Subroutine: ActiveSelect
Input: set of classifiers $\{h_1, h_2, \ldots, h_N\}$, label budget m, sequence of unlabeled examples $\mathcal{U}$
Output: classifier $\hat{h}$

For each $j, k \in \{1, 2, \ldots, N\}$ s.t. $j < k$,
  Let R_{jk} be the first $\left\lfloor \frac{m}{j(N-j)\ln(eN)} \right\rfloor$ points in $\mathcal{U} \cap \{x : h_j(x) \neq h_k(x)\}$ (if such values exist)
  Request the labels for R_{jk} and let Q_{jk} be the resulting set of labeled examples
  Let $m_{kj} = \mathrm{er}_{Q_{jk}}(h_k)$
Return $h_{\hat{k}}$, where $\hat{k} = \max\left\{k \in \{1, \ldots, N\} : \max_{j<k} m_{kj} \leq 7/12\right\}$

Meta-Algorithm 1 is stated as a function of three types of estimated probabilities: namely,

$$\hat{P}_m\left(S \in \mathcal{X}^{k-1} : V \text{ shatters } S \cup \{x\} \middle| V \text{ shatters } S\right),$$
$$\hat{P}_m\left(S \in \mathcal{X}^{k-1} : V[(x, -y)] \text{ does not shatter } S \middle| V \text{ shatters } S\right),$$
$$\text{and } \hat{P}_m\left(x : \hat{P}\left(S \in \mathcal{X}^{k-1} : V \text{ shatters } S \cup \{x\} \middle| V \text{ shatters } S\right) \geq 1/2\right).$$

These can be defined in a variety of ways to make this a universal activizer for $\mathbb{C}$. Generally, the only requirement seems to be that they converge to the appropriate respective probabilities at a sufficiently fast rate. For the theorem stated below regarding Meta-Algorithm 1, we will take the specific definitions stated in Appendix B.1.

Meta-Algorithm 1 requests labels in three batches: one to initially prune down the version space V, a second one to construct the labeled samples $\mathcal{L}_k$, and a third batch to select among the $d+1$ classifiers $\mathcal{A}_p(\mathcal{L}_k)$ in the ActiveSelect subroutine. As before, the choice of the number of (unlabeled) examples to process in the second batch guarantees (by a Chernoff bound) that the "$t < \lfloor 2n/3 \rfloor$" constraint in Step 6 is redundant. The mechanism for requesting labels in the second batch is motivated by the reasoning outlined above, using the shatterable sets S to split V into 2^{k-1} subsets, each of which approximates the target with high probability (for large n), and then checking whether the new point x is in the regions of disagreement for all 2^{k-1} subsets (by testing shatterability of $S \cup \{x\}$). To increase confidence in this test, we use many such S sets, and let them vote on whether or not to request the label (Step 6). As mentioned, if x is not in the region of disagreement for one of these 2^{k-1} subsets (call it V'), the agreed-upon label y has the property that $V[(x,-y)]$ does not shatter S (since $V[(x,-y)]$ does not intersect with V', which represents one of the 2^{k-1} labelings required to shatter S). Therefore, we infer that this label y is the correct label of x, and again we vote over many such S sets to increase confidence in this choice (Step 8). As mentioned, this reasoning leads to correctly inferred labels in Step 8 as long as n is sufficiently large *and* $\mathcal{P}^{k-1}(S \in \mathcal{X}^{k-1} : V \text{ shatters } S) \nrightarrow 0$. In particular, we are primarily interested in the largest value of k for which this reasoning holds, since this is the value at which the probability of requesting a label (Step 7) shrinks to zero as $n \to \infty$. However, since we typically cannot predict a priori what this largest valid k value will be (as it is target-dependent), we try all $d+1$ values of k, to generate $d+1$ hypotheses, and then use a simple pairwise testing procedure to select among them; note that we need at most try $d+1$ values, since V definitely cannot shatter any $S \in \mathcal{X}^{d+1}$. We will see that the ActiveSelect subroutine is guaranteed to select a classifier with error rate never significantly larger than the best among the classifiers given to it (say within a factor of 2, with high probability). Therefore, in the present context, we need only consider whether some k has a set $\mathcal{L}_k$ with correct labels *and* $|\mathcal{L}_k| \gg n$.

4.2 Examples

In the next subsection, we state a general result for Meta-Algorithm 1. But first, to illustrate how this procedure operates, we walk through its behavior on our usual examples; as we did for the examples of Meta-Algorithm 0, to simplify the explanation, for now we will ignore the fact that the $\hat{P}_m$ values are estimates, as well as the "$t < \lfloor 2n/3 \rfloor$" constraint of Step 6, and the issue of effectiveness of ActiveSelect; in the proofs of the general results below, we will show that these issues do not fundamentally change the analysis. For now, we merely focus on showing that some k has $\mathcal{L}_k$ correctly labeled and $|\mathcal{L}_k| \gg n$.

For threshold classifiers (Example 1), we have $d = 1$. In this case, the $k = 1$ round of the algorithm is essentially identical to Meta-Algorithm 0 (recall our conventions that $\mathcal{X}^0 = \{\varnothing\}$, $\mathcal{P}(\mathcal{X}^0) = 1$, and V shatters $\varnothing$ iff $V \neq \{\}$), and we therefore have $|\mathcal{L}_1| \gg n$, as discussed previously, so that Meta-Algorithm 1 is a universal activizer for threshold classifiers.

Next consider interval classifiers (Example 2), with $\mathcal{P}$ uniform on $[0,1]$; in this case, we have $d = 2$. If $f = h_{[a,b]}$ for $a < b$, then again the $k = 1$ round behaves essentially the same as Meta-Algorithm 0, and since we have seen $\mathcal{P}(\partial h_{[a,b]}) = 0$ in this case, we have $|\mathcal{L}_1| \gg n$. However, the behavior becomes far more interesting when $f = h_{[a,a]}$, which was precisely the case that prevented

Meta-Algorithm 0 from improving over passive learning. In this case, as we know from above, the $k=1$ round will have $|\mathcal{L}_1| = O(n)$, so that we need to consider larger values of k to identify improvements. In this case, the $k=2$ round behaves as follows. With probability 1, the initial $\lfloor n/3 \rfloor$ labels used to define V will all be negative. Thus, V is precisely the set of intervals that do not contain any of the initial $\lfloor n/3 \rfloor$ points. Now consider any $S = \{x_1\} \in \mathcal{X}^1$, with x_1 not equal to any of these initial $\lfloor n/3 \rfloor$ points, and consider any $x \notin \{x_1, X_1, \ldots, X_{\lfloor n/3 \rfloor}\}$. First note that V shatters S, since we can optionally put a small interval around x_1 using an element of V. If there is a point x' among the initial $\lfloor n/3 \rfloor$ *between* x and x_1, then any $h_{[a,b]} \in V$ with $x \in [a,b]$ cannot also have $x_1 \in [a,b]$, as it would also contain the observed negative point between them. Thus, V does *not* shatter $\{x_1, x\} = S \cup \{x\}$, so that this S will vote to infer (rather than request) the label of x in Step 6. Furthermore, we see that $V[(x,+1)]$ does not shatter S, while $V[(x,-1)]$ does shatter S, so that this S would also vote for the label $\hat{y} = -1$ in Step 8. For sufficiently large n, with high probability, any given x not equal one of the initial $\lfloor n/3 \rfloor$ should have *most* (probability at least $1 - O(n^{-1}\log n)$) of the possible x_1 values separated from it by at least one of the initial $\lfloor n/3 \rfloor$ points, so that the outcome of the vote in Step 6 will be a decision to infer (not request) the label, and the vote in Step 8 will be for -1. Since, with probability one, every $X_m \neq a$, we have every $Y_m = -1$, so that every point in $\mathcal{L}_2$ is labeled correctly. This also indicates that, for sufficiently large n, we have $\mathcal{P}(x : \mathcal{P}^1(S \in \mathcal{X}^1 : V \text{ shatters } S \cup \{x\} | V \text{ shatters } S) \geq 1/2) = 0$, so that the size of $\mathcal{L}_2$ is only limited by the precision of estimation in $\hat{P}_{m_n}$ in Step 3. Thus, as long as we implement $\hat{P}_{m_n}$ so that its value is at most $o(1)$ larger than the true probability, we can guarantee $|\mathcal{L}_2| \gg n$.

The unions of i intervals example (Example 3), again under $\mathcal{P}$ uniform on $[0,1]$, is slightly more involved; in this case, the appropriate value of k to consider for any given target depends on the minimum number of intervals necessary to represent the target function (up to zero-probability differences). If j intervals are required for this, then the appropriate value is $k = i - j + 1$. Specifically, suppose the target is minimally representable as a union of $j \in \{1, \ldots, i\}$ intervals of nonzero width: $[z_1, z_2] \cup [z_3, z_4] \cup \cdots \cup [z_{2j-1}, z_{2j}]$: that is, $z_1 < z_2 < \ldots < z_{2j-1} < z_{2j}$. Every target in $\mathbb{C}$ has distance zero to some classifier of this type, and will agree with that classifier on all samples with probability one, so we lose no generality by assuming all j intervals have nonzero width. Then consider any $x \in (0,1)$ and $S = \{x_1, \ldots, x_{i-j}\} \in \mathcal{X}^{i-j}$ such that, between any pair of elements of $S \cup \{x\} \cup \{z_1, \ldots, z_{2j}\}$, there is at least one of the initial $\lfloor n/3 \rfloor$ points, and none of $S \cup \{x\}$ are themselves equal to any of those initial points. First note that V shatters S, since for any x_ℓ not in one of the $[z_{2p-1}, z_{2p}]$ intervals (i.e., negative), we may optionally add an interval $[x_\ell, x_\ell]$ while staying in V, and for any x_ℓ in one of the $[z_{2p-1}, z_{2p}]$ intervals (i.e., positive), we may optionally split $[z_{2p-1}, z_{2p}]$ into two intervals to barely exclude the point x_ℓ (and a small neighborhood around it), by adding at most one interval to the representation; thus, in total we need to add at most $i-j$ intervals to the representation, so that the largest number of intervals used by any of these 2^{i-j} classifiers involved in shattering is i, as required; furthermore, note that one of these 2^{i-j} classifiers actually requires i intervals. Now for any such x and S as above, since one of the 2^{i-j} classifiers in V used to shatter S requires i intervals to represent it, and x is separated from each element of $S \cup \{z_1, \ldots, z_{2j}\}$ by a labeled example, we see that V cannot shatter $S \cup \{x\}$. Furthermore, if $f(x) = y$, then any labeled example to the immediate left or right of x is also labeled y, and in particular among the 2^{i-j} classifiers h from V that shatter S, the one h that requires i intervals to represent must also have $h(x) = y$, so that $V[(x,-y)]$ does not shatter S. Thus, any set S satisfying this separation property will vote to infer (rather than request) the label of x in Step 6, and will vote for the label $f(x)$ in Step 8. Furthermore, for sufficiently large n, for any given x separated from $\{z_1, \ldots, z_{2j}\}$ by $\{X_1, \ldots, X_{\lfloor n/3 \rfloor}\}$, with high probability most of the sets $S \in \mathcal{X}^{i-j}$ will satisfy this pairwise separation property, and therefore so will most of the shatterable sets $S \in \mathcal{X}^{i-j}$, so that the overall outcome of the votes will favor inferring the label of x, and in particular inferring the label $f(x)$ for x. On the other hand, for x not satisfying this property (i.e., not separated from some z_p by any of the initial $\lfloor n/3 \rfloor$ examples), for any set S as above, V *can* shatter $S \cup \{x\}$, since we can optionally increase or decrease this z_p to include or exclude x from the associated interval, in addition to optionally adding the extra intervals to shatter S; therefore, by the same reasoning as above, for sufficiently large n, any such x *will* satisfy the condition in Step 6, and thus have its label requested. Thus, for sufficiently large n, every example in $\mathcal{L}_{i-j+1}$ will be labeled correctly. Finally, note that with probability one, the set of points x separated from each of the z_p values by at least one of the $\lfloor n/3 \rfloor$ initial points has probability approaching one as $n \to \infty$, so that again we have $|\mathcal{L}_{i-j+1}| \gg n$.

The above examples give some intuition about the operation of this procedure. Next, we turn to general results showing that this type of improvement generally holds.

4.3 General Results on Activized Learning

Returning to the abstract setting, we have the following general theorem, representing one of the main results of this paper. Its proof is included in Appendix B.

Theorem 6 *For any VC class $\mathbb{C}$, Meta-Algorithm 1 is a universal activizer for $\mathbb{C}$.*

This result is interesting both for its strength and generality. Recall that it means that given any passive learning algorithm $\mathcal{A}_p$, the active learning algorithm obtained by providing $\mathcal{A}_p$ as input to Meta-Algorithm 1 achieves a label complexity that strongly dominates that of $\mathcal{A}_p$ for all nontrivial distributions $\mathcal{P}$ and target functions $f \in \mathbb{C}$. Results of this type were not previously known. The specific technical advance over existing results (namely, those of Balcan, Hanneke, and Vaughan, 2010) is the fact that Meta-Algorithm 1 has no direct dependence on the distribution $\mathcal{P}$; as mentioned earlier, the (very different) approach proposed by Balcan, Hanneke, and Vaughan (2010) has a strong direct dependence on the distribution, to the extent that the distribution-dependence in that approach cannot be removed by merely replacing certain calculations with data-dependent estimators (as we did in Meta-Algorithm 1). In the proof, we actually show a somewhat more general result: namely, that Meta-Algorithm 1 achieves these asymptotic improvements for any target function f in the *closure* of $\mathbb{C}$ (i.e., any f such that $\forall r > 0, \mathrm{B}(f,r) \neq \emptyset$).

The following corollary is one concrete implication of Theorem 6.

Corollary 7 *For any VC class $\mathbb{C}$, there exists an active learning algorithm achieving a label complexity Λ_a such that, for all target functions $f \in \mathbb{C}$ and distributions $\mathcal{P}$,*

$$\Lambda_a(\varepsilon, f, \mathcal{P}) = o(1/\varepsilon).$$

Proof The *one-inclusion graph* passive learning algorithm of Haussler, Littlestone, and Warmuth (1994) is known to achieve label complexity at most d/ε, for every target function $f \in \mathbb{C}$ and distribution $\mathcal{P}$. Thus, Theorem 6 implies that the (Meta-Algorithm 1)-activized one-inclusion graph algorithm satisfies the claim. ■

As a byproduct, Theorem 6 also establishes the basic fact that there *exist* activizers. In some sense, this observation opens up a new realm for exploration: namely, characterizing the *properties* that activizers can possess. This topic includes a vast array of questions, many of which deal with whether activizers are capable of *preserving* various properties of the given passive algorithm (e.g., margin-based dimension-independence, minimaxity, admissibility, etc.). Section 7 describes a variety of enticing questions of this type. In the sections below, we will consider quantifying how large the gap in label complexity between the given passive learning algorithm and the resulting activized algorithm can be. We will additionally study the effects of label noise on the possibility of activized learning.

4.4 Implementation and Efficiency

Meta-Algorithm 1 typically also has certain desirable efficiency guarantees. Specifically, suppose that for any m labeled examples Q, there is an algorithm with $\mathrm{poly}(d \cdot m)$ running time that finds some $h \in \mathbb{C}$ with $\mathrm{er}_Q(h) = 0$ if one exists, and otherwise returns a value indicating that no such h exists in $\mathbb{C}$; for many concept spaces there are known methods with this capability (e.g., linear or polynomial separators, rectangles, k-DNF) (Khachiyan, 1979; Karmarkar, 1984; Valiant, 1984; Kearns and Vazirani, 1994), while for others this is known to be hard (e.g., k-term DNF, bounded-size decision trees) (Pitt and Valiant, 1988; Alekhnovich, Braverman, Feldman, Klivans, and Pitassi, 2004). Given such a subroutine, we can create an efficient implementation of the main body of Meta-Algorithm 1. Specifically, rather than explicitly representing V in Step 1, we can simply store the set $Q_0 = \{(X_1,Y_1),\ldots,(X_{m_n},Y_{m_n})\}$. Then for any step in the algorithm where we need to test whether V shatters a set R, we can simply try all $2^{|R|}$ possible labelings of R, and for each one temporarily add these $|R|$ additional labeled examples to Q_0 and check whether there is an $h \in \mathbb{C}$ consistent with all of the labels. At first, it might seem that these 2^k evaluations would be prohibitive; however, supposing $\hat{P}_{m_n}$ is implemented so that it is $\Omega(1/\mathrm{poly}(n))$ (as it is in Appendix B.1), note that the loop beginning at Step 5 executes a nonzero number of times only if $n/\hat{\Delta}^{(k)} > 2^k$, so that $2^k \leq \mathrm{poly}(n)$; we can easily add a condition that skips the step of calculating $\hat{\Delta}^{(k)}$ if 2^k exceeds this $\mathrm{poly}(n)$ lower bound on $n/\hat{\Delta}^{(k)}$, so that even those shatterability tests can be skipped in this case. Thus, for the actual occurrences of it in the algorithm, testing whether V shatters R requires only $\mathrm{poly}(n) \cdot \mathrm{poly}(d \cdot (|Q_0| + |R|))$ time. The total number of times this test is performed in calculating $\hat{\Delta}^{(k)}$ (from Appendix B.1) is itself only $\mathrm{poly}(n)$, and the number of iterations of the loop in Step 5 is at most $n/\hat{\Delta}^{(k)} = \mathrm{poly}(n)$. Determining the label $\hat{y}$ in Step 8 can be performed in a similar fashion. So in general, the total running time of the main body of Meta-Algorithm 1 is $\mathrm{poly}(d \cdot n)$.

The only remaining question is the efficiency of the final step. Of course, we can require $\mathcal{A}_p$ to have running time polynomial in the size of its input set (and d). But beyond this, we must consider the efficiency of the ActiveSelect subroutine. This actually turns out to have some subtleties involved. The way it is stated above is simple and elegant, but not always efficient. Specifically, we have no a priori bound on the number of unlabeled examples the algorithm must process before finding a point X_m where $h_j(X_m) \neq h_k(X_m)$. Indeed, if $\mathcal{P}(x : h_j(x) \neq h_k(x)) = 0$, we may effectively need to examine the entire infinite sequence of X_m values to determine this. Fortunately, these problems can be corrected without difficulty, simply by truncating the search at a predetermined number of points. Specifically, rather than taking the next $\lfloor m/\binom{N}{2} \rfloor$ examples for which h_j and h_k disagree, simply restrict ourselves to at most this number, or at most the number of such points among the next M unlabeled examples. In Appendix B, we show that ActiveSelect, as originally stated, has a high-probability $(1 - \exp\{-\Omega(m)\})$ guarantee that the classifier it selects has error rate at most twice the best of the N it is given. With the modification to truncate the search at M unlabeled examples, this guarantee is increased to $\min_k \operatorname{er}(h_k) + \max\{\operatorname{er}(h_k), m/M\}$. For the concrete guarantee of Corollary 7, it suffices to take $M \gg m^2$. However, to guarantee the modified ActiveSelect can still be used in Meta-Algorithm 1 while maintaining (the stronger) Theorem 6, we need M at least as big as $\Omega(\min\{\exp\{m^c\}, m/\min_k \operatorname{er}(h_k)\})$, for any constant $c > 0$. In general, if we have a $1/\operatorname{poly}(n)$ lower bound on the error rate of the classifier produced by $\mathcal{A}_p$ for a given number of labeled examples as input, we can set M as above using this lower bound in place of $\min_k \operatorname{er}(h_k)$, resulting in an efficient version of ActiveSelect that still guarantees Theorem 6. However, it is presently not known whether there always exist universal activizers for $\mathbb{C}$ that are efficient (either $\operatorname{poly}(d \cdot n)$ or $\operatorname{poly}(d/\varepsilon)$ running time) when the above assumptions on efficiency of $\mathcal{A}_p$ and finding $h \in \mathbb{C}$ with $\operatorname{er}_Q(h) = 0$ hold.

5. The Magnitudes of Improvements

In the previous section, we saw that we can always improve the label complexity of a passive learning algorithm by activizing it. However, there remains the question of how large the gap is between the passive algorithm's label complexity and the activized algorithm's label complexity. In the present section, we refine the above procedures to take greater advantage of the sequential nature of active learning. For each, we characterize the improvements it achieves relative to any given passive algorithm.

As a byproduct, this provides concise sufficient conditions for *exponential* gains, addressing an open problem of Balcan, Hanneke, and Vaughan (2010). Specifically, consider the following definition, essentially similar to one explored by Balcan, Hanneke, and Vaughan (2010).

Definition 8 *For a concept space $\mathbb{C}$ and distribution $\mathcal{P}$, we say that $(\mathbb{C}, \mathcal{P})$ is* learnable at an exponential rate *if there exists an active learning algorithm achieving label complexity Λ such that, $\forall f \in \mathbb{C}$, $\Lambda(\varepsilon, f, \mathcal{P}) \in \operatorname{Polylog}(1/\varepsilon)$. We further say $\mathbb{C}$ is learnable at an exponential rate if there exists an active learning algorithm achieving label complexity Λ such that, for all distributions $\mathcal{P}$ and all $f \in \mathbb{C}$, $\Lambda(\varepsilon, f, \mathcal{P}) \in \operatorname{Polylog}(1/\varepsilon)$.*

5.1 The Label Complexity of Disagreement-Based Active Learning

As before, to establish a foundation to build upon, we begin by studying the label complexity gains achievable by disagreement-based active learning. From above, we already know that disagreement-based active learning is not sufficient to achieve the best possible gains; but as before, it will serve as a suitable starting place to gain intuition for how we might approach the problem of improving Meta-Algorithm 1 and quantifying the improvements achievable over passive learning by the resulting more sophisticated methods.

The upper bounds on the label complexity of disagreement-based learning in this subsection are essentially already known and available in the published literature (though in a slightly less general form). Specifically, we review (a modified version of) the method of Cohn, Atlas, and Ladner (1994), referred to as Meta-Algorithm 2 below, which was historically the original disagreement-based active learning algorithm. We then state the known results on the label complexities achievable by this method, in terms of a quantity known as the disagreement coefficient; that result is due to Hanneke (2011, 2007b). We further provide a novel lower bound on the label complexity of this method, again in terms of the disagreement coefficient; in particular, this shows that the stated upper bounds represent a fairly tight analysis of this method.

5.1.1 THE CAL ACTIVE LEARNING ALGORITHM

To begin, we consider the following simple disagreement-based method, typically referred to as CAL after its discoverers Cohn, Atlas, and Ladner (1994), though the version here is slightly modified compared to the original (see below). It essentially represents a refinement of Meta-Algorithm 0 to take greater advantage of the sequential aspect of active learning. That is, rather than requesting only two batches of labels, as in Meta-Algorithm 0, this method updates the version space after every label request, thus focusing the region of disagreement (and therefore the region in which it requests labels) after each label request.

Meta-Algorithm 2
Input: passive algorithm $\mathcal{A}_p$, label budget n
Output: classifier $\hat{h}$

---

0. $V \leftarrow \mathbb{C}$, $t \leftarrow 0$, $m \leftarrow 0$, $\mathcal{L} \leftarrow \{\}$
1. While $t < \lceil n/2 \rceil$ and $m \leq 2^n$
2. $m \leftarrow m+1$
3.   If $X_m \in \mathrm{DIS}(V)$
4.   Request the label Y_m of X_m and let $t \leftarrow t+1$
5.   Let $V \leftarrow V[(X_m, Y_m)]$
6. Let $\hat{\Delta} \leftarrow \hat{P}_m(\mathrm{DIS}(V))$
7. Do $\lfloor n/(6\hat{\Delta}) \rfloor$ times
8. $m \leftarrow m+1$
9.   If $X_m \in \mathrm{DIS}(V)$ and $t < n$
10.   Request the label Y_m of X_m and let $\hat{y} \leftarrow Y_m$ and $t \leftarrow t+1$
11.   Else let $\hat{y} = h(X_m)$ for an arbitrary $h \in V$
12.   Let $\mathcal{L} \leftarrow \mathcal{L} \cup \{(X_m, \hat{y})\}$ and $V \leftarrow V[(X_m, \hat{y})]$
13. Return $\mathcal{A}_p(\mathcal{L})$

The procedure is specified in terms of an estimator $\hat{P}_m$; for our purposes, we define this as in (13) of Appendix B.1 (with $k = 1$ there). Every example X_m added to the set $\mathcal{L}$ in Step 12 either has its label requested (Step 10) or inferred (Step 11). By the same Chernoff bound argument mentioned for the previous methods, we are guaranteed (with high probability) that the "$t < n$" constraint in Step 9 is always satisfied when $X_m \in \mathrm{DIS}(V)$. Since we assume $f \in \mathbb{C}$, an inductive argument shows that we will always have $f \in V$ as well; thus, every label requested *or* inferred will agree with f, and therefore the labels in $\mathcal{L}$ are all correct.

As with Meta-Algorithm 0, this method has two stages to it: one in which we focus on reducing the version space V, and a second in which we focus on constructing a set of labeled examples to feed into the passive algorithm. The original algorithm of Cohn, Atlas, and Ladner (1994) essentially used only the first stage, and simply returned any classifier in V after exhausting its budget for label requests. Here we have added the second stage (Steps 6-13) so that we can guarantee a certain conditional independence (given $|\mathcal{L}|$) among the examples fed into the passive algorithm, which is important for the general results (Theorem 10 below). Hanneke (2011) showed that the original (simpler) algorithm achieves the (less general) label complexity bound of Corollary 11 below.

5.1.2 EXAMPLES

Not surprisingly, by essentially the same argument as Meta-Algorithm 0, one can show Meta-Algorithm 2 satisfies the claim in Theorem 5. That is, Meta-Algorithm 2 is a universal activizer for $\mathbb{C}$ if and only if $\mathcal{P}(\partial f) = 0$ for every $\mathcal{P}$ and $f \in \mathbb{C}$. However, there are further results known on the label complexity achieved by Meta-Algorithm 2. Specifically, to illustrate the types of improvements achievable by Meta-Algorithm 2, consider our usual toy examples; as before, to simplify the explanation, for these examples we ignore the fact that $\hat{P}_m$ is only an estimate, as well as the "$t < n$" constraint in Step 9 (both of which will be addressed in the general results below).

First, consider threshold classifiers (Example 1) under a uniform $\mathcal{P}$ on $[0,1]$, and suppose $f = h_z \in \mathbb{C}$. Suppose the given passive algorithm has label complexity Λ_p. To get expected error at most ε in Meta-Algorithm 2, it suffices to have $|\mathcal{L}| \geq \Lambda_p(\varepsilon/2, f, \mathcal{P})$ with probability at least $1-\varepsilon/2$. Starting from any particular V set obtained in the algorithm, call it V_0, the set $\mathrm{DIS}(V_0)$ is simply the region between the largest negative example observed so far (say z_ℓ) and the smallest positive example observed so far (say z_r). With probability at least $1-\varepsilon/n$, at least one of the next $O(\log(n/\varepsilon))$ examples in this $[z_\ell, z_r]$ region will be in $[z_\ell + (1/3)(z_r - z_\ell), z_r - (1/3)(z_r - z_\ell)]$, so that after processing that example, we definitely have $\mathcal{P}(\mathrm{DIS}(V)) \leq (2/3)\mathcal{P}(\mathrm{DIS}(V_0))$. Thus, upon reaching Step 6, since we have made $n/2$ label requests, a union bound implies that with probability $1-\varepsilon/2$, we have $\mathcal{P}(\mathrm{DIS}(V)) \leq \exp\{-\Omega(n/\log(n/\varepsilon))\}$, and therefore $|\mathcal{L}| \geq \exp\{\Omega(n/\log(n/\varepsilon))\}$. Thus, for some value $\Lambda_a(\varepsilon, f, \mathcal{P}) = O(\log(\Lambda_p(\varepsilon/2, f, \mathcal{P}))\log(\log(\Lambda_p(\varepsilon/2, f, \mathcal{P}))/\varepsilon))$, any $n \geq \Lambda_a(\varepsilon, f, \mathcal{P})$ gives $|\mathcal{L}| \geq \Lambda_p(\varepsilon/2, f, \mathcal{P})$ with probability at least $1-\varepsilon/2$, so that the activized algorithm achieves label complexity $\Lambda_a(\varepsilon, f, \mathcal{P}) \in \mathrm{Polylog}(\Lambda_p(\varepsilon/2, f, \mathcal{P})/\varepsilon)$.

Consider also the intervals problem (Example 2) under a uniform $\mathcal{P}$ on $[0,1]$, and suppose $f = h_{[a,b]} \in \mathbb{C}$, for $b > a$. In this case, as with any disagreement-based algorithm, until the algorithm observes the first positive example (i.e., the first $X_m \in [a,b]$), it will request the label of every example (see the reasoning above for Meta-Algorithm 0). However, at every time after observing this first positive point, say x, the region $\mathrm{DIS}(V)$ is restricted to the region between the largest negative point less than x and smallest positive point, and the region between the largest positive point and the smallest negative point larger than x. For each of these two regions, the same arguments used for the threshold problem above can be applied to show that, with probability $1-O(\varepsilon)$, the region of disagreement is reduced by at least a constant fraction every $O(\log(n/\varepsilon))$ label requests, so that $|\mathcal{L}| \geq \exp\{\Omega(n/\log(n/\varepsilon))\}$. Thus, again the label complexity is of the form $O(\log(\Lambda_p(\varepsilon/2, f, \mathcal{P}))\log(\log(\Lambda_p(\varepsilon/2, f, \mathcal{P}))/\varepsilon))$, which is $\mathrm{Polylog}(\Lambda_p(\varepsilon/2, f, \mathcal{P})/\varepsilon)$, though this time there is a significant (additive) target-dependent term (roughly $\propto \frac{1}{b-a}\log(1/\varepsilon)$), accounting for the length of the initial phase before observing any positive examples. On the other hand, as with *any* disagreement-based algorithm, when $f = h_{[a,a]}$, because the algorithm never observes a positive example, it requests the label of every example it considers; in this case, by the same argument given for Meta-Algorithm 0, upon reaching Step 6 we have $\mathcal{P}(\mathrm{DIS}(V)) = 1$, so that $|\mathcal{L}| = O(n)$, and we observe no improvements for some passive algorithms $\mathcal{A}_p$.

A similar analysis can be performed for unions of i intervals under $\mathcal{P}$ uniform on $[0,1]$. In that case, we find that any $h_{\mathbf{z}} \in \mathbb{C}$ not representable (up to zero-probability differences) by a union of $i-1$ or fewer intervals allows for the exponential improvements of the type observed in the previous two examples; this time, the phase of exponentially decreasing $\mathcal{P}(\mathrm{DIS}(V))$ only occurs after observing an example in each of the i intervals and each of the $i-1$ negative regions separating the intervals, resulting in an additive term of roughly $\propto \frac{1}{\min_{1 \leq j < 2i} z_{j+1} - z_j} \log(i/\varepsilon)$ in the label complexity. However, any $h_{\mathbf{z}} \in \mathbb{C}$ representable (up to zero-probability differences) by a union of $i-1$ or fewer intervals has $\mathcal{P}(\partial h_{\mathbf{z}}) = 1$, which means $|\mathcal{L}| = O(n)$, and therefore (as with any disagreement-based algorithm) Meta-Algorithm 2 will not provide improvements for some passive algorithms $\mathcal{A}_p$.

5.1.3 THE DISAGREEMENT COEFFICIENT

Toward generalizing the arguments from the above examples, consider the following definition of Hanneke (2007b).

Definition 9 *For $\varepsilon \geq 0$, the* disagreement coefficient *of a classifier f with respect to a concept space $\mathbb{C}$ under a distribution $\mathcal{P}$ is defined as*

$$\theta_f(\varepsilon) = 1 \vee \sup_{r > \varepsilon} \frac{\mathcal{P}\left(\mathrm{DIS}(\mathrm{B}(f,r))\right)}{r}.$$

Also abbreviate $\theta_f = \theta_f(0)$.

Informally, the disagreement coefficient describes the rate of collapse of the region of disagreement, relative to the distance from f. It has been useful in characterizing the label complexities achieved by several disagreement-based active learning algorithms (Hanneke, 2007b, 2011; Dasgupta, Hsu, and Monteleoni, 2007; Beygelzimer, Dasgupta, and Langford, 2009; Wang, 2009; Koltchinskii, 2010; Beygelzimer, Hsu, Langford, and Zhang, 2010), and itself has been studied and bounded for various families of learning problems (Hanneke, 2007b, 2011; Balcan, Hanneke, and Vaughan, 2010; Friedman, 2009; Beygelzimer, Dasgupta, and Langford, 2009; Mahalanabis, 2011; Wang, 2011). See the paper of Hanneke (2011) for a detailed discussion of the disagreement coefficient, including its relationships to several related quantities, as well as a variety of general properties that it satisfies. In particular, below we use the fact that, for any constant $c \in [1,\infty)$, $\theta_f(\varepsilon) \leq \theta_f(\varepsilon/c) \leq c\theta_f(\varepsilon)$. Also note that $\mathcal{P}(\partial f) = 0$ if and only if $\theta_f(\varepsilon) = o(1/\varepsilon)$. See the papers of Friedman (2009) and Mahalanabis (2011) for some general conditions on $\mathbb{C}$ and $\mathcal{P}$, under which every $f \in \mathbb{C}$ has $\theta_f < \infty$, which (as we explain below) has particularly interesting implications for active learning (Hanneke, 2007b, 2011).

To build intuition about the behavior of the disagreement coefficient, we briefly go through its calculation for our usual toy examples from above. The first two of these calculations are taken from Hanneke (2007b), and the last is from Balcan, Hanneke, and Vaughan (2010). First, consider the thresholds problem (Example 1), and for simplicity suppose the distribution $\mathcal{P}$ is uniform on $[0,1]$. In this case, as in Section 3.2, $\mathrm{B}(h_z,r) = \{h_{z'} \in \mathbb{C} : |z'-z| \leq r\}$, and $\mathrm{DIS}(\mathrm{B}(h_z,r)) \subseteq [z-r, z+r)$ with equality for sufficiently small r. Therefore, $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_z,r))) \leq 2r$ (with equality for small r), and $\theta_{h_z}(\varepsilon) \leq 2$ with equality for sufficiently small ε. In particular, $\theta_{h_z} = 2$.

On the other hand, consider the intervals problem (Example 2), again under $\mathcal{P}$ uniform on $[0,1]$. This time, for $h_{[a,b]} \in \mathbb{C}$ with $b-a>0$, we have for $0<r<b-a$, $\mathrm{B}(h_{[a,b]},r) = \{h_{[a',b']} \in \mathbb{C} : |a-a'|+|b-b'| \leq r\}$, $\mathrm{DIS}(\mathrm{B}(h_{[a,b]},r)) \subseteq [a-r,a+r) \cup (b-r,b+r]$, and $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{[a,b]},r))) \leq 4r$ (with equality for sufficiently small r). But for $0 < b-a \leq r$, we have $\mathrm{B}(h_{[a,b]},r) \supseteq \{h_{[a',a']} : a' \in (0,1)\}$, so that $\mathrm{DIS}(\mathrm{B}(h_{[a,b]},r)) = (0,1)$ and $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{[a,b]},r))) = 1$. Thus, we generally have $\theta_{h_{[a,b]}}(\varepsilon) \leq \max\left\{\frac{1}{b-a}, 4\right\}$, with equality for sufficiently small ε. However, this last reasoning also indicates $\forall r > 0, \mathrm{B}(h_{[a,a]},r) \supseteq \{h_{[a',a']} : a' \in (0,1)\}$, so that $\mathrm{DIS}(\mathrm{B}(h_{[a,a]},r)) = (0,1)$ and $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{[a,a]},r))) = 1$; therefore, $\theta_{h_{[a,a]}}(\varepsilon) = \frac{1}{\varepsilon}$, the largest possible value for the disagreement coefficient; in particular, this also means $\theta_{h_{[a,a]}} = \infty$.

Finally, consider the unions of i intervals problem (Example 3), again under $\mathcal{P}$ uniform on $[0,1]$. First take any $h_{\mathbf{z}} \in \mathbb{C}$ such that any $h_{\mathbf{z}'} \in \mathbb{C}$ representable as a union of $i-1$ intervals has $\mathcal{P}(\{x : h_{\mathbf{z}}(x) \neq h_{\mathbf{z}'}(x)\}) > 0$. Then for $0 < r < \min_{1\leq j<2i} z_{j+1} - z_j$, $\mathrm{B}(h_{\mathbf{z}}, r) = \{h_{\mathbf{z}'} \in \mathbb{C} : \sum_{1\leq j\leq 2i} |z_j - z'_j| \leq r\}$, so that $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}}, r))) \leq 4ir$, with equality for sufficiently small r. For $r > \min_{1\leq j<2i} z_{j+1} - z_j$, $\mathrm{B}(h_{\mathbf{z}}, r)$ contains a set of classifiers that flips the labels (compared to $h_{\mathbf{z}}$) in that smallest region and uses the resulting extra interval to disagree with $h_{\mathbf{z}}$ on a tiny region at an arbitrary location (either by encompassing some point with a small interval, or by splitting an interval into two intervals separated by a small gap). Thus, $\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}}, r)) = (0,1)$, and $\mathcal{P}(\mathrm{DIS}(h_{\mathbf{z}}, r)) = 1$. So in total, $\theta_{h_{\mathbf{z}}}(\varepsilon) \leq \max\left\{\frac{1}{\min_{1\leq j<2i} z_{j+1}-z_j}, 4i\right\}$, with equality for sufficiently small ε. On the other hand, if $h_{\mathbf{z}} \in \mathbb{C}$ can be represented by a union of $i-1$ (or fewer) intervals, then we can use the extra interval to disagree with $h_{\mathbf{z}}$ on a tiny region at an arbitrary location, while still remaining in $\mathrm{B}(h_{\mathbf{z}}, r)$, so that $\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}}, r)) = (0,1)$, $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h_{\mathbf{z}}, r))) = 1$, and $\theta_{h_{\mathbf{z}}}(\varepsilon) = \frac{1}{\varepsilon}$; in particular, in this case we have $\theta_{h_{\mathbf{z}}} = \infty$.

5.1.4 General Upper Bounds on the Label Complexity of Meta-Algorithm 2

As mentioned, the disagreement coefficient has implications for the label complexities achievable by disagreement-based active learning. The intuitive reason for this is that, as the number of label requests increases, the *diameter* of the version space shrinks at a predictable rate. The disagreement coefficient then relates the diameter of the version space to the size of its region of disagreement, which in turn describes the probability of requesting a label. Thus, the expected frequency of label requests in the data sequence decreases at a predictable rate related to the disagreement coefficient, so that $|\mathcal{L}|$ in Meta-Algorithm 2 can be lower bounded by a function of the disagreement coefficient. Specifically, the following result was essentially established by Hanneke (2011, 2007b), though actually the result below is slightly more general than the original.

Theorem 10 *For any VC class $\mathbb{C}$, and any passive learning algorithm $\mathcal{A}_p$ achieving label complexity Λ_p, the active learning algorithm obtained by applying Meta-Algorithm 2 with $\mathcal{A}_p$ as input achieves a label complexity Λ_a that, for any distribution $\mathcal{P}$ and classifier $f \in \mathbb{C}$, satisfies*

$$\Lambda_a(\varepsilon, f, \mathcal{P}) = O\left(\theta_f\left(\Lambda_p(\varepsilon/2, f, \mathcal{P})^{-1}\right)\log^2 \frac{\Lambda_p(\varepsilon/2, f, \mathcal{P})}{\varepsilon}\right).$$

The proof of Theorem 10 is similar to the original result of Hanneke (2011, 2007b), with only minor modifications to account for using $\mathcal{A}_p$ instead of returning an arbitrary element of V. The formal details are implicit in the proof of Theorem 16 below (since Meta-Algorithm 2 is essentially identical to the $k=1$ round of Meta-Algorithm 3, defined below). We also have the following simple corollaries.

Corollary 11 *For any VC class $\mathbb{C}$, there exists a passive learning algorithm $\mathcal{A}_p$ such that, for every $f \in \mathbb{C}$ and distribution $\mathcal{P}$, the active learning algorithm obtained by applying Meta-Algorithm 2 with $\mathcal{A}_p$ as input achieves label complexity*

$$\Lambda_a(\varepsilon, f, \mathcal{P}) = O\left(\theta_f(\varepsilon)\log^2(1/\varepsilon)\right).$$

Proof The one-inclusion graph algorithm of Haussler, Littlestone, and Warmuth (1994) is a passive learning algorithm achieving label complexity $\Lambda_p(\varepsilon, f, \mathcal{P}) \le d/\varepsilon$. Plugging this into Theorem 10, using the fact that $\theta_f(\varepsilon/2d) \le 2d\theta_f(\varepsilon)$, and simplifying, we arrive at the result. In fact, we will see in the proof of Theorem 16 that incurring this extra constant factor of d is not actually necessary. ■

Corollary 12 *For any VC class $\mathbb{C}$ and distribution $\mathcal{P}$, if $\forall f \in \mathbb{C}$, $\theta_f < \infty$, then $(\mathbb{C}, \mathcal{P})$ is learnable at an exponential rate. If this is true for all $\mathcal{P}$, then $\mathbb{C}$ is learnable at an exponential rate.*

Proof The first claim follows directly from Corollary 11, since $\theta_f(\varepsilon) \le \theta_f$. The second claim then follows from the fact that Meta-Algorithm 2 is adaptive to $\mathcal{P}$ (has no direct dependence on $\mathcal{P}$ except via the data). ■

Aside from the disagreement coefficient and Λ_p terms, the other constant factors hidden in the big-O in Theorem 10 are only $\mathbb{C}$-dependent (i.e., independent of f and $\mathcal{P}$). As mentioned, if we are only interested in achieving the label complexity bound of Corollary 11, we can obtain this result more directly by the simpler original algorithm of Cohn, Atlas, and Ladner (1994) via the analysis of Hanneke (2011, 2007b).

5.1.5 General Lower Bounds on the Label Complexity of Meta-Algorithm 2

It is also possible to prove a kind of *lower bound* on the label complexity of Meta-Algorithm 2 in terms of the disagreement coefficient, so that the dependence on the disagreement coefficient in Theorem 10 is unavoidable. Specifically, there are two simple observations that intuitively explain the possibility of such lower bounds. The first observation is that the expected number of label requests Meta-Algorithm 2 makes among the first $\lceil 1/\varepsilon \rceil$ unlabeled examples is at least $\theta_f(\varepsilon)/2$ (assuming it does not halt first). Similarly, the second observation is that, to arrive at a region of disagreement with expected probability mass less than $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,\varepsilon)))/2$, Meta-Algorithm 2 requires a budget n of size at least $\theta_f(\varepsilon)/2$. These observations are formalized in Appendix C as Lemmas 47 and 48. The relevance of these observations in the context of deriving lower bounds based on the disagreement coefficient is clear. In particular, we can use the latter of these insights to arrive at the following theorem, which essentially complements Theorem 10, showing that it cannot generally be improved beyond reducing the constants and logarithmic factors, without altering the algorithm or introducing additional $\mathcal{A}_p$-dependent quantities in the label complexity bound. The proof is included in Appendix C.

Theorem 13 *For any set of classifiers $\mathbb{C}$, $f \in \mathbb{C}$, distribution $\mathcal{P}$, and nonincreasing function $\lambda : (0,1) \to \mathbb{N}$, there exists a passive learning algorithm $\mathcal{A}_p$ achieving a label complexity Λ_p with $\Lambda_p(\varepsilon, f, \mathcal{P}) = \lambda(\varepsilon)$ for all $\varepsilon > 0$, such that if Meta-Algorithm 2, with $\mathcal{A}_p$ as its argument, achieves label complexity Λ_a, then*

$$\Lambda_a(\varepsilon, f, \mathcal{P}) = \Omega\left(\theta_f\left(\Lambda_p(2\varepsilon, f, \mathcal{P})^{-1}\right)\right).$$

Recall that there are many natural learning problems for which $\theta_f = \infty$, and indeed where $\theta_f(\varepsilon) = \Omega(1/\varepsilon)$: for instance, intervals with $f = h_{[a,a]}$ under uniform $\mathcal{P}$, or unions of i intervals under uniform $\mathcal{P}$ with f representable as $i-1$ or fewer intervals. Thus, since we have just seen that the improvements gained by disagreement-based methods are well-characterized by the disagreement coefficient, if we would like to achieve exponential improvements over passive learning for these problems, we will need to move beyond these disagreement-based methods. In the subsections that follow, we will use an alternative algorithm and analysis, and prove a general result that is always at least as good as Theorem 10 (in a big-O sense), and often significantly better (in a little-o sense). In particular, it leads to a sufficient condition for learnability at an exponential rate, strictly more general than that of Corollary 12.

5.2 An Improved Activizer

In this subsection, we define a new active learning method based on shattering, as in Meta-Algorithm 1, but which also takes fuller advantage of the sequential aspect of active learning, as in Meta-Algorithm 2. We will see that this algorithm can be analyzed in a manner analogous to the disagreement coefficient analysis of Meta-Algorithm 2, leading to a new and often dramatically-improved label complexity bound. Specifically, consider the following meta-algorithm.

Meta-Algorithm 3
Input: passive algorithm $\mathcal{A}_p$, label budget n
Output: classifier $\hat{h}$

$V \leftarrow V_0 = \mathbb{C}$, $T_0 \leftarrow \lceil 2n/3 \rceil$, $t \leftarrow 0$, $m \leftarrow 0$
For $k = 1, 2, \ldots, d+1$
  Let $\mathcal{L}_k \leftarrow \{\}$, $T_k \leftarrow T_{k-1} - t$, and let $t \leftarrow 0$
  While $t < \lceil T_k/4 \rceil$ and $m \leq k \cdot 2^n$
    $m \leftarrow m+1$
    If $\hat{P}_m\left(S \in \mathcal{X}^{k-1} : V \text{ shatters } S \cup \{X_m\} | V \text{ shatters } S\right) \geq 1/2$
      Request the label Y_m of X_m, and let $\hat{y} \leftarrow Y_m$ and $t \leftarrow t+1$
    Else let $\hat{y} \leftarrow \underset{y \in \{-1,+1\}}{\operatorname{argmax}} \hat{P}_m\left(S \in \mathcal{X}^{k-1} : V[(X_m, -y)] \text{ does not shatter } S | V \text{ shatters } S\right)$
    Let $V \leftarrow V_m = V_{m-1}[(X_m, \hat{y})]$
  $\hat{\Delta}^{(k)} \leftarrow \hat{P}_m\left(x : \hat{P}\left(S \in \mathcal{X}^{k-1} : V \text{ shatters } S \cup \{x\} | V \text{ shatters } S\right) \geq 1/2\right)$
  Do $\lfloor T_k/(3\hat{\Delta}^{(k)}) \rfloor$ times
    $m \leftarrow m+1$
    If $\hat{P}_m\left(S \in \mathcal{X}^{k-1} : V \text{ shatters } S \cup \{X_m\} | V \text{ shatters } S\right) \geq 1/2$ and $t < \lfloor 3T_k/4 \rfloor$
      Request the label Y_m of X_m, and let $\hat{y} \leftarrow Y_m$ and $t \leftarrow t+1$
    Else, let $\hat{y} \leftarrow \underset{y \in \{-1,+1\}}{\operatorname{argmax}} \hat{P}_m\left(S \in \mathcal{X}^{k-1} : V[(X_m, -y)] \text{ does not shatter } S | V \text{ shatters } S\right)$
    Let $\mathcal{L}_k \leftarrow \mathcal{L}_k \cup \{(X_m, \hat{y})\}$ and $V \leftarrow V_m = V_{m-1}[(X_m, \hat{y})]$
Return ActiveSelect$(\{\mathcal{A}_p(\mathcal{L}_1), \mathcal{A}_p(\mathcal{L}_2), \ldots, \mathcal{A}_p(\mathcal{L}_{d+1})\}, \lfloor n/3 \rfloor, \{X_{m+1}, X_{m+2}, \ldots\})$

As before, the procedure is specified in terms of estimators $\hat{P}_m$. Again, these can be defined in a variety of ways, as long as they converge (at a fast enough rate) to their respective true probabilities. For the results below, we will use the definitions given in Appendix B.1: that is, the same definitions used in Meta-Algorithm 1. Following the same argument as for Meta-Algorithm 1, one can show that Meta-Algorithm 3 is a universal activizer for $\mathbb{C}$, for any VC class $\mathbb{C}$. However, we can also obtain more detailed results in terms of a generalization of the disagreement coefficient given below.

As with Meta-Algorithm 1, this procedure has three main components: one in which we focus on reducing the version space V, one in which we focus on collecting a (conditionally) i.i.d. sample to feed into $\mathcal{A}_p$, and one in which we select from among the $d+1$ executions of $\mathcal{A}_p$. However, unlike Meta-Algorithm 1, here the first stage is also broken up based on the value of k, so that each k has its own first and second stages, rather than sharing a single first stage. Again, the choice of the number of (unlabeled) examples processed in each second stage guarantees (by a Chernoff bound) that the "$t < \lfloor 3T_k/4 \rfloor$" constraint in Step 12 is redundant. Depending on the type of label complexity result we wish to prove, this multistage architecture is sometimes avoidable. In particular, as with Corollary 11 above, to directly achieve the label complexity bound in Corollary 17 below, we can use a much simpler approach that replaces Steps 9-16, instead simply returning an arbitrary element of V upon termination.

Within each value of k, Meta-Algorithm 3 behaves analogous to Meta-Algorithm 2, requesting the label of an example only if it cannot infer the label from known information, and updating the version space V after every label request; however, unlike Meta-Algorithm 2, for values of $k > 1$, the mechanism for inferring a label is based on shatterable sets, as in Meta-Algorithm 1, and is motivated by the same argument of splitting V into subsets containing arbitrarily good classifiers (see the discussion in Section 4.1). Also unlike Meta-Algorithm 2, even the inferred labels can be used to reduce the set V (Steps 8 and 15), since they are not only correct but also potentially informative in the sense that $x \in \mathrm{DIS}(V)$. As with Meta-Algorithm 1, the key to obtaining improvement guarantees is that some value of k has $|\mathcal{L}_k| \gg n$, while maintaining that all of the labels in $\mathcal{L}_k$ are correct; ActiveSelect then guarantees the overall performance is not too much worse than that obtained by $\mathcal{A}_p(\mathcal{L}_k)$ for this value of k.

To build intuition about the behavior of Meta-Algorithm 3, let us consider our usual toy examples, again under a uniform distribution $\mathcal{P}$ on $[0,1]$; as before, for simplicity we ignore the fact that $\hat{P}_m$ is only an estimate, as well as the constraint on t in Step 12 and the effectiveness of ActiveSelect, all of which will be addressed in the general analysis. First, for the behavior of the algorithm for thresholds and nonzero-width intervals, we may simply refer to the discussion of Meta-Algorithm 2, since the $k=1$ round of Meta-Algorithm 3 is essentially identical to Meta-Algorithm 2; in this case, we have already seen that $|\mathcal{L}_1|$ grows as $\exp\{\Omega(n/\log(n/\varepsilon))\}$ for thresholds, and does so for nonzero-width intervals after some initial period of slow growth related to the width of the target interval (i.e., the period before finding the first positive example). As with Meta-Algorithm 1, for zero-width intervals, we must look to the $k=2$ round of Meta-Algorithm 3 to find improvements. Also as with Meta-Algorithm 1, for sufficiently large n, every X_m processed in the $k=2$ round will have its label inferred (correctly) in Step 7 or 14 (i.e., it does not request any labels). But this means we reach Step 9 with $m = 2 \cdot 2^n + 1$; furthermore, in these circumstances the definition of $\hat{P}_m$ from Appendix B.1 guarantees (for sufficiently large n) that $\hat{\Delta}^{(2)} = 2/m$, so that $|\mathcal{L}_2| \propto n \cdot m = \Omega(n \cdot 2^n)$. Thus, we expect the label complexity gains to be *exponentially improved* compared to $\mathcal{A}_p$.

For a more involved example, consider unions of 2 intervals (Example 3), under uniform $\mathcal{P}$ on $[0,1]$, and suppose $f = h_{(a,b,a,b)}$ for $b - a > 0$; that is, the target function is representable as a single nonzero-width interval $[a,b] \subset (0,1)$. As we have seen, $\partial f = (0,1)$ in this case, so that disagreement-based methods are ineffective at improving over passive. This also means the $k=1$ round of Meta-Algorithm 3 will not provide improvements (i.e., $|\mathcal{L}_1| = O(n)$). However, consider the $k=2$ round. As discussed in Section 4.2, for sufficiently large n, after the first round ($k=1$) the set V is such that any label we infer in the $k=2$ round will be correct. Thus, it suffices to determine how large the set $\mathcal{L}_2$ becomes. By the same reasoning as in Section 4.2, for sufficiently large n, the examples X_m whose labels are requested in Step 6 are precisely those *not* separated from both a and b by at least one of the $m-1$ examples already processed (since V is consistent with the labels of all $m-1$ of those examples). But this is the same set of points Meta-Algorithm 2 would query for the *intervals* example in Section 5.1; thus, the same argument used there implies that in this problem we have $|\mathcal{L}_2| \geq \exp\{\Omega(n/\log(n/\varepsilon))\}$ with probability $1-\varepsilon/2$, which means we should expect a label complexity of $O(\log(\Lambda_p(\varepsilon/2, f, \mathcal{P}))\log(\log(\Lambda_p(\varepsilon/2, f, \mathcal{P}))/\varepsilon))$, where Λ_p is the label complexity of $\mathcal{A}_p$. For the case $f = h_{(a,a,a,a)}$, $k=3$ is the relevant round, and the analysis goes similarly to the $h_{[a,a]}$ scenario for intervals above. Unions of $i > 2$ intervals can be studied analogously, with the appropriate value of k to analyze being determined by the number of intervals required to represent the target up to zero-probability differences (see the discussion in Section 4.2).

5.3 Beyond the Disagreement Coefficient

In this subsection, we introduce a new quantity, a generalization of the disagreement coefficient, which we will later use to provide a general characterization of the improvements achievable by Meta-Algorithm 3, analogous to how the disagreement coefficient characterized the improvements achievable by Meta-Algorithm 2 in Theorem 10. First, let us define the following generalization of the disagreement core.

Definition 14 *For an integer $k \geq 0$, define the* k-dimensional shatter core *of a classifier f with respect to a set of classifiers $\mathcal{H}$ and probability measure P as*

$$\partial^k_{\mathcal{H},P} f = \lim_{r \to 0} \left\{ S \in \mathcal{X}^k : \mathrm{B}_{\mathcal{H},P}(f,r) \text{ shatters } S \right\}.$$

As before, when $P = \mathcal{P}$, and $\mathcal{P}$ is clear from the context, we will abbreviate $\partial^k_{\mathcal{H}} f = \partial^k_{\mathcal{H},\mathcal{P}} f$, and when we also intend $\mathcal{H} = \mathbb{C}$, the *full* concept space, and $\mathbb{C}$ is clearly defined in the given context, we further abbreviate $\partial^k f = \partial^k_{\mathbb{C}} f = \partial^k_{\mathbb{C},\mathcal{P}} f$. We have the following definition, which will play a key role in the label complexity bounds below.

Definition 15 *For any concept space $\mathbb{C}$, distribution $\mathcal{P}$, and classifier f, $\forall k \in \mathbb{N}$, $\forall \varepsilon \geq 0$, define*

$$\theta^{(k)}_f(\varepsilon) = 1 \vee \sup_{r > \varepsilon} \frac{\mathcal{P}^k\left(S \in \mathcal{X}^k : \mathrm{B}(f,r) \text{ shatters } S\right)}{r}.$$

Then define

$$\tilde{d}_f = \min\left\{ k \in \mathbb{N} : \mathcal{P}^k\left(\partial^k f\right) = 0 \right\}$$

and

$$\tilde{\theta}_f(\varepsilon) = \theta^{(\tilde{d}_f)}_f(\varepsilon).$$

Also abbreviate $\theta^{(k)}_f = \theta^{(k)}_f(0)$ and $\tilde{\theta}_f = \tilde{\theta}_f(0)$.

We might refer to the quantity $\theta^{(k)}_f(\varepsilon)$ as the order-k (or k-dimensional) disagreement coefficient, as it represents a direct generalization of the disagreement coefficient $\theta_f(\varepsilon)$. However, rather than merely measuring the rate of collapse of the probability of *disagreement* (one-dimensional shatterability), $\theta^{(k)}_f(\varepsilon)$ measures the rate of collapse of the probability of *k-dimensional shatterability*. In particular, we have $\tilde{\theta}_f(\varepsilon) = \theta^{(\tilde{d}_f)}_f(\varepsilon) \leq \theta^{(1)}_f(\varepsilon) = \theta_f(\varepsilon)$, so that this new quantity is never larger than the disagreement coefficient. However, unlike the disagreement coefficient, we *always* have $\tilde{\theta}_f(\varepsilon) = o(1/\varepsilon)$ for VC classes $\mathbb{C}$. In fact, we could equivalently define $\tilde{\theta}_f(\varepsilon)$ as the value of $\theta_f^{(k)}(\varepsilon)$ for the smallest k with $\theta_f^{(k)}(\varepsilon) = o(1/\varepsilon)$. Additionally, we will see below that there are many interesting cases where $\theta_f = \infty$ (even $\theta_f(\varepsilon) = \Omega(1/\varepsilon)$) but $\tilde{\theta}_f < \infty$ (e.g., intervals with a zero-width target, or unions of i intervals where the target is representable as a union of $i-1$ or fewer intervals). As was the case for θ_f, we will see that showing $\tilde{\theta}_f < \infty$ for a given learning problem has interesting implications for the label complexity of active learning (Corollary 18 below). In the process, we have also defined the quantity $\tilde{d}_f$, which may itself be of independent interest in the asymptotic analysis of learning in general. For VC classes, $\tilde{d}_f$ always exists, and in fact is at most $d+1$ (since $\mathbb{C}$ cannot shatter any $d+1$ points). When $d = \infty$, the quantity $\tilde{d}_f$ might not be defined (or defined as ∞), in which case $\tilde{\theta}_f(\varepsilon)$ is also not defined; in this work we restrict our discussion to VC classes, so that this issue never comes up; Section 7 discusses possible extensions to classes of infinite VC dimension.

We should mention that the restriction of $\tilde{\theta}_f(\varepsilon) \geq 1$ in the definition is only for convenience, as it simplifies the theorem statements and proofs below. It is not fundamental to the definition, and can be removed (at the expense of slightly more complicated theorem statements). In fact, this only makes a difference to the value of $\tilde{\theta}_f(\varepsilon)$ in some (seemingly unusual) degenerate cases. The same is true of $\theta_f(\varepsilon)$ in Definition 9.

The process of calculating $\tilde{\theta}_f(\varepsilon)$ is quite similar to that for the disagreement coefficient; we are interested in describing $\mathrm{B}(f,r)$, and specifically the variety of behaviors of elements of $\mathrm{B}(f,r)$ on points in $\mathcal{X}$, in this case with respect to shattering. To illustrate the calculation of $\tilde{\theta}_f(\varepsilon)$, consider our usual toy examples, again under $\mathcal{P}$ uniform on $[0,1]$. For the thresholds example (Example 1), we have $\tilde{d}_f = 1$, so that $\tilde{\theta}_f(\varepsilon) = \theta_f^{(1)}(\varepsilon) = \theta_f(\varepsilon)$, which we have seen is equal 2 for small ε. Similarly, for the intervals example (Example 2), any $f = h_{[a,b]} \in \mathbb{C}$ with $b-a > 0$ has $\tilde{d}_f = 1$, so that $\tilde{\theta}_f(\varepsilon) = \theta_f^{(1)}(\varepsilon) = \theta_f(\varepsilon)$, which for sufficiently small ε, is equal $\max\left\{\frac{1}{b-a}, 4\right\}$. Thus, for these two examples, $\tilde{\theta}_f(\varepsilon) = \theta_f(\varepsilon)$. However, continuing the intervals example, consider $f = h_{[a,a]} \in \mathbb{C}$. In this case, we have seen $\partial^1 f = \partial f = (0,1)$, so that $\mathcal{P}(\partial^1 f) = 1 > 0$. For any $x_1, x_2 \in (0,1)$ with $0 < |x_1 - x_2| \leq r$, $\mathrm{B}(f,r)$ can shatter (x_1, x_2), specifically using the classifiers $\{h_{[x_1,x_2]}, h_{[x_1,x_1]}, h_{[x_2,x_2]}, h_{[x_3,x_3]}\}$ for any $x_3 \in (0,1) \setminus \{x_1, x_2\}$. However, for any $x_1, x_2 \in (0,1)$ with $|x_1 - x_2| > r$, no element of $\mathrm{B}(f,r)$ classifies both as $+1$ (as it would need width greater than r, and thus would have distance from $h_{[a,a]}$ greater than r). Therefore, $\{S \in \mathcal{X}^2 : \mathrm{B}(f,r) \text{ shatters } S\} = \{(x_1,x_2) \in (0,1)^2 : 0 < |x_1 - x_2| \leq r\}$; this latter set has probability $2r(1-r) + r^2 = (2-r) \cdot r$, which shrinks to 0 as $r \to 0$. Therefore, $\tilde{d}_f = 2$. Furthermore, this shows $\tilde{\theta}_f(\varepsilon) = \theta_f^{(2)}(\varepsilon) = \sup_{r > \varepsilon}(2-r) = 2 - \varepsilon \leq 2$. Contrasting this with $\theta_f(\varepsilon) = 1/\varepsilon$, we see $\tilde{\theta}_f(\varepsilon)$ is significantly smaller than the disagreement coefficient; in particular, $\tilde{\theta}_f = 2 < \infty$, while $\theta_f = \infty$.

Consider also the space of unions of i intervals (Example 3) under $\mathcal{P}$ uniform on $[0,1]$. In this case, we have already seen that, for any $f = h_{\mathbf{z}} \in \mathbb{C}$ not representable (up to zero-probability differences) by a union of $i-1$ or fewer intervals, we have $\mathcal{P}(\partial^1 f) = \mathcal{P}(\partial f) = 0$, so that $\tilde{d}_f = 1$, and $\tilde{\theta}_f = \theta_f^{(1)} = \theta_f = \max\left\{\frac{1}{\min\limits_{1 \leq p < 2i} z_{p+1} - z_p}, 4i\right\}$. To generalize this, suppose $f = h_{\mathbf{z}}$ is minimally representable as a union of any number $j \leq i$ of intervals of nonzero width: $[z_1, z_2] \cup [z_3, z_4] \cup \cdots \cup [z_{2j-1}, z_{2j}]$, with $0 < z_1 < z_2 < \cdots < z_{2j} < 1$. For our purposes, this is fully general, since every element of $\mathbb{C}$ has distance zero to some $h_{\mathbf{z}}$ of this type, and $\tilde{\theta}_h = \tilde{\theta}_{h'}$ for any h, h' with $\mathcal{P}(x : h(x) \neq h'(x)) = 0$. Now for any $k < i - j + 1$, and any $S = (x_1, \ldots, x_k) \in \mathcal{X}^k$ with all elements distinct, the set $\mathrm{B}(f,r)$ can shatter S, as follows. Begin with the intervals $[z_{2p-1}, z_{2p}]$ as above, and modify the classifier in the following way for each labeling of S. For any of the x_ℓ values we wish to label $+1$, if it is already in an interval $[z_{2p-1}, z_{2p}]$, we do nothing; if it is not in one of the $[z_{2p-1}, z_{2p}]$ intervals, we add the interval $[x_\ell, x_\ell]$ to the classifier. For any of the x_ℓ values we wish to label -1, if it is not in any interval $[z_{2p-1}, z_{2p}]$, we do nothing; if it is in some interval $[z_{2p-1}, z_{2p}]$, we split the interval by setting to -1 the labels in a small region $(x_\ell - \gamma, x_\ell + \gamma)$, for $\gamma < r/k$ chosen small enough so that $(x_\ell - \gamma, x_\ell + \gamma)$ does not contain any other element of S. These operations add at most k new intervals to the minimal representation of the classifier as a union of intervals, which therefore has at most $j+k \leq i$ intervals. Furthermore, the classifier disagrees with f on a set of size at most r, so that it is contained in $\mathrm{B}(f,r)$. We therefore have $\mathcal{P}^k(S \in \mathcal{X}^k : \mathrm{B}(f,r) \text{ shatters } S) = 1$. However, note that for $0 < r < \min\limits_{1 \leq p < 2j} z_{p+1} - z_p$, for any k and $S \in \mathcal{X}^k$ with all elements of $S \cup \{z_p : 1 \leq p \leq 2j\}$ separated by a distance greater than r, classifying the points in S opposite to f while remaining r-close to f requires us to increase to a minimum of $j+k$ intervals. Thus, for $k = i-j+1$, any $S = (x_1, \ldots, x_k) \in \mathcal{X}^k$ with $\min\limits_{y_1, y_2 \in S \cup \{z_p\}_p : y_1 \neq y_2} |y_1 - y_2| > r$ is *not* shatterable by $\mathrm{B}(f,r)$. We therefore have $\{S \in \mathcal{X}^k : \mathrm{B}(f,r) \text{ shatters } S\} \subseteq \left\{S \in \mathcal{X}^k : \min\limits_{y_1, y_2 \in S \cup \{z_p\}_p : y_1 \neq y_2} |y_1 - y_2| \leq r\right\}$. For $r < \min\limits_{1 \leq p < 2j} z_{p+1} - z_p$, we can bound the probability of this latter set by considering sampling the points x_ℓ sequentially; the probability the ℓ^{th} point is within r of one of $x_1, \ldots, x_{\ell-1}, z_1, \ldots, z_{2j}$ is at most $2r(2j+\ell-1)$, so (by a union bound) the probability any of the k points $x_1, \ldots, x_k$ is within r of any other or any of $z_1, \ldots, z_{2j}$ is at most $\sum_{\ell=1}^{k} 2r(2j+\ell-1) = 2r\left(2jk + \binom{k}{2}\right) = (1+i-j)(i+3j)r$. Since this approaches zero as $r \to 0$, we have $\tilde{d}_f = i-j+1$. Furthermore, this analysis shows $\tilde{\theta}_f = \theta_f^{(i-j+1)} \leq \max\left\{\frac{1}{\min\limits_{1 \leq p < 2j} z_{p+1} - z_p}, (1+i-j)(i+3j)\right\}$. In fact, careful further inspection reveals that this upper bound is tight (i.e., this is the exact value of $\tilde{\theta}_f$). Recalling that $\theta_f(\varepsilon) = 1/\varepsilon$ for $j < i$, we see that again $\tilde{\theta}_f(\varepsilon)$ is significantly smaller than the disagreement coefficient; in particular, $\tilde{\theta}_f < \infty$ while $\theta_f = \infty$.

Of course, for the quantity $\tilde{\theta}_f(\varepsilon)$ to be truly useful, we need to be able to describe its behavior for families of learning problems beyond these simple toy problems. Fortunately, as with the disagreement coefficient, for learning problems with simple "geometric" interpretations, one can typically bound the value of $\tilde{\theta}_f$ without too much difficulty. For instance, consider $\mathcal{X}$ the surface of a unit hypersphere in p-dimensional Euclidean space (with $p \geq 3$), with $\mathcal{P}$ uniform on $\mathcal{X}$, and $\mathbb{C}$ the space of linear separators: $\mathbb{C} = \{h_{\mathbf{w},b}(\mathbf{x}) = \mathbb{1}^{\pm}_{[0,\infty)}(\mathbf{w} \cdot \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}\}$. Balcan, Hanneke, and Vaughan (2010) proved that $(\mathbb{C}, \mathcal{P})$ is learnable at an exponential rate, by a specialized argument for this space. In the process, they established that for any $f \in \mathbb{C}$ with $\mathcal{P}(x : f(x) = +1) \in (0,1)$, $\theta_f < \infty$; in fact, a similar argument shows $\theta_f \leq 4\pi\sqrt{p}/\min_y \mathcal{P}(x : f(x) = y)$. Thus, in this case, $\tilde{d}_f = 1$, and $\tilde{\theta}_f = \theta_f < \infty$. However, consider $f \in \mathbb{C}$ with $\mathcal{P}(x : f(x) = y) = 1$ for some $y \in \{-1, +1\}$. In this case, every $h \in \mathbb{C}$ with $\mathcal{P}(x : h(x) = -y) \leq r$ has $\mathcal{P}(x : h(x) \neq f(x)) \leq r$ and is therefore contained in $\mathrm{B}(f,r)$. In particular, for any $x \in \mathcal{X}$, there is such an h that disagrees with f on only a small spherical cap containing x, so that $\mathrm{DIS}(\mathrm{B}(f,r)) = \mathcal{X}$ for all $r > 0$. But this means $\partial f = \mathcal{X}$, which implies $\theta_f(\varepsilon) = 1/\varepsilon$ and $\tilde{d}_f > 1$. However, let us examine the value of $\theta_f^{(2)}$. Let $A_p = \frac{2\pi^{p/2}}{\Gamma\left(\frac{p}{2}\right)}$ denote the surface area of the unit sphere in $\mathbb{R}^p$, and let $C_p(z) = \frac{1}{2}A_p I_{2z-z^2}\left(\frac{p-1}{2}, \frac{1}{2}\right)$ denote the surface area of a spherical cap of height $z \in (0,1)$ (Li, 2011), where $I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1} \mathrm{d}t$

is the regularized incomplete beta function. In particular, since $\sqrt{\frac{p}{12}} \leq \frac{\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{p-1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \leq \frac{1}{2}\sqrt{p-2}$, the probability mass $\frac{C_p(z)}{A_p} = \frac{1}{2}\frac{\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{p-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}\int_0^{2z-z^2} t^{\frac{p-3}{2}}(1-t)^{-\frac{1}{2}}\mathrm{d}t$ contained in a spherical cap of height z satisfies

$$\frac{C_p(z)}{A_p} \geq \frac{1}{2}\sqrt{\frac{p}{12}}\int_0^{2z-z^2} t^{\frac{p-3}{2}}\mathrm{d}t = \sqrt{\frac{p}{12}}\frac{(2z-z^2)^{\frac{p-1}{2}}}{p-1} \geq \frac{(2z-z^2)^{\frac{p-1}{2}}}{\sqrt{12p}}, \tag{2}$$

and letting $\bar{z} = \min\{z, 1/2\}$, also satisfies

$$\begin{aligned}\frac{C_p(z)}{A_p} &\leq \frac{2C_p(\bar{z})}{A_p} \leq \frac{1}{2}\sqrt{p-2}\int_0^{2\bar{z}-\bar{z}^2} t^{\frac{p-3}{2}}(1-t)^{-\frac{1}{2}}\mathrm{d}t \\ &\leq \sqrt{p-2}\int_0^{2z-z^2} t^{\frac{p-3}{2}}\mathrm{d}t = \frac{2\sqrt{p-2}}{p-1}(2z-z^2)^{\frac{p-1}{2}} \leq \frac{(2z-z^2)^{\frac{p-1}{2}}}{\sqrt{p/6}} \leq \frac{(2z)^{\frac{p-1}{2}}}{\sqrt{p/6}}. \end{aligned}\tag{3}$$

Consider any linear separator $h \in \mathrm{B}(f,r)$ for $r < 1/2$, and let $z(h)$ denote the height of the spherical cap where $h(x) = -y$. Then (2) indicates the probability of this region is at least $\frac{(2z(h)-z(h)^2)^{\frac{p-1}{2}}}{\sqrt{12p}}$. Since $h \in \mathrm{B}(f,r)$, we know this probability mass is at most r, and we therefore have $2z(h) - z(h)^2 \leq \left(\sqrt{12pr}\right)^{\frac{2}{p-1}}$. Now for any $x_1 \in \mathcal{X}$, the set of $x_2 \in \mathcal{X}$ for which $\mathrm{B}(f,r)$ shatters (x_1,x_2) is equivalent to the set $\mathrm{DIS}(\{h \in \mathrm{B}(f,r) : h(x_1) = -y\})$. But if $h(x_1) = -y$, then x_1 is in the aforementioned spherical cap associated with h. A little trigonometry reveals that, for any spherical cap of height $z(h)$, any two points on the surface of this cap are within distance $2\sqrt{2z(h)-z(h)^2} \leq 2\left(\sqrt{12pr}\right)^{\frac{1}{p-1}}$ of each other. Thus, for any point x_2 further than $2\left(\sqrt{12pr}\right)^{\frac{1}{p-1}}$ from x_1, it must be outside the spherical cap associated with h, which means $h(x_2) = y$. But this is true for every $h \in \mathrm{B}(f,r)$ with $h(x_1) = -y$, so that $\mathrm{DIS}(\{h \in \mathrm{B}(f,r) : h(x_1) = -y\})$ is contained in the spherical cap of all elements of $\mathcal{X}$ within distance $2\left(\sqrt{12pr}\right)^{\frac{1}{p-1}}$ of x_1; a little more trigonometry reveals that the height of this spherical cap is $2\left(\sqrt{12pr}\right)^{\frac{2}{p-1}}$. Then (3) indicates the probability mass in this region is at most $\frac{2^{p-1}\sqrt{12pr}}{\sqrt{p/6}} = 2^p\sqrt{18}r$. Thus, $\mathcal{P}^2((x_1,x_2) : \mathrm{B}(f,r) \text{ shatters } (x_1,x_2)) = \int \mathcal{P}(\mathrm{DIS}(\{h \in \mathrm{B}(f,r) : h(x_1) = -y\}))\mathcal{P}(\mathrm{d}x_1) \leq 2^p\sqrt{18}r$. In particular, since this approaches zero as $r \to 0$, we have $\tilde{d}_f = 2$. This also shows that $\tilde{\theta}_f = \theta_f^{(2)} \leq 2^p\sqrt{18}$, a finite constant (albeit a rather large one). Following similar reasoning, using the opposite inequalities as appropriate, and taking r sufficiently small, one can also show $\tilde{\theta}_f \geq 2^p/(12\sqrt{2})$.

5.4 Bounds on the Label Complexity of Activized Learning

We have seen above that in the context of several examples, Meta-Algorithm 3 can offer significant advantages in label complexity over any given passive learning algorithm, and indeed also over disagreement-based active learning in many cases. In this subsection, we present a general result characterizing the magnitudes of these improvements over passive learning, in terms of $\tilde{\theta}_f(\varepsilon)$. Specifically, we have the following general theorem, along with two immediate corollaries. The proof is included in Appendix D.

Theorem 16 *For any VC class $\mathbb{C}$, and any passive learning algorithm $\mathcal{A}_p$ achieving label complexity Λ_p, the (Meta-Algorithm 3)-activized $\mathcal{A}_p$ algorithm achieves a label complexity Λ_a that, for any distribution $\mathcal{P}$ and classifier $f \in \mathbb{C}$, satisfies*

$$\Lambda_a(\varepsilon, f, \mathcal{P}) = O\left(\tilde{\theta}_f\left(\Lambda_p(\varepsilon/4, f, \mathcal{P})^{-1}\right) \log^2 \frac{\Lambda_p(\varepsilon/4, f, \mathcal{P})}{\varepsilon}\right).$$

Corollary 17 *For any VC class $\mathbb{C}$, there exists a passive learning algorithm $\mathcal{A}_p$ such that, the (Meta-Algorithm 3)-activized $\mathcal{A}_p$ algorithm achieves a label complexity Λ_a that, for any distribution $\mathcal{P}$ and classifier $f \in \mathbb{C}$, satisfies*

$$\Lambda_a(\varepsilon, f, \mathcal{P}) = O\left(\tilde{\theta}_f(\varepsilon) \log^2(1/\varepsilon)\right).$$

Proof The one-inclusion graph algorithm of Haussler, Littlestone, and Warmuth (1994) is a passive learning algorithm achieving label complexity $\Lambda_p(\varepsilon, f, \mathcal{P}) \leq d/\varepsilon$. Plugging this into Theorem 16, using the fact that $\tilde{\theta}_f(\varepsilon/4d) \leq 4d\tilde{\theta}_f(\varepsilon)$, and simplifying, we arrive at the result. In fact, we will see in the proof of Theorem 16 that incurring this extra constant factor of d is not actually necessary. ■

Corollary 18 *For any VC class $\mathbb{C}$ and distribution $\mathcal{P}$, if $\forall f \in \mathbb{C}, \tilde{\theta}_f < \infty$, then $(\mathbb{C}, \mathcal{P})$ is learnable at an exponential rate. If this is true for all $\mathcal{P}$, then $\mathbb{C}$ is learnable at an exponential rate.*

Proof The first claim follows directly from Corollary 17, since $\tilde{\theta}_f(\varepsilon) \leq \tilde{\theta}_f$. The second claim then follows from the fact that Meta-Algorithm 3 is adaptive to $\mathcal{P}$ (has no direct dependence on $\mathcal{P}$ except via the data). ■

Actually, in the proof we arrive at a somewhat more general result, in that the bound of Theorem 16 actually holds for any target function f in the "closure" of $\mathbb{C}$: that is, any f such that $\forall r > 0, \mathrm{B}(f, r) \neq \emptyset$. As previously mentioned, if our goal is only to obtain the label complexity bound of Corollary 17 by a direct approach, then we can use a simpler procedure (which cuts out Steps 9-16, instead returning an arbitrary element of V), analogous to how the analysis of the original algorithm of Cohn, Atlas, and Ladner (1994) by Hanneke (2011) obtains the label complexity bound of Corollary 11 (see also Algorithm 5 below). However, the general result of Theorem 16 is interesting in that it applies to any passive algorithm.

Inspecting the proof, we see that it is also possible to state a result that separates the probability of success from the achieved error rate, similar to the PAC model of Valiant (1984) and the analysis of active learning by Balcan, Hanneke, and Vaughan (2010). Specifically, suppose $\mathcal{A}_p$ is a passive learning algorithm such that, $\forall \varepsilon, \delta \in (0,1)$, there is a value $\lambda(\varepsilon, \delta, f, \mathcal{P}) \in \mathbb{N}$ such that $\forall n \geq \lambda(\varepsilon, \delta, f, \mathcal{P})$, $\mathbb{P}(\mathrm{er}(\mathcal{A}_p(\mathcal{Z}_n)) > \varepsilon) \leq \delta$. Suppose $\hat{h}_n$ is the classifier returned by the (Meta-Algorithm 3)-activized $\mathcal{A}_p$ with label budget n. Then for some $(\mathbb{C}, \mathcal{P}, f)$-dependent constant $c \in [1, \infty)$, $\forall \varepsilon, \delta \in (0, e^{-3})$, letting $\lambda = \lambda(\varepsilon/2, \delta/2, f, \mathcal{P})$,

$$\forall n \geq c\tilde{\theta}_f\left(\lambda^{-1}\right) \log^2(\lambda/\delta), \ \mathbb{P}\left(\mathrm{er}\left(\hat{h}_n\right) > \varepsilon\right) \leq \delta.$$

For instance, if $\mathcal{A}_p$ is an empirical risk minimization algorithm, then this is $\propto \tilde{\theta}_f(\varepsilon)\mathrm{polylog}\left(\frac{1}{\varepsilon\delta}\right)$.

5.5 Limitations and Potential Improvements

Theorem 16 and its corollaries represent significant improvements over most known results for the label complexity of active learning, and in particular over Theorem 10 and its corollaries. As for whether this also represents the best possible label complexity gains achievable by any active learning algorithm, the answer is mixed. As with most algorithms and analyses, Meta-Algorithm 3, Theorem 16, and corollaries, represent one set of solutions in a spectrum that trades strength of performance guarantees with simplicity. As such, there are several possible modifications one might make, which could potentially improve the performance guarantees. Here we sketch a few such possibilities. This subsection can be skipped by the casual reader without loss of continuity.

Even with Meta-Algorithm 3 as-is, various improvements to the bound of Theorem 16 should be possible, simply by being more careful in the analysis. For instance, as mentioned, Meta-Algorithm 3 is a *universal activizer* for any VC class $\mathbb{C}$, so in particular we know that whenever $\tilde{\theta}_f(\varepsilon) \neq o\left(1/\left(\varepsilon \log^2(1/\varepsilon)\right)\right)$, the above bound is not tight (see the work of Balcan, Hanneke, and Vaughan, 2010 for a construction leading to such $\tilde{\theta}_f(\varepsilon)$ values), and indeed any bound of the form $\tilde{\theta}_f(\varepsilon)\mathrm{polylog}(1/\varepsilon)$ will not be tight in some cases of this type. A more refined analysis may close this gap.

Another type of potential improvement is in the constant factors. Specifically, in the case when $\tilde{\theta}_f < \infty$, if we are only interested in *asymptotic* label complexity guarantees in Corollary 17, we can replace "$\sup_{r>0}$" in Definition 15 with "$\limsup_{r\to 0}$," which can sometimes be significantly smaller and/or easier to study. This is true for the disagreement coefficient in Corollary 11 as well. Additionally, the proof (in Appendix D) reveals that there are significant $(\mathbb{C}, \mathcal{P}, f)$-dependent constant factors other than $\tilde{\theta}_f(\varepsilon)$, and it is quite likely that these can be improved by a more careful analysis of Meta-Algorithm 3 (or in some cases, possibly an improved definition of the estimators $\hat{P}_m$).

However, even with such refinements to improve the results, the approach of using $\tilde{\theta}_f$ to prove learnability at an exponential rate has limits. For instance, it is known that any *countable* $\mathbb{C}$ is learnable at an exponential rate (Balcan, Hanneke, and Vaughan, 2010). However, there are countable VC classes $\mathbb{C}$ for which $\tilde{\theta}_f = \infty$ for some elements of $\mathbb{C}$ (e.g., take the tree-paths concept space of Balcan, Hanneke, and Vaughan (2010), except instead of all infinite-depth paths from the root, take all of the finite-depth paths from the root, but keep one infinite-depth path f; for this modified space $\mathbb{C}$, which is countable, every $h \in \mathbb{C}$ has $\tilde{d}_h = 1$, and for that one infinite-depth f we have $\tilde{\theta}_f = \infty$).

Inspecting the proof reveals that it is possible to make the results slightly sharper by replacing $\tilde{\theta}_f(r_0)$ (for $r_0 = \Lambda_p(\varepsilon/4, f, \mathcal{P})^{-1}$) with a somewhat more complicated quantity: namely,

$$\min_{k<\tilde{d}_f} \sup_{r>r_0} r^{-1} \cdot \mathcal{P}\left(x \in \mathcal{X} : \mathcal{P}^k\left(S \in \mathcal{X}^k : \mathrm{B}(f,r) \text{ shatters } S \cup \{x\}\right) \geq \mathbb{P}\left(\partial^k f\right)/16\right). \tag{4}$$

This quantity can be bounded in terms of $\tilde{\theta}_f(r_0)$ via Markov's inequality, but is sometimes smaller.

As for improving Meta-Algorithm 3 itself, there are several possibilities. One immediate improvement one can make is to replace the condition in Steps 5 and 12 by $\min_{1\leq j\leq k} \hat{P}_m(S \in \mathcal{X}^{j-1} : V \text{ shatters } S \cup \{X_m\} | V \text{ shatters } S) \geq 1/2$, likewise replacing the corresponding quantity in Step 9, and substituting in Steps 7 and 14 the quantity $\max_{1\leq j\leq k} \hat{P}_m(S \in \mathcal{X}^{j-1} : V[(X_m, -y)] \text{ does not shatter } S | V \text{ shatters } S)$; in particular, the results stated for Meta-Algorithm 3 remain valid with this substitution, requiring only minor modifications to the proofs. However, it is not clear what gains in theoretical guarantees this achieves.

Additionally, there are various quantities in this procedure that can be altered almost arbitrarily, allowing room for fine-tuning. Specifically, the $2/3$ in Step 0 and $1/3$ in Step 16 can be set to arbitrary constants summing to 1. Likewise, the $1/4$ in Step 3, $1/3$ in Step 10, and $3/4$ in Step 12 can be changed to any constants in $(0,1)$, possibly depending on k, such that the sum of the first two is strictly less than the third. Also, the $1/2$ in Steps 5, 9, and 12 can be set to any constant in $(0,1)$. Furthermore, the $k \cdot 2^n$ in Step 3 only prevents infinite looping, and can be set to any function growing superlinearly in n, though to get the largest possible improvements it should at least grow exponentially in n; typically, *any* active learning algorithm capable of exponential improvements over reasonable passive learning algorithms will require access to a number of unlabeled examples exponential in n, and Meta-Algorithm 3 is no exception to this.

One major issue in the design of the procedure is an inherent trade-off between the achieved label complexity and the number of unlabeled examples used by the algorithm. This is noteworthy both because of the practical concerns of gathering such large quantities of unlabeled data, and also for computational efficiency reasons. In contrast to disagreement-based methods, the design of the estimators used in Meta-Algorithm 3 introduces such a trade-off, though in contrast to the splitting index analysis of Dasgupta (2005), the trade-off here seems only in the constant factors. The choice of these $\hat{P}_m$ estimators, both in their definition in Appendix B.1, and indeed in the very quantities they estimate, is such that we can (if desired) limit the number of unlabeled examples the main body of the algorithm uses (the actual number it needs to achieve Theorem 16 can be extracted from the proofs in Appendix D.1). However, if the number of unlabeled examples used by the algorithm is not a limiting factor, we can suggest more effective quantities. Specifically, following the original motivation for using shatterable sets, we might consider a greedily-constructed distribution over the set $\{S \in \mathcal{X}^j : V \text{ shatters } S, 1 \le j < k, \text{ and either } j = k-1 \text{ or } \mathcal{P}(s : V \text{ shatters } S \cup \{s\}) = 0\}$. We can construct the distribution implicitly, via the following generative model. First we set $S = \{\}$. Then repeat the following. If $|S| = k-1$ or $\mathcal{P}(s \in \mathcal{X} : V \text{ shatters } S \cup \{s\}) = 0$, output S; otherwise, sample s according to the conditional distribution of X given that V shatters $S \cup \{X\}$. If we denote this distribution (over S) as $\tilde{\mathcal{P}}_k$, then replacing the estimator $\hat{P}_m\left(S \in \mathcal{X}^{k-1} : V \text{ shatters } S \cup \{X_m\} | V \text{ shatters } S\right)$ in Meta-Algorithm 3 with an appropriately constructed estimator of $\tilde{\mathcal{P}}_k\left(S : V \text{ shatters } S \cup \{X_m\}\right)$ (and similarly replacing the other estimators) can lead to some improvements in the constant factors of the label complexity. However, such a modification can also dramatically increase the number of unlabeled examples required by the algorithm, since rejection-sampling to get a point from the conditional distribution of X given V shatters $S \cup \{X\}$ can be costly, as can determining whether $\mathcal{P}(s \in \mathcal{X} : V \text{ shatters } S \cup \{s\}) \approx 0$.

Unlike Meta-Algorithm 1, there remain serious efficiency concerns about Meta-Algorithm 3. If we knew the value of $\tilde{d}_f$ and $\tilde{d}_f \le c \log_2(d)$ for some constant c, then we could potentially design an efficient version of Meta-Algorithm 3 still achieving Corollary 17. Specifically, suppose we can find a classifier in $\mathbb{C}$ consistent with any given sample, or determine that no such classifier exists, in time polynomial in the sample size (and d), and also that $\mathcal{A}_p$ efficiently returns a classifier in $\mathbb{C}$ consistent with the sample it is given. Then restricting the loop of Step 1 to those $k \le \tilde{d}_f$ and returning $\mathcal{A}_p(\mathcal{L}_{\tilde{d}_f})$, the algorithm becomes efficient, in the sense that with high probability, its running time is $\text{poly}(d/\varepsilon)$, where ε is the error rate guarantee from inverting the label complexity at the value of n given to the algorithm. To be clear, in some cases we may obtain values $m \propto \exp\{\Omega(n)\}$, but the error rate guaranteed by $\mathcal{A}_p$ is $\tilde{O}(1/m)$ in these cases, so that we still have m polynomial in d/ε. However, in the absence of this access to $\tilde{d}_f$, the values of $k > \tilde{d}_f$ in Meta-Algorithm 3 may reach values of m much larger than $\text{poly}(d/\varepsilon)$, since the error rates obtained from these $\mathcal{A}_p(\mathcal{L}_k)$

evaluations are not guaranteed to be better than the $\mathcal{A}_p(\mathcal{L}_{\tilde{d}_f})$ evaluations, and yet we may have $|\mathcal{L}_k| \gg |\mathcal{L}_{\tilde{d}_f}|$. Thus, there remains a challenging problem of obtaining the results above (Theorem 16 and Corollary 17) via an efficient algorithm, adaptive to the value of $\tilde{d}_f$.

6. Toward Agnostic Activized Learning

The previous sections addressed learning in the *realizable* case, where there is a perfect classifier $f \in \mathbb{C}$ (i.e., $\mathrm{er}(f) = 0$). To move beyond these scenarios, to problems in which f is not a perfect classifier (i.e., stochastic labels) or not well-approximated by $\mathbb{C}$, requires a change in technique to make the algorithms more robust to such issues. As we will see in Section 6.2, the results we can prove in this more general setting are not quite as strong as those of the previous sections, but in some ways they are more interesting, both from a practical perspective, as we expect real learning problems to involve imperfect teachers or underspecified instance representations, and also from a theoretical perspective, as the class of problems addressed is significantly more general than those encompassed by the realizable case above.

In this context, we will be largely interested in more general versions of the same types of questions as above, such as whether one can activize a given passive learning algorithm, in this case guaranteeing strictly improved label complexities for all nontrivial joint distributions over $\mathcal{X} \times \{-1,+1\}$. In Section 6.3, we present a general conjecture regarding this type of strong domination. To approach such questions, we will explore techniques for making the above algorithms robust to label noise. Specifically, we will use a natural generalization of a technique developed for noise-robust disagreement-based active learning. Toward this end, as well as for the sake of comparison, we will review the known techniques and results for disagreement-based agnostic active learning in Section 6.5. We then extend these techniques in Section 6.6 to develop a new type of agnostic active learning algorithm, based on shatterable sets, which relates to the disagreement-based agnostic active learning algorithms in a way analogous to how Meta-Algorithm 3 relates to Meta-Algorithm 2. Furthermore, we present a bound on the label complexities achieved by this method, representing a natural generalization of both Corollary 17 and the known results on disagreement-based agnostic active learning (Hanneke, 2011).

Although we present several new results, in some sense this section is less about what we know and more about what we do not yet know. As such, we will focus less on presenting a complete and elegant theory, and more on identifying potentially promising directions for exploration. In particular, Section 6.8 sketches out some interesting directions, which could potentially lead to a resolution of the aforementioned general conjecture from Section 6.3.

6.1 Definitions and Notation

In this setting, there is a joint distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1,+1\}$, with marginal distribution $\mathcal{P}$ on $\mathcal{X}$. For any classifier h, we denote by $\mathrm{er}(h) = \mathcal{P}_{XY}((x,y) : h(x) \neq y)$. Also, denote by $\nu^*(\mathcal{P}_{XY}) = \inf_{h:\mathcal{X}\to\{-1,+1\}} \mathrm{er}(h)$ the *Bayes error rate*, or simply ν^* when $\mathcal{P}_{XY}$ is clear from the context; also define the conditional label distribution $\eta(x;\mathcal{P}_{XY}) = \mathbb{P}(Y = +1|X = x)$, where $(X,Y) \sim \mathcal{P}_{XY}$, or $\eta(x) = \eta(x;\mathcal{P}_{XY})$ when $\mathcal{P}_{XY}$ is clear from the context. For a given concept space $\mathbb{C}$, denote $\nu(\mathbb{C};\mathcal{P}_{XY}) = \inf_{h\in\mathbb{C}} \mathrm{er}(h)$, called the *noise rate* of $\mathbb{C}$; when $\mathbb{C}$ and/or $\mathcal{P}_{XY}$ is clear from the context, we may abbreviate $\nu = \nu(\mathbb{C}) = \nu(\mathbb{C};\mathcal{P}_{XY})$. For $\mathcal{H} \subseteq \mathbb{C}$, the *diameter* is defined as $\mathrm{diam}(\mathcal{H};\mathcal{P}) = \sup_{h_1,h_2\in\mathcal{H}} \mathcal{P}(x : h_1(x) \neq$

$h_2(x)$). Also, for any $\varepsilon > 0$, define the ε-minimal set $\mathbb{C}(\varepsilon;\mathcal{P}_{XY}) = \{h \in \mathbb{C} : \mathrm{er}(h) \leq \nu + \varepsilon\}$. For any set of classifiers $\mathcal{H}$, define the *closure*, denoted $\mathrm{cl}(\mathcal{H};\mathcal{P})$, as the set of all measurable $h : \mathcal{X} \to \{-1,+1\}$ such that $\forall r > 0, \mathrm{B}_{\mathcal{H},\mathcal{P}}(h,r) \neq \emptyset$. When $\mathcal{P}_{XY}$ is clear from the context, we will simply refer to $\mathbb{C}(\varepsilon) = \mathbb{C}(\varepsilon;\mathcal{P}_{XY})$, and when $\mathcal{P}$ is clear, we write $\mathrm{diam}(\mathcal{H}) = \mathrm{diam}(\mathcal{H};\mathcal{P})$ and $\mathrm{cl}(\mathcal{H}) = \mathrm{cl}(\mathcal{H};\mathcal{P})$.

In the noisy setting, rather than being a *perfect* classifier, we will let f denote an arbitrary element of $\mathrm{cl}(\mathbb{C};\mathcal{P})$ with $\mathrm{er}(f) = \nu(\mathbb{C};\mathcal{P}_{XY})$: that is, $f \in \bigcap_{\varepsilon>0} \mathrm{cl}\left(\mathbb{C}(\varepsilon;\mathcal{P}_{XY});\mathcal{P}\right)$. Such a classifier must exist, since $\mathrm{cl}(\mathbb{C})$ is *compact* in the pseudo-metric $\rho(h,g) = \int |h-g| \mathrm{d}\mathcal{P} \propto \mathcal{P}(x : h(x) \neq g(x))$ (in the usual sense of the equivalence classes being compact in the ρ-induced metric). This can be seen by recalling that $\mathbb{C}$ is totally bounded (Haussler, 1992), and thus so is $\mathrm{cl}(\mathbb{C})$, and that $\mathrm{cl}(\mathbb{C})$ is a closed subset of $\mathcal{L}^1(\mathcal{P})$, which is complete (Dudley, 2002), so $\mathrm{cl}(\mathbb{C})$ is also complete (Munkres, 2000). Total boundedness and completeness together imply compactness (Munkres, 2000), and this implies the existence of f since monotone sequences of nonempty closed subsets of a compact space have a nonempty limit set (Munkres, 2000).

As before, in the learning problem there is a sequence $\mathcal{Z} = \{(X_1,Y_1),(X_2,Y_2),\ldots\}$, where the (X_i,Y_i) are independent and identically distributed, and we denote by $\mathcal{Z}_m = \{(X_i,Y_i)\}_{i=1}^m$. As before, the $X_i \sim \mathcal{P}$, but rather than having each Y_i value determined as a function of X_i, instead we have each pair $(X_i,Y_i) \sim \mathcal{P}_{XY}$. The learning protocol is defined identically as above; that is, the algorithm has direct access to the X_i values, but must request the Y_i (label) values one at a time, sequentially, and can request at most n total labels, where n is a budget provided as input to the algorithm. The label complexity is now defined just as before (Definition 1), but generalized by replacing $(f,\mathcal{P})$ with the joint distribution $\mathcal{P}_{XY}$. Specifically, we have the following formal definition, which will be used throughout this section (and the corresponding appendices).

Definition 19 *An active learning algorithm $\mathcal{A}$ achieves label complexity $\Lambda(\cdot,\cdot)$ if, for any joint distribution $\mathcal{P}_{XY}$, for any $\varepsilon \in (0,1)$ and any integer $n \geq \Lambda(\varepsilon,\mathcal{P}_{XY})$, we have $\mathbb{E}\left[\mathrm{er}\left(\mathcal{A}(n)\right)\right] \leq \varepsilon$.*

However, because there may not be any classifier with error rate less than any arbitrary $\varepsilon \in (0,1)$, our objective changes here to achieving error rate at most $\nu + \varepsilon$ for any given $\varepsilon \in (0,1)$. Thus, we are interested in the quantity $\Lambda(\nu+\varepsilon,\mathcal{P}_{XY})$, and will be particularly interested in this quantity's asymptotic dependence on ε, as $\varepsilon \to 0$. In particular, $\Lambda(\varepsilon,\mathcal{P}_{XY})$ may often be infinite for $\varepsilon < \nu$.

The label complexity for passive learning can be generalized analogously, again replacing $(f,\mathcal{P})$ by $\mathcal{P}_{XY}$ in Definition 2 as follows.

Definition 20 *A passive learning algorithm $\mathcal{A}$ achieves label complexity $\Lambda(\cdot,\cdot)$ if, for any joint distribution $\mathcal{P}_{XY}$, for any $\varepsilon \in (0,1)$ and any integer $n \geq \Lambda(\varepsilon,\mathcal{P}_{XY})$, we have $\mathbb{E}\left[\mathrm{er}\left(\mathcal{A}\left(\mathcal{Z}_n\right)\right)\right] \leq \varepsilon$.*

For any label complexity Λ in the agnostic case, define the set $\mathrm{Nontrivial}(\Lambda;\mathbb{C})$ as the set of all distributions $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1,+1\}$ such that $\forall \varepsilon > 0, \Lambda(\nu+\varepsilon,\mathcal{P}_{XY}) < \infty$, and $\forall g \in \mathrm{Polylog}(1/\varepsilon)$, $\Lambda(\nu+\varepsilon,\mathcal{P}_{XY}) = \omega(g(\varepsilon))$. In this context, we can define an *activizer* for a given passive algorithm as follows.

Definition 21 *We say an active meta-algorithm $\mathcal{A}_a$* activizes *a passive algorithm $\mathcal{A}_p$ for $\mathbb{C}$ in the agnostic case if the following holds. For any label complexity Λ_p achieved by $\mathcal{A}_p$, the active learning algorithm $\mathcal{A}_a(\mathcal{A}_p,\cdot)$ achieves a label complexity Λ_a such that, for every distribution $\mathcal{P}_{XY} \in \mathrm{Nontrivial}(\Lambda_p;\mathbb{C})$, there exists a constant $c \in [1,\infty)$ such that*

$$\Lambda_a(\nu+c\varepsilon,\mathcal{P}_{XY}) = o\left(\Lambda_p(\nu+\varepsilon,\mathcal{P}_{XY})\right).$$

In this case, $\mathcal{A}_a$ is called an activizer *for $\mathcal{A}_p$ with respect to $\mathbb{C}$ in the agnostic case, and the active learning algorithm $\mathcal{A}_a(\mathcal{A}_p, \cdot)$ is called the $\mathcal{A}_a$*-activized *$\mathcal{A}_p$.*

6.2 A Negative Result

First, the bad news: we cannot generally hope for universal activizers for VC classes in the agnostic case. In fact, there even exist passive algorithms that *cannot be activized*, even by any specialized active learning algorithm.

Specifically, consider again Example 1, where $\mathcal{X} = [0,1]$ and $\mathbb{C}$ is the class of threshold classifiers, and let $\check{\mathcal{A}}_p$ be a passive learning algorithm that behaves as follows. Given n points $\mathcal{Z}_n = \{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$, $\check{\mathcal{A}}_p(\mathcal{Z}_n)$ returns the classifier $h_{\hat{z}} \in \mathbb{C}$, where $\hat{z} = \frac{1-2\hat{\eta}_0}{1-\hat{\eta}_0}$ and $\hat{\eta}_0 = \left(\frac{|\{i\in\{1,\ldots,n\}:X_i=0,Y_i=+1\}|}{|\{i\in\{1,\ldots,n\}:X_i=0\}|} \vee \frac{1}{8}\right) \wedge \frac{3}{8}$, taking $\hat{\eta}_0 = 1/8$ if $\{i \in \{1,\ldots,n\} : X_i = 0\} = \emptyset$. For most distributions $\mathcal{P}_{XY}$, this algorithm clearly would not behave "reasonably," in that its error rate would be quite large; in particular, in the realizable case, the algorithm's worst-case expected error rate does not converge to zero as $n \to \infty$. However, for certain distributions $\mathcal{P}_{XY}$ engineered specifically for this algorithm, it has near-optimal behavior in a strong sense. Specifically, we have the following result, the proof of which is included in Appendix E.1.

Theorem 22 *There is no activizer for $\check{\mathcal{A}}_p$ with respect to the space of threshold classifiers in the agnostic case.*

Recall that threshold classifiers were, in some sense, one of the simplest scenarios for activized learning in the realizable case. Also, since threshold-like problems are embedded in most "geometric" concept spaces, this indicates we should generally not expect there to exist activizers for arbitrary passive algorithms in the agnostic case. However, this leaves open the question of whether certain families of passive learning algorithms can be activized in the agnostic case, a topic we turn to next.

6.3 A Conjecture: Activized Empirical Risk Minimization

The counterexample above is interesting, in that it exposes the limits on generality in the agnostic setting. However, the passive algorithm that cannot be activized there is in many ways not very reasonable, in that it has suboptimal worst-case expected excess error rate (among other deficiencies). It may therefore be more interesting to ask whether some family of "reasonable" passive learning algorithms can be activized in the agnostic case. It seems that, unlike $\check{\mathcal{A}}_p$ above, certain passive learning algorithms should not have too peculiar a dependence on the label noise, so that they use Y_i to help determine $f(X_i)$ and that is all. In such cases, any Y_i value for which we can already infer the value $f(X_i)$ should simply be ignored as redundant information, so that we needn't request such values. While this discussion is admittedly vague, consider the following formal conjecture.

Recall that an *empirical risk minimization* algorithm for $\mathbb{C}$ is a type of passive learning algorithm $\mathcal{A}$, characterized by the fact that for any set $\mathcal{L} \in \bigcup_m (\mathcal{X} \times \{-1,+1\})^m$, $\mathcal{A}(\mathcal{L}) \in \operatorname*{argmin}_{h\in\mathbb{C}} \mathrm{er}_{\mathcal{L}}(h)$.

Conjecture 23 *For any VC class, there exists an active meta-algorithm $\mathcal{A}_a$ and an empirical risk minimization algorithm $\mathcal{A}_p$ for $\mathbb{C}$ such that $\mathcal{A}_a$* activizes *$\mathcal{A}_p$ for $\mathbb{C}$ in the agnostic case.*

Resolution of this conjecture would be interesting for a variety of reasons. If the conjecture is correct, it means that the vast (and growing) literature on the label complexity of empirical risk minimization has direct implications for the potential performance of active learning under the same conditions. We might also expect activized empirical risk minimization to be quite effective in practical applications.

While this conjecture remains open at this time, the remainder of this section might be viewed as partial evidence in its favor, as we show that active learning is able to achieve improvements over the known bounds on the label complexity of passive learning in many cases.

6.4 Low Noise Conditions

In the subsections below, we will be interested in stating bounds on the label complexity of active learning, analogous to those of Theorem 10 and Theorem 16, but for learning with label noise. As in the realizable case, we should expect such bounds to have some explicit dependence on the distribution $\mathcal{P}_{XY}$. Initially, one might hope that we could state interesting label complexity bounds purely in terms of a simple quantity such as $\nu(\mathbb{C};\mathcal{P}_{XY})$. However, it is known that any label complexity bound for a nontrivial $\mathbb{C}$ (for either passive or active) depending on $\mathcal{P}_{XY}$ only via $\nu(\mathbb{C};\mathcal{P}_{XY})$ will be $\Omega\left(\varepsilon^{-2}\right)$ when $\nu(\mathbb{C};\mathcal{P}_{XY}) > 0$ (Kääriäinen, 2006). Since passive learning can achieve a $\mathcal{P}_{XY}$-independent $O\left(\varepsilon^{-2}\right)$ label complexity bound for any VC class (Alexander, 1984), we will need to discuss label complexity bounds that depend on $\mathcal{P}_{XY}$ via more detailed quantities than merely $\nu(\mathbb{C};\mathcal{P}_{XY})$ if we are to characterize the improvements of active learning over passive.

In this subsection, we review an index commonly used to describe certain properties of $\mathcal{P}_{XY}$ relative to $\mathbb{C}$: namely, the Mammen-Tsybakov margin conditions (Mammen and Tsybakov, 1999; Tsybakov, 2004; Koltchinskii, 2006). Specifically, we have the following formal condition from Koltchinskii (2006).

Condition 1 *There exist constants* $\mu,\kappa \in [1,\infty)$ *such that* $\forall\varepsilon > 0$, $\operatorname{diam}(\mathbb{C}(\varepsilon;\mathcal{P}_{XY});\mathcal{P}) \leq \mu\cdot\varepsilon^{\frac{1}{\kappa}}$.

This condition has recently been studied in depth in the passive learning literature, as it can be used to characterize scenarios where the label complexity of passive learning is between the worst-case $\Theta(1/\varepsilon^2)$ and the realizable case $\Theta(1/\varepsilon)$ (e.g., Mammen and Tsybakov, 1999; Tsybakov, 2004; Massart and Nédélec, 2006; Koltchinskii, 2006). The condition can equivalently be stated as

$$\exists\mu' \in (0,1], \kappa \in [1,\infty) \text{ s.t. } \forall h \in \mathbb{C}, \operatorname{er}(h) - \nu(\mathbb{C};\mathcal{P}_{XY}) \geq \mu'\cdot\mathcal{P}(x : h(x) \neq f(x))^{\kappa}.$$

The condition is implied by a variety of interesting special cases. For instance, it is satisfied when $\nu(\mathbb{C};\mathcal{P}_{XY}) = \nu^*(\mathcal{P}_{XY})$ and

$$\exists\mu'',\alpha \in (0,\infty) \text{ s.t. } \forall\varepsilon > 0, \mathcal{P}(x : |\eta(x;\mathcal{P}_{XY}) - 1/2| \leq \varepsilon) \leq \mu''\cdot\varepsilon^{\alpha},$$

where κ and μ are functions of α and μ'' (Mammen and Tsybakov, 1999; Tsybakov, 2004); in particular, $\kappa = (1+\alpha)/\alpha$. This can intuitively be interpreted as saying that very noisy points are relatively rare. Special cases of this condition have also been studied in depth; for instance, *bounded noise* conditions, wherein $\nu(\mathbb{C};\mathcal{P}_{XY}) = \nu^*(\mathcal{P}_{XY})$ and $\forall x, |\eta(x;\mathcal{P}_{XY}) - 1/2| > c$ for some constant $c > 0$ (e.g., Giné and Koltchinskii, 2006; Massart and Nédélec, 2006), are a special case of Condition 1 with $\kappa = 1$.

Condition 1 can be interpreted in a variety of ways, depending on the context. For instance, in certain concept spaces with a geometric interpretation, it can often be realized as a kind of *large* *margin* condition, under some condition relating the noisiness of a point's label to its distance from the optimal decision surface. That is, if the magnitude of noise $(1/2 - |\eta(x;\mathcal{P}_{XY}) - 1/2|)$ for a given point depends inversely on its distance from the optimal decision surface, so that points closer to the decision surface have noisier labels, a small value of κ in Condition 1 will occur if the distribution $\mathcal{P}$ has *low density* near the optimal decision surface (assuming $\nu(\mathbb{C};\mathcal{P}_{XY}) = \nu^*(\mathcal{P}_{XY})$) (e.g., Dekel, Gentile, and Sridharan, 2010). On the other hand, when there is *high* density near the optimal decision surface, the value of κ may be determined by how quickly $\eta(x;\mathcal{P}_{XY})$ changes as x approaches the decision boundary (Castro and Nowak, 2008). See the works of Mammen and Tsybakov (1999), Tsybakov (2004), Koltchinskii (2006), Massart and Nédélec (2006), Castro and Nowak (2008), Dekel, Gentile, and Sridharan (2010) and Bartlett, Jordan, and McAuliffe (2006) for further interpretations of Condition 1.

In the context of passive learning, one natural method to study is that of *empirical risk minimization*. Recall that a passive learning algorithm $\mathcal{A}$ is called an empirical risk minimization algorithm for $\mathbb{C}$ if it returns a classifier from $\mathbb{C}$ making the minimum number of mistakes on the labeled sample it is given as input. It is known that for any VC class $\mathbb{C}$, for any $\mathcal{P}_{XY}$ satisfying Condition 1 for finite μ and κ, every empirical risk minimization algorithm for $\mathbb{C}$ achieves a label complexity

$$\Lambda(\nu+\varepsilon,\mathcal{P}_{XY}) = O\left(\varepsilon^{\frac{1}{\kappa}-2} \cdot \log\frac{1}{\varepsilon}\right). \tag{5}$$

This follows from the works of Koltchinskii (2006) and Massart and Nédélec (2006). Furthermore, for nontrivial concept spaces, one can show that $\inf_\Lambda \sup_{\mathcal{P}_{XY}} \Lambda(\nu+\varepsilon;\mathcal{P}_{XY}) = \Omega\left(\varepsilon^{\frac{1}{\kappa}-2}\right)$, where the supremum ranges over all $\mathcal{P}_{XY}$ satisfying Condition 1 for the given μ and κ values, and the infimum ranges over all label complexities achievable by passive learning algorithms (Castro and Nowak, 2008; Hanneke, 2011); that is, the bound (5) cannot be significantly improved by any passive algorithm, without allowing the label complexity to have a more refined dependence on $\mathcal{P}_{XY}$ than afforded by Condition 1.

In the context of active learning, a variety of results are presently known, which in some cases show improvements over (5). Specifically, for any VC class $\mathbb{C}$ and any $\mathcal{P}_{XY}$ satisfying Condition 1, a certain noise-robust disagreement-based active learning algorithm achieves label complexity

$$\Lambda(\nu+\varepsilon,\mathcal{P}_{XY}) = O\left(\theta_f\left(\varepsilon^{\frac{1}{\kappa}}\right) \cdot \varepsilon^{\frac{2}{\kappa}-2} \cdot \log^2\frac{1}{\varepsilon}\right).$$

This general result was established by Hanneke (2011) (analyzing the algorithm of Dasgupta, Hsu, and Monteleoni, 2007), generalizing earlier $\mathbb{C}$-specific results by Castro and Nowak (2008) and Balcan, Broder, and Zhang (2007), and was later simplified and refined in some cases by Koltchinskii (2010). Comparing this to (5), when $\theta_f < \infty$ this is an improvement over passive learning by a factor of $\varepsilon^{\frac{1}{\kappa}} \cdot \log(1/\varepsilon)$. Note that this generalizes the label complexity bound of Corollary 11 above, since the realizable case entails Condition 1 with $\kappa = \mu/2 = 1$. It is also known that this type of improvement is essentially the best we can hope for when we describe $\mathcal{P}_{XY}$ purely in terms of the parameters of Condition 1. Specifically, for any nontrivial concept space $\mathbb{C}$, $\inf_\Lambda \sup_{\mathcal{P}_{XY}} \Lambda(\nu+\varepsilon,\mathcal{P}_{XY}) = \Omega\left(\max\left\{\varepsilon^{\frac{2}{\kappa}-2}, \log\frac{1}{\varepsilon}\right\}\right)$, where the supremum ranges over all $\mathcal{P}_{XY}$ satisfying Condition 1 for the given μ and κ values, and the infimum ranges over all label complexities achievable by active learning algorithms (Hanneke, 2011; Castro and Nowak, 2008).

In the following subsection, we review the established techniques and results for disagreement-based agnostic active learning; the algorithm presented here is slightly different from that originally analyzed by Hanneke (2011), but the label complexity bounds of Hanneke (2011) hold for this new algorithm as well. We follow this in Section 6.7 with a new agnostic active learning method that goes beyond disagreement-based learning, again generalizing the notion of disagreement to the notion of shatterability; this can be viewed as analogous to the generalization of Meta-Algorithm 2 represented by Meta-Algorithm 3, and as in that case the resulting label complexity bound replaces $\theta_f(\cdot)$ with $\tilde{\theta}_f(\cdot)$.

For both passive and active learning, results under Condition 1 are also known for more general scenarios than VC classes: namely, under entropy conditions (Mammen and Tsybakov, 1999; Tsybakov, 2004; Koltchinskii, 2006, 2011; Massart and Nédélec, 2006; Castro and Nowak, 2008; Hanneke, 2011; Koltchinskii, 2010). For a nonparametric class known as *boundary fragments*, Castro and Nowak (2008) find that active learning sometimes offers advantages over passive learning, under a special case of Condition 1. Furthermore, Hanneke (2011) shows a general result on the label complexity achievable by disagreement-based agnostic active learning, which sometimes exhibits an improved dependence on the parameters of Condition 1 under conditions on the disagreement coefficient and certain entropy conditions for $(\mathbb{C}, \mathcal{P})$ (see also Koltchinskii, 2010). These results will not play a role in the discussion below, as in the present work we restrict ourselves strictly to VC classes, leaving more general results for future investigations.

6.5 Disagreement-Based Agnostic Active Learning

Unlike the realizable case, here in the agnostic case we cannot eliminate a classifier from the version space after making merely a single mistake, since even the best classifier is potentially imperfect. Rather, we take a collection of samples with labels, and eliminate those classifiers making significantly more mistakes relative to some others in the version space. This is the basic idea underlying most of the known agnostic active learning algorithms, including those discussed in the present work. The precise meaning of "significantly more," sufficient to guarantee the version space always contains some good classifier, is typically determined by established bounds on the deviation of excess empirical error rates from excess true error rates, taken from the passive learning literature.

The following disagreement-based algorithm is slightly different from any in the existing literature, but is similar in style to a method of Beygelzimer, Dasgupta, and Langford (2009); it also bares resemblance to the algorithms of Koltchinskii (2010); Dasgupta, Hsu, and Monteleoni (2007); Balcan, Beygelzimer, and Langford (2006a, 2009). It should be considered as representative of the family of disagreement-based agnostic active learning algorithms, and all results below concerning it have analogous results for variants of these other disagreement-based methods.

Algorithm 4 Input: label budget n, confidence parameter δ Output: classifier $\hat{h}$
0. $m \leftarrow 0$, $i \leftarrow 0$, $V_0 \leftarrow \mathbb{C}$, $\mathcal{L}_1 \leftarrow \{\}$ 1. While $t < n$ and $m \leq 2^n$ 2. $\quad m \leftarrow m+1$ 3. $\quad$ If $X_m \in \mathrm{DIS}(V_i)$ 4. $\quad\quad$ Request the label Y_m of X_m, and let $\mathcal{L}_{i+1} \leftarrow \mathcal{L}_{i+1} \cup \{(X_m, Y_m)\}$ and $t \leftarrow t+1$ 5. $\quad$ Else let $\hat{y}$ be the label agreed upon by classifiers in V_i, and $\mathcal{L}_{i+1} \leftarrow \mathcal{L}_{i+1} \cup \{(X_m, \hat{y})\}$ 6. $\quad$ If $m = 2^{i+1}$ 7. $\quad\quad V_{i+1} \leftarrow \left\{ h \in V_i : \mathrm{er}_{\mathcal{L}_{i+1}}(h) - \min_{h' \in V_i} \mathrm{er}_{\mathcal{L}_{i+1}}(h') \leq \hat{U}_{i+1}(V_i, \delta) \right\}$ 8. $\quad\quad i \leftarrow i+1$, and then $\mathcal{L}_{i+1} \leftarrow \{\}$ 9. Return any $\hat{h} \in V_i$

The algorithm is specified in terms of an estimator, $\hat{U}_i$. The definition of $\hat{U}_i$ should typically be based on generalization bounds known for passive learning. Inspired by the work of Koltchinskii (2006) and applications thereof in active learning (Hanneke, 2011; Koltchinskii, 2010), we will take a definition of $\hat{U}_i$ based on a data-dependent Rademacher complexity, as follows. Let $\xi_1, \xi_2, \ldots$ denote a sequence of independent Rademacher random variables (i.e., uniform in $\{-1, +1\}$), also independent from all other random variables in the algorithm (i.e., $\mathcal{Z}$). Then for any set $\mathcal{H} \subseteq \mathbb{C}$, define

$$\hat{R}_i(\mathcal{H}) = \sup_{h_1, h_2 \in \mathcal{H}} 2^{-i} \sum_{m=2^{i-1}+1}^{2^i} \xi_m \cdot (h_1(X_m) - h_2(X_m)),$$

$$\hat{D}_i(\mathcal{H}) = \sup_{h_1, h_2 \in \mathcal{H}} 2^{-i} \sum_{m=2^{i-1}+1}^{2^i} |h_1(X_m) - h_2(X_m)|,$$

$$\hat{U}_i(\mathcal{H}, \delta) = 12\hat{R}_i(\mathcal{H}) + 34\sqrt{\hat{D}_i(\mathcal{H}) \frac{\ln(32i^2/\delta)}{2^{i-1}}} + \frac{752 \ln(32i^2/\delta)}{2^{i-1}}. \tag{6}$$

Algorithm 4 operates by repeatedly doubling the sample size $|\mathcal{L}_{i+1}|$, while only requesting the labels of the points in the region of disagreement of the version space. Each time it doubles the size of the sample $\mathcal{L}_{i+1}$, it updates the version space by eliminating any classifiers that make significantly more mistakes on $\mathcal{L}_{i+1}$ relative to others in the version space. Since the labels of the examples we infer in Step 5 are agreed upon by all elements of the version space, the *difference* of empirical error rates in Step 7 is identical to the difference of empirical error rates under the *true* labels. This allows us to use established results on deviations of excess empirical error rates from excess true error rates to judge suboptimality of some of the classifiers in the version space in Step 7, thus reducing the version space.

As with Meta-Algorithm 2, for computational feasibility, the sets V_i and $\mathrm{DIS}(V_i)$ in Algorithm 4 can be represented implicitly by a set of constraints imposed by previous rounds of the loop. Also, the update to $\mathcal{L}_{i+1}$ in Step 5 is included only to make Step 7 somewhat simpler or more intuitive; it can be be removed without altering the behavior of the algorithm, as long as we compensate by multiplying $\mathrm{er}_{\mathcal{L}_{i+1}}$ by an appropriate renormalization constant in Step 7: namely, $2^{-i}|\mathcal{L}_{i+1}|$.

We have the following result about the label complexity of Algorithm 4; it is representative of the type of theorem one can prove about disagreement-based active learning under Condition 1.

Lemma 24 *Let $\mathbb{C}$ be a VC class and suppose the joint distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1,+1\}$ satisfies Condition 1 for finite parameters μ and κ. There is a $(\mathbb{C},\mathcal{P}_{XY})$-dependent constant $c \in (0,\infty)$ such that, for any $\varepsilon, \delta \in (0, e^{-3})$, and any integer*

$$n \geq c \cdot \theta_f\left(\varepsilon^{\frac{1}{\kappa}}\right) \cdot \varepsilon^{\frac{2}{\kappa}-2} \cdot \log^2 \frac{1}{\varepsilon\delta},$$

if $\hat{h}_n$ is the output of Algorithm 4 when run with label budget n and confidence parameter δ, then on an event of probability at least $1-\delta$,

$$\mathrm{er}\left(\hat{h}_n\right) \leq \nu + \varepsilon.$$

The proof of this result is essentially similar to the proof by Hanneke (2011), combined with some simplifying ideas from Koltchinskii (2010). It is also implicit in the proof of Lemma 26 below (by replacing "$\tilde{d}_f$" with "1" in the proof). The details are omitted. This result leads immediately to the following implication concerning the label complexity.

Theorem 25 *Let $\mathbb{C}$ be a VC class and suppose the joint distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1,+1\}$ satisfies Condition 1 for finite parameters $\mu, \kappa \in (1,\infty)$. With an appropriate (n,κ)-dependent setting of δ, Algorithm 4 achieves a label complexity Λ_a with*

$$\Lambda_a(\nu+\varepsilon, \mathcal{P}_{XY}) = O\left(\theta_f\left(\varepsilon^{\frac{1}{\kappa}}\right) \cdot \varepsilon^{\frac{2}{\kappa}-2} \cdot \log^2 \frac{1}{\varepsilon}\right).$$

Proof Taking $\delta = n^{-\frac{\kappa}{2\kappa-2}}$, the result follows by simple algebra. ■

We should note that it is possible to design a kind of wrapper to adaptively determine an appropriate δ value, so that the algorithm achieves the label complexity guarantee of Theorem 25 without requiring any explicit dependence on the noise parameter κ. Specifically, one can use an idea similar to the model selection procedure of Hanneke (2011) for this purpose. However, as our focus in this work is on moving beyond disagreement-based active learning, we do not include the details of such a procedure here.

Note that Theorem 25 represents an improvement over the known results for passive learning (namely, (5)) whenever $\theta_f(\varepsilon)$ is small, and in particular this gap can be large when $\theta_f < \infty$. The results of Lemma 24 and Theorem 25 represent the state-of-the-art (up to logarithmic factors) in our understanding of the label complexity of agnostic active learning for VC classes. Thus, any significant improvement over these would advance our understanding of the fundamental capabilities of active learning in the presence of label noise. Next, we provide such an improvement.

6.6 A New Type of Agnostic Active Learning Algorithm Based on Shatterable Sets

Algorithm 4 and Theorem 25 represent natural extensions of Meta-Algorithm 2 and Theorem 10 to the agnostic setting. As such, they not only benefit from the advantages of those methods (small $\theta_f(\varepsilon)$ implies improved label complexity), but also suffer the same disadvantages ($\mathcal{P}(\partial f) > 0$ implies no strong improvements over passive). It is therefore natural to investigate whether the improvements offered by Meta-Algorithm 3 and the corresponding Theorem 16 can be extended to the agnostic setting in a similar way. In particular, as was possible for Theorem 16 with respect to Theorem 10, we might wonder whether it is possible to replace $\theta_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$ in Theorem 25 with $\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$ by a modification of Algorithm 4 analogous to the modification of Meta-Algorithm 2 embodied in Meta-Algorithm 3. As we have seen, $\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$ is often significantly smaller in its asymptotic dependence on ε, compared to $\theta_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$, in many cases even bounded by a finite constant when $\theta_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$ is not. This would therefore represent a significant improvement over the known results for active learning under Condition 1. Toward this end, consider the following algorithm.

Algorithm 5
Input: label budget n, confidence parameter δ
Output: classifier $\hat{h}$

---

0. $m \leftarrow 0$, $i_0 \leftarrow 0$, $V_0 \leftarrow \mathbb{C}$
1. For $k = 1, 2, \ldots, d+1$
2. $\quad t \leftarrow 0$, $i_k \leftarrow i_{k-1}$, $m \leftarrow 2^{i_k}$, $V_{i_k+1} \leftarrow V_{i_k}$, $\mathcal{L}_{i_k+1} \leftarrow \{\}$
3. $\quad$ While $t < \lfloor 2^{-k} n \rfloor$ and $m \leq k \cdot 2^n$
4. $\quad\quad m \leftarrow m+1$
5. $\quad\quad$ If $\hat{P}_{4m}\left(S \in \mathcal{X}^{k-1} : V_{i_k+1} \text{ shatters } S \cup \{X_m\} | V_{i_k+1} \text{ shatters } S\right) \geq 1/2$
6. $\quad\quad\quad$ Request the label Y_m of X_m, and let $\mathcal{L}_{i_k+1} \leftarrow \mathcal{L}_{i_k+1} \cup \{(X_m, Y_m)\}$ and $t \leftarrow t+1$
7. $\quad\quad$ Else $\hat{y} \leftarrow \underset{y \in \{-1,+1\}}{\operatorname{argmax}} \hat{P}_{4m}\left(S \in \mathcal{X}^{k-1} : V_{i_k+1}[(X_m, -y)] \text{ does not shatter } S | V_{i_k+1} \text{ shatters } S\right)$
8. $\quad\quad\quad \mathcal{L}_{i_k+1} \leftarrow \mathcal{L}_{i_k+1} \cup \{(X_m, \hat{y})\}$ and $V_{i_k+1} \leftarrow V_{i_k+1}[(X_m, \hat{y})]$
9. $\quad\quad$ If $m = 2^{i_k+1}$
10. $\quad\quad\quad V_{i_k+1} \leftarrow \left\{ h \in V_{i_k+1} : \operatorname{er}_{\mathcal{L}_{i_k+1}}(h) - \min_{h' \in V_{i_k+1}} \operatorname{er}_{\mathcal{L}_{i_k+1}}(h') \leq \hat{U}_{i_k+1}(V_{i_k}, \delta) \right\}$
11. $\quad\quad\quad i_k \leftarrow i_k + 1$, then $V_{i_k+1} \leftarrow V_{i_k}$, and $\mathcal{L}_{i_k+1} \leftarrow \{\}$
12. Return any $\hat{h} \in V_{i_{d+1}+1}$

For the argmax in Step 7, we break ties in favor of a $\hat{y}$ value with $V_{i_k+1}[(X_m, \hat{y})] \neq \emptyset$ to maintain the invariant that $V_{i_k+1} \neq \emptyset$ (see the proof of Lemma 59); when both y values satisfy this, we may break ties arbitrarily. The procedure is specified in terms of several estimators. The $\hat{P}_{4m}$ estimators, as usual, are defined in Appendix B.1. For $\hat{U}_i$, we again use the definition (6) above, based on a data-dependent Rademacher complexity.

Algorithm 5 is largely based on the same principles as Algorithm 4, combined with Meta-Algorithm 3. As in Algorithm 4, the algorithm proceeds by repeatedly doubling the size of a labeled sample $\mathcal{L}_{i+1}$, while only requesting a subset of the labels in $\mathcal{L}_{i+1}$, inferring the others. As before, it updates the version space every time it doubles the size of the sample $\mathcal{L}_{i+1}$, and the update eliminates classifiers from the version space that make significantly more mistakes on $\mathcal{L}_{i+1}$ compared to others in the version space. In Algorithm 4, this is guaranteed to be effective, since the classifiers in the version space agree on all of the inferred labels, so that the differences of empirical error rates remain equal to the *true* differences of empirical error rates (i.e., under the true Y_m labels for all elements of $\mathcal{L}_{i+1}$); thus, the established results from the passive learning literature bounding the deviations of excess empirical error rates from excess true error rates can be applied, showing that this does not eliminate the best classifiers. In Algorithm 5, the situation is somewhat more subtle, but the principle remains the same. In this case, we *enforce* that the classifiers in the version space agree on the inferred labels in $\mathcal{L}_{i+1}$ by explicitly removing the disagreeing classifiers in Step 8. Thus, as long as Step 8 does not eliminate all of the good classifiers, then neither will Step 10. To argue that Step 8 does not eliminate all good classifiers, we appeal to the same reasoning as for Meta-Algorithm 1 and Meta-Algorithm 3. That is, for $k \leq \tilde{d}_f$ and sufficiently large n, as long as there exist good classifiers in the version space, the labels $\hat{y}$ inferred in Step 7 will agree with some good classifiers, and thus Step 8 will not eliminate all good classifiers. However, for $k > \tilde{d}_f$, the labels $\hat{y}$ in Step 7 have no such guarantees, so that we are only guaranteed that *some* classifier in the version space is not eliminated. Thus, determining guarantees on the error rate of this algorithm hinges on bounding the worst excess error rate among all classifiers in the version space at the conclusion of the $k = \tilde{d}_f$ round. This is essentially determined by the size of $\mathcal{L}_{i_k}$ at the conclusion of that round, which itself is largely determined by how frequently the algorithm requests labels during this $k = \tilde{d}_f$ round. Thus, once again the analysis rests on bounding the rate at which the frequency of label requests shrinks in the $k = \tilde{d}_f$ round, which determines the rate of growth of $|\mathcal{L}_{i_k}|$, and thus the final guarantee on the excess error rate.

As before, for computational feasibility, we can maintain the sets V_i implicitly as a set of constraints imposed by the previous updates, so that we may perform the various calculations required for the estimators $\hat{P}$ as constrained optimizations. Also, the update to $\mathcal{L}_{i_k+1}$ in Step 8 is merely included to make the algorithm statement and the proofs somewhat more elegant; it can be omitted, as long as we compensate with an appropriate renormalization of the $\mathrm{er}_{\mathcal{L}_{i_k+1}}$ values in Step 10 (i.e., multiplying by $2^{-i_k}|\mathcal{L}_{i_k+1}|$). Additionally, the same potential improvements we proposed in Section 5.5 for Meta-Algorithm 3 can be made to Algorithm 5 as well, again with only minor modifications to the proofs. We should note that Algorithm 5 is certainly not the only reasonable way to extend Meta-Algorithm 3 to the agnostic setting. For instance, another natural extension of Meta-Algorithm 1 to the agnostic setting, based on a completely different idea, appears in the author's doctoral dissertation (Hanneke, 2009b); that method can be improved in a natural way to take advantage of the sequential aspect of active learning, yielding an agnostic extension of Meta-Algorithm 3 differing from Algorithm 5 in several interesting ways (see the discussion in Section 6.8 below).

In the next subsection, we will see that the label complexities achieved by Algorithm 5 are often significantly better than the known results for passive learning. In fact, they are often significantly better than the presently-known results for any *active* learning algorithms in the published literature.

6.7 Improved Label Complexity Bounds for Active Learning with Noise

Under Condition 1, we can extend Lemma 24 and Theorem 25 in an analogous way to how Theorem 16 extends Theorem 10. Specifically, we have the following result, the proof of which is included in Appendix E.2.

Lemma 26 *Let $\mathbb{C}$ be a VC class and suppose the joint distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1,+1\}$ satisfies Condition 1 for finite parameters μ and κ. There is a $(\mathbb{C}, \mathcal{P}_{XY})$-dependent constant $c \in (0,\infty)$ such that, for any $\varepsilon, \delta \in (0, e^{-3})$, and any integer*

$$n \geq c \cdot \tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right) \cdot \varepsilon^{\frac{2}{\kappa}-2} \cdot \log^2 \frac{1}{\varepsilon\delta},$$

if $\hat{h}_n$ is the output of Algorithm 5 when run with label budget n and confidence parameter δ, then on an event of probability at least $1-\delta$,

$$\mathrm{er}\left(\hat{h}_n\right) \leq \nu + \varepsilon.$$

This has the following implication for the label complexity of Algorithm 5.

Theorem 27 *Let $\mathbb{C}$ be a VC class and suppose the joint distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1,+1\}$ satisfies Condition 1 for finite parameters $\mu, \kappa \in (1,\infty)$. With an appropriate (n,κ)-dependent setting of δ, Algorithm 5 achieves a label complexity Λ_a with*

$$\Lambda_a(\nu+\varepsilon, \mathcal{P}_{XY}) = O\left(\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right) \cdot \varepsilon^{\frac{2}{\kappa}-2} \cdot \log^2 \frac{1}{\varepsilon}\right).$$

Proof Taking $\delta = n^{-\frac{\kappa}{2\kappa-2}}$, the result follows by simple algebra. ■

Theorem 27 represents an interesting generalization beyond the realizable case, and beyond the disagreement coefficient analysis. Note that if $\tilde{\theta}_f(\varepsilon) = o\left(\varepsilon^{-1} \log^{-2}(1/\varepsilon)\right)$, Theorem 27 represents an improvement over the known results for passive learning (Massart and Nédélec, 2006). As we always have $\tilde{\theta}_f(\varepsilon) = o(1/\varepsilon)$, we should typically expect such improvements for all but the most extreme learning problems. Recall that $\theta_f(\varepsilon)$ is often *not* $o(1/\varepsilon)$, so that Theorem 27 is often a much stronger statement than Theorem 25. In particular, this is a significant improvement over the known results for passive learning whenever $\tilde{\theta}_f < \infty$, and an equally significant improvement over Theorem 25 whenever $\tilde{\theta}_f < \infty$ but $\theta_f(\varepsilon) = \Omega(1/\varepsilon)$ (see above for examples of this). However, note that unlike Meta-Algorithm 3, Algorithm 5 is *not* an activizer. Indeed, it is not clear (to the author) how to modify the algorithm to make it a universal activizer for $\mathbb{C}$ (even for the realizable case), while maintaining the guarantees of Theorem 27.

As with Theorem 16 and Corollary 17, Algorithm 5 and Theorem 27 can potentially be improved in a variety of ways, as outlined in Section 5.5. In particular, Theorem 27 can be made slightly sharper in some cases by replacing $\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$ with the sometimes-smaller (though more complicated) quantity (4) (with $r_0 = \varepsilon^{\frac{1}{\kappa}}$).

6.8 Beyond Condition 1

While Theorem 27 represents an improvement over the known results for agnostic active learning, Condition 1 is not fully general, and disallows many important and interesting scenarios. In particular, one key property of Condition 1, heavily exploited in the label complexity proofs for both passive learning and disagreement-based active learning, is that it implies $\mathrm{diam}(\mathbb{C}(\varepsilon)) \to 0$ as $\varepsilon \to 0$. In scenarios where this shrinking diameter condition is not satisfied, the existing proofs of (5) for passive learning break down, and furthermore, the disagreement-based algorithms themselves cease to give significant improvements over passive learning, for essentially the same reasons leading to the "only if" part of Theorem 5 (i.e., the sampling region never focuses beyond some nonzero-probability region). Even more alarming (at first glance) is the fact that this same problem can sometimes be observed for the $k = \tilde{d}_f$ round of Algorithm 5; that is, $\mathcal{P}\left(x : \mathcal{P}^{\tilde{d}_f-1}(S \in \mathcal{X}^{\tilde{d}_f-1} : V_{i_{\tilde{d}_f}+1} \text{ shatters } S \cup \{x\} | V_{i_{\tilde{d}_f}+1} \text{ shatters } S) \geq 1/2\right)$ is no longer guaranteed to approach 0 as the budget n increases (as it *does* when $\mathrm{diam}(\mathbb{C}(\varepsilon)) \to 0$). Thus, if we wish to approach an understanding of improvements achievable by active learning in general, we must come to terms with scenarios where $\mathrm{diam}(\mathbb{C}(\varepsilon))$ does not shrink to zero.

Interestingly, it seems that $\mathrm{diam}(\mathbb{C}(\varepsilon)) \nrightarrow 0$ might not be a problem for some algorithms based on shatterable sets, such as Algorithm 5. In particular, Algorithm 5 appears to continue exhibiting reasonable behavior in such scenarios. That is, even if there is a nonshrinking probability that the query condition in Step 5 is satisfied for $k = \tilde{d}_f$, on any given sequence $\mathcal{Z}$ there must be *some* smallest value of k for which this probability *does* shrink as $n \to \infty$. For this value of k, we should expect to observe good behavior from the algorithm, in that (for sufficiently large n) the inferred labels in Step 7 will tend to agree with *some* classifier $f \in \mathrm{cl}(\mathbb{C})$ with $\mathrm{er}(f) = \nu(\mathbb{C};\mathcal{P}_{XY})$. Thus, the algorithm addresses the problem of multiple optimal classifiers by effectively *selecting* one of the optimal classifiers.

To illustrate this phenomenon, consider learning with respect to the space of threshold classifiers (Example 1) with $\mathcal{P}$ uniform in $[0,1]$, and let $(X,Y) \sim \mathcal{P}_{XY}$ satisfy $\mathbb{P}(Y = +1|X) = 0$ for $X < 1/3$, $\mathbb{P}(Y = +1|X) = 1/2$ for $1/3 \leq X < 2/3$, and $\mathbb{P}(Y = +1|X) = 1$ for $2/3 \leq X$. As we know from above, $\tilde{d}_f = 1$ here. However, in this scenario we have $\mathrm{DIS}(\mathbb{C}(\varepsilon)) \to [1/3,2/3]$ as $\varepsilon \to 0$. Thus, Algorithm 4 never focuses its queries beyond a constant fraction of $\mathcal{X}$, and therefore cannot improve over certain passive learning algorithms in terms of the asymptotic dependence of its label complexity on ε (assuming a worst-case choice of $\hat{h}$ in Step 9). However, for $k = 2$ in Algorithm 5, every X_m will be assigned a label $\hat{y}$ in Step 7 (since no two points are shattered); furthermore, for sufficiently large n we have (with high probability) $\mathrm{DIS}(V_{i_1})$ not too much larger than $[1/3,2/3]$, so that most points in $\mathrm{DIS}(V_{i_1})$ can be labeled either $+1$ or -1 by some optimal classifier. For us, this has two implications. First, one can show that with very high probability, the $S \in [1/3,2/3]^1$ will dominate the votes for $\hat{y}$ in Step 7 (for all m processed while $k = 2$), so that the $\hat{y}$ inferred for any $X_m \notin [1/3,2/3]$ will agree with all of the optimal classifiers. Second, the inferred labels $\hat{y}$ for $X_m \in [1/3,2/3]$ will definitely agree with *some* optimal classifier. Since we also impose the $h(X_m) = \hat{y}$ constraint for V_{i_2+1} in Step 8, the inferred $\hat{y}$ labels must all be consistent with the *same* optimal classifier, so that V_{i_2+1} will quickly converge to within a small neighborhood around that classifier, without any further label requests. Note, however, that the particular optimal classifier the algorithm converges to will be a random variable, determined by the particular sequence of data points processed by the algorithm; thus, it cannot be determined a priori, which significantly complicates any general attempt to analyze the label complexity achieved by the algorithm for arbitrary $\mathbb{C}$ and $\mathcal{P}_{XY}$. In particular, for some $\mathbb{C}$ and $\mathcal{P}_{XY}$, even this minimal k for which convergence occurs may be a nondeterministic random variable. At this time, it is not entirely clear how general this phenomenon is (i.e., Algorithm 5 providing improvements over certain passive algorithms even for distributions with $\mathrm{diam}(\mathbb{C}(\varepsilon)) \nrightarrow 0$), nor how to characterize the label complexity achieved by Algorithm 5 in general settings where $\mathrm{diam}(\mathbb{C}(\varepsilon)) \nrightarrow 0$.

However, as mentioned earlier, there are other natural ways to generalize Meta-Algorithm 3 to handle noise, some of which have more predictable behavior. In particular, the original thesis work of Hanneke (2009b) explores a technique for active learning, which unlike Algorithm 5, only uses the *requested* labels, not the inferred labels, and as a consequence never eliminates any optimal classifier from V. Because of this fact, the sampling region for each k converges to a predictable limiting region, so that we have an accurate *a priori* characterization of the algorithm's behavior. However, it is not immediately clear (to the author) whether this alternative technique might lead to a method achieving results similar to Theorem 27.

To get a better understanding of the scenario where $\mathrm{diam}(\mathbb{C}(\varepsilon)) \nrightarrow 0$, it will be helpful to partition the distributions into two distinct categories, which we will refer to as the *benign noise* case and the *misspecified model* case. The $\mathcal{P}_{XY}$ in the benign noise case are characterized by the property that $\nu(\mathbb{C};\mathcal{P}_{XY}) = \nu^*(\mathcal{P}_{XY})$; this is in some ways similar to the realizable case, in that $\mathbb{C}$ can approximate an optimal classifier, except that the labels are stochastic. In the benign noise case, the only reason $\operatorname{diam}(\mathbb{C}(\varepsilon))$ would not shrink to zero is if there is a nonzero probability set of points x with $\eta(x) = 1/2$; that is, there are at least two classifiers achieving the Bayes error rate, and they are at nonzero distance from each other, which must mean they disagree on some points that have equal probability of either label occurring. In contrast, the misspecified model case is characterized by $\nu(\mathbb{C};\mathcal{P}_{XY}) > \nu^*(\mathcal{P}_{XY})$. In this case, if the diameter does not shrink, it is because of the existence of two classifiers $h_1, h_2 \in \operatorname{cl}(\mathbb{C})$ achieving error rate $\nu(\mathbb{C};\mathcal{P}_{XY})$, with $\mathcal{P}(x : h_1(x) \neq h_2(x)) > 0$. However, unlike above, since they do not achieve the Bayes error rate, it is possible that a significant fraction of the set of points they disagree on may have $\eta(x) \neq 1/2$.

Intuitively, the benign noise case is relatively easier for active learning, since the noisy points that prevent $\operatorname{diam}(\mathbb{C}(\varepsilon))$ from shrinking can essentially be assigned arbitrary labels, as in the thresholds example above. For instance, as in Algorithm 5, we could assign a label to points in this region and discard any classifiers inconsistent with the label, confident that we have kept at least one optimal classifier. Another possibility is simply to ignore the points in this region, since in the end they are inconsequential for the excess error rate of the classifier we return; in some sense, this is the strategy taken by the method of Hanneke (2009b).

In contrast, the misspecified model case intuitively makes the active learning problem more difficult. For instance, if h_1 and h_2 in $\operatorname{cl}(\mathbb{C})$ both have error rate $\nu(\mathbb{C};\mathcal{P}_{XY})$, the original method of Hanneke (2009b) has the possibility of inferring the label $h_2(x)$ for some point x when in fact $h_1(x)$ is better for that particular x, and vice versa for the points x where $h_2(x)$ would be better, thus getting the worst of both and potentially doubling the error rate in the process. Algorithm 5 may fare better in this case, since imposing the inferred label $\hat{y}$ as a constraint in Step 8 effectively *selects* one of h_1 or h_2, and discards the other one. As before, whether Algorithm 5 selects h_1 or h_2 will generally depend on the particular data sequence $\mathcal{Z}$, which therefore makes any a priori analysis of the label complexity more challenging.

Interestingly, it turns out that, for the purpose of exploring Conjecture 23, we can circumvent all of these issues by noting that there is a trivial solution to the misspecified model case. Specifically, since in our present context we are only interested in the label complexity for achieving error rate better than $\nu + \varepsilon$, we can simply turn to any algorithm that asymptotically achieves an error rate strictly better than ν (e.g., Devroye et al., 1996), in which case the algorithm should require only a finite constant number of labels to achieve an expected error rate better than ν. To make the algorithm effective for the general case, we simply split our budget in three: one part for an active learning algorithm, such as Algorithm 5, for the benign noise case, one part for the method above handling the misspecified model case, and one part to select among their outputs. The full details of such a procedure are specified in Appendix E.3, along with a proof of its performance guarantees, which are summarized as follows.

Theorem 28 *Fix any concept space $\mathbb{C}$. Suppose there exists an active learning algorithm $\mathcal{A}_a$ achieving a label complexity Λ_a. Then there exists an active learning algorithm $\mathcal{A}'_a$ achieving a label complexity Λ'_a such that, for any distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1,+1\}$, there exists a function $\lambda(\varepsilon) \in \operatorname{Polylog}(1/\varepsilon)$ such that*

$$\Lambda'_a(\nu+\varepsilon,\mathcal{P}_{XY}) \leq \begin{cases} \max\{2\Lambda_a(\nu+\varepsilon/2,\mathcal{P}_{XY}),\lambda(\varepsilon)\}, & \textit{in the benign noise case} \\ \lambda(\varepsilon), & \textit{in the misspecified model case} \end{cases}.$$

The main point of Theorem 28 is that, for our purposes, we can safely ignore the misspecified model case (as its solution is a trivial extension), and focus entirely on the performance of algorithms for the benign noise case. In particular, for any label complexity Λ_p, every $\mathcal{P}_{XY} \in \text{Nontrivial}(\Lambda_p;\mathbb{C})$ in the misspecified model case has $\Lambda'_a(\nu+\varepsilon,\mathcal{P}_{XY}) = o(\Lambda_p(\nu+\varepsilon,\mathcal{P}_{XY}))$, for Λ'_a as in Theorem 28. Thus, if there exists an active meta-algorithm achieving the strong improvement guarantees of an activizer for some passive learning algorithm $\mathcal{A}_p$ (Definition 21) for all distributions $\mathcal{P}_{XY}$ in the benign noise case, then there exists an activizer for $\mathcal{A}_p$ with respect to $\mathbb{C}$ in the agnostic case.

7. Open Problems

In some sense, this work raises more questions than it answers. Here, we list several problems that remain open at this time. Resolving any of these problems would make a significant contribution to our understanding of the fundamental capabilities of active learning.

- We have established the existence of universal activizers for VC classes in the realizable case. However, we have not made any serious attempt to characterize the properties that such activizers can possess. In particular, as mentioned, it would be interesting to know whether activizers exist that *preserve* certain favorable properties of the given passive learning algorithm. For instance, we know that some passive learning algorithms (say, for linear separators) achieve a label complexity that is independent of the dimensionality of the space $\mathcal{X}$, under a large margin condition on f and $\mathcal{P}$ (Balcan, Blum, and Vempala, 2006b). Is there an activizer for such algorithms that preserves this large-margin-based dimension-independence in the label complexity? Similarly, there are passive algorithms whose label complexity has a weak dependence on dimensionality, due to sparsity considerations (Bunea, Tsybakov, and Wegkamp, 2009; Wang and Shen, 2007). Is there an activizer for these algorithms that preserves this sparsity-based weak dependence on dimension? Is there an activizer that preserves adaptiveness to the dimension of the manifold to which $\mathcal{P}$ is restricted? What about an activizer that is *sparsistent* (Rocha, Wang, and Yu, 2009), given any sparsistent passive learning algorithm as input? Is there an activizer that preserves admissibility, in that given any admissible passive learning algorithm, the activized algorithm is an admissible active learning algorithm? Is there an activizer that, given any minimax optimal passive learning algorithm as input, produces a minimax optimal active learning algorithm? What about preserving other notions of optimality, or other properties?

- There may be some waste in the above activizers, since the label requests used in their initial phase (reducing the version space) are not used by the passive algorithm to produce the final classifier. This guarantees the examples fed into the passive algorithm are conditionally independent given the number of examples. Intuitively, this seems necessary for the general results, since any dependence among the examples fed to the passive algorithm could influence its label complexity. However, it is not clear (to the author) how dramatic this effect can be, nor whether a simpler strategy (e.g., slightly randomizing the budget of label requests) might yield a similar effect while allowing a single-stage approach where all labels are used in the passive algorithm. It seems intuitively clear that some special types of passive algorithms should be able to use the full set of examples, from both phases, while still maintaining the strict improvements guaranteed in the main theorems above. What general properties must such passive algorithms possess?

- As previously mentioned, the vast majority of empirically-tested *heuristic* active learning algorithms in the published literature are designed in a reduction style, using a well-known passive learning algorithm as a subroutine, constructing sets of labeled examples and feeding them into the passive learning algorithm at various points in the execution of the active learning algorithm (e.g., Abe and Mamitsuka, 1998; McCallum and Nigam, 1998; Schohn and Cohn, 2000; Campbell, Cristianini, and Smola, 2000; Tong and Koller, 2001; Roy and McCallum, 2001; Muslea, Minton, and Knoblock, 2002; Lindenbaum, Markovitch, and Rusakov, 2004; Mitra, Murthy, and Pal, 2004; Roth and Small, 2006; Schein and Ungar, 2007; Har-Peled, Roth, and Zimak, 2007; Beygelzimer, Dasgupta, and Langford, 2009). However, rather than including some examples whose labels are requested and other examples whose labels are *inferred* in the sets of labeled examples given to the passive learning algorithm (as in our rigorous methods above), these heuristic methods typically only input to the passive algorithm the examples whose labels were *requested*. We should expect that meta-algorithms of this type could not be *universal* activizers for $\mathbb{C}$, but perhaps there do exist meta-algorithms of this type that are activizers for every passive learning algorithm of some special type. What are some general conditions on the passive learning algorithm so that some meta-algorithm of this type (i.e., feeding in only the *requested* labels) can activize every passive learning algorithm satisfying those conditions?

- As discussed earlier, the definition of "activizer" is based on a trade-off between the strength of claimed improvements for nontrivial scenarios, and ease of analysis within the framework. There are two natural questions regarding the possibility of stronger notions of "activizer." In Definition 3 we allow a constant factor c loss in the ε argument of the label complexity. In most scenarios, this loss is inconsequential (e.g., typically $\Lambda_p(\varepsilon/c, f, \mathcal{P}) = O(\Lambda_p(\varepsilon, f, \mathcal{P}))$), but one can construct scenarios where it does make a difference. In our proofs, we see that it is possible to achieve $c = 3$; in fact, a careful inspection of the proofs reveals we can even get $c = (1 + o(1))$, a function of ε, converging to 1. However, whether there exist universal activizers for every VC class that have $c = 1$ remains an open question.

  A second question regards our notion of "nontrivial problems." In Definition 3, we have chosen to think of any target and distribution with label complexity growing faster than $\mathrm{Polylog}(1/\varepsilon)$ as *nontrivial*, and do not require the activized algorithm to improve over the underlying passive algorithm for scenarios that are trivial for the passive algorithm. As mentioned, Definition 3 does have implications for the label complexities of these problems, as the label complexity of the activized algorithm will improve over every nontrivial upper bound on the label complexity of the passive algorithm. However, in order to allow for various operations in the meta-algorithm that may introduce additive $\mathrm{Polylog}(1/\varepsilon)$ terms due to exponentially small failure probabilities, such as the test that selects among hypotheses in ActiveSelect, we do not require the activized algorithm to achieve the same *order* of label complexity in trivial scenarios. For instance, there may be cases in which a passive algorithm achieves $O(1)$ label complexity for a particular $(f, \mathcal{P})$, but its activized counterpart has $\Theta(\log(1/\varepsilon))$ label complexity. The intention is to define a framework that focuses on nontrivial scenarios, where passive learning uses prohibitively many labels, rather than one that requires us to obsess over extra additive logarithmic terms. Nonetheless, there is a question of whether these losses in the label complexities of trivial problems are necessary to gain these improvements in the label complexities of nontrivial problems.

There is also the question of how much the definition of "nontrivial" can be relaxed. Specifically, we have the following question: to what extent can we relax the notion of "nontrivial" in Definition 3, while still maintaining the existence of universal activizers for VC classes? We see from our proofs that we can at least replace $\mathrm{Polylog}(1/\varepsilon)$ with $O(\log(1/\varepsilon))$. However, it is not clear whether we can go further than this in the realizable case (e.g., to say "nontrivial" means $\omega(1)$). When there is noise, it is clear that we cannot relax the notion of "nontrivial" beyond replacing $\mathrm{Polylog}(1/\varepsilon)$ with $O(\log(1/\varepsilon))$. Specifically, whenever $\mathrm{DIS}(\mathbb{C}) \neq \emptyset$, for any label complexity Λ_a achieved by an active learning algorithm, there must be some $\mathcal{P}_{XY}$ with $\Lambda_a(\nu+\varepsilon, \mathcal{P}_{XY}) = \Omega(\log(1/\varepsilon))$, even with the support of $\mathcal{P}$ restricted to a *single point* $x \in \mathrm{DIS}(\mathbb{C})$; the proof of this is via a reduction from sequential hypothesis testing for whether a coin has bias α or $1-\alpha$, for some $\alpha \in (0, 1/2)$. Since passive learning via empirical risk minimization can achieve label complexity $\Lambda_p(\nu+\varepsilon, \mathcal{P}_{XY}) = O(\log(1/\varepsilon))$ whenever the support of $\mathcal{P}$ is restricted to a single point, we cannot further relax the notion of "nontrivial," while preserving the possibility of a positive outcome for Conjecture 23. It is interesting to note that this entire issue vanishes if we are only interested in methods that achieve error at most ε with probability at least $1-\delta$, where $\delta \in (0,1)$ is some acceptable constant failure probability, as in the work of Balcan, Hanneke, and Vaughan (2010); in this case, we can simply take "nontrivial" to mean $\omega(1)$ label complexity, and both Meta-Algorithm 1 and Meta-Algorithm 3 remain universal activizers for $\mathbb{C}$ under this alternative definition, and achieve $O(1)$ label complexity in trivial scenarios.

- Another interesting question concerns efficiency. Suppose there exists an algorithm to find an element of $\mathbb{C}$ consistent with any labeled sequence $\mathcal{L}$ in time polynomial in $|\mathcal{L}|$ and d, and that $\mathcal{A}_p(\mathcal{L})$ has running time polynomial in $|\mathcal{L}|$ and d. Under these conditions, is there an activizer for $\mathcal{A}_p$ capable of achieving an error rate smaller than any ε in running time polynomial in $1/\varepsilon$ and d, given some appropriately large budget n? Recall that if we knew the value of $\tilde{d}_f$ and $\tilde{d}_f \leq c \log d$, then Meta-Algorithm 1 could be made efficient, as discussed above. Therefore, this question is largely focused on the issue of adapting to the value of $\tilde{d}_f$. Another related question is whether there is an efficient active learning algorithm achieving the label complexity bound of Corollary 7 or Corollary 17.

- One question that comes up in the results above is the minimum number of *batches* of label requests necessary for a universal activizer for $\mathbb{C}$. In Meta-Algorithm 0 and Theorem 5, we saw that sometimes two batches are sufficient: one to reduce the version space, and another to construct the labeled sample by requesting only those points in the region of disagreement. We certainly cannot use fewer than two batches in a universal activizer for any nontrivial concept space, so that this represents the minimum. However, to get a universal activizer for *every* concept space, we increased the number of batches to *three* in Meta-Algorithm 1. The question is whether this increase is really necessary. Is there always a universal activizer using only *two* batches of label requests, for every VC class $\mathbb{C}$?

- For some $\mathbb{C}$, the learning process in the above methods might be viewed in two components: one component that performs active learning as usual (say, disagreement-based) under the assumption that the target function is very simple, and another component that searches for signs that the target function is in fact more complex. Thus, for some natural classes such as linear separators, it would be interesting to find simpler, more specialized methods, which explicitly execute these two components. For instance, for the first component, we might consider the usual margin-based active learning methods, which query near a current guess of the separator (Dasgupta, Kalai, and Monteleoni, 2005, 2009; Balcan, Broder, and Zhang, 2007), except that we bias toward simple hypotheses via a regularization penalty in the optimization that defines how we update the separator in response to a query. The second component might then be a simple random search for points whose correct classification requires larger values of the regularization term.

- Can we construct universal activizers for some concept spaces with infinite VC dimension? What about under some constraints on the distribution $\mathcal{P}$ or $\mathcal{P}_{XY}$ (e.g., the usual entropy conditions)? It seems we can still run Meta-Algorithm 1, Meta-Algorithm 3, and Algorithm 5 in this case, except we should increase the number of rounds (values of k) as a function of n; this may continue to have reasonable behavior even in some cases where $\tilde{d}_f = \infty$, especially when $\mathcal{P}^k(\partial^k f) \to 0$ as $k \to \infty$. However, it is not clear whether they will continue to guarantee the strict improvements over passive learning in the realizable case, nor what label complexity guarantees they will achieve. One specific question is whether there is a method always achieving label complexity $o\left(\varepsilon^{\frac{1-\rho}{\kappa}-2}\right)$, where ρ is from the entropy conditions (van der Vaart and Wellner, 1996) and κ is from Condition 1. This would be an improvement over the known results for passive learning (Mammen and Tsybakov, 1999; Tsybakov, 2004; Koltchinskii, 2006). Another related question is whether we can improve over the known results for active learning in these scenarios. Specifically, Hanneke (2011) proved a bound of $\tilde{O}\left(\theta_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\varepsilon^{\frac{2-\rho}{\kappa}-2}\right)$ on the label complexity of a certain disagreement-based active learning method, under entropy conditions and Condition 1. Do there exist active learning methods achieving asymptotically smaller label complexities than this, in particular improving the $\theta_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$ factor? The quantity $\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$ is no longer defined when $\tilde{d}_f = \infty$, so this might not be a direct extension of Theorem 27, but we could perhaps use the sequence of $\theta_f^{(k)}\left(\varepsilon^{\frac{1}{\kappa}}\right)$ values in some other way to replace $\theta_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$ in this case.

- Generalizing the previous question, we might even be so bold as to ask whether there exists a universal activizer for the space of *all* classifiers. Let us refer to such a method as a *universal activizer* (in general). The present work shows that there is a universal activizer for every VC class. Furthermore, Lemma 34 implies that, for any sequence $\mathbb{C}_1, \mathbb{C}_2, \ldots$ of concept spaces for which there exist universal activizers, there also exists a universal activizer for $\bigcup_{i=1}^{\infty} \mathbb{C}_i$: namely, the method that runs each of the activizers for $\mathbb{C}_i$ with respective budgets $\lfloor 3n/(\pi i)^2 \rfloor$, for $i = 1, 2, \ldots, \lfloor \sqrt{3n}/\pi \rfloor$, producing hypotheses $h_1, \ldots, h_{\lfloor \sqrt{3n}/\pi \rfloor}$, then returns the value of ActiveSelect$(\{h_1, \ldots, h_{\lfloor \sqrt{3n}/\pi \rfloor}\}, \lceil n/2 \rceil, \{X_M, X_{M+1}, \ldots\})$, where M is larger than any index accessed by these $\lfloor \sqrt{3n}/\pi \rfloor$ activizers. In fact, the proof of Theorem 6 entails that the $o(\Lambda_p(\varepsilon, f, \mathcal{P}))$ guarantee holds for f in the *closure* $\mathrm{cl}(\mathbb{C})$ of any VC class $\mathbb{C}$. Combined with the above trick, it follows that we can achieve the $o(\Lambda_p(\varepsilon, f, \mathcal{P}))$ strong improvement guarantee over passive learning for all f in $\bigcup_{i=1}^{\infty} \mathrm{cl}(\mathbb{C}_i)$, where the $\mathbb{C}_i$ sets are VC classes. We can always construct a sequence of VC classes $\mathbb{C}_1, \mathbb{C}_2, \ldots$ such that $\mathrm{cl}\left(\bigcup_{i=1}^{\infty} \mathbb{C}_i\right)$ is the set of all classifiers. However, $\bigcup_{i=1}^{\infty} \mathrm{cl}(\mathbb{C}_i)$ is generally not the same as $\mathrm{cl}\left(\bigcup_{i=1}^{\infty} \mathbb{C}_i\right)$, so that achieving $\Lambda_a(c\varepsilon, f, \mathcal{P}) = o(\Lambda_p(\varepsilon, f, \mathcal{P}))$ for all $f \in \bigcup_{i=1}^{\infty} \mathrm{cl}(\mathbb{C}_i)$ does not necessarily guarantee the same for all $f \in \mathrm{cl}\left(\bigcup_{i=1}^{\infty} \mathbb{C}_i\right)$. Thus, constructing a general universal activizer would be a nontrivial extension of the present work, and the fundamental question of the existence (or nonexistence) of such meta-algorithms remains a fascinating open question.

- There is also a question about generalizing this approach to label spaces other than $\{-1,+1\}$, and possibly other loss functions. It should be straightforward to extend these results to the setting of multiclass classification. However, it is not clear what the implications would be for general structured prediction problems, where the label space may be quite large (even infinite), and the loss function involves a notion of *distance* between labels. From a practical perspective, this question is particularly interesting, since problems with more complicated label spaces are often the scenarios where active learning is most needed, as it takes substantial time or effort to label each example. At this time, there are no published theoretical results on the label complexity improvements achievable for general structured prediction problems.
- All of the claims in this work also hold when $\mathcal{A}_p$ is a *semi-supervised* passive learning algorithm, simply by withholding a set of unlabeled data points in a preprocessing step, and feeding them into the passive algorithm along with the labeled set generated by the activizer. However, it is not clear whether further claims are possible when activizing a semi-supervised algorithm, for instance by taking into account specific details of the learning bias used by the particular semi-supervised algorithm (e.g., a cluster assumption).
- The splitting index analysis of Dasgupta (2005) has the interesting feature of characterizing a *trade-off* between the number of label requests and the number of unlabeled examples used by the active learning algorithm. In the present work, we do not characterize any such trade-off. Indeed, the algorithms do not really have any parameter to adjust the number of unlabeled examples they use (aside from the precision of the $\hat{P}$ estimators), so that they simply use as many as they need and then halt. This is true in both the realizable case and in the agnostic case. It would be interesting to try to modify these algorithms and their analysis so that, when there are more unlabeled examples available than would be used by the above methods, the algorithms can take advantage of this in a way that can be reflected in improved label complexity bounds, and when there are fewer unlabeled examples available, the algorithms can alter their behavior to compensate for this, at the cost of an increased label complexity. This would be interesting both for the realizable and agnostic cases. In fact, in the agnostic case, there are no known methods that exhibit this type of trade-off.
- Finally, as mentioned in the previous section, there is a serious question concerning what types of algorithms can be activized in the agnostic case, and how large the improvements in label complexity will be. In particular, Conjecture 23 hypothesizes that for any VC class, we can activize some empirical risk minimization algorithm in the agnostic case. Resolving this conjecture (either positively or negatively) should significantly advance our understanding of the capabilities of active learning compared to passive learning.

Acknowledgments

I am grateful to Nina Balcan, Rui Castro, Sanjoy Dasgupta, Carlos Guestrin, Vladimir Koltchinskii, John Langford, Rob Nowak, Larry Wasserman, and Eric Xing for insightful discussions. I also thank the anonymous referees for several thoughtful comments and suggestions.

Appendix A. Proofs Related to Section 3: Disagreement-Based Learning

The following result follows from a theorem of Anthony and Bartlett (1999), based on the classic results of Vapnik (1982) (with slightly better constant factors); see also the work of Blumer, Ehrenfeucht, Haussler, and Warmuth (1989).

Lemma 29 *For any VC class $\mathbb{C}$, $m \in \mathbb{N}$, and classifier f such that $\forall r > 0, \mathrm{B}(f,r) \neq \emptyset$, let $V_m^\star = \{h \in \mathbb{C} : \forall i \leq m, h(X_i) = f(X_i)\}$; for any $\delta \in (0,1)$, there is an event $H_m(\delta)$ with $\mathbb{P}(H_m(\delta)) \geq 1-\delta$ such that, on $H_m(\delta)$, $V_m^\star \subseteq \mathrm{B}(f, \phi(m;\delta))$, where*

$$\phi(m;\delta) = 2\frac{d \ln \frac{2e\max\{m,d\}}{d} + \ln(2/\delta)}{m}.$$

A fact we will use repeatedly is that, for any $N(\varepsilon) = \omega(\log(1/\varepsilon))$, we have $\phi(N(\varepsilon);\varepsilon) = o(1)$.

Lemma 30 *For $\hat{P}_n(\mathrm{DIS}(V))$ from (1), on an event J_n with $\mathbb{P}(J_n) \geq 1 - 2\cdot\exp\{-n/4\}$,*

$$\max\{\mathcal{P}(\mathrm{DIS}(V)), 4/n\} \leq \hat{P}_n(\mathrm{DIS}(V)) \leq \max\{4\mathcal{P}(\mathrm{DIS}(V)), 8/n\}.$$

Proof Note that the sequence $\mathcal{U}_n$ from (1) is independent from both V and $\mathcal{L}$. By a Chernoff bound, on an event J_n with $\mathbb{P}(J_n) \geq 1-2\cdot\exp\{-n/4\}$,

$$\mathcal{P}(\mathrm{DIS}(V)) > 2/n \implies \frac{\mathcal{P}(\mathrm{DIS}(V))}{\frac{1}{n^2}\sum_{x\in\mathcal{U}_n} \mathbb{1}_{\mathrm{DIS}(V)}(x)} \in [1/2, 2],$$
$$\text{and } \mathcal{P}(\mathrm{DIS}(V)) \leq 2/n \implies \frac{1}{n^2}\sum_{x\in\mathcal{U}_n} \mathbb{1}_{\mathrm{DIS}(V)}(x) \leq 4/n.$$

This immediately implies the stated result. ■

Lemma 31 *Let $\lambda : (0,1) \to (0,\infty)$ and $L : \mathbb{N}\times(0,1) \to [0,\infty)$ be s.t. $\lambda(\varepsilon) = \omega(1)$, $L(1,\varepsilon) = 0$ and $L(n,\varepsilon) \to \infty$ as $n \to \infty$ for every $\varepsilon \in (0,1)$, and for any $\mathbb{N}$-valued $N(\varepsilon) = \omega(\lambda(\varepsilon))$, $L(N(\varepsilon),\varepsilon) = \omega(N(\varepsilon))$. Let $L^{-1}(m;\varepsilon) = \max\{n \in \mathbb{N} : L(n,\varepsilon) < m\}$ for every $m \in (0,\infty)$. Then for any $\Lambda : (0,1) \to (0,\infty)$ with $\Lambda(\varepsilon) = \omega(\lambda(\varepsilon))$, we have $L^{-1}(\Lambda(\varepsilon);\varepsilon) = o(\Lambda(\varepsilon))$.*

Proof First note that L^{-1} is well-defined and finite, due to the facts that $L(n,\varepsilon)$ can be 0 and is diverging in n. Let $\Lambda(\varepsilon) = \omega(\lambda(\varepsilon))$. It is fairly straightforward to show $L^{-1}(\Lambda(\varepsilon);\varepsilon) \neq \Omega(\Lambda(\varepsilon))$, but the stronger $o(\Lambda(\varepsilon))$ result takes slightly more work. Let $\bar{L}(n,\varepsilon) = \min\left\{L(n,\varepsilon), n^2/\lambda(\varepsilon)\right\}$ for every $n \in \mathbb{N}$ and $\varepsilon \in (0,1)$, and let $\bar{L}^{-1}(m;\varepsilon) = \max\{n \in \mathbb{N} : \bar{L}(n,\varepsilon) < m\}$. We will first prove the result for $\bar{L}$.

Note that by definition of $\bar{L}^{-1}$, we know

$$\left(\bar{L}^{-1}(\Lambda(\varepsilon);\varepsilon)+1\right)^2/\lambda(\varepsilon) \geq \bar{L}\left(\bar{L}^{-1}(\Lambda(\varepsilon);\varepsilon)+1,\varepsilon\right) \geq \Lambda(\varepsilon) = \omega(\lambda(\varepsilon)),$$

which implies $\bar{L}^{-1}(\Lambda(\varepsilon);\varepsilon) = \omega(\lambda(\varepsilon))$. But, by definition of $\bar{L}^{-1}$ and the condition on L,

$$\Lambda(\varepsilon) > \bar{L}\left(\bar{L}^{-1}(\Lambda(\varepsilon);\varepsilon),\varepsilon\right) = \omega\left(\bar{L}^{-1}(\Lambda(\varepsilon);\varepsilon)\right).$$

Since $\bar{L}^{-1}(m;\varepsilon) \geq L^{-1}(m;\varepsilon)$ for all $m > 0$, this implies $\Lambda(\varepsilon) = \omega\left(L^{-1}(\Lambda(\varepsilon);\varepsilon)\right)$, or equivalently $L^{-1}(\Lambda(\varepsilon);\varepsilon) = o(\Lambda(\varepsilon))$. ■

Lemma 32 *For any VC class $\mathbb{C}$ and passive algorithm $\mathcal{A}_p$, if $\mathcal{A}_p$ achieves label complexity Λ_p, then Meta-Algorithm 0, with $\mathcal{A}_p$ as its argument, achieves a label complexity Λ_a such that, for every $f \in \mathbb{C}$ and distribution $\mathcal{P}$ over $\mathcal{X}$, if $\mathcal{P}(\partial_{\mathbb{C},\mathcal{P}} f) = 0$ and $\infty > \Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon))$, then $\Lambda_a(2\varepsilon, f, \mathcal{P}) = o(\Lambda_p(\varepsilon, f, \mathcal{P}))$.*

Proof This proof follows similar lines to a proof of a related result of Balcan, Hanneke, and Vaughan (2010). Suppose $\mathcal{A}_p$ achieves a label complexity Λ_p, and that $f \in \mathbb{C}$ and distribution $\mathcal{P}$ satisfy $\infty > \Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon))$ and $\mathcal{P}(\partial_{\mathbb{C},\mathcal{P}} f) = 0$. Let $\varepsilon \in (0,1)$. For $n \in \mathbb{N}$, let $\Delta_n(\varepsilon) = \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f, \phi(\lfloor n/2 \rfloor; \varepsilon/2))))$, $L(n;\varepsilon) = \lfloor n / \max\{32/n, 16\Delta_n(\varepsilon)\} \rfloor$, and for $m \in (0,\infty)$ let $L^{-1}(m;\varepsilon) = \max\{n \in \mathbb{N} : L(n;\varepsilon) < m\}$. Suppose

$$n \geq \max\left\{12\ln(6/\varepsilon), 1 + L^{-1}\left(\Lambda_p(\varepsilon, f, \mathcal{P}); \varepsilon\right)\right\}.$$

Consider running Meta-Algorithm 0 with $\mathcal{A}_p$ and n as arguments, while f is the target function and $\mathcal{P}$ is the data distribution. Let V and $\mathcal{L}$ be as in Meta-Algorithm 0, and let $\hat{h}_n = \mathcal{A}_p(\mathcal{L})$ denote the classifier returned at the end.

By Lemma 29, on the event $H_{\lfloor n/2 \rfloor}(\varepsilon/2)$, $V \subseteq \mathrm{B}(f, \phi(\lfloor n/2 \rfloor; \varepsilon/2))$, so that $\mathcal{P}(\mathrm{DIS}(V)) \leq \Delta_n(\varepsilon)$. Letting $\mathcal{U} = \{X_{\lfloor n/2 \rfloor + 1}, \ldots, X_{\lfloor n/2 \rfloor + \lfloor n/(4\hat{\Delta}) \rfloor}\}$, by Lemma 30, on $H_{\lfloor n/2 \rfloor}(\varepsilon/2) \cap J_n$ we have

$$\lfloor n / \max\{32/n, 16\Delta_n(\varepsilon)\} \rfloor \leq |\mathcal{U}| \leq \lfloor n / \max\{4\mathcal{P}(\mathrm{DIS}(V)), 16/n\} \rfloor. \tag{7}$$

By a Chernoff bound, for an event K_n with $\mathbb{P}(K_n) \geq 1 - \exp\{-n/12\}$, on $H_{\lfloor n/2 \rfloor}(\varepsilon/2) \cap J_n \cap K_n$, $|\mathcal{U} \cap \mathrm{DIS}(V)| \leq 2\mathcal{P}(\mathrm{DIS}(V)) \cdot \lfloor n / \max\{4\mathcal{P}(\mathrm{DIS}(V)), 16/n\} \rfloor \leq \lceil n/2 \rceil$. Defining the event $G_n(\varepsilon) = H_{\lfloor n/2 \rfloor}(\varepsilon/2) \cap J_n \cap K_n$, we see that on $G_n(\varepsilon)$, every time $X_m \in \mathrm{DIS}(V)$ in Step 5 of Meta-Algorithm 0, we have $t < n$; therefore, since $f \in V$ implies that the inferred labels in Step 6 are correct as well, we have that on $G_n(\varepsilon)$,

$$\forall (x, \hat{y}) \in \mathcal{L}, \hat{y} = f(x). \tag{8}$$

Noting that

$$\mathbb{P}(G_n(\varepsilon)^c) \leq \mathbb{P}\left(H_{\lfloor n/2 \rfloor}(\varepsilon/2)^c\right) + \mathbb{P}(J_n^c) + \mathbb{P}(K_n^c) \leq \varepsilon/2 + 2 \cdot \exp\{-n/4\} + \exp\{-n/12\} \leq \varepsilon,$$

we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\right] \\
&\leq \mathbb{E}\left[\mathbb{1}_{G_n(\varepsilon)} \mathbb{1}\left[|\mathcal{L}| \geq \Lambda_p(\varepsilon, f, \mathcal{P})\right] \mathrm{er}\left(\hat{h}_n\right)\right] + \mathbb{P}\left(G_n(\varepsilon) \cap \{|\mathcal{L}| < \Lambda_p(\varepsilon, f, \mathcal{P})\}\right) + \mathbb{P}\left(G_n(\varepsilon)^c\right) \\
&\leq \mathbb{E}\left[\mathbb{1}_{G_n(\varepsilon)} \mathbb{1}\left[|\mathcal{L}| \geq \Lambda_p(\varepsilon, f, \mathcal{P})\right] \mathrm{er}\left(\mathcal{A}_p(\mathcal{L})\right)\right] + \mathbb{P}\left(G_n(\varepsilon) \cap \{|\mathcal{L}| < \Lambda_p(\varepsilon, f, \mathcal{P})\}\right) + \varepsilon.
\end{aligned} \tag{9}$$

On $G_n(\varepsilon)$, (7) implies $|\mathcal{L}| \geq L(n;\varepsilon)$, and we chose n large enough so that $L(n;\varepsilon) \geq \Lambda_p(\varepsilon, f, \mathcal{P})$. Thus, the second term in (9) is zero, and we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\right] &\leq \mathbb{E}\left[\mathbb{1}_{G_n(\varepsilon)} \mathbb{1}\left[|\mathcal{L}| \geq \Lambda_p(\varepsilon, f, \mathcal{P})\right] \mathrm{er}\left(\mathcal{A}_p\left(\mathcal{L}\right)\right)\right] + \varepsilon \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{G_n(\varepsilon)} \mathrm{er}\left(\mathcal{A}_p\left(\mathcal{L}\right)\right) \middle| |\mathcal{L}|\right] \mathbb{1}\left[|\mathcal{L}| \geq \Lambda_p(\varepsilon, f, \mathcal{P})\right]\right] + \varepsilon.
\end{aligned} \tag{10}$$

For any $\ell \in \mathbb{N}$ with $\mathbb{P}(|\mathcal{L}| = \ell) > 0$, the conditional of $\mathcal{U} | \{|\mathcal{U}| = \ell\}$ is a product distribution $\mathcal{P}^\ell$; that is, the samples in $\mathcal{U}$ are conditionally independent and identically distributed with distribution $\mathcal{P}$, which is the same as the distribution of $\{X_1, X_2, \ldots, X_\ell\}$. Therefore, for any such ℓ with $\ell \geq \Lambda_p(\varepsilon, f, \mathcal{P})$, by (8) we have

$$\mathbb{E}\left[\mathbb{1}_{G_n(\varepsilon)} \operatorname{er}\left(\mathcal{A}_p\left(\mathcal{L}\right)\right) \middle| \{|\mathcal{L}| = \ell\}\right] \leq \mathbb{E}\left[\operatorname{er}\left(\mathcal{A}_p\left(\mathcal{Z}_\ell\right)\right)\right] \leq \varepsilon.$$

In particular, this means (10) is at most 2ε. This implies Meta-Algorithm 0, with $\mathcal{A}_p$ as its argument, achieves a label complexity Λ_a such that

$$\Lambda_a(2\varepsilon, f, \mathcal{P}) \leq \max\left\{12\ln(6/\varepsilon), 1 + L^{-1}\left(\Lambda_p(\varepsilon, f, \mathcal{P}); \varepsilon\right)\right\}.$$

Since $\Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon)) \Rightarrow 12\ln(6/\varepsilon) = o(\Lambda_p(\varepsilon, f, \mathcal{P}))$, it remains only to show that $L^{-1}\left(\Lambda_p(\varepsilon, f, \mathcal{P}); \varepsilon\right) = o(\Lambda_p(\varepsilon, f, \mathcal{P}))$. Note that $\forall \varepsilon \in (0,1)$, $L(1;\varepsilon) = 0$ and $L(n;\varepsilon)$ is diverging in n. Furthermore, by the assumption $\mathcal{P}(\partial_{\mathbb{C},\mathcal{P}} f) = 0$, we know that for any $N(\varepsilon) = \omega(\log(1/\varepsilon))$, we have $\Delta_{N(\varepsilon)}(\varepsilon) = o(1)$ (by continuity of probability measures), which implies $L(N(\varepsilon);\varepsilon) = \omega(N(\varepsilon))$. Thus, since $\Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon))$, Lemma 31 implies $L^{-1}\left(\Lambda_p(\varepsilon, f, \mathcal{P}); \varepsilon\right) = o\left(\Lambda_p(\varepsilon, f, \mathcal{P})\right)$, as desired. ∎

Lemma 33 *For any VC class $\mathbb{C}$, target function $f \in \mathbb{C}$, and distribution $\mathcal{P}$, if $\mathcal{P}(\partial_{\mathbb{C},\mathcal{P}} f) > 0$, then there exists a passive learning algorithm $\mathcal{A}_p$ achieving a label complexity Λ_p such that $(f, \mathcal{P}) \in$* Nontrivial(Λ_p)*, and for any label complexity Λ_a achieved by running Meta-Algorithm 0 with $\mathcal{A}_p$ as its argument, and any constant $c \in (0, \infty)$,*

$$\Lambda_a(c\varepsilon, f, \mathcal{P}) \neq o(\Lambda_p(\varepsilon, f, \mathcal{P})).$$

Proof The proof can be broken down into three essential claims. First, it follows from Lemma 35 below that, on an event H' of probability one, $\mathcal{P}(\partial_V f) \geq \mathcal{P}(\partial_{\mathbb{C}} f)$; since $\mathcal{P}(\mathrm{DIS}(V)) \geq \mathcal{P}(\partial_V f)$, we have $\mathcal{P}(\mathrm{DIS}(V)) \geq \mathcal{P}(\partial_{\mathbb{C}} f)$ on H'.

The second claim is that on $H' \cap J_n$, $|\mathcal{L}| = O(n)$. This follows from Lemma 30 and our first claim by noting that, on $H' \cap J_n$, $|\mathcal{L}| = \lfloor n/(4\hat{\Delta}) \rfloor \leq n/(4\mathcal{P}(\mathrm{DIS}(V))) \leq n/(4\mathcal{P}(\partial_{\mathbb{C}} f))$.

Finally, we construct a passive algorithm $\mathcal{A}_p$ whose label complexity is not significantly improved when $|\mathcal{L}| = O(n)$. There is a fairly obvious randomized $\mathcal{A}_p$ with this property (simply returning $-f$ with probability $1/|\mathcal{L}|$, and otherwise f); however, we can even satisfy the property with a deterministic $\mathcal{A}_p$, as follows. Let $\mathcal{H}_f = \{h_i\}_{i=1}^{\infty}$ be any sequence of classifiers (not necessarily in $\mathbb{C}$) with $0 < \mathcal{P}(x : h_i(x) \neq f(x))$ strictly decreasing to 0, (say with $h_1 = -f$). We know such a sequence must exist since $\mathcal{P}(\partial_{\mathbb{C}} f) > 0$. Now define, for nonempty S,

$$\mathcal{A}_p(S) = \operatorname*{argmin}_{h_i \in \mathcal{H}_f} \mathcal{P}(x : h_i(x) \neq f(x)) + 2\mathbb{1}_{[0,1/|S|)}(\mathcal{P}(x : h_i(x) \neq f(x))).$$

$\mathcal{A}_p$ is constructed so that, in the special case that this particular f is the target function and this particular $\mathcal{P}$ is the data distribution, $\mathcal{A}_p(S)$ returns the $h_i \in \mathcal{H}_f$ with minimal $\operatorname{er}(h_i)$ such that $\operatorname{er}(h_i) \geq 1/|S|$. For completeness, let $\mathcal{A}_p(\emptyset) = h_1$. Define $\varepsilon_i = \operatorname{er}(h_i) = \mathcal{P}(x : h_i(x) \neq f(x))$.

Now let $\hat{h}_n$ be the returned classifier from running Meta-Algorithm 0 with $\mathcal{A}_p$ and n as inputs, let Λ_p be the (minimal) label complexity achieved by $\mathcal{A}_p$, and let Λ_a be the (minimal) label complexity achieved by Meta-Algorithm 0 with $\mathcal{A}_p$ as input. Take any $c \in (0, \infty)$, and i sufficiently large so that $\varepsilon_{i-1} < 1/2$. Then we know that for any $\varepsilon \in [\varepsilon_i, \varepsilon_{i-1})$, $\Lambda_p(\varepsilon, f, \mathcal{P}) = \lceil 1/\varepsilon_i \rceil$. In particular,

$\Lambda_p(\varepsilon, f, \mathcal{P}) \geq 1/\varepsilon$, so that $(f, \mathcal{P}) \in \text{Nontrivial}(\Lambda_p)$. Also, by Markov's inequality and the above results on $|\mathcal{L}|$,

$$\begin{aligned}\mathbb{E}[\text{er}(\hat{h}_n)] \geq \mathbb{E}\left[\frac{1}{|\mathcal{L}|}\right] &\geq \frac{4\mathcal{P}(\partial_{\mathbb{C}} f)}{n}\mathbb{P}\left(\frac{1}{|\mathcal{L}|} > \frac{4\mathcal{P}(\partial_{\mathbb{C}} f)}{n}\right) \\ &\geq \frac{4\mathcal{P}(\partial_{\mathbb{C}} f)}{n}\mathbb{P}(H' \cap J_n) \geq \frac{4\mathcal{P}(\partial_{\mathbb{C}} f)}{n}\left(1 - 2 \cdot \exp\{-n/4\}\right).\end{aligned}$$

This implies that for $4\ln(4) < n < \frac{2\mathcal{P}(\partial_{\mathbb{C}} f)}{c\varepsilon_i}$, we have $\mathbb{E}\left[\text{er}(\hat{h}_n)\right] > c\varepsilon_i$, so that for all sufficiently large i,

$$\Lambda_a(c\varepsilon_i, f, \mathcal{P}) \geq \frac{2\mathcal{P}(\partial_{\mathbb{C}} f)}{c\varepsilon_i} \geq \frac{\mathcal{P}(\partial_{\mathbb{C}} f)}{c}\left\lceil\frac{1}{\varepsilon_i}\right\rceil = \frac{\mathcal{P}(\partial_{\mathbb{C}} f)}{c}\Lambda_p(\varepsilon_i, f, \mathcal{P}).$$

Since this happens for all sufficiently large i, and thus for arbitrarily small ε_i values, we have

$$\Lambda_a(c\varepsilon, f, \mathcal{P}) \neq o\left(\Lambda_p(\varepsilon, f, \mathcal{P})\right).$$

■

Proof [Theorem 5] Theorem 5 now follows directly from Lemmas 32 and 33, corresponding to the "if" and "only if" parts of the claim, respectively. ■

Appendix B. Proofs Related to Section 4: Basic Activizer

In this section, we provide detailed definitions, lemmas and proofs related to Meta-Algorithm 1.

In fact, we will develop slightly more general results here. Specifically, we fix an arbitrary constant $\gamma \in (0,1)$, and will prove the result for a family of meta-algorithms parameterized by the value γ, used as the threshold in Steps 3 and 6 of Meta-Algorithm 1, which were set to $1/2$ above to simplify the algorithm. Thus, setting $\gamma = 1/2$ in the statements below will give the stated theorem.

Throughout this section, we will assume $\mathbb{C}$ is a VC class with VC dimension d, and let $\mathcal{P}$ denote the (arbitrary) marginal distribution of X_i ($\forall i$). We also fix an arbitrary classifier $f \in \text{cl}(\mathbb{C})$, where (as in Section 6) $\text{cl}(\mathbb{C}) = \{h : \forall r > 0, \text{B}(h,r) \neq \emptyset\}$ denotes the closure of $\mathbb{C}$. In the present context, f corresponds to the target function when running Meta-Algorithm 1. Thus, we will study the behavior of Meta-Algorithm 1 for this fixed f and $\mathcal{P}$; since they are chosen arbitrarily, to establish Theorem 6 it will suffice to prove that for any passive $\mathcal{A}_p$, Meta-Algorithm 1 with $\mathcal{A}_p$ as input achieves superior label complexity compared to $\mathcal{A}_p$ for this f and $\mathcal{P}$. In fact, because here we only assume $f \in \text{cl}(\mathbb{C})$ (rather than $f \in \mathbb{C}$), we actually end up proving a slightly more general version of Theorem 6. But more importantly, this relaxation to $\text{cl}(\mathbb{C})$ will also make the lemmas developed below more useful for subsequent proofs: namely, those in Appendix E.2. For this same reason, many of the lemmas of this section are substantially more general than is necessary for the proof of Theorem 6; the more general versions will be used in the proofs of results in later sections.

For any $m \in \mathbb{N}$, we define $V_m^\star = \{h \in \mathbb{C} : \forall i \leq m, h(X_i) = f(X_i)\}$. Additionally, for $\mathcal{H} \subseteq \mathbb{C}$, and an integer $k \geq 0$, we will adopt the notation

$$\begin{aligned}\mathcal{S}^k(\mathcal{H}) &= \left\{S \in \mathcal{X}^k : \mathcal{H} \text{ shatters } S\right\}, \\ \bar{\mathcal{S}}^k(\mathcal{H}) &= \mathcal{X}^k \setminus \mathcal{S}^k(\mathcal{H}),\end{aligned}$$

and as in Section 5, we define the k-dimensional shatter core of f with respect to $\mathcal{H}$ (and $\mathcal{P}$) as

$$\partial_{\mathcal{H}}^{k} f = \lim_{r \to 0} \mathcal{S}^{k}\left(\mathrm{B}_{\mathcal{H}}(f,r)\right),$$

and further define

$$\bar{\partial}_{\mathcal{H}}^{k} f = \mathcal{X}^{k} \setminus \partial_{\mathcal{H}}^{k} f.$$

Also as in Section 5, define

$$\tilde{d}_f = \min\left\{k \in \mathbb{N} : \mathcal{P}^{k}\left(\partial_{\mathbb{C}}^{k} f\right) = 0\right\}.$$

For convenience, we also define the abbreviation

$$\tilde{\delta}_f = \mathcal{P}^{\tilde{d}_f - 1}\left(\partial_{\mathbb{C}}^{\tilde{d}_f - 1} f\right).$$

Also, recall that we are using the convention that $\mathcal{X}^0 = \{\varnothing\}$, $\mathcal{P}^0(\mathcal{X}^0) = 1$, and we say a set of classifiers $\mathcal{H}$ shatters $\varnothing$ iff $\mathcal{H} \neq \{\}$. In particular, $\mathcal{S}^0(\mathcal{H}) \neq \{\}$ iff $\mathcal{H} \neq \{\}$, and $\partial_{\mathcal{H}}^{0} f \neq \{\}$ iff $\inf_{h \in \mathcal{H}} \mathcal{P}(x : h(x) \neq f(x)) = 0$. For any measurable sets $S_1, S_2 \subseteq \mathcal{X}^k$ with $\mathcal{P}^k(S_2) > 0$, as usual we define $\mathcal{P}^k(S_1|S_2) = \mathcal{P}^k(S_1 \cap S_2)/\mathcal{P}^k(S_2)$; in the situation where $\mathcal{P}^k(S_2) = 0$, it will be convenient to define $\mathcal{P}^k(S_1|S_2) = 0$. We use the definition of $\operatorname{er}(h)$ from above, and additionally define the *conditional* error rate $\operatorname{er}(h|S) = \mathcal{P}(\{x : h(x) \neq f(x)\}|S)$ for any measurable $S \subseteq \mathcal{X}$. We also adopt the usual short-hand for equalities and inequalities involving conditional expectations and probabilities given random variables, wherein for instance, we write $\mathbb{E}[X|Y] = Z$ to mean that there is a version of $\mathbb{E}[X|Y]$ that is everywhere equal to Z, so that in particular, any version of $\mathbb{E}[X|Y]$ equals Z almost everywhere (see, e.g., Ash and Doléans-Dade, 2000).

B.1 Definition of Estimators for Meta-Algorithm 1

While the estimated probabilities used in Meta-Algorithm 1 can be defined in a variety of ways to make it a universal activizer for $\mathbb{C}$, in the statement of Theorem 6 above and proof thereof below, we take the following specific definitions. After the definition, we discuss alternative possibilities.

Though it is a slight twist on the formal model, it will greatly simplify our discussion below to suppose we have access to two independent sequences of i.i.d. unlabeled examples $W_1 = \{w_1, w_2, \ldots\}$ and $W_2 = \{w'_1, w'_2, \ldots\}$, also independent from the main sequence $\{X_1, X_2, \ldots\}$, with $w_i, w'_i \sim \mathcal{P}$. Since the data sequence $\{X_1, X_2, \ldots\}$ is i.i.d., this is distributionally equivalent to supposing we partition the data sequence in a preprocessing step, into three subsequences, alternatingly assigning each data point to either $\mathcal{Z}'_X$, W_1, or W_2. Then, if we suppose $\mathcal{Z}'_X = \{X'_1, X'_2, \ldots\}$, and we replace all references to X_i with X'_i in the algorithms and results, we obtain the equivalent statements holding for the model as originally stated. Thus, supposing the existence of these W_i sequences simply serves to simplify notation, and does not represent a further assumption on top of the previously stated framework.

For each $k \geq 2$, we partition W_2 into subsets of size $k-1$, as follows. For $i \in \mathbb{N}$, let

$$S_i^{(k)} = \{w'_{1+(i-1)(k-1)}, \ldots, w'_{i(k-1)}\}.$$

We define the $\hat{P}_m$ estimators in terms of three types of functions, defined below. For any $\mathcal{H} \subseteq \mathbb{C}$, $x \in \mathcal{X}$, $y \in \{-1,+1\}$, $m \in \mathbb{N}$, we define

$$\hat{P}_m\left(S \in \mathcal{X}^{k-1} : \mathcal{H} \text{ shatters } S \cup \{x\} | \mathcal{H} \text{ shatters } S\right) = \hat{\Delta}_m^{(k)}(x, W_2, \mathcal{H}), \tag{11}$$

$$\hat{P}_m\left(S \in \mathcal{X}^{k-1} : \mathcal{H}[(x,-y)] \text{ does not shatter } S | \mathcal{H} \text{ shatters } S\right) = \hat{\Gamma}_m^{(k)}(x, y, W_2, \mathcal{H}), \tag{12}$$

$$\hat{P}_m\left(x : \hat{P}\left(S \in \mathcal{X}^{k-1} : \mathcal{H} \text{ shatters } S \cup \{x\} | \mathcal{H} \text{ shatters } S\right) \geq \gamma\right) = \hat{\Delta}_m^{(k)}(W_1, W_2, \mathcal{H}). \tag{13}$$

The quantities $\hat{\Delta}_m^{(k)}(x, W_2, \mathcal{H})$, $\hat{\Gamma}_m^{(k)}(x, y, W_2, \mathcal{H})$, and $\hat{\Delta}_m^{(k)}(W_1, W_2, \mathcal{H})$ are specified as follows.

For $k = 1$, $\hat{\Gamma}_m^{(1)}(x, y, W_2, \mathcal{H})$ is simply an indicator for whether every $h \in \mathcal{H}$ has $h(x) = y$, while $\hat{\Delta}_m^{(1)}(x, W_2, \mathcal{H})$ is an indicator for whether $x \in \mathrm{DIS}(\mathcal{H})$. Formally, they are defined as follows.

$$\hat{\Gamma}_m^{(1)}(x, y, W_2, \mathcal{H}) = \mathbb{1}_{\bigcap\limits_{h \in \mathcal{H}} \{h(x)\}}(y).$$

$$\hat{\Delta}_m^{(1)}(x, W_2, \mathcal{H}) = \mathbb{1}_{\mathrm{DIS}(\mathcal{H})}(x).$$

For $k \geq 2$, we first define

$$M_m^{(k)}(\mathcal{H}) = \max\left\{1, \sum_{i=1}^{m^3} \mathbb{1}_{\mathcal{S}^{k-1}(\mathcal{H})}\left(S_i^{(k)}\right)\right\}.$$

Then we take the following definitions for $\hat{\Gamma}^{(k)}$ and $\hat{\Delta}^{(k)}$.

$$\hat{\Gamma}_m^{(k)}(x, y, W_2, \mathcal{H}) = \frac{1}{M_m^{(k)}(\mathcal{H})} \sum_{i=1}^{m^3} \mathbb{1}_{\bar{\mathcal{S}}^{k-1}(\mathcal{H}[(x,-y)])}\left(S_i^{(k)}\right) \mathbb{1}_{\mathcal{S}^{k-1}(\mathcal{H})}\left(S_i^{(k)}\right). \tag{14}$$

$$\hat{\Delta}_m^{(k)}(x, W_2, \mathcal{H}) = \frac{1}{M_m^{(k)}(\mathcal{H})} \sum_{i=1}^{m^3} \mathbb{1}_{\mathcal{S}^{k}(\mathcal{H})}\left(S_i^{(k)} \cup \{x\}\right). \tag{15}$$

For the remaining estimator, for any k we generally define

$$\hat{\Delta}_m^{(k)}(W_1, W_2, \mathcal{H}) = \frac{2}{m} + \frac{1}{m^3} \sum_{i=1}^{m^3} \mathbb{1}_{[\gamma/4, \infty)}\left(\hat{\Delta}_m^{(k)}(w_i, W_2, \mathcal{H})\right).$$

The above definitions will be used in the proofs below. However, there are certainly viable alternative definitions one can consider, some of which may have interesting theoretical properties. In general, one has the same sorts of trade-offs present whenever estimating a conditional probability. For instance, we could replace "m^3" in (14) and (15) by $\min\left\{\ell \in \mathbb{N} : M_\ell^{(k)}(\mathcal{H}) = m^3\right\}$, and then normalize by m^3 instead of $M_m^{(k)}(\mathcal{H})$; this would give us m^3 samples from the conditional distribution with which to estimate the conditional probability. The advantages of this approach would be its simplicity or elegance, and possibly some improvement in the constant factors in the label complexity bounds below. On the other hand, the drawback of this alternative definition would be that we do not know a priori how many unlabeled samples we will need to process in order to calculate it;

indeed, for some values of k and $\mathcal{H}$, we expect $\mathcal{P}^{k-1}\left(\mathcal{S}^{k-1}(\mathcal{H})\right) = 0$, so that $M_\ell^{(k)}(\mathcal{H})$ is bounded, and we might technically need to examine the entire sequence to distinguish this case from the case of very small $\mathcal{P}^{k-1}\left(\mathcal{S}^{k-1}(\mathcal{H})\right)$. Of course, these practical issues can be addressed with small modifications, but only at the expense of complicating the analysis, thus losing the elegance factor. For these reasons, we have opted for the slightly looser and less elegant, but more practical, definitions above in (14) and (15).

B.2 Proof of Theorem 6

At a high level, the structure of the proof is the following. The primary components of the proof are three lemmas: 34, 37, and 38. Setting aside, for a moment, the fact that we are using the $\hat{P}_m$ estimators rather than the actual probability values they estimate, Lemma 38 indicates that the number of data points in $\mathcal{L}_{\tilde{d}_f}$ grows superlinearly in n (the number of label requests), while Lemma 37 guarantees that the labels of these points are correct, and Lemma 34 tells us that the classifier returned in the end is never much worse than $\mathcal{A}_p(\mathcal{L}_{\tilde{d}_f})$. These three factors combine to prove the result. The rest of the proof is composed of supporting lemmas and details regarding the $\hat{P}_m$ estimators. Specifically, Lemmas 35 and 36 serve a supporting role, with the purpose of showing that the set of V-shatterable k-tuples converges to the k-dimensional shatter core (up to probability-zero differences). The other lemmas below (39–45) are needed primarily to extend the above basic idea to the actual scenario where the $\hat{P}_m$ estimators are used as surrogates for the probability values. Additionally, a sub-case of Lemma 45 is needed in order to guarantee the label request budget will not be reached prematurely. Again, in many cases we prove a more general lemma than is required for its use in the proof of Theorem 6; these more general results will be needed in subsequent proofs: namely, in the proofs of Theorem 16 and Lemma 26.

We begin with a lemma concerning the ActiveSelect subroutine.

Lemma 34 *For any $k^*, M, N \in \mathbb{N}$ with $k^* \leq N$, and N classifiers $\{h_1, h_2, \ldots, h_N\}$ (themselves possibly random variables independent from $\{X_M, X_{M+1}, \ldots\}$), a call to* ActiveSelect($\{h_1, h_2, \ldots, h_N\}$, m, $\{X_M, X_{M+1}, \ldots\}$) *makes at most m label requests, and if $h_{\hat{k}}$ is the classifier it returns, then with probability at least $1 - eN \cdot \exp\{-m/(72k^*N\ln(eN))\}$, we have* $\mathrm{er}(h_{\hat{k}}) \leq 2\,\mathrm{er}(h_{k^*})$.

Proof This proof is essentially identical to a similar result of Balcan, Hanneke, and Vaughan (2010), but is included here for completeness.

Let $M_k = \left\lfloor \frac{m}{k(N-k)\ln(eN)} \right\rfloor$. First note that the total number of label requests in ActiveSelect is at most m, since summing up the sizes of the batches of label requests made in all executions of Step 2 yields

$$\sum_{j=1}^{N-1} \sum_{k=j+1}^{N} \left\lfloor \frac{m}{j(N-j)\ln(eN)} \right\rfloor \leq \sum_{j=1}^{N-1} \frac{m}{j\ln(eN)} \leq m.$$

Let $k^{**} = \mathrm{argmin}_{k \in \{1,\ldots,k^*\}} \mathrm{er}(h_k)$. For any $j \in \{1, 2, \ldots, k^{**}-1\}$ with $\mathcal{P}(x : h_j(x) \neq h_{k^{**}}(x)) > 0$, the law of large numbers implies that with probability one $|\{X_M, X_{M+1}, \ldots\} \cap \{x : h_j(x) \neq h_{k^{**}}(x)\}| \geq M_j$, and since $\mathrm{er}(h_{k^{**}}|\{x : h_j(x) \neq h_{k^{**}}(x)\}) \leq 1/2$, Hoeffding's inequality implies that

$$\mathbb{P}\left(m_{k^{**}j} > 7/12\right) \leq \exp\left\{-M_j/72\right\} \leq \exp\left\{1 - m/(72k^*N\ln(eN))\right\}.$$

A union bound implies

$$\mathbb{P}\left(\max_{j<k^{**}} m_{k^{**}j} > 7/12\right) \leq k^{**} \cdot \exp\left\{1 - m/\left(72k^{*}N\ln(eN)\right)\right\}.$$

In particular, note that when $\max_{j<k^{**}} m_{k^{**}j} \leq 7/12$, we must have $\hat{k} \geq k^{**}$.

Now suppose $j \in \{k^{**}+1,\ldots,N\}$ has $\mathrm{er}(h_j) > 2\,\mathrm{er}(h_{k^{**}})$. In particular, this implies $\mathrm{er}(h_j|\{x : h_{k^{**}}(x) \neq h_j(x)\}) > 2/3$ and $\mathcal{P}(x : h_j(x) \neq h_{k^{**}}(x)) > 0$, which again means (with probability one) $|\{X_M, X_{M+1},\ldots\} \cap \{x : h_j(x) \neq h_{k^{**}}(x)\}| \geq M_{k^{**}}$. By Hoeffding's inequality, we have that

$$\mathbb{P}\left(m_{jk^{**}} \leq 7/12\right) \leq \exp\left\{-M_{k^{**}}/72\right\} \leq \exp\left\{1 - m/\left(72k^{*}N\ln(eN)\right)\right\}.$$

By a union bound, we have that

$$\mathbb{P}\left(\exists j > k^{**} : \mathrm{er}(h_j) > 2\,\mathrm{er}(h_{k^{**}}) \text{ and } m_{jk^{**}} \leq 7/12\right) \leq (N - k^{**}) \cdot \exp\left\{1 - m/\left(72k^{*}N\ln(eN)\right)\right\}.$$

In particular, when $\hat{k} \geq k^{**}$, and $m_{jk^{**}} > 7/12$ for all $j > k^{**}$ with $\mathrm{er}(h_j) > 2\,\mathrm{er}(h_{k^{**}})$, it must be true that $\mathrm{er}(h_{\hat{k}}) \leq 2\,\mathrm{er}(h_{k^{**}}) \leq 2\,\mathrm{er}(h_{k^{*}})$.

So, by a union bound, with probability $\geq 1 - eN \cdot \exp\left\{-m/\left(72k^{*}N\ln(eN)\right)\right\}$, the $\hat{k}$ chosen by ActiveSelect has $\mathrm{er}(h_{\hat{k}}) \leq 2\,\mathrm{er}(h_{k^{*}})$. ■

The next two lemmas describe the limiting behavior of $\mathcal{S}^k(V_m^{\star})$. In particular, we see that its limiting value is precisely $\partial_{\mathbb{C}}^k f$ (up to zero-probability differences). Lemma 35 establishes that $\mathcal{S}^k(V_m^{\star})$ does not decrease below $\partial_{\mathbb{C}}^k f$ (except for a zero-probability set), and Lemma 36 establishes that its limit is not larger than $\partial_{\mathbb{C}}^k f$ (again, except for a zero-probability set).

Lemma 35 *There is an event H' with $\mathbb{P}(H') = 1$ such that on H', $\forall m \in \mathbb{N}$, $\forall k \in \{0,\ldots,\tilde{d}_f - 1\}$, for any $\mathcal{H}$ with $V_m^{\star} \subseteq \mathcal{H} \subseteq \mathbb{C}$,*

$$\mathcal{P}^k\left(\mathcal{S}^k(\mathcal{H})\middle|\partial_{\mathbb{C}}^k f\right) = \mathcal{P}^k\left(\partial_{\mathcal{H}}^k f\middle|\partial_{\mathbb{C}}^k f\right) = 1,$$

and

$$\forall i \in \mathbb{N}, \mathbb{1}_{\partial_{\mathcal{H}}^k f}\left(S_i^{(k+1)}\right) = \mathbb{1}_{\partial_{\mathbb{C}}^k f}\left(S_i^{(k+1)}\right).$$

Also, on H', every such $\mathcal{H}$ has $\mathcal{P}^k\left(\partial_{\mathcal{H}}^k f\right) = \mathcal{P}^k\left(\partial_{\mathbb{C}}^k f\right)$, and $M_\ell^{(k)}(\mathcal{H}) \to \infty$ as $\ell \to \infty$.

Proof We will show the first claim for the set $V_m^{\star}$, and the result will then hold for $\mathcal{H}$ by monotonicity. In particular, we will show this for any fixed $k \in \{0,\ldots,\tilde{d}_f - 1\}$ and $m \in \mathbb{N}$, and the existence of H' then holds by a union bound. Fix any set $S \in \partial_{\mathbb{C}}^k f$. Suppose $\mathrm{B}_{V_m^{\star}}(f,r)$ does not shatter S for some $r > 0$. There is an infinite sequence of sets $\{\{h_1^{(i)}, h_2^{(i)},\ldots,h_{2^k}^{(i)}\}\}_i$ with $\forall j \leq 2^k$, $\mathcal{P}(x : h_j^{(i)}(x) \neq f(x)) \downarrow 0$, such that each $\{h_1^{(i)},\ldots,h_{2^k}^{(i)}\} \subseteq \mathrm{B}(f,r)$ and shatters S. Since $\mathrm{B}_{V_m^{\star}}(f,r)$ does not shatter S,

$$1 = \inf_i \mathbb{1}\left[\exists j : h_j^{(i)} \notin \mathrm{B}_{V_m^{\star}}(f,r)\right] = \inf_i \mathbb{1}\left[\exists j \leq 2^k, \ell \leq m : h_j^{(i)}(X_\ell) \neq f(X_\ell)\right].$$

But

$$\mathbb{P}\left(\inf_i \mathbb{1}\left[\exists j \leq 2^k, \ell \leq m : h_j^{(i)}(X_\ell) \neq f(X_\ell)\right] = 1\right) \leq \inf_i \mathbb{P}\left(\exists j \leq 2^k, \ell \leq m : h_j^{(i)}(X_\ell) \neq f(X_\ell)\right)$$
$$\leq \lim_{i\to\infty} \sum_{j\leq 2^k} m\mathcal{P}\left(x : h_j^{(i)}(x) \neq f(x)\right) = \sum_{j\leq 2^k} m \lim_{i\to\infty} \mathcal{P}\left(x : h_j^{(i)}(x) \neq f(x)\right) = 0,$$

where the second inequality follows by a union bound. Therefore, $\forall r > 0$, $\mathbb{P}\left(S \notin \mathcal{S}^k\left(\mathrm{B}_{V_m^\star}(f,r)\right)\right) = 0$. Furthermore, since $\bar{\mathcal{S}}^k\left(\mathrm{B}_{V_m^\star}(f,r)\right)$ is monotonic in r, the dominated convergence theorem gives us that

$$\mathbb{P}\left(S \notin \partial^k_{V_m^\star} f\right) = \mathbb{E}\left[\lim_{r\to 0} \mathbb{1}_{\bar{\mathcal{S}}^k(\mathrm{B}_{V_m^\star}(f,r))}(S)\right] = \lim_{r\to 0} \mathbb{P}\left(S \notin \mathcal{S}^k\left(\mathrm{B}_{V_m^\star}(f,r)\right)\right) = 0.$$

This implies that (letting $\mathbf{S} \sim \mathcal{P}^k$ be independent from $V_m^\star$)

$$\begin{aligned}
\mathbb{P}\left(\mathcal{P}^k\left(\bar{\partial}^k_{V_m^\star} f \middle| \partial^k_{\mathbb{C}} f\right) > 0\right) &= \mathbb{P}\left(\mathcal{P}^k\left(\bar{\partial}^k_{V_m^\star} f \cap \partial^k_{\mathbb{C}} f\right) > 0\right) \\
&= \lim_{\xi\to 0} \mathbb{P}\left(\mathcal{P}^k\left(\bar{\partial}^k_{V_m^\star} f \cap \partial^k_{\mathbb{C}} f\right) > \xi\right) \\
&\leq \lim_{\xi\to 0} \frac{1}{\xi}\mathbb{E}\left[\mathcal{P}^k\left(\bar{\partial}^k_{V_m^\star} f \cap \partial^k_{\mathbb{C}} f\right)\right] && \text{(Markov)} \\
&= \lim_{\xi\to 0} \frac{1}{\xi}\mathbb{E}\left[\mathbb{1}_{\partial^k_{\mathbb{C}} f}(\mathbf{S})\mathbb{P}\left(\mathbf{S} \notin \partial^k_{V_m^\star} f \middle| \mathbf{S}\right)\right] && \text{(Fubini)} \\
&= \lim_{\xi\to 0} 0 = 0.
\end{aligned}$$

This establishes the first claim for $V_m^\star$, on an event of probability 1, and monotonicity extends the claim to any $\mathcal{H} \supseteq V_m^\star$. Also note that, on this event,

$$\mathcal{P}^k\left(\partial^k_{\mathcal{H}} f\right) \geq \mathcal{P}^k\left(\partial^k_{\mathcal{H}} f \cap \partial^k_{\mathbb{C}} f\right) = \mathcal{P}^k\left(\partial^k_{\mathcal{H}} f \middle| \partial^k_{\mathbb{C}} f\right)\mathcal{P}^k\left(\partial^k_{\mathbb{C}} f\right) = \mathcal{P}^k\left(\partial^k_{\mathbb{C}} f\right),$$

where the last equality follows from the first claim. Noting that for $\mathcal{H} \subseteq \mathbb{C}$, $\partial^k_{\mathcal{H}} f \subseteq \partial^k_{\mathbb{C}} f$, we must have

$$\mathcal{P}^k\left(\partial^k_{\mathcal{H}} f\right) = \mathcal{P}^k\left(\partial^k_{\mathbb{C}} f\right).$$

This establishes the third claim. From the first claim, for any given value of $i \in \mathbb{N}$ the second claim holds for $S_i^{(k+1)}$ (with $\mathcal{H} = V_m^\star$) on an additional event of probability 1; taking a union bound over all $i \in \mathbb{N}$ extends this claim to every $S_i^{(k)}$ on an event of probability 1. Monotonicity then implies

$$\mathbb{1}_{\partial^k_{\mathbb{C}} f}\left(S_i^{(k+1)}\right) = \mathbb{1}_{\partial^k_{V_m^\star} f}\left(S_i^{(k+1)}\right) \leq \mathbb{1}_{\partial^k_{\mathcal{H}} f}\left(S_i^{(k+1)}\right) \leq \mathbb{1}_{\partial^k_{\mathbb{C}} f}\left(S_i^{(k+1)}\right),$$

extending the result to general $\mathcal{H}$. Also, as $k < \tilde{d}_f$, we know $\mathcal{P}^k\left(\partial^k_{\mathbb{C}} f\right) > 0$, and since we also know $V_m^\star$ is independent from W_2, the strong law of large numbers implies the final claim (for $V_m^\star$) on an additional event of probability 1; again, monotonicity extends this claim to any $\mathcal{H} \supseteq V_m^\star$. Intersecting the above events over values $m \in \mathbb{N}$ and $k < \tilde{d}_f$ gives the event H', and as each of the above events has probability 1 and there are countably many such events, a union bound implies $\mathbb{P}(H') = 1$. ■

Note that one specific implication of Lemma 35, obtained by taking $k=0$, is that on H', $V_m^\star \neq \emptyset$ (even if $f \in \mathrm{cl}(\mathbb{C}) \setminus \mathbb{C}$). This is because, for $f \in \mathrm{cl}(\mathbb{C})$, we have $\partial_{\mathbb{C}}^0 f = \mathcal{X}^0$ so that $\mathcal{P}^0\left(\partial_{\mathbb{C}}^0 f\right) = 1$, which means $\mathcal{P}^0\left(\partial_{V_m^\star}^0 f\right) = 1$ (on H'), so that we must have $\partial_{V_m^\star}^0 f = \mathcal{X}^0$, which implies $V_m^\star \neq \emptyset$. In particular, this also means $f \in \mathrm{cl}\left(V_m^\star\right)$.

Lemma 36 *There is a monotonic function $q(r) = o(1)$ (as $r \to 0$) such that, on event H', for any $k \in \left\{0,\ldots,\tilde{d}_f - 1\right\}$, $m \in \mathbb{N}$, $r > 0$, and set $\mathcal{H}$ such that $V_m^\star \subseteq \mathcal{H} \subseteq \mathrm{B}(f,r)$,*

$$\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \middle| \mathcal{S}^k(\mathcal{H})\right) \leq q(r).$$

In particular, for $\tau \in \mathbb{N}$ and $\delta > 0$, on $H_\tau(\delta) \cap H'$ (where $H_\tau(\delta)$ is from Lemma 29), every $m \geq \tau$ and $k \in \left\{0,\ldots,\tilde{d}_f - 1\right\}$ has $\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \middle| \mathcal{S}^k\left(V_m^\star\right)\right) \leq q(\phi(\tau;\delta))$.

Proof Fix any $k \in \left\{0,\ldots,\tilde{d}_f - 1\right\}$. By Lemma 35, we know that on event H',

$$\begin{aligned}
\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \middle| \mathcal{S}^k(\mathcal{H})\right) &= \frac{\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \cap \mathcal{S}^k(\mathcal{H})\right)}{\mathcal{P}^k\left(\mathcal{S}^k(\mathcal{H})\right)} \leq \frac{\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \cap \mathcal{S}^k(\mathcal{H})\right)}{\mathcal{P}^k\left(\partial_{\mathcal{H}}^k f\right)} \\
&= \frac{\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \cap \mathcal{S}^k(\mathcal{H})\right)}{\mathcal{P}^k\left(\partial_{\mathbb{C}}^k f\right)} \leq \frac{\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \cap \mathcal{S}^k(\mathrm{B}(f,r))\right)}{\mathcal{P}^k\left(\partial_{\mathbb{C}}^k f\right)}.
\end{aligned}$$

Define $q_k(r)$ as this latter quantity. Since $\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \cap \mathcal{S}^k(\mathrm{B}(f,r))\right)$ is monotonic in r,

$$\lim_{r \to 0} \frac{\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \cap \mathcal{S}^k(\mathrm{B}(f,r))\right)}{\mathcal{P}^k\left(\partial_{\mathbb{C}}^k f\right)} = \frac{\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \cap \lim\limits_{r \to 0} \mathcal{S}^k(\mathrm{B}(f,r))\right)}{\mathcal{P}^k\left(\partial_{\mathbb{C}}^k f\right)} = \frac{\mathcal{P}^k\left(\bar{\partial}_{\mathbb{C}}^k f \cap \partial_{\mathbb{C}}^k f\right)}{\mathcal{P}^k\left(\partial_{\mathbb{C}}^k f\right)} = 0.$$

This proves $q_k(r) = o(1)$. Defining

$$q(r) = \max\left\{q_k(r) : k \in \left\{0,1,\ldots,\tilde{d}_f - 1\right\}\right\} = o(1)$$

completes the proof of the first claim.

For the final claim, simply recall that by Lemma 29, on $H_\tau(\delta)$, every $m \geq \tau$ has $V_m^\star \subseteq V_\tau^\star \subseteq \mathrm{B}(f,\phi(\tau;\delta))$. ■

Lemma 37 *For $\zeta \in (0,1)$, define*

$$r_\zeta = \sup\{r \in (0,1) : q(r) < \zeta\}/2.$$

On H', $\forall k \in \left\{0,\ldots,\tilde{d}_f - 1\right\}$, $\forall \zeta \in (0,1)$, $\forall m \in \mathbb{N}$, for any set $\mathcal{H}$ such that $V_m^\star \subseteq \mathcal{H} \subseteq \mathrm{B}(f,r_\zeta)$,

$$\begin{aligned}
&\mathcal{P}\left(x : \mathcal{P}^k\left(\bar{\mathcal{S}}^k(\mathcal{H}[(x,f(x))]) \middle| \mathcal{S}^k(\mathcal{H})\right) > \zeta\right) \\
&\qquad = \mathcal{P}\left(x : \mathcal{P}^k\left(\bar{\mathcal{S}}^k(\mathcal{H}[(x,f(x))]) \middle| \partial_{\mathcal{H}}^k f\right) > \zeta\right) = 0. \qquad (16)
\end{aligned}$$

In particular, for $\delta \in (0,1)$, defining $\tau(\zeta;\delta) = \min\left\{\tau \in \mathbb{N} : \sup\limits_{m \geq \tau} \phi(m;\delta) \leq r_\zeta\right\}$, $\forall \tau \geq \tau(\zeta;\delta)$, and $\forall m \geq \tau$, on $H_\tau(\delta) \cap H'$, (16) holds for $\mathcal{H} = V_m^\star$.

Proof Fix $k, m, \mathcal{H}$ as described above, and suppose $q = \mathcal{P}^k\left(\bar{\partial}^k_{\mathbb{C}} f | \mathcal{S}^k(\mathcal{H})\right) < \zeta$; by Lemma 36, this happens on H'. Since, $\partial^k_{\mathcal{H}} f \subseteq \mathcal{S}^k(\mathcal{H})$, we have that $\forall x \in \mathcal{X}$,

$$\begin{aligned}\mathcal{P}^k\left(\bar{\mathcal{S}}^k\left(\mathcal{H}[(x,f(x))]\right)\middle|\mathcal{S}^k(\mathcal{H})\right) &= \mathcal{P}^k\left(\bar{\mathcal{S}}^k\left(\mathcal{H}[(x,f(x))]\right)\middle|\partial^k_{\mathcal{H}} f\right)\mathcal{P}^k\left(\partial^k_{\mathcal{H}} f\middle|\mathcal{S}^k(\mathcal{H})\right)\\ &\quad + \mathcal{P}^k\left(\bar{\mathcal{S}}^k\left(\mathcal{H}[(x,f(x))]\right)\middle|\mathcal{S}^k(\mathcal{H})\cap\bar{\partial}^k_{\mathcal{H}} f\right)\mathcal{P}^k\left(\bar{\partial}^k_{\mathcal{H}} f\middle|\mathcal{S}^k(\mathcal{H})\right).\end{aligned}$$

Since all probability values are bounded by 1, we have

$$\mathcal{P}^k\left(\bar{\mathcal{S}}^k\left(\mathcal{H}[(x,f(x))]\right)\middle|\mathcal{S}^k(\mathcal{H})\right) \leq \mathcal{P}^k\left(\bar{\mathcal{S}}^k\left(\mathcal{H}[(x,f(x))]\right)\middle|\partial^k_{\mathcal{H}} f\right) + \mathcal{P}^k\left(\bar{\partial}^k_{\mathcal{H}} f\middle|\mathcal{S}^k(\mathcal{H})\right). \tag{17}$$

Isolating the right-most term in (17), by basic properties of probabilities we have

$$\begin{aligned}&\mathcal{P}^k\left(\bar{\partial}^k_{\mathcal{H}} f\middle|\mathcal{S}^k(\mathcal{H})\right)\\ &= \mathcal{P}^k\left(\bar{\partial}^k_{\mathcal{H}} f\middle|\mathcal{S}^k(\mathcal{H})\cap\bar{\partial}^k_{\mathbb{C}} f\right)\mathcal{P}^k\left(\bar{\partial}^k_{\mathbb{C}} f\middle|\mathcal{S}^k(\mathcal{H})\right) + \mathcal{P}^k\left(\bar{\partial}^k_{\mathcal{H}} f\middle|\mathcal{S}^k(\mathcal{H})\cap\partial^k_{\mathbb{C}} f\right)\mathcal{P}^k\left(\partial^k_{\mathbb{C}} f\middle|\mathcal{S}^k(\mathcal{H})\right)\\ &\leq \mathcal{P}^k\left(\bar{\partial}^k_{\mathbb{C}} f\middle|\mathcal{S}^k(\mathcal{H})\right) + \mathcal{P}^k\left(\bar{\partial}^k_{\mathcal{H}} f\middle|\mathcal{S}^k(\mathcal{H})\cap\partial^k_{\mathbb{C}} f\right).\end{aligned} \tag{18}$$

By assumption, the left term in (18) equals q. Examining the right term in (18), we see that

$$\begin{aligned}\mathcal{P}^k\left(\bar{\partial}^k_{\mathcal{H}} f\middle|\mathcal{S}^k(\mathcal{H})\cap\partial^k_{\mathbb{C}} f\right) &= \mathcal{P}^k\left(\mathcal{S}^k(\mathcal{H})\cap\bar{\partial}^k_{\mathcal{H}} f\middle|\partial^k_{\mathbb{C}} f\right)/\mathcal{P}^k\left(\mathcal{S}^k(\mathcal{H})\middle|\partial^k_{\mathbb{C}} f\right)\\ &\leq \mathcal{P}^k\left(\bar{\partial}^k_{\mathcal{H}} f\middle|\partial^k_{\mathbb{C}} f\right)/\mathcal{P}^k\left(\partial^k_{\mathcal{H}} f\middle|\partial^k_{\mathbb{C}} f\right).\end{aligned} \tag{19}$$

By Lemma 35, on H' the denominator in (19) is 1 and the numerator is 0. Thus, combining this fact with (17) and (18), we have that on H',

$$\mathcal{P}\left(x : \mathcal{P}^k\left(\bar{\mathcal{S}}^k(\mathcal{H}[(x,f(x))])\middle|\mathcal{S}^k(\mathcal{H})\right) > \zeta\right) \leq \mathcal{P}\left(x : \mathcal{P}^k\left(\bar{\mathcal{S}}^k(\mathcal{H}[(x,f(x))])\middle|\partial^k_{\mathcal{H}} f\right) > \zeta - q\right). \tag{20}$$

Note that proving the right side of (20) equals zero will suffice to establish the result, since it upper bounds *both* the first expression of (16) (as just established) *and* the second expression of (16) (by monotonicity of measures). Letting $X \sim \mathcal{P}$ be independent from the other random variables $(\mathcal{Z}, W_1, W_2)$, by Markov's inequality, the right side of (20) is at most

$$\frac{1}{\zeta - q}\mathbb{E}\left[\mathcal{P}^k\left(\bar{\mathcal{S}}^k\left(\mathcal{H}[(X,f(X))]\right)\middle|\partial^k_{\mathcal{H}} f\right)\middle|\mathcal{H}\right] = \frac{\mathbb{E}\left[\mathcal{P}^k\left(\bar{\mathcal{S}}^k\left(\mathcal{H}[(X,f(X))]\right)\cap\partial^k_{\mathcal{H}} f\right)\middle|\mathcal{H}\right]}{(\zeta - q)\mathcal{P}^k\left(\partial^k_{\mathcal{H}} f\right)},$$

and by Fubini's theorem, this is (letting $\mathbf{S} \sim \mathcal{P}^k$ be independent from the other random variables)

$$\frac{\mathbb{E}\left[\mathbb{1}_{\partial^k_{\mathcal{H}} f}(\mathbf{S})\mathcal{P}\left(x : \mathbf{S} \notin \mathcal{S}^k\left(\mathcal{H}[(x,f(x))]\right)\right)\middle|\mathcal{H}\right]}{(\zeta - q)\mathcal{P}^k\left(\partial^k_{\mathcal{H}} f\right)}.$$

Lemma 35 implies this equals

$$\frac{\mathbb{E}\left[\mathbb{1}_{\partial^k_{\mathcal{H}} f}(\mathbf{S})\mathcal{P}\left(x : \mathbf{S} \notin \mathcal{S}^k\left(\mathcal{H}[(x,f(x))]\right)\right)\middle|\mathcal{H}\right]}{(\zeta - q)\mathcal{P}^k\left(\partial^k_{\mathbb{C}} f\right)}. \tag{21}$$

For any fixed $S \in \partial_{\mathcal{H}}^k f$, there is an infinite sequence of sets $\left\{\left\{h_1^{(i)}, h_2^{(i)}, \ldots, h_{2^k}^{(i)}\right\}\right\}_{i\in\mathbb{N}}$ with $\forall j \leq 2^k$, $\mathcal{P}\left(x : h_j^{(i)}(x) \neq f(x)\right) \downarrow 0$, such that each $\left\{h_1^{(i)}, \ldots, h_{2^k}^{(i)}\right\} \subseteq \mathcal{H}$ and shatters S. If $\mathcal{H}[(x, f(x))]$ does not shatter S, then

$$1 = \inf_i \mathbb{1}\left[\exists j : h_j^{(i)} \notin \mathcal{H}[(x, f(x))]\right] = \inf_i \mathbb{1}\left[\exists j : h_j^{(i)}(x) \neq f(x)\right].$$

In particular,

$$\begin{aligned}
&\mathcal{P}\left(x : S \notin \mathcal{S}^k\left(\mathcal{H}[(x, f(x))]\right)\right) \leq \mathcal{P}\left(x : \inf_i \mathbb{1}\left[\exists j : h_j^{(i)}(x) \neq f(x)\right] = 1\right) \\
&= \mathcal{P}\left(\bigcap_i \left\{x : \exists j : h_j^{(i)}(x) \neq f(x)\right\}\right) \leq \inf_i \mathcal{P}\left(x : \exists j \text{ s.t. } h_j^{(i)}(x) \neq f(x)\right) \\
&\leq \lim_{i\to\infty} \sum_{j \leq 2^k} \mathcal{P}\left(x : h_j^{(i)}(x) \neq f(x)\right) = \sum_{j\leq 2^k} \lim_{i\to\infty} \mathcal{P}\left(x : h_j^{(i)}(x) \neq f(x)\right) = 0.
\end{aligned}$$

Thus (21) is zero, which establishes the result.

The final claim is then implied by Lemma 29 and monotonicity of $V_m^\star$ in m: that is, on $H_\tau(\delta)$, $V_m^\star \subseteq V_\tau^\star \subseteq \mathrm{B}(f, \phi(\tau; \delta)) \subseteq \mathrm{B}(f, r_\zeta)$. ∎

Lemma 38 *For any $\zeta \in (0,1)$, there are values $\left\{\Delta_n^{(\zeta)}(\varepsilon) : n \in \mathbb{N}, \varepsilon \in (0,1)\right\}$ such that, for any $n \in \mathbb{N}$ and $\varepsilon > 0$, on event $H_{\lfloor n/3 \rfloor}(\varepsilon/2) \cap H'$, letting $V = V_{\lfloor n/3\rfloor}^\star$,*

$$\mathcal{P}\left(x : \mathcal{P}^{\tilde{d}_f - 1}\left(S \in \mathcal{X}^{\tilde{d}_f - 1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \mathcal{S}^{\tilde{d}_f - 1}(V)\right) \geq \zeta\right) \leq \Delta_n^{(\zeta)}(\varepsilon),$$

and for any $\mathbb{N}$-valued $N(\varepsilon) = \omega(\log(1/\varepsilon))$, $\Delta_{N(\varepsilon)}^{(\zeta)}(\varepsilon) = o(1)$.

Proof Throughout, we suppose the event $H_{\lfloor n/3\rfloor}(\varepsilon/2) \cap H'$, and fix some $\zeta \in (0,1)$. We have $\forall x$,

$$\begin{aligned}
&\mathcal{P}^{\tilde{d}_f - 1}\left(S \in \mathcal{X}^{\tilde{d}_f - 1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \mathcal{S}^{\tilde{d}_f - 1}(V)\right) \\
&= \mathcal{P}^{\tilde{d}_f - 1}\left(S \in \mathcal{X}^{\tilde{d}_f - 1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \mathcal{S}^{\tilde{d}_f - 1}(V) \cap \partial_{\mathbb{C}}^{\tilde{d}_f - 1} f\right) \mathcal{P}^{\tilde{d}_f - 1}\left(\partial_{\mathbb{C}}^{\tilde{d}_f - 1} f \middle| \mathcal{S}^{\tilde{d}_f - 1}(V)\right) \\
&+ \mathcal{P}^{\tilde{d}_f - 1}\left(S \in \mathcal{X}^{\tilde{d}_f - 1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \mathcal{S}^{\tilde{d}_f - 1}(V) \cap \bar{\partial}_{\mathbb{C}}^{\tilde{d}_f - 1} f\right) \mathcal{P}^{\tilde{d}_f - 1}\left(\bar{\partial}_{\mathbb{C}}^{\tilde{d}_f - 1} f \middle| \mathcal{S}^{\tilde{d}_f - 1}(V)\right) \\
&\leq \mathcal{P}^{\tilde{d}_f - 1}\left(S \in \mathcal{X}^{\tilde{d}_f - 1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \mathcal{S}^{\tilde{d}_f - 1}(V) \cap \partial_{\mathbb{C}}^{\tilde{d}_f - 1} f\right) + \mathcal{P}^{\tilde{d}_f - 1}\left(\bar{\partial}_{\mathbb{C}}^{\tilde{d}_f - 1} f \middle| \mathcal{S}^{\tilde{d}_f - 1}(V)\right). \quad (22)
\end{aligned}$$

By Lemma 35, the left term in (22) equals

$$\begin{aligned}
&\mathcal{P}^{\tilde{d}_f - 1}\left(S \in \mathcal{X}^{\tilde{d}_f - 1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \mathcal{S}^{\tilde{d}_f - 1}(V) \cap \partial_{\mathbb{C}}^{\tilde{d}_f - 1} f\right) \mathcal{P}^{\tilde{d}_f - 1}\left(\mathcal{S}^{\tilde{d}_f - 1}(V) \middle| \partial_{\mathbb{C}}^{\tilde{d}_f - 1} f\right) \\
&= \mathcal{P}^{\tilde{d}_f - 1}\left(S \in \mathcal{X}^{\tilde{d}_f - 1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \partial_{\mathbb{C}}^{\tilde{d}_f - 1} f\right),
\end{aligned}$$

and by Lemma 36, the right term in (22) is at most $q(\phi(\lfloor n/3 \rfloor;\varepsilon/2))$. Thus, we have

$$\begin{aligned}
&\mathcal{P}\left(x : \mathcal{P}^{\tilde{d}_f-1}\left(S \in \mathcal{X}^{\tilde{d}_f-1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \mathcal{S}^{\tilde{d}_f-1}(V)\right) \geq \zeta\right) \\
&\leq \mathcal{P}\left(x : \mathcal{P}^{\tilde{d}_f-1}\left(S \in \mathcal{X}^{\tilde{d}_f-1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \partial_{\mathbb{C}}^{\tilde{d}_f-1} f\right) \geq \zeta - q(\phi(\lfloor n/3 \rfloor;\varepsilon/2))\right). \qquad (23)
\end{aligned}$$

For $n < 3\tau(\zeta/2;\varepsilon/2)$ (for $\tau(\cdot;\cdot)$ defined in Lemma 37), we define $\Delta_n^{(\zeta)}(\varepsilon) = 1$. Otherwise, suppose $n \geq 3\tau(\zeta/2;\varepsilon/2)$, so that $q(\phi(\lfloor n/3 \rfloor;\varepsilon/2)) < \zeta/2$, and thus (23) is at most

$$\mathcal{P}\left(x : \mathcal{P}^{\tilde{d}_f-1}\left(S \in \mathcal{X}^{\tilde{d}_f-1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}(V) \middle| \partial_{\mathbb{C}}^{\tilde{d}_f-1} f\right) \geq \zeta/2\right).$$

By Lemma 29, this is at most

$$\mathcal{P}\left(x : \mathcal{P}^{\tilde{d}_f-1}\left(S \in \mathcal{X}^{\tilde{d}_f-1} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}\left(\mathrm{B}(f,\phi(\lfloor n/3 \rfloor;\varepsilon/2))\right) \middle| \partial_{\mathbb{C}}^{\tilde{d}_f-1} f\right) \geq \zeta/2\right).$$

Letting $X \sim \mathcal{P}$, by Markov's inequality this is at most

$$\begin{aligned}
&\frac{2}{\zeta}\mathbb{E}\left[\mathcal{P}^{\tilde{d}_f-1}\left(S \in \mathcal{X}^{\tilde{d}_f-1} : S \cup \{X\} \in \mathcal{S}^{\tilde{d}_f}\left(\mathrm{B}(f,\phi(\lfloor n/3 \rfloor;\varepsilon/2))\right) \middle| \partial_{\mathbb{C}}^{\tilde{d}_f-1} f\right)\right] \\
&= \frac{2}{\zeta\tilde{\delta}_f}\mathcal{P}^{\tilde{d}_f}\left(S \cup \{x\} \in \mathcal{X}^{\tilde{d}_f} : S \cup \{x\} \in \mathcal{S}^{\tilde{d}_f}\left(\mathrm{B}(f,\phi(\lfloor n/3 \rfloor;\varepsilon/2))\right) \text{ and } S \in \partial_{\mathbb{C}}^{\tilde{d}_f-1} f\right) \\
&\leq \frac{2}{\zeta\tilde{\delta}_f}\mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}\left(\mathrm{B}(f,\phi(\lfloor n/3 \rfloor;\varepsilon/2))\right)\right). \qquad (24)
\end{aligned}$$

Thus, defining $\Delta_n^{(\zeta)}(\varepsilon)$ as (24) for $n \geq 3\tau(\zeta/2;\varepsilon/2)$ establishes the first claim.

It remains only to prove the second claim. Let $N(\varepsilon) = \omega(\log(1/\varepsilon))$. Since $\tau(\zeta/2;\varepsilon/2) \leq \left\lceil \frac{4}{r_{\zeta/2}}\left(d\ln\left(\frac{4e}{r_{\zeta/2}}\right) + \ln\left(\frac{4}{\varepsilon}\right)\right)\right\rceil = O(\log(1/\varepsilon))$, we have that for all sufficiently small $\varepsilon > 0$, $N(\varepsilon) \geq 3\tau(\zeta/2;\varepsilon/2)$, so that $\Delta_{N(\varepsilon)}^{(\zeta)}(\varepsilon)$ equals (24) (with $n = N(\varepsilon)$). Furthermore, since $\tilde{\delta}_f > 0$, while $\mathcal{P}^{\tilde{d}_f}\left(\partial_{\mathbb{C}}^{\tilde{d}_f} f\right) = 0$, and $\phi(\lfloor N(\varepsilon)/3 \rfloor;\varepsilon/2) = o(1)$, by continuity of probability measures we know (24) is $o(1)$ when $n = N(\varepsilon)$, so that we generally have $\Delta_{N(\varepsilon)}^{(\zeta)}(\varepsilon) = o(1)$. ■

For any $m \in \mathbb{N}$, define

$$\tilde{M}(m) = m^3\tilde{\delta}_f/2.$$

Lemma 39 *There is a $(\mathbb{C},\mathcal{P},f)$-dependent constant $c^{(i)} \in (0,\infty)$ such that, for any $\tau \in \mathbb{N}$ there is an event $H_\tau^{(i)} \subseteq H'$ with*

$$\mathbb{P}\left(H_\tau^{(i)}\right) \geq 1 - c^{(i)} \cdot \exp\left\{-\tilde{M}(\tau)/4\right\}$$

such that on $H_\tau^{(i)}$, if $\tilde{d}_f \geq 2$, then $\forall k \in \left\{2,\ldots,\tilde{d}_f\right\}$, $\forall m \geq \tau$, $\forall \ell \in \mathbb{N}$, for any set $\mathcal{H}$ such that $V_\ell^\star \subseteq \mathcal{H} \subseteq \mathbb{C}$,

$$M_m^{(k)}(\mathcal{H}) \geq \tilde{M}(m).$$

Proof On H', Lemma 35 implies every $\mathbb{1}_{\mathcal{S}^{k-1}(\mathcal{H})}\left(S_i^{(k)}\right) \geq \mathbb{1}_{\partial_{\mathcal{H}}^{k-1} f}\left(S_i^{(k)}\right) = \mathbb{1}_{\partial_{\mathbb{C}}^{k-1} f}\left(S_i^{(k)}\right)$, so we focus on showing $\left|\left\{S_i^{(k)} : i \leq m^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right| \geq \tilde{M}(m)$ on an appropriate event. We know

$$\begin{aligned}
&\mathbb{P}\left(\forall k \in \left\{2, \ldots, \tilde{d}_f\right\}, \forall m \geq \tau, \left|\left\{S_i^{(k)} : i \leq m^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right| \geq \tilde{M}(m)\right) \\
&= 1 - \mathbb{P}\left(\exists k \in \left\{2, \ldots, \tilde{d}_f\right\}, m \geq \tau : \left|\left\{S_i^{(k)} : i \leq m^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right| < \tilde{M}(m)\right) \\
&\geq 1 - \sum_{m \geq \tau} \sum_{k=2}^{\tilde{d}_f} \mathbb{P}\left(\left|\left\{S_i^{(k)} : i \leq m^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right| < \tilde{M}(m)\right),
\end{aligned}$$

where the last line follows by a union bound. Thus, we will focus on bounding

$$\sum_{m \geq \tau} \sum_{k=2}^{\tilde{d}_f} \mathbb{P}\left(\left|\left\{S_i^{(k)} : i \leq m^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right| < \tilde{M}(m)\right). \tag{25}$$

Fix any $k \in \left\{2, \ldots, \tilde{d}_f\right\}$, and integer $m \geq \tau$. Since

$$\mathbb{E}\left[\left|\left\{S_i^{(k)} : i \leq m^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right|\right] = \mathcal{P}^{k-1}\left(\partial_{\mathbb{C}}^{k-1} f\right) m^3 \geq \tilde{\delta}_f m^3,$$

a Chernoff bound implies that

$$\begin{aligned}
\mathbb{P}\left(\left|\left\{S_i^{(k)} : i \leq m^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right| < \tilde{M}(m)\right) &\leq \exp\left\{-m^3 \mathcal{P}^{k-1}\left(\partial_{\mathbb{C}}^{k-1} f\right)/8\right\} \\
&\leq \exp\left\{-m^3 \tilde{\delta}_f/8\right\}.
\end{aligned}$$

Thus, we have that (25) is at most

$$\begin{aligned}
\sum_{m \geq \tau} \sum_{k=2}^{\tilde{d}_f} \exp\left\{-m^3 \tilde{\delta}_f/8\right\} &\leq \sum_{m \geq \tau} \tilde{d}_f \cdot \exp\left\{-m^3 \tilde{\delta}_f/8\right\} \leq \sum_{m \geq \tau^3} \tilde{d}_f \cdot \exp\left\{-m \tilde{\delta}_f/8\right\} \\
&\leq \tilde{d}_f \cdot \exp\left\{-\tilde{M}(\tau)/4\right\} + \tilde{d}_f \cdot \int_{\tau^3}^{\infty} \exp\left\{-x \tilde{\delta}_f/8\right\} \mathrm{d}x \\
&= \tilde{d}_f \cdot \left(1 + 8/\tilde{\delta}_f\right) \cdot \exp\left\{-\tilde{M}(\tau)/4\right\} \\
&\leq \left(9 \tilde{d}_f/\tilde{\delta}_f\right) \cdot \exp\left\{-\tilde{M}(\tau)/4\right\}.
\end{aligned}$$

Note that since $\mathbb{P}(H') = 1$, defining

$$H_\tau^{(i)} = \left\{\forall k \in \left\{2, \ldots, \tilde{d}_f\right\}, \forall m \geq \tau, \left|\left\{S_i^{(k)} : i \leq m^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right| \geq \tilde{M}(m)\right\} \cap H'$$

has the required properties. ∎

Lemma 40 *For any $\tau \in \mathbb{N}$, there is an event $G_\tau^{(i)}$ with*

$$\mathbb{P}\left(H_\tau^{(i)} \setminus G_\tau^{(i)}\right) \leq \left(121\tilde{d}_f/\tilde{\delta}_f\right) \cdot \exp\left\{-\tilde{M}(\tau)/60\right\}$$

such that, on $G_\tau^{(i)}$, if $\tilde{d}_f \geq 2$, then for every integer $s \geq \tau$ and $k \in \left\{2, \ldots, \tilde{d}_f\right\}$, $\forall r \in \left(0, r_{1/6}\right]$,

$$M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right) \leq (3/2)\left|\left\{S_i^{(k)} : i \leq s^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right|.$$

Proof Fix integers $s \geq \tau$ and $k \in \left\{2, \ldots, \tilde{d}_f\right\}$, and let $r = r_{1/6}$. Define the set $\hat{\mathcal{S}}^{k-1} = \left\{S_i^{(k)} : i \leq s^3\right\} \cap \mathcal{S}^{k-1}\left(\mathrm{B}\left(f,r\right)\right)$. Note $\left|\hat{\mathcal{S}}^{k-1}\right| = M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)$ and the elements of $\hat{\mathcal{S}}^{k-1}$ are conditionally i.i.d. given $M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)$, each with conditional distribution equivalent to the conditional distribution of $S_1^{(k)}$ given $\left\{S_1^{(k)} \in \mathcal{S}^{k-1}\left(\mathrm{B}\left(f,r\right)\right)\right\}$. In particular,

$$\mathbb{E}\left[\left|\hat{\mathcal{S}}^{k-1} \cap \partial_{\mathbb{C}}^{k-1} f\right| \middle| M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)\right] = \mathcal{P}^{k-1}\left(\partial_{\mathbb{C}}^{k-1} f \middle| \mathcal{S}^{k-1}\left(\mathrm{B}\left(f,r\right)\right)\right) M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right).$$

Define the event

$$G_\tau^{(i)}(k,s) = \left\{\left|\hat{\mathcal{S}}^{k-1}\right| \leq (3/2)\left|\hat{\mathcal{S}}^{k-1} \cap \partial_{\mathbb{C}}^{k-1} f\right|\right\}.$$

By Lemma 36 (indeed by definition of $q(r)$ and $r_{1/6}$) we have

$$\begin{aligned}
&1 - \mathbb{P}\left(G_\tau^{(i)}(k,s) \middle| M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)\right) \\
&= \mathbb{P}\left(\left|\hat{\mathcal{S}}^{k-1} \cap \partial_{\mathbb{C}}^{k-1} f\right| < (2/3) M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right) \middle| M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)\right) \\
&\leq \mathbb{P}\left(\left|\hat{\mathcal{S}}^{k-1} \cap \partial_{\mathbb{C}}^{k-1} f\right| < (4/5)\left(1 - q(r)\right) M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right) \middle| M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)\right) \\
&\leq \mathbb{P}\left(\left|\hat{\mathcal{S}}^{k-1} \cap \partial_{\mathbb{C}}^{k-1} f\right| < (4/5)\mathcal{P}^{k-1}\left(\partial_{\mathbb{C}}^{k-1} f \middle| \mathcal{S}^{k-1}\left(\mathrm{B}\left(f,r\right)\right)\right) M_s^{(k)}(\mathrm{B}\left(f,r\right)) \middle| M_s^{(k)}(\mathrm{B}\left(f,r\right))\right). \quad (26)
\end{aligned}$$

By a Chernoff bound, (26) is at most

$$\begin{aligned}
&\exp\left\{-M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)\mathcal{P}^{k-1}\left(\partial_{\mathbb{C}}^{k-1} f \middle| \mathcal{S}^{k-1}\left(\mathrm{B}\left(f,r\right)\right)\right)/50\right\} \\
&\qquad \leq \exp\left\{-M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)\left(1 - q(r)\right)/50\right\} \leq \exp\left\{-M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)/60\right\}.
\end{aligned}$$

Thus, by Lemma 39,

$$\begin{aligned}
&\mathbb{P}\left(H_\tau^{(i)} \setminus G_\tau^{(i)}(k,s)\right) \leq \mathbb{P}\left(\left\{M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right) \geq \tilde{M}(s)\right\} \setminus G_\tau^{(i)}(k,s)\right) \\
&= \mathbb{E}\left[\left(1 - \mathbb{P}\left(G_\tau^{(i)}(k,s) \middle| M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)\right)\right) \mathbb{1}_{[\tilde{M}(s),\infty)}\left(M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)\right)\right] \\
&\leq \mathbb{E}\left[\exp\left\{-M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)/60\right\} \mathbb{1}_{[\tilde{M}(s),\infty)}\left(M_s^{(k)}\left(\mathrm{B}\left(f,r\right)\right)\right)\right] \leq \exp\left\{-\tilde{M}(s)/60\right\}.
\end{aligned}$$

Now defining $G_\tau^{(i)} = \bigcap_{s\geq\tau}\bigcap_{k=2}^{\tilde{d}_f} G_\tau^{(i)}(k,s)$, a union bound implies

$$\begin{aligned}
\mathbb{P}\left(H_\tau^{(i)} \setminus G_\tau^{(i)}\right) &\leq \sum_{s\geq\tau} \tilde{d}_f \cdot \exp\left\{-\tilde{M}(s)/60\right\} \\
&\leq \tilde{d}_f\left(\exp\left\{-\tilde{M}(\tau)/60\right\} + \int_{\tau^3}^{\infty} \exp\left\{-x\tilde{\delta}_f/120\right\} \mathrm{d}x\right) \\
&= \tilde{d}_f\left(1+120/\tilde{\delta}_f\right)\cdot\exp\left\{-\tilde{M}(\tau)/60\right\} \\
&\leq \left(121\tilde{d}_f/\tilde{\delta}_f\right)\cdot\exp\left\{-\tilde{M}(\tau)/60\right\}.
\end{aligned}$$

This completes the proof for $r = r_{1/6}$. Monotonicity extends the result to any $r \in \left(0, r_{1/6}\right]$. ■

Lemma 41 *There exist $(\mathbb{C},\mathcal{P},f,\gamma)$-dependent constants $\tau^* \in \mathbb{N}$ and $c^{(ii)} \in (0,\infty)$ such that, for any integer $\tau \geq \tau^*$, there is an event $H_\tau^{(ii)} \subseteq G_\tau^{(i)}$ with*

$$\mathbb{P}\left(H_\tau^{(i)} \setminus H_\tau^{(ii)}\right) \leq c^{(ii)} \cdot \exp\left\{-\tilde{M}(\tau)^{1/3}/60\right\} \tag{27}$$

such that, on $H_\tau^{(i)} \cap H_\tau^{(ii)}$, $\forall s,m,\ell,k \in \mathbb{N}$ with $\ell < m$ and $k \leq \tilde{d}_f$, for any set of classifiers $\mathcal{H}$ with $V_\ell^\star \subseteq \mathcal{H}$, if either $k=1$, or $s \geq \tau$ and $\mathcal{H} \subseteq \mathrm{B}(f, r_{(1-\gamma)/6})$, then

$$\hat{\Delta}_s^{(k)}(X_m, W_2, \mathcal{H}) < \gamma \implies \hat{\Gamma}_s^{(k)}(X_m, -f(X_m), W_2, \mathcal{H}) < \hat{\Gamma}_s^{(k)}(X_m, f(X_m), W_2, \mathcal{H}).$$

In particular, for $\delta \in (0,1)$ and $\tau \geq \max\{\tau((1-\gamma)/6;\delta), \tau^*\}$, on $H_\tau(\delta) \cap H_\tau^{(i)} \cap H_\tau^{(ii)}$, this is true for $\mathcal{H} = V_\ell^\star$ for every $k,\ell,m,s \in \mathbb{N}$ satisfying $\tau \leq \ell < m$, $\tau \leq s$, and $k \leq \tilde{d}_f$.

Proof Let $\tau^* = (6/(1-\gamma))\cdot\left(2/\tilde{\delta}_f\right)^{1/3}$, and consider any $\tau,k,\ell,m,s,\mathcal{H}$ as described above. If $k=1$, the result clearly holds. In particular, Lemma 35 implies that on $H_\tau^{(i)}$, $\mathcal{H}[(X_m, f(X_m))] \supseteq V_m^\star \neq \emptyset$, so that some $h \in \mathcal{H}$ has $h(X_m) = f(X_m)$, and therefore

$$\hat{\Gamma}_s^{(1)}(X_m, -f(X_m), W_2, \mathcal{H}) = \mathbb{1}_{\bigcap\limits_{h\in\mathcal{H}}\{h(X_m)\}}(-f(X_m)) = 0,$$

and since $\hat{\Delta}_s^{(1)}(X_m, W_2, \mathcal{H}) = \mathbb{1}_{\mathrm{DIS}(\mathcal{H})}(X_m)$, if $\hat{\Delta}_s^{(1)}(X_m, W_2, \mathcal{H}) < \gamma$, then since $\gamma < 1$ we have $X_m \notin \mathrm{DIS}(\mathcal{H})$, so that

$$\hat{\Gamma}_s^{(1)}(X_m, f(X_m), W_2, \mathcal{H}) = \mathbb{1}_{\bigcap\limits_{h\in\mathcal{H}}\{h(X_m)\}}(f(X_m)) = 1.$$

Otherwise, suppose $2 \le k \le \tilde{d}_f$. Note that on $H_\tau^{(i)} \cap G_\tau^{(i)}$, $\forall m \in \mathbb{N}$, and any $\mathcal{H}$ with $V_\ell^\star \subseteq \mathcal{H} \subseteq \mathrm{B}(f, r_{(1-\gamma)/6})$ for some $\ell \in \mathbb{N}$,

$$\begin{aligned}
&\hat{\Gamma}_s^{(k)}(X_m, -f(X_m), W_2, \mathcal{H}) \\
&= \frac{1}{M_s^{(k)}(\mathcal{H})} \sum_{i=1}^{s^3} \mathbb{1}_{\bar{\mathcal{S}}^{k-1}(\mathcal{H}[(X_m, f(X_m))])}\left(S_i^{(k)}\right) \mathbb{1}_{\mathcal{S}^{k-1}(\mathcal{H})}\left(S_i^{(k)}\right) \\
&\le \frac{1}{\left|\left\{S_i^{(k)} : i \le s^3\right\} \cap \partial_{\mathcal{H}}^{k-1} f\right|} \sum_{i=1}^{s^3} \mathbb{1}_{\bar{\mathcal{S}}^{k-1}(V_m^\star)}\left(S_i^{(k)}\right) \mathbb{1}_{\mathcal{S}^{k-1}\left(\mathrm{B}(f, r_{(1-\gamma)/6})\right)}\left(S_i^{(k)}\right) \qquad \text{(monotonicity)} \\
&\le \frac{1}{\left|\left\{S_i^{(k)} : i \le s^3\right\} \cap \partial_{\mathcal{H}}^{k-1} f\right|} \sum_{i=1}^{s^3} \mathbb{1}_{\bar{\partial}_{V_m^\star}^{k-1} f}\left(S_i^{(k)}\right) \mathbb{1}_{\mathcal{S}^{k-1}\left(\mathrm{B}(f, r_{(1-\gamma)/6})\right)}\left(S_i^{(k)}\right) \qquad \text{(monotonicity)} \\
&= \frac{1}{\left|\left\{S_i^{(k)} : i \le s^3\right\} \cap \partial_{\mathbb{C}}^{k-1} f\right|} \sum_{i=1}^{s^3} \mathbb{1}_{\bar{\partial}_{\mathbb{C}}^{k-1} f}\left(S_i^{(k)}\right) \mathbb{1}_{\mathcal{S}^{k-1}\left(\mathrm{B}(f, r_{(1-\gamma)/6})\right)}\left(S_i^{(k)}\right) \qquad \text{(Lemma 35)} \\
&\le \frac{3}{2M_s^{(k)}(\mathrm{B}(f, r_{(1-\gamma)/6}))} \sum_{i=1}^{s^3} \mathbb{1}_{\bar{\partial}_{\mathbb{C}}^{k-1} f}\left(S_i^{(k)}\right) \mathbb{1}_{\mathcal{S}^{k-1}\left(\mathrm{B}(f, r_{(1-\gamma)/6})\right)}\left(S_i^{(k)}\right). \qquad \text{(Lemma 40)}
\end{aligned}$$

For brevity, let $\hat{\Gamma}$ denote this last quantity, and let $M_{ks} = M_s^{(k)}\left(\mathrm{B}\left(f, r_{(1-\gamma)/6}\right)\right)$. By Hoeffding's inequality, we have

$$\mathbb{P}\left((2/3)\hat{\Gamma} > \mathcal{P}^{k-1}\left(\bar{\partial}_{\mathbb{C}}^{k-1} f \middle| \mathcal{S}^{k-1}\left(\mathrm{B}\left(f, r_{(1-\gamma)/6}\right)\right)\right) + M_{ks}^{-1/3} \middle| M_{ks}\right) \le \exp\left\{-2M_{ks}^{1/3}\right\}.$$

Thus, by Lemmas 36, 39, and 40,

$$\begin{aligned}
&\mathbb{P}\left(\left\{(2/3)\hat{\Gamma}_s^{(k)}(X_m, -f(X_m), W_2, \mathcal{H}) > q\left(r_{(1-\gamma)/6}\right) + \tilde{M}(s)^{-1/3}\right\} \cap H_\tau^{(i)} \cap G_\tau^{(i)}\right) \\
&\le \mathbb{P}\left(\left\{(2/3)\hat{\Gamma} > \mathcal{P}^{k-1}\left(\bar{\partial}_{\mathbb{C}}^{k-1} f \middle| \mathcal{S}^{k-1}\left(\mathrm{B}\left(f, r_{(1-\gamma)/6}\right)\right)\right) + \tilde{M}(s)^{-1/3}\right\} \cap H_\tau^{(i)}\right) \\
&\le \mathbb{P}\left(\left\{(2/3)\hat{\Gamma} > \mathcal{P}^{k-1}\left(\bar{\partial}_{\mathbb{C}}^{k-1} f \middle| \mathcal{S}^{k-1}\left(\mathrm{B}\left(f, r_{(1-\gamma)/6}\right)\right)\right) + M_{ks}^{-1/3}\right\} \cap \left\{M_{ks} \ge \tilde{M}(s)\right\}\right) \\
&= \mathbb{E}\left[\mathbb{P}\left((2/3)\hat{\Gamma} > \mathcal{P}^{k-1}\left(\bar{\partial}_{\mathbb{C}}^{k-1} f \middle| \mathcal{S}^{k-1}\left(\mathrm{B}\left(f, r_{(1-\gamma)/6}\right)\right)\right) + M_{ks}^{-1/3} \middle| M_{ks}\right) \mathbb{1}_{[\tilde{M}(s), \infty)}(M_{ks})\right] \\
&\le \mathbb{E}\left[\exp\left\{-2M_{ks}^{1/3}\right\} \mathbb{1}_{[\tilde{M}(s), \infty)}(M_{ks})\right] \le \exp\left\{-2\tilde{M}(s)^{1/3}\right\}.
\end{aligned}$$

Thus, there is an event $H_\tau^{(ii)}(k,s)$ with $\mathbb{P}\left(H_\tau^{(i)} \cap G_\tau^{(ii)} \setminus H_\tau^{(ii)}(k,s)\right) \le \exp\left\{-2\tilde{M}(s)^{1/3}\right\}$ such that

$$\hat{\Gamma}_s^{(k)}(X_m, -f(X_m), W_2, \mathcal{H}) \le (3/2)\left(q\left(r_{(1-\gamma)/6}\right) + \tilde{M}(s)^{-1/3}\right)$$

holds for these particular values of k and s.

To extend to the full range of values, we simply take $H_\tau^{(ii)} = G_\tau^{(i)} \cap \bigcap_{s\geq\tau} \bigcap_{k\leq\tilde{d}_f} H_\tau^{(ii)}(k,s)$. Since $\tau \geq (2/\tilde{\delta}_f)^{1/3}$, we have $\tilde{M}(\tau) \geq 1$, so a union bound implies

$$\begin{aligned}
&\mathbb{P}\left(H_\tau^{(i)} \cap G_\tau^{(i)} \setminus H_\tau^{(ii)}\right) \leq \sum_{s\geq\tau} \tilde{d}_f \cdot \exp\left\{-2\tilde{M}(s)^{1/3}\right\} \\
&\leq \tilde{d}_f \cdot \left(\exp\left\{-2\tilde{M}(\tau)^{1/3}\right\} + \int_\tau^\infty \exp\left\{-2\tilde{M}(x)^{1/3}\right\} \mathrm{d}x\right) \\
&= \tilde{d}_f \left(1 + 2^{-2/3}\tilde{\delta}_f^{-1/3}\right) \cdot \exp\left\{-2\tilde{M}(\tau)^{1/3}\right\} \leq 2\tilde{d}_f\tilde{\delta}_f^{-1/3} \cdot \exp\left\{-2\tilde{M}(\tau)^{1/3}\right\}.
\end{aligned}$$

Then Lemma 40 and a union bound imply

$$\begin{aligned}
\mathbb{P}\left(H_\tau^{(i)} \setminus H_\tau^{(ii)}\right) &\leq 2\tilde{d}_f\tilde{\delta}_f^{-1/3} \cdot \exp\left\{-2\tilde{M}(\tau)^{1/3}\right\} + 121\tilde{d}_f\tilde{\delta}_f^{-1} \cdot \exp\left\{-\tilde{M}(\tau)/60\right\} \\
&\leq 123\tilde{d}_f\tilde{\delta}_f^{-1} \cdot \exp\left\{-\tilde{M}(\tau)^{1/3}/60\right\}.
\end{aligned}$$

On $H_\tau^{(i)} \cap H_\tau^{(ii)}$, every such s, m, ℓ, k and $\mathcal{H}$ satisfy

$$\begin{aligned}
\hat{\Gamma}_s^{(k)}\left(X_m, -f(X_m), W_2, \mathcal{H}\right) &\leq (3/2)\left(q(r_{(1-\gamma)/6}) + \tilde{M}(s)^{-1/3}\right) \\
&< (3/2)\left((1-\gamma)/6 + (1-\gamma)/6\right) = (1-\gamma)/2, \qquad (28)
\end{aligned}$$

where the second inequality follows by definition of $r_{(1-\gamma)/6}$ and $s \geq \tau \geq \tau^*$.

If $\hat{\Delta}_s^{(k)}\left(X_m, W_2, \mathcal{H}\right) < \gamma$, then

$$1 - \gamma < 1 - \hat{\Delta}_s^{(k)}\left(X_m, W_2, \mathcal{H}\right) = \frac{1}{M_s^{(k)}(\mathcal{H})} \sum_{i=1}^{s^3} \mathbb{1}_{\mathcal{S}^{k-1}(\mathcal{H})}\left(S_i^{(k)}\right) \mathbb{1}_{\bar{\mathcal{S}}^k(\mathcal{H})}\left(S_i^{(k)} \cup \{X_m\}\right). \qquad (29)$$

Finally, noting that we always have

$$\mathbb{1}_{\bar{\mathcal{S}}^k(\mathcal{H})}\left(S_i^{(k)} \cup \{X_m\}\right) \leq \mathbb{1}_{\bar{\mathcal{S}}^{k-1}(\mathcal{H}[(X_m, f(X_m))])}\left(S_i^{(k)}\right) + \mathbb{1}_{\bar{\mathcal{S}}^{k-1}(\mathcal{H}[(X_m, -f(X_m))])}\left(S_i^{(k)}\right),$$

we have that, on the event $H_\tau^{(i)} \cap H_\tau^{(ii)}$, if $\hat{\Delta}_s^{(k)}\left(X_m, W_2, \mathcal{H}\right) < \gamma$, then

$$\begin{aligned}
&\hat{\Gamma}_s^{(k)}\left(X_m, -f(X_m), W_2, \mathcal{H}\right) \\
&< (1-\gamma)/2 = -(1-\gamma)/2 + (1-\gamma) && \text{by (28)} \\
&< -(1-\gamma)/2 + \frac{1}{M_s^{(k)}(\mathcal{H})} \sum_{i=1}^{s^3} \mathbb{1}_{\mathcal{S}^{k-1}(\mathcal{H})}\left(S_i^{(k)}\right) \mathbb{1}_{\bar{\mathcal{S}}^k(\mathcal{H})}\left(S_i^{(k)} \cup \{X_m\}\right) && \text{by (29)} \\
&\leq -(1-\gamma)/2 + \frac{1}{M_s^{(k)}(\mathcal{H})} \sum_{i=1}^{s^3} \mathbb{1}_{\mathcal{S}^{k-1}(\mathcal{H})}\left(S_i^{(k)}\right) \mathbb{1}_{\bar{\mathcal{S}}^{k-1}(\mathcal{H}[(X_m, f(X_m))])}\left(S_i^{(k)}\right) \\
&\qquad + \frac{1}{M_s^{(k)}(\mathcal{H})} \sum_{i=1}^{s^3} \mathbb{1}_{\mathcal{S}^{k-1}(\mathcal{H})}\left(S_i^{(k)}\right) \mathbb{1}_{\bar{\mathcal{S}}^{k-1}(\mathcal{H}[(X_m, -f(X_m))])}\left(S_i^{(k)}\right) \\
&= -(1-\gamma)/2 + \hat{\Gamma}_s^{(k)}\left(X_m, -f(X_m), W_2, \mathcal{H}\right) + \hat{\Gamma}_s^{(k)}\left(X_m, f(X_m), W_2, \mathcal{H}\right) \\
&< \hat{\Gamma}_s^{(k)}\left(X_m, f(X_m), W_2, \mathcal{H}\right). && \text{by (28)}
\end{aligned}$$

The final claim in the lemma statement is then implied by Lemma 29, since we have $V_\ell^\star \subseteq V_\tau^\star \subseteq \mathrm{B}\left(f, \phi(\tau;\delta)\right) \subseteq \mathrm{B}\left(f, r_{(1-\gamma)/6}\right)$ on $H_\tau(\delta)$. ∎

For any $k, \ell, m \in \mathbb{N}$, and any $x \in \mathcal{X}$, define

$$\begin{aligned} \hat{p}_x(k,\ell,m) &= \hat{\Delta}_m^{(k)}\left(x, W_2, V_\ell^\star\right) \\ p_x(k,\ell) &= \mathcal{P}^{k-1}\left(S \in \mathcal{X}^{k-1} : S \cup \{x\} \in \mathcal{S}^k\left(V_\ell^\star\right) \middle| \mathcal{S}^{k-1}\left(V_\ell^\star\right)\right). \end{aligned}$$

Lemma 42 *For any $\zeta \in (0,1)$, there is a $(\mathbb{C}, \mathcal{P}, f, \zeta)$-dependent constant $c^{(iii)}(\zeta) \in (0,\infty)$ such that, for any $\tau \in \mathbb{N}$, there is an event $H_\tau^{(iii)}(\zeta)$ with*

$$\mathbb{P}\left(H_\tau^{(i)} \setminus H_\tau^{(iii)}(\zeta)\right) \leq c^{(iii)}(\zeta) \cdot \exp\left\{-\zeta^2 \tilde{M}(\tau)\right\}$$

such that on $H_\tau^{(i)} \cap H_\tau^{(iii)}(\zeta)$, $\forall k, \ell, m \in \mathbb{N}$ with $\tau \leq \ell \leq m$ and $k \leq \tilde{d}_f$, for any $x \in \mathcal{X}$,

$$\mathcal{P}\left(x : |p_x(k,\ell) - \hat{p}_x(k,\ell,m)| > \zeta\right) \leq \exp\left\{-\zeta^2 \tilde{M}(m)\right\}.$$

Proof Fix any $k, \ell, m \in \mathbb{N}$ with $\tau \leq \ell \leq m$ and $k \leq \tilde{d}_f$. Recall our convention that $\mathcal{X}^0 = \{\varnothing\}$ and $\mathcal{P}^0\left(\mathcal{X}^0\right) = 1$; thus, if $k = 1$, $\hat{p}_x(k,\ell,m) = \mathbb{1}_{\mathrm{DIS}\left(V_\ell^\star\right)}(x) = \mathbb{1}_{\mathcal{S}^1\left(V_\ell^\star\right)}(x) = p_x(k,\ell)$, so the result clearly holds for $k = 1$.

For the remaining case, suppose $2 \leq k \leq \tilde{d}_f$. To simplify notation, let $\tilde{m} = M_m^{(k)}\left(V_\ell^\star\right)$, $X = X_{\ell+1}$, $p_x = p_x(k,\ell)$ and $\hat{p}_x = \hat{p}_x(k,\ell,m)$. Consider the event

$$H^{(iii)}(k,\ell,m,\zeta) = \left\{\mathcal{P}\left(x : |p_x - \hat{p}_x| > \zeta\right) \leq \exp\left\{-\zeta^2 \tilde{M}(m)\right\}\right\}.$$

We have

$$\begin{aligned} &\mathbb{P}\left(H_\tau^{(i)} \setminus H^{(iii)}(k,\ell,m,\zeta) \middle| V_\ell^\star\right) &(30)\\ &\leq \mathbb{P}\left(\left\{\tilde{m} \geq \tilde{M}(m)\right\} \setminus H^{(iii)}(k,\ell,m,\zeta) \middle| V_\ell^\star\right) \text{ (by Lemma 39)} \\ &= \mathbb{P}\left(\left\{\tilde{m} \geq \tilde{M}(m)\right\} \cap \left\{\mathbb{P}\left(e^{s\tilde{m}|p_X - \hat{p}_X|} > e^{s\tilde{m}\zeta} \middle| W_2, V_\ell^\star\right) > e^{-\zeta^2 \tilde{M}(m)}\right\} \middle| V_\ell^\star\right), &(31) \end{aligned}$$

for any value $s > 0$. Proceeding as in Chernoff's bounding technique, by Markov's inequality (31) is at most

$$\begin{aligned} &\mathbb{P}\left(\left\{\tilde{m} \geq \tilde{M}(m)\right\} \cap \left\{e^{-s\tilde{m}\zeta} \mathbb{E}\left[e^{s\tilde{m}|p_X - \hat{p}_X|} \middle| W_2, V_\ell^\star\right] > e^{-\zeta^2 \tilde{M}(m)}\right\} \middle| V_\ell^\star\right) \\ &\leq \mathbb{P}\left(\left\{\tilde{m} \geq \tilde{M}(m)\right\} \cap \left\{e^{-s\tilde{m}\zeta} \mathbb{E}\left[e^{s\tilde{m}(p_X - \hat{p}_X)} + e^{s\tilde{m}(\hat{p}_X - p_X)} \middle| W_2, V_\ell^\star\right] > e^{-\zeta^2 \tilde{M}(m)}\right\} \middle| V_\ell^\star\right) \\ &= \mathbb{E}\left[\mathbb{1}_{[\tilde{M}(m),\infty)}(\tilde{m}) \mathbb{P}\left(e^{-s\tilde{m}\zeta} \mathbb{E}\left[e^{s\tilde{m}(p_X - \hat{p}_X)} + e^{s\tilde{m}(\hat{p}_X - p_X)} \middle| W_2, V_\ell^\star\right] > e^{-\zeta^2 \tilde{M}(m)} \middle| \tilde{m}, V_\ell^\star\right) \middle| V_\ell^\star\right] \end{aligned}$$

By Markov's inequality, this is at most

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{[\tilde{M}(m),\infty)}(\tilde{m})\,e^{\zeta^2\tilde{M}(m)}\mathbb{E}\left[e^{-s\tilde{m}\zeta}\mathbb{E}\left[e^{s\tilde{m}(p_X-\hat{p}_X)}+e^{s\tilde{m}(\hat{p}_X-p_X)}\middle|W_2,V_\ell^\star\right]\middle|\tilde{m},V_\ell^\star\right]\middle|V_\ell^\star\right]\\
&=\mathbb{E}\left[\mathbb{1}_{[\tilde{M}(m),\infty)}(\tilde{m})\,e^{\zeta^2\tilde{M}(m)}e^{-s\tilde{m}\zeta}\mathbb{E}\left[e^{s\tilde{m}(p_X-\hat{p}_X)}+e^{s\tilde{m}(\hat{p}_X-p_X)}\middle|\tilde{m},V_\ell^\star\right]\middle|V_\ell^\star\right]\\
&=\mathbb{E}\left[\mathbb{1}_{[\tilde{M}(m),\infty)}(\tilde{m})\,e^{\zeta^2\tilde{M}(m)}e^{-s\tilde{m}\zeta}\mathbb{E}\left[\mathbb{E}\left[e^{s\tilde{m}(p_X-\hat{p}_X)}+e^{s\tilde{m}(\hat{p}_X-p_X)}\middle|X,\tilde{m},V_\ell^\star\right]\middle|\tilde{m},V_\ell^\star\right]\middle|V_\ell^\star\right]. \qquad (32)
\end{aligned}$$

The conditional distribution of $\tilde{m}\hat{p}_X$ given $(X,\tilde{m},V_\ell^\star)$ is Binomial$(\tilde{m},p_X)$, so letting $\{\mathbf{B}_j(p_X)\}_{j=1}^{\infty}$ denote a sequence of random variables, conditionally independent given $(X,\tilde{m},V_\ell^\star)$, with the conditional distribution of each $\mathbf{B}_j(p_X)$ being Bernoulli(p_X) given $(X,\tilde{m},V_\ell^\star)$, we have

$$\begin{aligned}
&\mathbb{E}\left[e^{s\tilde{m}(p_X-\hat{p}_X)}+e^{s\tilde{m}(\hat{p}_X-p_X)}\middle|X,\tilde{m},V_\ell^\star\right]\\
&=\mathbb{E}\left[e^{s\tilde{m}(p_X-\hat{p}_X)}\middle|X,\tilde{m},V_\ell^\star\right]+\mathbb{E}\left[e^{s\tilde{m}(\hat{p}_X-p_X)}\middle|X,\tilde{m},V_\ell^\star\right]\\
&=\mathbb{E}\left[\prod_{i=1}^{\tilde{m}}e^{s(p_X-\mathbf{B}_i(p_X))}\middle|X,\tilde{m},V_\ell^\star\right]+\mathbb{E}\left[\prod_{i=1}^{\tilde{m}}e^{s(\mathbf{B}_i(p_X)-p_X)}\middle|X,\tilde{m},V_\ell^\star\right]\\
&=\mathbb{E}\left[e^{s(p_X-\mathbf{B}_1(p_X))}\middle|X,\tilde{m},V_\ell^\star\right]^{\tilde{m}}+\mathbb{E}\left[e^{s(\mathbf{B}_1(p_X)-p_X)}\middle|X,\tilde{m},V_\ell^\star\right]^{\tilde{m}}. \qquad (33)
\end{aligned}$$

It is known that for $\mathbf{B}\sim$ Bernoulli(p), $\mathbb{E}\left[e^{s(\mathbf{B}-p)}\right]$ and $\mathbb{E}\left[e^{s(p-\mathbf{B})}\right]$ are at most $e^{s^2/8}$ (see, e.g., Lemma 8.1 of Devroye, Györfi, and Lugosi, 1996). Thus, taking $s=4\zeta$, (33) is at most $2e^{2\tilde{m}\zeta^2}$, and (32) is at most

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{[\tilde{M}(m),\infty)}(\tilde{m})\,2e^{\zeta^2\tilde{M}(m)}e^{-4\tilde{m}\zeta^2}e^{2\tilde{m}\zeta^2}\middle|V_\ell^\star\right]&=\mathbb{E}\left[\mathbb{1}_{[\tilde{M}(m),\infty)}(\tilde{m})\,2e^{\zeta^2\tilde{M}(m)}e^{-2\tilde{m}\zeta^2}\middle|V_\ell^\star\right]\\
&\leq 2\exp\left\{-\zeta^2\tilde{M}(m)\right\}.
\end{aligned}$$

Since this bound holds for (30), the law of total probability implies

$$\mathbb{P}\left(H_\tau^{(i)}\setminus H^{(iii)}(k,\ell,m,\zeta)\right)=\mathbb{E}\left[\mathbb{P}\left(H_\tau^{(i)}\setminus H^{(iii)}(k,\ell,m,\zeta)\middle|V_\ell^\star\right)\right]\leq 2\cdot\exp\left\{-\zeta^2\tilde{M}(m)\right\}.$$

Defining $H_{\tau}^{(iii)}(\zeta) = \bigcap_{\ell \geq \tau} \bigcap_{m \geq \ell} \bigcap_{k=2}^{\tilde{d}_f} H^{(iii)}(k,\ell,m,\zeta)$, we have the required property for the claimed ranges of k, ℓ and m, and a union bound implies

$$\begin{aligned}
\mathbb{P}\left(H_{\tau}^{(i)} \setminus H_{\tau}^{(iii)}(\zeta)\right) &\leq \sum_{\ell \geq \tau} \sum_{m \geq \ell} 2\tilde{d}_f \cdot \exp\left\{-\zeta^2 \tilde{M}(m)\right\} \\
&\leq 2\tilde{d}_f \cdot \sum_{\ell \geq \tau} \left(\exp\left\{-\zeta^2 \tilde{M}(\ell)\right\} + \int_{\ell^3}^{\infty} \exp\left\{-x\zeta^2 \tilde{\delta}_f/2\right\} \mathrm{d}x \right) \\
&= 2\tilde{d}_f \cdot \sum_{\ell \geq \tau} \left(1 + 2\zeta^{-2}\tilde{\delta}_f^{-1}\right) \cdot \exp\left\{-\zeta^2 \tilde{M}(\ell)\right\} \\
&\leq 2\tilde{d}_f \cdot \left(1 + 2\zeta^{-2}\tilde{\delta}_f^{-1}\right) \cdot \left(\exp\left\{-\zeta^2 \tilde{M}(\tau)\right\} + \int_{\tau^3}^{\infty} \exp\left\{-x\zeta^2 \tilde{\delta}_f/2\right\} \mathrm{d}x \right) \\
&= 2\tilde{d}_f \cdot \left(1 + 2\zeta^{-2}\tilde{\delta}_f^{-1}\right)^2 \cdot \exp\left\{-\zeta^2 \tilde{M}(\tau)\right\} \\
&\leq 18\tilde{d}_f \zeta^{-4} \tilde{\delta}_f^{-2} \cdot \exp\left\{-\zeta^2 \tilde{M}(\tau)\right\}.
\end{aligned}$$

■

For $k,\ell,m \in \mathbb{N}$ and $\zeta \in (0,1)$, define

$$\bar{p}_{\zeta}(k,\ell,m) = \mathcal{P}\left(x : \hat{p}_x(k,\ell,m) \geq \zeta\right). \tag{34}$$

Lemma 43 *For any $\alpha,\zeta,\delta \in (0,1)$, $\beta \in \left(0, 1-\sqrt{\alpha}\right]$, and integer $\tau \geq \tau(\beta;\delta)$, on $H_{\tau}(\delta) \cap H_{\tau}^{(i)} \cap H_{\tau}^{(iii)}(\beta\zeta)$, for any $k,\ell,\ell',m \in \mathbb{N}$ with $\tau \leq \ell \leq \ell' \leq m$ and $k \leq \tilde{d}_f$,*

$$\bar{p}_{\zeta}(k,\ell',m) \leq \mathcal{P}\left(x : p_x(k,\ell) \geq \alpha\zeta\right) + \exp\left\{-\beta^2\zeta^2\tilde{M}(m)\right\}. \tag{35}$$

Proof Fix any $\alpha,\zeta,\delta \in (0,1)$, $\beta \in \left(0, 1-\sqrt{\alpha}\right]$, $\tau,k,\ell,\ell',m \in \mathbb{N}$ with $\tau(\beta;\delta) \leq \tau \leq \ell \leq \ell' \leq m$ and $k \leq \tilde{d}_f$.

If $k = 1$, the result clearly holds. In particular, we have

$$\bar{p}_{\zeta}(1,\ell',m) = \mathcal{P}\left(\mathrm{DIS}\left(V_{\ell'}^{\star}\right)\right) \leq \mathcal{P}\left(\mathrm{DIS}\left(V_{\ell}^{\star}\right)\right) = \mathcal{P}\left(x : p_x(1,\ell) \geq \alpha\zeta\right).$$

Otherwise, suppose $2 \leq k \leq \tilde{d}_f$. By a union bound,

$$\begin{aligned}
\bar{p}_{\zeta}(k,\ell',m) &= \mathcal{P}\left(x : \hat{p}_x(k,\ell',m) \geq \zeta\right) \\
&\leq \mathcal{P}\left(x : p_x(k,\ell') \geq \sqrt{\alpha}\zeta\right) + \mathcal{P}\left(x : \left|p_x(k,\ell') - \hat{p}_x(k,\ell',m)\right| > (1-\sqrt{\alpha})\zeta\right).
\end{aligned} \tag{36}$$

Since

$$\mathcal{P}\left(x : \left|p_x(k,\ell') - \hat{p}_x(k,\ell',m)\right| > (1-\sqrt{\alpha})\zeta\right) \leq \mathcal{P}\left(x : \left|p_x(k,\ell') - \hat{p}_x(k,\ell',m)\right| > \beta\zeta\right),$$

Lemma 42 implies that, on $H_{\tau}^{(i)} \cap H_{\tau}^{(iii)}(\beta\zeta)$,

$$\mathcal{P}\left(x : \left|p_x(k,\ell') - \hat{p}_x(k,\ell',m)\right| > (1-\sqrt{\alpha})\zeta\right) \leq \exp\left\{-\beta^2\zeta^2\tilde{M}(m)\right\}. \tag{37}$$

It remains only to examine the first term on the right side of (36). For this, if $\mathcal{P}^{k-1}\left(\mathcal{S}^{k-1}\left(V^{\star}_{\ell'}\right)\right)=0$, then the first term is 0 by our aforementioned convention, and thus (35) holds; otherwise, since

$$\forall x\in\mathcal{X}, \left\{S\in\mathcal{X}^{k-1}:S\cup\{x\}\in\mathcal{S}^{k}\left(V^{\star}_{\ell'}\right)\right\}\subseteq\mathcal{S}^{k-1}\left(V^{\star}_{\ell'}\right),$$

we have

$$\mathcal{P}\left(x:p_x(k,\ell')\geq\sqrt{\alpha}\zeta\right)=\mathcal{P}\left(x:\mathcal{P}^{k-1}\left(S\in\mathcal{X}^{k-1}:S\cup\{x\}\in\mathcal{S}^{k}\left(V^{\star}_{\ell'}\right)\middle|\mathcal{S}^{k-1}\left(V^{\star}_{\ell'}\right)\right)\geq\sqrt{\alpha}\zeta\right)$$
$$=\mathcal{P}\left(x:\mathcal{P}^{k-1}\left(S\in\mathcal{X}^{k-1}:S\cup\{x\}\in\mathcal{S}^{k}\left(V^{\star}_{\ell'}\right)\right)\geq\sqrt{\alpha}\zeta\mathcal{P}^{k-1}\left(\mathcal{S}^{k-1}\left(V^{\star}_{\ell'}\right)\right)\right). \tag{38}$$

By Lemma 35 and monotonicity, on $H^{(i)}_{\tau}\subseteq H'$, (38) is at most

$$\mathcal{P}\left(x:\mathcal{P}^{k-1}\left(S\in\mathcal{X}^{k-1}:S\cup\{x\}\in\mathcal{S}^{k}\left(V^{\star}_{\ell'}\right)\right)\geq\sqrt{\alpha}\zeta\mathcal{P}^{k-1}\left(\partial^{k-1}_{\mathbb{C}}f\right)\right),$$

and monotonicity implies this is at most

$$\mathcal{P}\left(x:\mathcal{P}^{k-1}\left(S\in\mathcal{X}^{k-1}:S\cup\{x\}\in\mathcal{S}^{k}\left(V^{\star}_{\ell}\right)\right)\geq\sqrt{\alpha}\zeta\mathcal{P}^{k-1}\left(\partial^{k-1}_{\mathbb{C}}f\right)\right). \tag{39}$$

By Lemma 36, for $\tau\geq\tau(\beta;\delta)$, on $H_{\tau}(\delta)\cap H^{(i)}_{\tau}$,

$$\mathcal{P}^{k-1}\left(\bar{\partial}^{k-1}_{\mathbb{C}}f\middle|\mathcal{S}^{k-1}\left(V^{\star}_{\ell}\right)\right)\leq q(\phi(\tau;\delta))<\beta\leq1-\sqrt{\alpha},$$

which implies

$$\mathcal{P}^{k-1}\left(\partial^{k-1}_{\mathbb{C}}f\right)\geq\mathcal{P}^{k-1}\left(\partial^{k-1}_{\mathbb{C}}f\cap\mathcal{S}^{k-1}\left(V^{\star}_{\ell}\right)\right)$$
$$=\left(1-\mathcal{P}^{k-1}\left(\bar{\partial}^{k-1}_{\mathbb{C}}f\middle|\mathcal{S}^{k-1}\left(V^{\star}_{\ell}\right)\right)\right)\mathcal{P}^{k-1}\left(\mathcal{S}^{k-1}\left(V^{\star}_{\ell}\right)\right)\geq\sqrt{\alpha}\mathcal{P}^{k-1}\left(\mathcal{S}^{k-1}\left(V^{\star}_{\ell}\right)\right).$$

Altogether, for $\tau\geq\tau(\beta;\delta)$, on $H_{\tau}(\delta)\cap H^{(i)}_{\tau}$, (39) is at most

$$\mathcal{P}\left(x:\mathcal{P}^{k-1}\left(S\in\mathcal{X}^{k-1}:S\cup\{x\}\in\mathcal{S}^{k}\left(V^{\star}_{\ell}\right)\right)\geq\alpha\zeta\mathcal{P}^{k-1}\left(\mathcal{S}^{k-1}\left(V^{\star}_{\ell}\right)\right)\right)=\mathcal{P}\left(x:p_x(k,\ell)\geq\alpha\zeta\right),$$

which, combined with (36) and (37), establishes (35). ■

Lemma 44 *There are events* $\left\{H^{(iv)}_{\tau}:\tau\in\mathbb{N}\right\}$ *with*

$$\mathbb{P}\left(H^{(iv)}_{\tau}\right)\geq1-3\tilde{d}_f\cdot\exp\{-2\tau\}$$

s.t. for any $\xi\in(0,\gamma/16]$, $\delta\in(0,1)$, *letting* $\tau^{(iv)}(\xi;\delta)=\max\left\{\tau(4\xi/\gamma;\delta),\left(\frac{4}{\tilde{\delta}_f\xi^2}\ln\left(\frac{4}{\tilde{\delta}_f\xi^2}\right)\right)^{1/3}\right\}$, *for any integer* $\tau\geq\tau^{(iv)}(\xi;\delta)$, *on* $H_{\tau}(\delta)\cap H^{(i)}_{\tau}\cap H^{(iii)}_{\tau}(\xi)\cap H^{(iv)}_{\tau}$, $\forall k\in\left\{1,\ldots,\tilde{d}_f\right\}$, $\forall\ell\in\mathbb{N}$ *with* $\ell\geq\tau$,

$$\mathcal{P}\Big(x:p_x(k,\ell)\geq\gamma/2\Big)+\exp\left\{-\gamma^2\tilde{M}(\ell)/256\right\}\leq\hat{\Delta}^{(k)}_{\ell}\left(W_1,W_2,V^{\star}_{\ell}\right) \tag{40}$$
$$\leq\mathcal{P}\left(x:p_x(k,\ell)\geq\gamma/8\right)+4\ell^{-1}. \tag{41}$$

Proof For any $k, \ell \in \mathbb{N}$, by Hoeffding's inequality and the law of total probability, on an event $G^{(iv)}(k,\ell)$ with $\mathbb{P}\left(G^{(iv)}(k,\ell)\right) \geq 1 - 2\exp\{-2\ell\}$, we have

$$\left| \bar{p}_{\gamma/4}(k,\ell,\ell) - \ell^{-3} \sum_{i=1}^{\ell^3} \mathbb{1}_{[\gamma/4,\infty)}\left(\hat{\Delta}_\ell^{(k)}(w_i, W_2, V_\ell^\star)\right) \right| \leq \ell^{-1}. \tag{42}$$

Define the event $H_\tau^{(iv)} = \bigcap_{\ell \geq \tau} \bigcap_{k=1}^{\tilde{d}_f} G^{(iv)}(k,\ell)$. By a union bound, we have

$$\begin{aligned}
1 - \mathbb{P}\left(H_\tau^{(iv)}\right) &\leq 2\tilde{d}_f \cdot \sum_{\ell \geq \tau} \exp\{-2\ell\} \\
&\leq 2\tilde{d}_f \cdot \left(\exp\{-2\tau\} + \int_\tau^\infty \exp\{-2x\}\,\mathrm{d}x\right) = 3\tilde{d}_f \cdot \exp\{-2\tau\}.
\end{aligned}$$

Now fix any $\ell \geq \tau$ and $k \in \left\{1, \ldots, \tilde{d}_f\right\}$. By a union bound,

$$\mathcal{P}(x : p_x(k,\ell) \geq \gamma/2) \leq \mathcal{P}(x : \hat{p}_x(k,\ell,\ell) \geq \gamma/4) + \mathcal{P}(x : |p_x(k,\ell) - \hat{p}_x(k,\ell,\ell)| > \gamma/4). \tag{43}$$

By Lemma 42, on $H_\tau^{(i)} \cap H_\tau^{(iii)}(\xi)$,

$$\mathcal{P}(x : |p_x(k,\ell) - \hat{p}_x(k,\ell,\ell)| > \gamma/4) \leq \mathcal{P}(x : |p_x(k,\ell) - \hat{p}_x(k,\ell,\ell)| > \xi) \leq \exp\left\{-\xi^2 \tilde{M}(\ell)\right\}. \tag{44}$$

Also, on $H_\tau^{(iv)}$, (42) implies

$$\begin{aligned}
\mathcal{P}(x : \hat{p}_x(k,\ell,\ell) \geq \gamma/4) &= \bar{p}_{\gamma/4}(k,\ell,\ell) \\
&\leq \ell^{-1} + \ell^{-3} \sum_{i=1}^{\ell^3} \mathbb{1}_{[\gamma/4,\infty)}\left(\hat{\Delta}_\ell^{(k)}(w_i, W_2, V_\ell^\star)\right) \\
&= \hat{\Delta}_\ell^{(k)}(W_1, W_2, V_\ell^\star) - \ell^{-1}.
\end{aligned} \tag{45}$$

Combining (43) with (44) and (45) yields

$$\mathcal{P}(x : p_x(k,\ell) \geq \gamma/2) \leq \hat{\Delta}_\ell^{(k)}(W_1, W_2, V_\ell^\star) - \ell^{-1} + \exp\left\{-\xi^2 \tilde{M}(\ell)\right\}. \tag{46}$$

For $\tau \geq \tau^{(iv)}(\xi;\delta)$, $\exp\left\{-\xi^2 \tilde{M}(\ell)\right\} - \ell^{-1} \leq -\exp\left\{-\gamma^2 \tilde{M}(\ell)/256\right\}$, so that (46) implies the first inequality of the lemma: namely (40).

For the second inequality (i.e., (41)), on $H_\tau^{(iv)}$, (42) implies we have

$$\hat{\Delta}_\ell^{(k)}(W_1, W_2, V_\ell^\star) \leq \bar{p}_{\gamma/4}(k,\ell,\ell) + 3\ell^{-1}. \tag{47}$$

Also, by Lemma 43 (with $\alpha = 1/2$, $\zeta = \gamma/4$, $\beta = \xi/\zeta < 1 - \sqrt{\alpha}$), for $\tau \geq \tau^{(iv)}(\xi;\delta)$, on $H_\tau(\delta) \cap H_\tau^{(i)} \cap H_\tau^{(iii)}(\xi)$,

$$\bar{p}_{\gamma/4}(k,\ell,\ell) \leq \mathcal{P}(x : p_x(k,\ell) \geq \gamma/8) + \exp\left\{-\xi^2 \tilde{M}(\ell)\right\}. \tag{48}$$

Thus, combining (47) with (48) yields

$$\hat{\Delta}_\ell^{(k)}(W_1, W_2, V_\ell^\star) \leq \mathcal{P}(x : p_x(k,\ell) \geq \gamma/8) + 3\ell^{-1} + \exp\left\{-\xi^2 \tilde{M}(\ell)\right\}.$$

For $\tau \geq \tau^{(iv)}(\xi;\delta)$, we have $\exp\left\{-\xi^2\tilde{M}(\ell)\right\} \leq \ell^{-1}$, which establishes (41). ■

For $n \in \mathbb{N}$ and $k \in \{1,\ldots,d+1\}$, define the set

$$\mathcal{U}_n^{(k)} = \left\{m_n+1,\ldots,m_n+\left\lfloor n/\left(6\cdot 2^k\hat{\Delta}_{m_n}^{(k)}(W_1,W_2,V)\right)\right\rfloor\right\},$$

where $m_n = \lfloor n/3 \rfloor$; $\mathcal{U}_n^{(k)}$ represents the set of indices processed in the inner loop of Meta-Algorithm 1 for the specified value of k.

Lemma 45 *There are $(f,\mathbb{C},\mathcal{P},\gamma)$-dependent constants $\hat{c}_1,\hat{c}_2 \in (0,\infty)$ such that, for any $\varepsilon \in (0,1)$ and integer $n \geq \hat{c}_1 \ln(\hat{c}_2/\varepsilon)$, on an event $\hat{H}_n(\varepsilon)$ with*

$$\mathbb{P}(\hat{H}_n(\varepsilon)) \geq 1-(3/4)\varepsilon, \tag{49}$$

we have, for $V = V^\star_{m_n}$,

$$\forall k \in \left\{1,\ldots,\tilde{d}_f\right\}, \left|\left\{m \in \mathcal{U}_n^{(k)} : \hat{\Delta}_m^{(k)}(X_m,W_2,V) \geq \gamma\right\}\right| \leq \left\lfloor n/\left(3\cdot 2^k\right)\right\rfloor, \tag{50}$$

$$\hat{\Delta}_{m_n}^{(\tilde{d}_f)}(W_1,W_2,V) \leq \Delta_n^{(\gamma/8)}(\varepsilon)+4m_n^{-1}, \tag{51}$$

and $\forall m \in \mathcal{U}_n^{(\tilde{d}_f)}$,

$$\hat{\Delta}_m^{(\tilde{d}_f)}(X_m,W_2,V) < \gamma \Rightarrow \hat{\Gamma}_m^{(\tilde{d}_f)}(X_m,-f(X_m),W_2,V) < \hat{\Gamma}_m^{(\tilde{d}_f)}(X_m,f(X_m),W_2,V). \tag{52}$$

Proof Suppose $n \geq \hat{c}_1 \ln(\hat{c}_2/\varepsilon)$, where

$$\begin{aligned}\hat{c}_1 &= \max\left\{\frac{2^{\tilde{d}_f+12}}{\tilde{\delta}_f\gamma^2},\frac{24}{r_{(1/16)}},\frac{24}{r_{(1-\gamma)/6}},3\tau^*\right\}\\ \text{and } \hat{c}_2 &= \max\left\{4\left(c^{(i)}+c^{(ii)}+c^{(iii)}(\gamma/16)+6\tilde{d}_f\right),4\left(\frac{4e}{r_{(1/16)}}\right)^d,4\left(\frac{4e}{r_{(1-\gamma)/6}}\right)^d\right\}.\end{aligned}$$

In particular, we have chosen $\hat{c}_1$ and $\hat{c}_2$ large enough so that

$$m_n \geq \max\left\{\tau(1/16;\varepsilon/2),\tau^{(iv)}(\gamma/16;\varepsilon/2),\tau((1-\gamma)/6;\varepsilon/2),\tau^*\right\}.$$

We begin with (50). By Lemmas 43 and 44, on the event

$$\hat{H}_n^{(1)}(\varepsilon) = H_{m_n}(\varepsilon/2)\cap H_{m_n}^{(i)}\cap H_{m_n}^{(iii)}(\gamma/16)\cap H_{m_n}^{(iv)},$$

$\forall m \in \mathcal{U}_n^{(k)}, \forall k \in \left\{1,\ldots,\tilde{d}_f\right\}$,

$$\begin{aligned}\bar{p}_\gamma(k,m_n,m) &\leq \mathcal{P}\left(x : p_x(k,m_n) \geq \gamma/2\right)+\exp\left\{-\gamma^2\tilde{M}(m)/256\right\}\\ &\leq \mathcal{P}\left(x : p_x(k,m_n) \geq \gamma/2\right)+\exp\left\{-\gamma^2\tilde{M}(m_n)/256\right\} \leq \hat{\Delta}_{m_n}^{(k)}(W_1,W_2,V).\end{aligned} \tag{53}$$

Recall that $\left\{X_m : m \in \mathcal{U}_n^{(k)}\right\}$ is a sample of size $\left\lfloor n/(6 \cdot 2^k \hat{\Delta}_{m_n}^{(k)}(W_1, W_2, V))\right\rfloor$, conditionally i.i.d. (given (W_1, W_2, V)) with conditional distributions $\mathcal{P}$. Thus, $\forall k \in \left\{1, \ldots, \tilde{d}_f\right\}$, on $\hat{H}_n^{(1)}(\varepsilon)$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\left\{m \in \mathcal{U}_n^{(k)} : \hat{\Delta}_m^{(k)}(X_m, W_2, V) \geq \gamma\right\}\right| > n/\left(3 \cdot 2^k\right) \middle| W_1, W_2, V\right) \\
&\leq \mathbb{P}\left(\left|\left\{m \in \mathcal{U}_n^{(k)} : \hat{\Delta}_m^{(k)}(X_m, W_2, V) \geq \gamma\right\}\right| > 2\left|\mathcal{U}_n^{(k)}\right| \hat{\Delta}_{m_n}^{(k)}(W_1, W_2, V) \middle| W_1, W_2, V\right) \\
&\leq \mathbb{P}\left(\mathbf{B}\left(|\mathcal{U}_n^{(k)}|, \hat{\Delta}_{m_n}^{(k)}(W_1, W_2, V)\right) > 2\left|\mathcal{U}_n^{(k)}\right| \hat{\Delta}_{m_n}^{(k)}(W_1, W_2, V) \middle| W_1, W_2, V\right), \qquad (54)
\end{aligned}$$

where this last inequality follows from (53), and $\mathbf{B}(u, p) \sim \text{Binomial}(u, p)$ is independent from W_1, W_2, V (for any fixed u and p). By a Chernoff bound, (54) is at most

$$\exp\left\{-\left\lfloor n/\left(6 \cdot 2^k \hat{\Delta}_{m_n}^{(k)}(W_1, W_2, V)\right)\right\rfloor \hat{\Delta}_{m_n}^{(k)}(W_1, W_2, V)/3\right\} \leq \exp\left\{1 - n/\left(18 \cdot 2^k\right)\right\}.$$

By the law of total probability and a union bound, there exists an event $\hat{H}_n^{(2)}$ with

$$\mathbb{P}\left(\hat{H}_n^{(1)}(\varepsilon) \setminus \hat{H}_n^{(2)}\right) \leq \tilde{d}_f \cdot \exp\left\{1 - n/\left(18 \cdot 2^{\tilde{d}_f}\right)\right\}$$

such that, on $\hat{H}_n^{(1)}(\varepsilon) \cap \hat{H}_n^{(2)}$, (50) holds.

Next, by Lemma 44, on $\hat{H}_n^{(1)}(\varepsilon)$,

$$\hat{\Delta}_{m_n}^{(\tilde{d}_f)}(W_1, W_2, V) \leq \mathcal{P}\left(x : p_x\left(\tilde{d}_f, m_n\right) \geq \gamma/8\right) + 4m_n^{-1},$$

and by Lemma 38, on $\hat{H}_n^{(1)}(\varepsilon)$, this is at most $\Delta_n^{(\gamma/8)}(\varepsilon) + 4m_n^{-1}$, which establishes (51).

Finally, Lemma 41 implies that on $\hat{H}_n^{(1)}(\varepsilon) \cap H_{m_n}^{(ii)}$, $\forall m \in \mathcal{U}_n^{(\tilde{d}_f)}$, (52) holds.

Thus, defining

$$\hat{H}_n(\varepsilon) = \hat{H}_n^{(1)}(\varepsilon) \cap \hat{H}_n^{(2)} \cap H_{m_n}^{(ii)},$$

it remains only to establish (49). By a union bound, we have

$$\begin{aligned}
1 - \mathbb{P}\left(\hat{H}_n\right) \leq\ & (1 - \mathbb{P}(H_{m_n}(\varepsilon/2))) + \left(1 - \mathbb{P}\left(H_{m_n}^{(i)}\right)\right) + \mathbb{P}\left(H_{m_n}^{(i)} \setminus H_{m_n}^{(ii)}\right) \\
& + \mathbb{P}\left(H_{m_n}^{(i)} \setminus H_{m_n}^{(iii)}(\gamma/16)\right) + \left(1 - \mathbb{P}\left(H_{m_n}^{(iv)}\right)\right) + \mathbb{P}\left(\hat{H}_n^{(1)}(\varepsilon) \setminus \hat{H}_n^{(2)}\right). \\
\leq\ & \varepsilon/2 + c^{(i)} \cdot \exp\left\{-\tilde{M}(m_n)/4\right\} + c^{(ii)} \cdot \exp\left\{-\tilde{M}(m_n)^{1/3}/60\right\} \\
& + c^{(iii)}(\gamma/16) \cdot \exp\left\{-\tilde{M}(m_n)\gamma^2/256\right\} + 3\tilde{d}_f \cdot \exp\left\{-2m_n\right\} \\
& + \tilde{d}_f \cdot \exp\left\{1 - n/\left(18 \cdot 2^{\tilde{d}_f}\right)\right\} \\
\leq\ & \varepsilon/2 + \left(c^{(i)} + c^{(ii)} + c^{(iii)}(\gamma/16) + 6\tilde{d}_f\right) \cdot \exp\left\{-n\tilde{\delta}_f \gamma^2 2^{-\tilde{d}_f - 12}\right\}. \qquad (55)
\end{aligned}$$

We have chosen n large enough so that (55) is at most $(3/4)\varepsilon$, which establishes (49). ■

The following result is a slightly stronger version of Theorem 6.

Lemma 46 *For any passive learning algorithm $\mathcal{A}_p$, if $\mathcal{A}_p$ achieves a label complexity Λ_p with $\infty > \Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon))$, then Meta-Algorithm 1, with $\mathcal{A}_p$ as its argument, achieves a label complexity Λ_a such that $\Lambda_a(3\varepsilon, f, \mathcal{P}) = o(\Lambda_p(\varepsilon, f, \mathcal{P}))$.*

Proof Suppose $\mathcal{A}_p$ achieves label complexity Λ_p with $\infty > \Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon))$. Let $\varepsilon \in (0,1)$, define $L(n;\varepsilon) = \left\lfloor n / \left(6 \cdot 2^{\tilde{d}_f} \left(\Delta_n^{(\gamma/8)}(\varepsilon) + 4m_n^{-1}\right)\right)\right\rfloor$ (for any $n \in \mathbb{N}$), and let $L^{-1}(m;\varepsilon) = \max\{n \in \mathbb{N} : L(n;\varepsilon) < m\}$ (for any $m \in (0,\infty)$). Define

$$c_1 = \max\left\{\hat{c}_1, 2 \cdot 6^3 (d+1)\tilde{d}_f \ln(e(d+1))\right\} \quad \text{and} \quad c_2 = \max\left\{\hat{c}_2, 4e(d+1)\right\},$$

and suppose

$$n \geq \max\left\{c_1 \ln(c_2/\varepsilon), 1 + L^{-1}\left(\Lambda_p(\varepsilon, f, \mathcal{P}); \varepsilon\right)\right\}.$$

Consider running Meta-Algorithm 1 with $\mathcal{A}_p$ and n as inputs, while f is the target function and $\mathcal{P}$ is the data distribution.

Letting $\hat{h}_n$ denote the classifier returned from Meta-Algorithm 1, Lemma 34 implies that on an event $\hat{E}_n$ with $\mathbb{P}(\hat{E}_n) \geq 1 - e(d+1) \cdot \exp\left\{-\lfloor n/3 \rfloor / (72\tilde{d}_f(d+1)\ln(e(d+1)))\right\} \geq 1 - \varepsilon/4$, we have

$$\operatorname{er}(\hat{h}_n) \leq 2 \operatorname{er}\left(\mathcal{A}_p\left(\mathcal{L}_{\tilde{d}_f}\right)\right).$$

By a union bound, the event $\hat{G}_n(\varepsilon) = \hat{E}_n \cap \hat{H}_n(\varepsilon)$ has $\mathbb{P}\left(\hat{G}_n(\varepsilon)\right) \geq 1 - \varepsilon$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\operatorname{er}\left(\hat{h}_n\right)\right] \leq\; & \mathbb{E}\left[\mathbb{1}_{\hat{G}_n(\varepsilon)} \mathbb{1}\left[|\mathcal{L}_{\tilde{d}_f}| \geq \Lambda_p(\varepsilon, f, \mathcal{P})\right] \operatorname{er}\left(\hat{h}_n\right)\right] \\
& + \mathbb{P}\left(\hat{G}_n(\varepsilon) \cap \left\{|\mathcal{L}_{\tilde{d}_f}| < \Lambda_p(\varepsilon, f, \mathcal{P})\right\}\right) + \mathbb{P}\left(\hat{G}_n(\varepsilon)^c\right) \\
\leq\; & \mathbb{E}\left[\mathbb{1}_{\hat{G}_n(\varepsilon)} \mathbb{1}\left[|\mathcal{L}_{\tilde{d}_f}| \geq \Lambda_p(\varepsilon, f, \mathcal{P})\right] 2 \operatorname{er}\left(\mathcal{A}_p\left(\mathcal{L}_{\tilde{d}_f}\right)\right)\right] \\
& + \mathbb{P}\left(\hat{G}_n(\varepsilon) \cap \left\{|\mathcal{L}_{\tilde{d}_f}| < \Lambda_p(\varepsilon, f, \mathcal{P})\right\}\right) + \varepsilon. \qquad (56)
\end{aligned}$$

On $\hat{G}_n(\varepsilon)$, (51) of Lemma 45 implies $|\mathcal{L}_{\tilde{d}_f}| \geq L(n;\varepsilon)$, and we chose n large enough so that $L(n;\varepsilon) \geq \Lambda_p(\varepsilon, f, \mathcal{P})$. Thus, the second term in (56) is zero, and we have

$$\begin{aligned}
\mathbb{E}\left[\operatorname{er}\left(\hat{h}_n\right)\right] &\leq 2 \cdot \mathbb{E}\left[\mathbb{1}_{\hat{G}_n(\varepsilon)} \mathbb{1}\left[|\mathcal{L}_{\tilde{d}_f}| \geq \Lambda_p(\varepsilon, f, \mathcal{P})\right] \operatorname{er}\left(\mathcal{A}_p\left(\mathcal{L}_{\tilde{d}_f}\right)\right)\right] + \varepsilon \\
&= 2 \cdot \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\hat{G}_n(\varepsilon)} \operatorname{er}\left(\mathcal{A}_p\left(\mathcal{L}_{\tilde{d}_f}\right)\right) \middle| |\mathcal{L}_{\tilde{d}_f}|\right] \mathbb{1}\left[|\mathcal{L}_{\tilde{d}_f}| \geq \Lambda_p(\varepsilon, f, \mathcal{P})\right]\right] + \varepsilon. \qquad (57)
\end{aligned}$$

Note that for any ℓ with $\mathbb{P}(|\mathcal{L}_{\tilde{d}_f}| = \ell) > 0$, the conditional distribution of $\left\{X_m : m \in \mathcal{U}_n^{(\tilde{d}_f)}\right\}$ given $\left\{|\mathcal{L}_{\tilde{d}_f}| = \ell\right\}$ is simply the product $\mathcal{P}^\ell$ (i.e., conditionally i.i.d.), which is the same as the distribution of $\{X_1, X_2, \ldots, X_\ell\}$. Furthermore, on $\hat{G}_n(\varepsilon)$, (50) implies that the $t < \lfloor 2n/3 \rfloor$ condition is always satisfied in Step 6 of Meta-Algorithm 1 while $k \leq \tilde{d}_f$, and (52) implies that the inferred labels from Step 8 for $k = \tilde{d}_f$ are all correct. Therefore, for any such ℓ with $\ell \geq \Lambda_p(\varepsilon, f, \mathcal{P})$, we have

$$\mathbb{E}\left[\mathbb{1}_{\hat{G}_n(\varepsilon)} \operatorname{er}\left(\mathcal{A}_p\left(\mathcal{L}_{\tilde{d}_f}\right)\right) \middle| \left\{|\mathcal{L}_{\tilde{d}_f}| = \ell\right\}\right] \leq \mathbb{E}\left[\operatorname{er}\left(\mathcal{A}_p\left(\mathcal{Z}_\ell\right)\right)\right] \leq \varepsilon.$$

In particular, this means (57) is at most 3ε. This implies that Meta-Algorithm 1, with $\mathcal{A}_p$ as its argument, achieves a label complexity Λ_a such that

$$\Lambda_a(3\varepsilon, f, \mathcal{P}) \leq \max\left\{c_1 \ln(c_2/\varepsilon), 1 + L^{-1}\left(\Lambda_p(\varepsilon, f, \mathcal{P}); \varepsilon\right)\right\}.$$

Since $\Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon)) \Rightarrow c_1 \ln(c_2/\varepsilon) = o\left(\Lambda_p(\varepsilon, f, \mathcal{P})\right)$, it remains only to show that $L^{-1}\left(\Lambda_p(\varepsilon, f, \mathcal{P}); \varepsilon\right) = o\left(\Lambda_p(\varepsilon, f, \mathcal{P})\right)$. Note that $\forall \varepsilon \in (0,1)$, $L(1;\varepsilon) = 0$ and $L(n;\varepsilon)$ is diverging in n. Furthermore, by Lemma 38, we know that for any $\mathbb{N}$-valued $N(\varepsilon) = \omega(\log(1/\varepsilon))$, we have $\Delta^{(\gamma/8)}_{N(\varepsilon)}(\varepsilon) = o(1)$, which implies $L(N(\varepsilon);\varepsilon) = \omega(N(\varepsilon))$. Thus, since $\Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon))$, Lemma 31 implies $L^{-1}\left(\Lambda_p(\varepsilon, f, \mathcal{P}); \varepsilon\right) = o\left(\Lambda_p(\varepsilon, f, \mathcal{P})\right)$, as desired.

This establishes the result for an arbitrary $\gamma \in (0,1)$. To specialize to the specific procedure stated as Meta-Algorithm 1, we simply take $\gamma = 1/2$. ■

Proof [Theorem 6] Theorem 6 now follows immediately from Lemma 46. Specifically, we have proven Lemma 46 for an arbitrary distribution $\mathcal{P}$ on $\mathcal{X}$, an arbitrary $f \in \mathrm{cl}(\mathbb{C})$, and an arbitrary passive algorithm $\mathcal{A}_p$. Therefore, it will certainly hold for every $\mathcal{P}$ and $f \in \mathbb{C}$, and since every $(f, \mathcal{P}) \in \mathrm{Nontrivial}(\Lambda_p)$ has $\infty > \Lambda_p(\varepsilon, f, \mathcal{P}) = \omega(\log(1/\varepsilon))$, the implication that Meta-Algorithm 1 activizes every passive algorithm $\mathcal{A}_p$ for $\mathbb{C}$ follows. ■

Careful examination of the proofs above reveals that the "3" in Lemma 46 can be set to any arbitrary constant strictly larger than 1, by an appropriate modification of the "7/12" threshold in ActiveSelect. In fact, if we were to replace Step 4 of ActiveSelect by instead selecting $\hat{k} = \mathrm{argmin}_k \max_{j \neq k} m_{kj}$ (where $m_{kj} = \mathrm{er}_{Q_{kj}}(h_k)$ when $k < j$), then we could even make this a certain $(1 + o(1))$ function of ε, at the expense of larger constant factors in Λ_a.

Appendix C. The Label Complexity of Meta-Algorithm 2

As mentioned, Theorem 10 is essentially implied by the details of the proof of Theorem 16 in Appendix D below. Here we present a proof of Theorem 13, along with two useful related lemmas. The first, Lemma 47, lower bounds the expected number of label requests Meta-Algorithm 2 would make while processing a given number of random unlabeled examples. The second, Lemma 48, bounds the amount by which each label request is expected to reduce the probability mass in the region of disagreement. Although we will only use Lemma 48 in our proof of Theorem 13, Lemma 47 may be of independent interest, as it provides additional insights into the behavior of disagreement based methods, as related to the disagreement coefficient, and is included for this reason.

Throughout, we fix an arbitrary class $\mathbb{C}$, a target function $f \in \mathbb{C}$, and a distribution $\mathcal{P}$, and we continue using the notational conventions of the proofs above, such as $V^\star_m = \{h \in \mathbb{C} : \forall i \leq m, h(X_i) = f(X_i)\}$ (with $V^\star_0 = \mathbb{C}$). Additionally, for $t \in \mathbb{N}$, define the random variable

$$M(t) = \min\left\{m \in \mathbb{N} : \sum_{\ell=1}^{m} \mathbb{1}_{\mathrm{DIS}\left(V^\star_{\ell-1}\right)}(X_\ell) = t\right\},$$

which represents the index of the t^{th} unlabeled example Meta-Algorithm 2 would request the label of (assuming it has not yet halted).

The two aforementioned lemmas are formally stated as follows.

Lemma 47 *For any $r \in (0,1)$ and $\ell \in \mathbb{N}$,*

$$\mathbb{E}\left[\mathcal{P}\left(\mathrm{DIS}\left(V_\ell^\star \cap \mathrm{B}\left(f,r\right)\right)\right)\right] \geq (1-r)^\ell \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}\left(f,r\right)\right)\right),$$

$$\text{and} \qquad \mathbb{E}\left[\sum_{m=1}^{\lceil 1/r \rceil} \mathbb{1}_{\mathrm{DIS}\left(V_{m-1}^\star \cap \mathrm{B}(f,r)\right)}\left(X_m\right)\right] \geq \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right)}{2r}.$$

Lemma 48 *For any $r \in (0,1)$ and $n \in \mathbb{N}$,*

$$\mathbb{E}\left[\mathcal{P}\left(\mathrm{DIS}\left(V_{M(n)}^\star \cap \mathrm{B}\left(f,r\right)\right)\right)\right] \geq \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right) - nr.$$

Note these results immediately imply that

$$\mathbb{E}\left[\sum_{m=1}^{\lceil 1/r \rceil} \mathbb{1}_{\mathrm{DIS}\left(V_{m-1}^\star\right)}\left(X_m\right)\right] \geq \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right)}{2r}$$

and

$$\mathbb{E}\left[\mathcal{P}\left(\mathrm{DIS}\left(V_{M(n)}^\star\right)\right)\right] \geq \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right) - nr,$$

which are then directly relevant to the expected number of label requests made by Meta-Algorithm 2 among the first m data points, and the probability Meta-Algorithm 2 requests the label of the next point, after already making n label requests, respectively.

Before proving these lemmas, let us first mention their relevance to the disagreement coefficient analysis. Specifically, for any $\varepsilon \in (0,r]$, we have

$$\mathbb{E}\left[\sum_{m=1}^{\lceil 1/\varepsilon \rceil} \mathbb{1}_{\mathrm{DIS}\left(V_{m-1}^\star\right)}\left(X_m\right)\right] \geq \mathbb{E}\left[\sum_{m=1}^{\lceil 1/r \rceil} \mathbb{1}_{\mathrm{DIS}\left(V_{m-1}^\star\right)}\left(X_m\right)\right] \geq \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right)}{2r}.$$

In particular, maximizing over $r > \varepsilon$, we have

$$\mathbb{E}\left[\sum_{m=1}^{\lceil 1/\varepsilon \rceil} \mathbb{1}_{\mathrm{DIS}\left(V_{m-1}^\star\right)}\left(X_m\right)\right] \geq \theta_f(\varepsilon)/2.$$

Thus, the expected number of label requests among the first $\lceil 1/\varepsilon \rceil$ unlabeled examples processed by Meta-Algorithm 2 is at least $\theta_f(\varepsilon)/2$ (assuming it does not halt first). Similarly, for any $\varepsilon \in (0,r]$, for any $n \leq \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))/(2r)$, Lemma 48 implies

$$\mathbb{E}\left[\mathcal{P}\left(\mathrm{DIS}\left(V_{M(n)}^\star\right)\right)\right] \geq \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right)/2 \geq \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,\varepsilon)\right)\right)/2.$$

Maximizing over $r > \varepsilon$, we see that

$$n \leq \theta_f(\varepsilon)/2 \implies \mathbb{E}\left[\mathcal{P}\left(\mathrm{DIS}\left(V_{M(n)}^\star\right)\right)\right] \geq \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,\varepsilon)\right)\right)/2.$$

In other words, for Meta-Algorithm 2 to arrive at a region of disagreement with expected probability mass less than $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,\varepsilon)))/2$ requires a budget n of at least $\theta_f(\varepsilon)/2$.

We now present proofs of Lemmas 47 and 48.

Proof [Lemma 47] Let $D_m = \mathrm{DIS}\left(V^\star_m \cap \mathrm{B}(f,r)\right)$. Since

$$\mathbb{E}\left[\sum_{m=1}^{\lceil 1/r \rceil} \mathbb{1}_{D_{m-1}}(X_m)\right] = \sum_{m=1}^{\lceil 1/r \rceil} \mathbb{E}\left[\mathbb{P}\left(X_m \in D_{m-1} \middle| V^\star_{m-1}\right)\right]$$
$$= \sum_{m=1}^{\lceil 1/r \rceil} \mathbb{E}\left[\mathcal{P}\left(D_{m-1}\right)\right], \tag{58}$$

we focus on lower bounding $\mathbb{E}[\mathcal{P}(D_m)]$ for $m \in \mathbb{N} \cup \{0\}$. Note that for any $x \in \mathrm{DIS}(\mathrm{B}(f,r))$, there exists some $h_x \in \mathrm{B}(f,r)$ with $h_x(x) \neq f(x)$, and if this $h_x \in V^\star_m$, then $x \in D_m$ as well. This means $\forall x, \mathbb{1}_{D_m}(x) \geq \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f,r))}(x) \cdot \mathbb{1}_{V^\star_m}(h_x) = \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f,r))}(x) \cdot \prod_{\ell=1}^{m} \mathbb{1}_{\mathrm{DIS}(\{h_x,f\})^c}(X_\ell)$. Therefore,

$$\mathbb{E}[\mathcal{P}(D_m)] = \mathbb{P}(X_{m+1} \in D_m) = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{D_m}(X_{m+1}) \middle| X_{m+1}\right]\right]$$
$$\geq \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f,r))}(X_{m+1}) \cdot \prod_{\ell=1}^{m} \mathbb{1}_{\mathrm{DIS}(\{h_{X_{m+1}},f\})^c}(X_\ell) \middle| X_{m+1}\right]\right]$$
$$= \mathbb{E}\left[\prod_{\ell=1}^{m} \mathbb{P}\left(h_{X_{m+1}}(X_\ell) = f(X_\ell) \middle| X_{m+1}\right) \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f,r))}(X_{m+1})\right] \tag{59}$$
$$\geq \mathbb{E}\left[(1-r)^m \mathbb{1}_{\mathrm{DIS}(\mathrm{B}(f,r))}(X_{m+1})\right] = (1-r)^m \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r))), \tag{60}$$

where the equality in (59) is by conditional independence of the $\mathbb{1}_{\mathrm{DIS}(\{h_{X_{m+1}},f\})^c}(X_\ell)$ indicators, given X_{m+1}, and the inequality in (60) is due to $h_{X_{m+1}} \in \mathrm{B}(f,r)$. This indicates (58) is at least

$$\sum_{m=1}^{\lceil 1/r \rceil} (1-r)^{m-1} \mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right) = \left(1 - (1-r)^{\lceil 1/r \rceil}\right) \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right)}{r}$$
$$\geq \left(1 - \frac{1}{e}\right) \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right)}{r} \geq \frac{\mathcal{P}\left(\mathrm{DIS}\left(\mathrm{B}(f,r)\right)\right)}{2r}.$$

■

Proof [Lemma 48] For each $m \in \mathbb{N} \cup \{0\}$, let $D_m = \mathrm{DIS}\left(\mathrm{B}(f,r) \cap V^\star_m\right)$. For convenience, let $M(0) = 0$. We prove the result by induction. We clearly have $\mathbb{E}\left[\mathcal{P}\left(D_{M(0)}\right)\right] = \mathbb{E}\left[\mathcal{P}\left(D_0\right)\right] = \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))$, which serves as our base case. Now fix any $n \in \mathbb{N}$ and take as the inductive hypothesis that

$$\mathbb{E}\left[\mathcal{P}\left(D_{M(n-1)}\right)\right] \geq \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r))) - (n-1)r.$$

As in the proof of Lemma 47, for any $x \in D_{M(n-1)}$, there exists $h_x \in \mathrm{B}(f,r) \cap V^\star_{M(n-1)}$ with $h_x(x) \neq f(x)$; unlike the proof of Lemma 47, here h_x is a random variable, determined by $V^\star_{M(n-1)}$. If h_x is also in $V^\star_{M(n)}$, then $x \in D_{M(n)}$ as well. Thus, $\forall x, \mathbb{1}_{D_{M(n)}}(x) \geq \mathbb{1}_{D_{M(n-1)}}(x) \cdot \mathbb{1}_{V^\star_{M(n)}}(h_x) = \mathbb{1}_{D_{M(n-1)}}(x) \cdot \mathbb{1}_{\mathrm{DIS}(\{h_x,f\})^c}(X_{M(n)})$, where this last equality is due to the fact that every $m \in \{M(n-1)+1, \ldots, M(n)-1\}$ has $X_m \notin \mathrm{DIS}\left(V^\star_{m-1}\right)$, so that in particular $h_x(X_m) = f(X_m)$. Therefore, letting $X \sim \mathcal{P}$ be independent of the data $\mathcal{Z}$,

$$\begin{aligned}\mathbb{E}\left[\mathcal{P}\left(D_{M(n)}\right)\right] = \mathbb{E}\left[\mathbb{1}_{D_{M(n)}}(X)\right] &\geq \mathbb{E}\left[\mathbb{1}_{D_{M(n-1)}}(X)\cdot\mathbb{1}_{\mathrm{DIS}(\{h_X,f\})^c}(X_{M(n)})\right] \\ &= \mathbb{E}\left[\mathbb{1}_{D_{M(n-1)}}(X)\cdot\mathbb{P}\left(h_X(X_{M(n)}) = f(X_{M(n)})\middle|X, V^\star_{M(n-1)}\right)\right]. \end{aligned} \tag{61}$$

The conditional distribution of $X_{M(n)}$ given $V^\star_{M(n-1)}$ is merely $\mathcal{P}$ but with support restricted to $\mathrm{DIS}\left(V^\star_{M(n-1)}\right)$ and renormalized to a probability measure: that is $\mathcal{P}\left(\cdot\middle|\mathrm{DIS}\left(V^\star_{M(n-1)}\right)\right)$. Thus, since any $x\in D_{M(n-1)}$ has $\mathrm{DIS}(\{h_x,f\})\subseteq\mathrm{DIS}\left(V^\star_{M(n-1)}\right)$, we have

$$\mathbb{P}\left(h_x(X_{M(n)})\neq f(X_{M(n)})\middle|V^\star_{M(n-1)}\right) = \frac{\mathcal{P}\left(\mathrm{DIS}(\{h_x,f\})\right)}{\mathcal{P}\left(\mathrm{DIS}\left(V^\star_{M(n-1)}\right)\right)} \leq \frac{r}{\mathcal{P}\left(D_{M(n-1)}\right)},$$

where the inequality follows from $h_x\in\mathrm{B}(f,r)$ and $D_{M(n-1)}\subseteq\mathrm{DIS}\left(V^\star_{M(n-1)}\right)$. Therefore, (61) is at least

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{D_{M(n-1)}}(X)\cdot\left(1-\frac{r}{\mathcal{P}(D_{M(n-1)})}\right)\right] \\ &\qquad = \mathbb{E}\left[\mathbb{P}\left(X\in D_{M(n-1)}\middle|D_{M(n-1)}\right)\cdot\left(1-\frac{r}{\mathcal{P}(D_{M(n-1)})}\right)\right] \\ &\qquad = \mathbb{E}\left[\mathcal{P}\left(D_{M(n-1)}\right)\cdot\left(1-\frac{r}{\mathcal{P}(D_{M(n-1)})}\right)\right] = \mathbb{E}\left[\mathcal{P}\left(D_{M(n-1)}\right)\right]-r.\end{aligned}$$

By the inductive hypothesis, this is at least $\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))-nr$. ■

With Lemma 48 in hand, we are ready for the proof of Theorem 13.

Proof [Theorem 13] Let $\mathbb{C}$, f, $\mathcal{P}$, and λ be as in the theorem statement. For $m\in\mathbb{N}$, let $\lambda^{-1}(m) = \inf\{\varepsilon>0:\lambda(\varepsilon)\leq m\}$, or 1 if this is not defined. We define $\mathcal{A}_p$ as a randomized algorithm such that, for $m\in\mathbb{N}$ and $\mathcal{L}\in(\mathcal{X}\times\{-1,+1\})^m$, $\mathcal{A}_p(\mathcal{L})$ returns f with probability $1-\lambda^{-1}(|\mathcal{L}|)$ and returns $-f$ with probability $\lambda^{-1}(|\mathcal{L}|)$ (independent of the contents of $\mathcal{L}$). Note that, for any integer $m\geq\lambda(\varepsilon)$, $\mathbb{E}\left[\mathrm{er}\left(\mathcal{A}_p\left(\mathcal{Z}_m\right)\right)\right] = \lambda^{-1}(m)\leq\lambda^{-1}(\lambda(\varepsilon))\leq\varepsilon$. Therefore, $\mathcal{A}_p$ achieves some label complexity Λ_p with $\Lambda_p(\varepsilon,f,\mathcal{P}) = \lambda(\varepsilon)$ for all $\varepsilon>0$.

If $\theta_f\left(\lambda(\varepsilon)^{-1}\right)\neq\omega(1)$, then monotonicity implies $\theta_f\left(\lambda(\varepsilon)^{-1}\right) = O(1)$, and since every label complexity Λ_a is $\Omega(1)$, the result clearly holds. Otherwise, suppose $\theta_f\left(\lambda(\varepsilon)^{-1}\right) = \omega(1)$; in particular, this means $\exists\varepsilon_0\in(0,1/2)$ such that $\theta_f\left(\lambda(2\varepsilon_0)^{-1}\right)\geq 12$. Fix any $\varepsilon\in(0,\varepsilon_0)$, let $r>\lambda(2\varepsilon)^{-1}$ be such that $\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))}{r}\geq\theta_f\left(\lambda(2\varepsilon)^{-1}\right)/2$, and let $n\in\mathbb{N}$ satisfy $n\leq\theta_f\left(\lambda(2\varepsilon)^{-1}\right)/4$.

Consider running Meta-Algorithm 2 with arguments $\mathcal{A}_p$ and n, and let $\hat{\mathcal{L}}$ denote the final value of the set $\mathcal{L}$, and let $\check{m}$ denote the value of m upon reaching Step 6. Note that any $m<\lambda(2\varepsilon)$ and $L\in(\mathcal{X}\times\{-1,+1\})^m$ has $\mathrm{er}\left(\mathcal{A}_p(L)\right) = \lambda^{-1}(m)\geq\inf\{\varepsilon'>0:\lambda(\varepsilon')<\lambda(2\varepsilon)\}\geq 2\varepsilon$. Therefore, we have

$$\begin{aligned}\mathbb{E}\left[\mathrm{er}\left(\mathcal{A}_p\left(\hat{\mathcal{L}}\right)\right)\right] &\geq 2\varepsilon\mathbb{P}\left(|\hat{\mathcal{L}}|<\lambda(2\varepsilon)\right) = 2\varepsilon\mathbb{P}\left(\left\lfloor n/\left(6\hat{\Delta}\right)\right\rfloor<\lambda(2\varepsilon)\right) \\ &= 2\varepsilon\mathbb{P}\left(\hat{\Delta}>\frac{n}{6\lambda(2\varepsilon)}\right) = 2\varepsilon\left(1-\mathbb{P}\left(\hat{\Delta}\leq\frac{n}{6\lambda(2\varepsilon)}\right)\right).\end{aligned} \tag{62}$$

Since $n \leq \theta_f\left(\lambda(2\varepsilon)^{-1}\right)/4 \leq \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))/(2r) < \lambda(2\varepsilon)\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))/2$, we have

$$\mathbb{P}\left(\hat{\Delta} \leq \frac{n}{6\lambda(2\varepsilon)}\right) \leq \mathbb{P}\left(\hat{\Delta} < \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))/12\right)$$
$$\leq \mathbb{P}\left(\left\{\mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{\check{m}}\right)\right) < \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))/12\right\} \cup \left\{\hat{\Delta} < \mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{\check{m}}\right)\right)\right\}\right). \quad (63)$$

Since $\check{m} \leq M(\lceil n/2 \rceil)$, monotonicity and a union bound imply this is at most

$$\mathbb{P}\left(\mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{M(\lceil n/2\rceil)}\right)\right) < \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))/12\right) + \mathbb{P}\left(\hat{\Delta} < \mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{\check{m}}\right)\right)\right). \quad (64)$$

Markov's inequality implies

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{M(\lceil n/2\rceil)}\right)\right) < \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))/12\right)\\
&= \mathbb{P}\left(\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r))) - \mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{M(\lceil n/2\rceil)}\right)\right) > \frac{11}{12}\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))\right)\\
&\leq \mathbb{P}\left(\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r))) - \mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{M(\lceil n/2\rceil)} \cap \mathrm{B}(f,r)\right)\right) > \frac{11}{12}\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))\right)\\
&\leq \frac{\mathbb{E}\left[\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r))) - \mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{M(\lceil n/2\rceil)} \cap \mathrm{B}(f,r)\right)\right)\right]}{\frac{11}{12}\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))}\\
&= \frac{12}{11}\left(1 - \frac{\mathbb{E}\left[\mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{M(\lceil n/2\rceil)} \cap \mathrm{B}(f,r)\right)\right)\right]}{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))}\right).
\end{aligned}$$

Lemma 48 implies this is at most $\frac{12}{11}\frac{\lceil n/2\rceil r}{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))} \leq \frac{12}{11}\left\lceil\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))}{4r}\right\rceil\frac{r}{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))}$. Since any $a \geq 3/2$ has $\lceil a \rceil \leq (3/2)a$, and $\theta_f\left(\lambda(2\varepsilon)^{-1}\right) \geq 12$ implies $\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))}{4r} \geq 3/2$, we have $\left\lceil\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))}{4r}\right\rceil \leq \frac{3}{8}\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))}{r}$, so that $\frac{12}{11}\left\lceil\frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))}{4r}\right\rceil\frac{r}{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))} \leq \frac{9}{22}$. Combining the above, we have

$$\mathbb{P}\left(\mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{M(\lceil n/2\rceil)}\right)\right) < \mathcal{P}(\mathrm{DIS}(\mathrm{B}(f,r)))/12\right) \leq \frac{9}{22}. \quad (65)$$

Examining the second term in (64), Hoeffding's inequality and the definition of $\hat{\Delta}$ from (13) imply

$$\mathbb{P}\left(\hat{\Delta} < \mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{\check{m}}\right)\right)\right) = \mathbb{E}\left[\mathbb{P}\left(\hat{\Delta} < \mathcal{P}\left(\mathrm{DIS}\left(V^{\star}_{\check{m}}\right)\right)\middle| V^{\star}_{\check{m}}, \check{m}\right)\right] \leq \mathbb{E}\left[e^{-8\check{m}}\right] \leq e^{-8} < 1/11. \quad (66)$$

Combining (62), (63), (64), (65), and (66) implies

$$\mathbb{E}\left[\mathrm{er}\left(\mathcal{A}_p\left(\hat{\mathcal{L}}\right)\right)\right] > 2\varepsilon\left(1 - \frac{9}{22} - \frac{1}{11}\right) = \varepsilon.$$

Thus, for any label complexity Λ_a achieved by running Meta-Algorithm 2 with $\mathcal{A}_p$ as its argument, we must have $\Lambda_a(\varepsilon, f, \mathcal{P}) > \theta_f\left(\lambda(2\varepsilon)^{-1}\right)/4$. Since this is true for all $\varepsilon \in (0, \varepsilon_0)$, this establishes the result. ∎

Appendix D. The Label Complexity of Meta-Algorithm 3

As in Appendix B, we will assume $\mathbb{C}$ is a fixed VC class, $\mathcal{P}$ is some arbitrary distribution, and $f \in \mathrm{cl}(\mathbb{C})$ is an arbitrary fixed function. We continue using the notation introduced above: in particular, $\mathcal{S}^k(\mathcal{H}) = \{S \in \mathcal{X}^k : \mathcal{H} \text{ shatters } S\}$, $\bar{\mathcal{S}}^k(\mathcal{H}) = \mathcal{X}^k \setminus \mathcal{S}^k(\mathcal{H})$, $\bar{\partial}^k_{\mathcal{H}} f = \mathcal{X}^k \setminus \partial^k_{\mathcal{H}} f$, and $\tilde{\delta}_f = \mathcal{P}^{\tilde{d}_f - 1}\left(\partial_{\mathbb{C}}^{\tilde{d}_f - 1} f\right)$. Also, as above, we will prove a more general result replacing the "1/2" in Steps 5, 9, and 12 of Meta-Algorithm 3 with an arbitrary value $\gamma \in (0,1)$; thus, the specific result for the stated algorithm will be obtained by taking $\gamma = 1/2$.

For the estimators $\hat{P}_m$ in Meta-Algorithm 3, we take precisely the same definitions as given in Appendix B.1 for the estimators in Meta-Algorithm 1. In particular, the quantities $\hat{\Delta}_m^{(k)}(x, W_2, \mathcal{H})$, $\hat{\Delta}_m^{(k)}(W_1, W_2, \mathcal{H})$, $\hat{\Gamma}_m^{(k)}(x, y, W_2, \mathcal{H})$, and $M_m^{(k)}(\mathcal{H})$ are all defined as in Appendix B.1, and the $\hat{P}_m$ estimators are again defined as in (11), (12) and (13).

Also, we sometimes refer to quantities defined above, such as $\bar{p}_\zeta(k, \ell, m)$ (defined in (34)), as well as the various events from the lemmas of the previous appendix, such as $H_\tau(\delta)$, H', $H_\tau^{(i)}$, $H_\tau^{(ii)}$, $H_\tau^{(iii)}(\zeta)$, $H_\tau^{(iv)}$, and $G_\tau^{(i)}$.

D.1 Proof of Theorem 16

Throughout the proof, we will make reference to the sets V_m defined in Meta-Algorithm 3. Also let $V^{(k)}$ denote the final value of V obtained for the specified value of k in Meta-Algorithm 3. Both V_m and $V^{(k)}$ are implicitly functions of the budget, n, given to Meta-Algorithm 3. As above, we continue to denote by $V_m^\star = \{h \in \mathbb{C} : \forall i \le m, h(X_m) = f(X_m)\}$. One important fact we will use repeatedly below is that if $V_m = V_m^\star$ for some m, then since Lemma 35 implies that $V_m^\star \neq \emptyset$ on H', we must have that all of the previous $\hat{y}$ values were consistent with f, which means that $\forall \ell \le m$, $V_\ell = V_\ell^\star$. In particular, if $V^{(k')} = V_m^\star$ for the largest m value obtained while $k = k'$ in Meta-Algorithm 3, then $V_\ell = V_\ell^\star$ for all ℓ obtained while $k \le k'$ in Meta-Algorithm 3.

Additionally, define $\tilde{m}_n = \lfloor n/24 \rfloor$, and note that the value $m = \lceil n/6 \rceil$ is obtained while $k = 1$ in Meta-Algorithm 3. We also define the following quantities, which we will show are typically equal to related quantities in Meta-Algorithm 3. Define $\hat{m}_0 = 0$, $T_0^\star = \lceil 2n/3 \rceil$, and $\hat{t}_0 = 0$, and for each $k \in \{1, \ldots, d+1\}$, inductively define

$$
\begin{aligned}
T_k^\star &= T_{k-1}^\star - \hat{t}_{k-1},\\
I_{mk}^\star &= \mathbb{1}_{[\gamma,\infty)}\left(\hat{\Delta}_m^{(k)}\left(X_m, W_2, V_{m-1}^\star\right)\right), \forall m \in \mathbb{N},\\
\check{m}_k &= \min\left\{ m \geq \hat{m}_{k-1} : \sum_{\ell=\hat{m}_{k-1}+1}^{m} I_{\ell k}^\star = \lceil T_k^\star/4 \rceil \right\} \cup \left\{\max\left\{k \cdot 2^n + 1, \hat{m}_{k-1}\right\}\right\},\\
\hat{m}_k &= \check{m}_k + \left\lfloor T_k^\star / \left(3\hat{\Delta}_{\check{m}_k}^{(k)}\left(W_1, W_2, V_{\check{m}_k}^\star\right)\right)\right\rfloor,\\
\check{\mathcal{U}}_k &= (\hat{m}_{k-1}, \check{m}_k] \cap \mathbb{N},\\
\hat{\mathcal{U}}_k &= (\check{m}_k, \hat{m}_k] \cap \mathbb{N},\\
C_{mk}^\star &= \mathbb{1}_{\left[0, \left\lfloor 3T_k^\star/4 \right\rfloor\right)}\left(\sum_{\ell=\hat{m}_{k-1}+1}^{m-1} I_{\ell k}^\star\right)\\
Q_k^\star &= \sum_{m \in \hat{\mathcal{U}}_k} I_{mk}^\star \cdot C_{mk}^\star,\\
\text{and } \hat{t}_k &= Q_k^\star + \sum_{m \in \check{\mathcal{U}}_k} I_{mk}^\star.
\end{aligned}
$$

The meaning of these values can be understood in the context of Meta-Algorithm 3, under the condition that $V_m = V_m^\star$ for values of m obtained for the respective value of k. Specifically, under this condition, $T_k^\star$ corresponds to T_k, $\hat{t}_k$ represents the final value t for round k, $\check{m}_k$ represents the value of m upon reaching Step 9 in round k, while $\hat{m}_k$ represents the value of m at the end of round k, $\check{\mathcal{U}}_k$ corresponds to the set of indices arrived at in Step 4 during round k, while $\hat{\mathcal{U}}_k$ corresponds to the set of indices arrived at in Step 11 during round k, for $m \in \check{\mathcal{U}}_k$, $I_{mk}^\star$ indicates whether the label of X_m is requested, while for $m \in \hat{\mathcal{U}}_k$, $I_{mk}^\star \cdot C_{mk}^\star$ indicates whether the label of X_m is requested. Finally $Q_k^\star$ corresponds to the number of label requests in Step 13 during round k. In particular, note $\check{m}_1 \geq \tilde{m}_n$.

Lemma 49 *For any $\tau \in \mathbb{N}$, on the event $H' \cap G_\tau^{(i)}$, $\forall k, \ell, m \in \mathbb{N}$ with $k \leq \tilde{d}_f$, $\forall x \in \mathcal{X}$, for any sets $\mathcal{H}$ and $\mathcal{H}'$ with $V_\ell^\star \subseteq \mathcal{H} \subseteq \mathcal{H}' \subseteq \mathrm{B}(f, r_{1/6})$, if either $k = 1$ or $m \geq \tau$, then*

$$\hat{\Delta}_m^{(k)}(x, W_2, \mathcal{H}) \leq (3/2)\hat{\Delta}_m^{(k)}\left(x, W_2, \mathcal{H}'\right).$$

In particular, for any $\delta \in (0,1)$ and $\tau \geq \tau(1/6; \delta)$, on $H' \cap H_\tau(\delta) \cap G_\tau^{(i)}$, $\forall k, \ell, \ell', m \in \mathbb{N}$ with $m \geq \tau$, $\ell \geq \ell' \geq \tau$, and $k \leq \tilde{d}_f$, $\forall x \in \mathcal{X}$, $\hat{\Delta}_m^{(k)}(x, W_2, V_\ell^\star) \leq (3/2)\hat{\Delta}_m^{(k)}\left(x, W_2, V_{\ell'}^\star\right)$.

Proof First note that $\forall m \in \mathbb{N}$, $\forall x \in \mathcal{X}$,

$$\hat{\Delta}_m^{(1)}(x, W_2, \mathcal{H}) = \mathbb{1}_{\mathrm{DIS}(\mathcal{H})}(x) \leq \mathbb{1}_{\mathrm{DIS}(\mathcal{H}')}(x) = \hat{\Delta}_m^{(1)}\left(x, W_2, \mathcal{H}'\right),$$

so the result holds for $k = 1$. Lemma 35, Lemma 40, and monotonicity of $M_m^{(k)}(\cdot)$ imply that on $H' \cap G_\tau^{(i)}$, for any $m \geq \tau$ and $k \in \left\{2, \ldots, \tilde{d}_f\right\}$,

$$M_m^{(k)}(\mathcal{H}) \geq \sum_{i=1}^{m^3} \mathbb{1}_{\partial_{\mathbb{C}}^{k-1} f}\left(S_i^{(k)}\right) \geq (2/3) M_m^{(k)}\left(\mathrm{B}(f, r_{1/6})\right) \geq (2/3) M_m^{(k)}\left(\mathcal{H}'\right),$$

so that $\forall x \in \mathcal{X}$,

$$\begin{aligned}
\hat{\Delta}_m^{(k)}(x, W_2, \mathcal{H}) &= M_m^{(k)}(\mathcal{H})^{-1} \sum_{i=1}^{m^3} \mathbb{1}_{\mathcal{S}^k(\mathcal{H})}\left(S_i^{(k)} \cup \{x\}\right) \\
&\leq M_m^{(k)}(\mathcal{H})^{-1} \sum_{i=1}^{m^3} \mathbb{1}_{\mathcal{S}^k(\mathcal{H}')}\left(S_i^{(k)} \cup \{x\}\right) \\
&\leq (3/2) M_m^{(k)}\left(\mathcal{H}'\right)^{-1} \sum_{i=1}^{m^3} \mathbb{1}_{\mathcal{S}^k(\mathcal{H}')}\left(S_i^{(k)} \cup \{x\}\right) = (3/2)\hat{\Delta}_m^{(k)}\left(x, W_2, \mathcal{H}'\right).
\end{aligned}$$

The final claim follows from Lemma 29. ■

Lemma 50 *For any $k \in \{1, \ldots, d+1\}$, if $n \geq 3 \cdot 4^{k-1}$, then $T_k^\star \geq 4^{1-k}(2n/3)$ and $\hat{t}_k \leq \lfloor 3T_k^\star/4 \rfloor$.*

Proof Recall $T_1^\star = \lceil 2n/3 \rceil \geq 2n/3$. If $n \geq 2$, we also have $\lfloor 3T_1^\star/4 \rfloor \geq \lceil T_1^\star/4 \rceil$, so that (due to the $C_{m1}^\star$ factors) $\hat{t}_1 \leq \lfloor 3T_1^\star/4 \rfloor$. For the purpose of induction, suppose some $k \in \{2, \ldots, d+1\}$ has $n \geq 3 \cdot 4^{k-1}$, $T_{k-1}^\star \geq 4^{2-k}(2n/3)$, and $\hat{t}_{k-1} \leq \lfloor 3T_{k-1}^\star/4 \rfloor$. Then $T_k^\star = T_{k-1}^\star - \hat{t}_{k-1} \geq T_{k-1}^\star/4 \geq 4^{1-k}(2n/3)$, and since $n \geq 3 \cdot 4^{k-1}$, we also have $\lfloor 3T_k^\star/4 \rfloor \geq \lceil T_k^\star/4 \rceil$, so that $\hat{t}_k \leq \lfloor 3T_k^\star/4 \rfloor$ (again, due to the $C_{mk}^\star$ factors). Thus, by induction, this holds for all $k \in \{1, \ldots, d+1\}$ with $n \geq 3 \cdot 4^{k-1}$. ■

The next lemma indicates that the "$t < \lfloor 3T_k/4 \rfloor$" constraint in Step 12 is redundant for $k \leq \tilde{d}_f$. It is similar to (50) in Lemma 45, but is made only slightly more complicated by the fact that the $\hat{\Delta}^{(k)}$ estimate is calculated in Step 9 based on a set V_m different from the ones used to decide whether or not to request a label in Step 12.

Lemma 51 *There exist $(\mathbb{C}, \mathcal{P}, f, \gamma)$-dependent constants $\tilde{c}_1^{(i)}, \tilde{c}_2^{(i)} \in [1, \infty)$ such that, for any $\delta \in (0,1)$, and any integer $n \geq \tilde{c}_1^{(i)} \ln\left(\tilde{c}_2^{(i)}/\delta\right)$, on an event*

$$\tilde{H}_n^{(i)}(\delta) \subseteq G_{\tilde{m}_n}^{(i)} \cap H_{\tilde{m}_n}(\delta) \cap H_{\tilde{m}_n}^{(i)} \cap H_{\tilde{m}_n}^{(iii)}(\gamma/16) \cap H_{\tilde{m}_n}^{(iv)}$$

with $\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\right) \geq 1 - 2\delta$, $\forall k \in \left\{1, \ldots, \tilde{d}_f\right\}$, $\hat{t}_k = \sum_{m=\hat{m}_{k-1}+1}^{\hat{m}_k} I_{mk}^\star \leq 3T_k^\star/4$.

Proof Define the constants

$$\tilde{c}_1^{(i)} = \max\left\{\tfrac{192d}{r_{(3/32)}}, \tfrac{3 \cdot 4^{\tilde{d}_f+6}}{\tilde{\delta}_f \gamma^2}\right\}, \quad \tilde{c}_2^{(i)} = \max\left\{\frac{8e}{r_{(3/32)}}, \left(c^{(i)} + c^{(iii)}(\gamma/16) + 125\tilde{d}_f\tilde{\delta}_f^{-1}\right)\right\},$$

and let $n^{(i)}(\delta) = \tilde{c}_1^{(i)} \ln\left(\tilde{c}_2^{(i)}/\delta\right)$. Fix any integer $n \geq n^{(i)}(\delta)$ and consider the event

$$\tilde{H}_n^{(1)}(\delta) = G_{\tilde{m}_n}^{(i)} \cap H_{\tilde{m}_n}(\delta) \cap H_{\tilde{m}_n}^{(i)} \cap H_{\tilde{m}_n}^{(iii)}(\gamma/16) \cap H_{\tilde{m}_n}^{(iv)}.$$

By Lemma 49 and the fact that $\check{m}_k \geq \tilde{m}_n$ for all $k \geq 1$, since $n \geq n^{(i)}(\delta) \geq 24\tau(1/6;\delta)$, on $\tilde{H}_n^{(1)}(\delta)$, $\forall k \in \left\{1,\ldots,\tilde{d}_f\right\}$, $\forall m \in \hat{\mathcal{U}}_k$,

$$\hat{\Delta}_m^{(k)}\left(X_m, W_2, V_{m-1}^{\star}\right) \leq (3/2)\hat{\Delta}_m^{(k)}\left(X_m, W_2, V_{\check{m}_k}^{\star}\right). \tag{67}$$

Now fix any $k \in \left\{1,\ldots,\tilde{d}_f\right\}$. Since $n \geq n^{(i)}(\delta) \geq 27 \cdot 4^{k-1}$, Lemma 50 implies $T_k^{\star} \geq 18$, which means $3T_k^{\star}/4 - \lceil T_k^{\star}/4 \rceil \geq 4T_k^{\star}/9$. Also note $\sum_{m \in \check{\mathcal{U}}_k} I_{mk}^{\star} \leq \lceil T_k^{\star}/4 \rceil$. Let $N_k = (4/3)\hat{\Delta}_{\check{m}_k}^{(k)}\left(W_1, W_2, V_{\check{m}_k}^{\star}\right)\left|\hat{\mathcal{U}}_k\right|$, and note that $\left|\hat{\mathcal{U}}_k\right| = \left\lfloor T_k^{\star}/\left(3\hat{\Delta}_{\check{m}_k}^{(k)}\left(W_1, W_2, V_{\check{m}_k}^{\star}\right)\right)\right\rfloor$, so that $N_k \leq (4/9)T_k^{\star}$. Thus, we have

$$\begin{aligned}
&\mathbb{P}\left(\tilde{H}_n^{(1)}(\delta) \cap \left\{\sum_{m=\hat{m}_{k-1}+1}^{\hat{m}_k} I_{mk}^{\star} > 3T_k^{\star}/4\right\}\right)\\
&\quad \leq \mathbb{P}\left(\tilde{H}_n^{(1)}(\delta) \cap \left\{\sum_{m \in \hat{\mathcal{U}}_k} I_{mk}^{\star} > 4T_k^{\star}/9\right\}\right) \leq \mathbb{P}\left(\tilde{H}_n^{(1)}(\delta) \cap \left\{\sum_{m \in \hat{\mathcal{U}}_k} I_{mk}^{\star} > N_k\right\}\right)\\
&\quad \leq \mathbb{P}\left(\tilde{H}_n^{(1)}(\delta) \cap \left\{\sum_{m \in \hat{\mathcal{U}}_k} \mathbb{1}_{[2\gamma/3,\infty)}\left(\hat{\Delta}_m^{(k)}\left(X_m, W_2, V_{\check{m}_k}^{\star}\right)\right) > N_k\right\}\right),
\end{aligned} \tag{68}$$

where this last inequality is by (67). To simplify notation, define $\tilde{Z}_k = \left(T_k^{\star}, \check{m}_k, W_1, W_2, V_{\check{m}_k}^{\star}\right)$. By Lemmas 43 and 44 (with $\beta = 3/32$, $\zeta = 2\gamma/3$, $\alpha = 3/4$, and $\xi = \gamma/16$), since $n \geq n^{(i)}(\delta) \geq 24 \cdot \max\left\{\tau^{(iv)}(\gamma/16;\delta), \tau(3/32;\delta)\right\}$, on $\tilde{H}_n^{(1)}(\delta)$, $\forall m \in \hat{\mathcal{U}}_k$,

$$\begin{aligned}
\bar{p}_{2\gamma/3}(k, \check{m}_k, m) &\leq \mathcal{P}\left(x : p_x(k, \check{m}_k) \geq \gamma/2\right) + \exp\left\{-\gamma^2 \tilde{M}(m)/256\right\}\\
&\leq \mathcal{P}\left(x : p_x(k, \check{m}_k) \geq \gamma/2\right) + \exp\left\{-\gamma^2 \tilde{M}(\check{m}_k)/256\right\}\\
&\leq \hat{\Delta}_{\check{m}_k}^{(k)}\left(W_1, W_2, V_{\check{m}_k}^{\star}\right).
\end{aligned}$$

Letting $\tilde{G}_n'(k)$ denote the event $\bar{p}_{2\gamma/3}(k, \check{m}_k, m) \leq \hat{\Delta}_{\check{m}_k}^{(k)}\left(W_1, W_2, V_{\check{m}_k}^{\star}\right)$, we see that $\tilde{G}_n'(k) \supseteq \tilde{H}_n^{(1)}(\delta)$. Thus, since the $\mathbb{1}_{[2\gamma/3,\infty)}\left(\hat{\Delta}_m^{(k)}\left(X_m, W_2, V_{\check{m}_k}^{\star}\right)\right)$ variables are conditionally independent given $\tilde{Z}_k$ for $m \in \hat{\mathcal{U}}_k$, each with respective conditional distribution Bernoulli$\left(\bar{p}_{2\gamma/3}(k, \check{m}_k, m)\right)$, the law of total probability and a Chernoff bound imply that (68) is at most

$$\begin{aligned}
&\mathbb{P}\left(\tilde{G}_n'(k) \cap \left\{\sum_{m \in \hat{\mathcal{U}}_k} \mathbb{1}_{[2\gamma/3,\infty)}\left(\hat{\Delta}_m^{(k)}\left(X_m, W_2, V_{\check{m}_k}^{\star}\right)\right) > N_k\right\}\right)\\
&= \mathbb{E}\left[\mathbb{P}\left(\sum_{m \in \hat{\mathcal{U}}_k} \mathbb{1}_{[2\gamma/3,\infty)}\left(\hat{\Delta}_m^{(k)}\left(X_m, W_2, V_{\check{m}_k}^{\star}\right)\right) > N_k \middle| \tilde{Z}_k\right) \cdot \mathbb{1}_{\tilde{G}_n'(k)}\right]\\
&\leq \mathbb{E}\left[\exp\left\{-\hat{\Delta}_{\check{m}_k}^{(k)}\left(W_1, W_2, V_{\check{m}_k}^{\star}\right)\left|\hat{\mathcal{U}}_k\right|/27\right\}\right] \leq \mathbb{E}\left[\exp\{-T_k^{\star}/162\}\right] \leq \exp\left\{-n/\left(243 \cdot 4^{k-1}\right)\right\},
\end{aligned}$$

where the last inequality is by Lemma 50. Thus, there exists $\tilde{G}_n(k)$ with $\mathbb{P}\left(\tilde{H}_n^{(1)}(\delta)\setminus\tilde{G}_n(k)\right) \leq \exp\left\{-n/\left(243\cdot 4^{k-1}\right)\right\}$ such that, on $\tilde{H}_n^{(1)}(\delta)\cap\tilde{G}_n(k)$, we have $\sum_{m=\hat{m}_{k-1}+1}^{\hat{m}_k} I^{\star}_{mk} \leq 3T^{\star}_k/4$. Defining $\tilde{H}_n^{(i)}(\delta) = \tilde{H}_n^{(1)}(\delta)\cap\bigcap_{k=1}^{\tilde{d}_f}\tilde{G}_n(k)$, a union bound implies

$$\mathbb{P}\left(\tilde{H}_n^{(1)}(\delta)\setminus\tilde{H}_n^{(i)}(\delta)\right) \leq \tilde{d}_f\cdot\exp\left\{-n/\left(243\cdot 4^{\tilde{d}_f-1}\right)\right\}, \tag{69}$$

and on $\tilde{H}_n^{(i)}(\delta)$, every $k\in\left\{1,\ldots,\tilde{d}_f\right\}$ has $\sum_{m=\hat{m}_{k-1}+1}^{\hat{m}_k} I^{\star}_{mk} \leq 3T^{\star}_k/4$. In particular, this means the $C^{\star}_{mk}$ factors are redundant in $Q^{\star}_k$, so that $\hat{t}_k = \sum_{m=\hat{m}_{k-1}+1}^{\hat{m}_k} I^{\star}_{mk}$.

To get the stated probability bound, a union bound implies that

$$\begin{aligned}
1-\mathbb{P}\left(\tilde{H}_n^{(1)}(\delta)\right) &\leq \left(1-\mathbb{P}\left(H_{\tilde{m}_n}(\delta)\right)\right) + \left(1-\mathbb{P}\left(H_{\tilde{m}_n}^{(i)}\right)\right) + \mathbb{P}\left(H_{\tilde{m}_n}^{(i)}\setminus H_{\tilde{m}_n}^{(iii)}(\gamma/16)\right)\\
&\quad + \left(1-\mathbb{P}\left(H_{\tilde{m}_n}^{(iv)}\right)\right) + \mathbb{P}\left(H_{\tilde{m}_n}^{(i)}\setminus G_{\tilde{m}_n}^{(i)}\right)\\
&\leq \delta + c^{(i)}\cdot\exp\left\{-\tilde{M}\left(\tilde{m}_n\right)/4\right\}\\
&\quad + c^{(iii)}(\gamma/16)\cdot\exp\left\{-\tilde{M}\left(\tilde{m}_n\right)\gamma^2/256\right\} + 3\tilde{d}_f\cdot\exp\left\{-2\tilde{m}_n\right\}\\
&\quad + 121\tilde{d}_f\tilde{\delta}_f^{-1}\cdot\exp\left\{-\tilde{M}\left(\tilde{m}_n\right)/60\right\}\\
&\leq \delta + \left(c^{(i)} + c^{(iii)}(\gamma/16) + 124\tilde{d}_f\tilde{\delta}_f^{-1}\right)\cdot\exp\left\{-\tilde{m}_n\tilde{\delta}_f\gamma^2/512\right\}.
\end{aligned} \tag{70}$$

Since $n \geq n^{(i)}(\delta) \geq 24$, we have $\tilde{m}_n \geq n/48$, so that summing (69) and (70) gives us

$$1-\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\right) \leq \delta + \left(c^{(i)} + c^{(iii)}(\gamma/16) + 125\tilde{d}_f\tilde{\delta}_f^{-1}\right)\cdot\exp\left\{-n\tilde{\delta}_f\gamma^2/\left(512\cdot 48\cdot 4^{\tilde{d}_f-1}\right)\right\}. \tag{71}$$

Finally, note that we have chosen $n^{(i)}(\delta)$ sufficiently large so that (71) is at most 2δ. ■

The next lemma indicates that the redundancy of the "$t < \lfloor 3T_k/4\rfloor$" constraint, just established in Lemma 51, implies that all $\hat{y}$ labels obtained while $k \leq \tilde{d}_f$ are consistent with the target function.

Lemma 52 *Consider running Meta-Algorithm 3 with a budget $n\in\mathbb{N}$, while f is the target function and $\mathcal{P}$ is the data distribution. There is an event $\tilde{H}_n^{(ii)}$ and $(\mathbb{C},\mathcal{P},f,\gamma)$-dependent constants $\tilde{c}_1^{(ii)},\tilde{c}_2^{(ii)}\in[1,\infty)$ such that, for any $\delta\in(0,1)$, if $n\geq\tilde{c}_1^{(ii)}\ln\left(\tilde{c}_2^{(ii)}/\delta\right)$, then $\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\setminus\tilde{H}_n^{(ii)}\right)\leq\delta$, and on $\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}$, we have $V^{(\tilde{d}_f)} = V_{\hat{m}_{\tilde{d}_f}} = V^{\star}_{\hat{m}_{\tilde{d}_f}}$.*

Proof Define $\tilde{c}_1^{(ii)} = \max\left\{\tilde{c}_1^{(i)}, \frac{192d}{r_{(1-\gamma)/6}}, \frac{2^{11}}{\tilde{\delta}_f^{1/3}}\right\}$, $\tilde{c}_2^{(ii)} = \max\left\{\tilde{c}_2^{(i)}, \frac{8e}{r_{(1-\gamma)/6}}, c^{(ii)}, \exp\left\{\tau^*\right\}\right\}$, let $n^{(ii)}(\delta) = \tilde{c}_1^{(ii)}\ln\left(\tilde{c}_2^{(ii)}/\delta\right)$, suppose $n\geq n^{(ii)}(\delta)$, and define the event $\tilde{H}_n^{(ii)} = H_{\tilde{m}_n}^{(ii)}$.

By Lemma 41, since $n\geq n^{(ii)}(\delta)\geq 24\cdot\max\left\{\tau((1-\gamma)/6;\delta),\tau^*\right\}$, on $\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}$, $\forall m\in\mathbb{N}$ and $k\in\left\{1,\ldots,\tilde{d}_f\right\}$ with either $k=1$ or $m>\tilde{m}_n$,

$$\hat{\Delta}_m^{(k)}\left(X_m,W_2,V^{\star}_{m-1}\right) < \gamma \Rightarrow \hat{\Gamma}_m^{(k)}\left(X_m,-f(X_m),W_2,V^{\star}_{m-1}\right) < \hat{\Gamma}_m^{(k)}\left(X_m,f(X_m),W_2,V^{\star}_{m-1}\right). \tag{72}$$

Recall that $\tilde{m}_n \le \min\{\lceil T_1/4 \rceil, 2^n\} = \lceil \lceil 2n/3 \rceil / 4 \rceil$. Therefore, $V_{\tilde{m}_n}$ is obtained purely by $\tilde{m}_n$ executions of Step 8 while $k = 1$. Thus, for every m obtained in Meta-Algorithm 3, either $k = 1$ or $m > \tilde{m}_n$. We now proceed by induction on m. We already know $V_0 = \mathbb{C} = V_0^\star$, so this serves as our base case. Now consider some value $m \in \mathbb{N}$ obtained in Meta-Algorithm 3 while $k \le \tilde{d}_f$, and suppose every $m' < m$ has $V_{m'} = V_{m'}^\star$. But this means that $T_k = T_k^\star$ and the value of t upon obtaining this particular m has $t \le \sum_{\ell=\hat{m}_{k-1}+1}^{m-1} I_{\ell k}^\star$. In particular, if $\hat{\Delta}_m^{(k)}(X_m, W_2, V_{m-1}) \ge \gamma$, then $I_{mk}^\star = 1$, so that $t < \sum_{\ell=\hat{m}_{k-1}+1}^{m} I_{mk}^\star$; by Lemma 51, on $\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)}$, $\sum_{\ell=\hat{m}_{k-1}+1}^{m} I_{mk}^\star \le \sum_{\ell=\hat{m}_{k-1}+1}^{\hat{m}_k} I_{mk}^\star \le 3T_k^\star/4$, so that $t < 3T_k^\star/4$, and therefore $\hat{y} = Y_m = f(X_m)$; this implies $V_m = V_m^\star$. On the other hand, on $\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)}$, if $\hat{\Delta}_m^{(k)}(X_m, W_2, V_{m-1}) < \gamma$, then (72) implies

$$\hat{y} = \underset{y \in \{-1,+1\}}{\operatorname{argmax}} \hat{\Gamma}_m^{(k)}(X_m, y, W_2, V_{m-1}) = f(X_m),$$

so that again $V_m = V_m^\star$. Thus, by the principle of induction, on $\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)}$, for every $m \in \mathbb{N}$ obtained while $k \le \tilde{d}_f$, we have $V_m = V_m^\star$; in particular, this implies $V^{(\tilde{d}_f)} = V_{\hat{m}_{\tilde{d}_f}} = V_{\hat{m}_{\tilde{d}_f}}^\star$. The bound on $\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta) \setminus \tilde{H}_n^{(ii)}\right)$ then follows from Lemma 41, as we have chosen $n^{(ii)}(\delta)$ sufficiently large so that (27) (with $\tau = \tilde{m}_n$) is at most δ. ■

Lemma 53 *Consider running Meta-Algorithm 3 with a budget $n \in \mathbb{N}$, while f is the target function and $\mathcal{P}$ is the data distribution. There exist $(\mathbb{C}, \mathcal{P}, f, \gamma)$-dependent constants $\tilde{c}_1^{(iii)}, \tilde{c}_2^{(iii)} \in [1, \infty)$ such that, for any $\delta \in (0, e^{-3})$, $\lambda \in [1, \infty)$, and $n \in \mathbb{N}$, there exists an event $\tilde{H}_n^{(iii)}(\delta, \lambda)$ having $\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)} \setminus \tilde{H}_n^{(iii)}(\delta, \lambda)\right) \le \delta$ with the property that, if*

$$n \ge \tilde{c}_1^{(iii)} \tilde{\theta}_f(d/\lambda) \ln^2\left(\frac{\tilde{c}_2^{(iii)} \lambda}{\delta}\right),$$

then on $\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)} \cap \tilde{H}_n^{(iii)}(\delta, \lambda)$, at the conclusion of Meta-Algorithm 3, $\left|\mathcal{L}_{\tilde{d}_f}\right| \ge \lambda$.

Proof Let $\tilde{c}_1^{(iii)} = \max\left\{\tilde{c}_1^{(i)}, \tilde{c}_1^{(ii)}, \frac{d \cdot \tilde{d}_f \cdot 4^{10+2\tilde{d}_f}}{\gamma^3 \tilde{\delta}_f^3}, \frac{192d}{r_{(3/32)}}\right\}$, $\tilde{c}_2^{(iii)} = \max\left\{\tilde{c}_2^{(i)}, \tilde{c}_2^{(ii)}, \frac{8e}{r_{(3/32)}}\right\}$, fix any $\delta \in (0, e^{-3})$, $\lambda \in [1, \infty)$, let $n^{(iii)}(\delta, \lambda) = \tilde{c}_1^{(iii)} \tilde{\theta}_f(d/\lambda) \ln^2(\tilde{c}_2^{(iii)} \lambda/\delta)$, and suppose $n \ge n^{(iii)}(\delta, \lambda)$.

Define a sequence $\ell_i = 2^i$ for integers $i \ge 0$, and let $\hat{\iota} = \left\lceil \log_2\left(4^{2+\tilde{d}_f} \lambda / \gamma \tilde{\delta}_f\right) \right\rceil$. Also define $\tilde{\phi}(m, \delta, \lambda) = \max\{\phi(m; \delta/2\hat{\iota}), d/\lambda\}$, where ϕ is defined in Lemma 29. Then define the events

$$\tilde{H}^{(3)}(\delta, \lambda) = \bigcap_{i=1}^{\hat{\iota}} H_{\ell_i}(\delta/2\hat{\iota}), \quad \tilde{H}_n^{(iii)}(\delta, \lambda) = \tilde{H}^{(3)}(\delta, \lambda) \cap \left\{\check{m}_{\tilde{d}_f} \ge \ell_{\hat{\iota}}\right\}.$$

Note that $\hat{\iota} \le n$, so that $\ell_{\hat{\iota}} \le 2^n$, and therefore the truncation in the definition of $\check{m}_{\tilde{d}_f}$, which enforces $\check{m}_{\tilde{d}_f} \le \max\left\{\tilde{d}_f \cdot 2^n + 1, \hat{m}_{k-1}\right\}$, will never be a factor in whether or not $\check{m}_{\tilde{d}_f} \ge \ell_{\hat{\iota}}$ is satisfied.

Since $n \ge n^{(iii)}(\lambda, \delta) \ge \tilde{c}_1^{(ii)} \ln\left(\tilde{c}_2^{(ii)}/\delta\right)$, Lemma 52 implies that on $\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)}$, $V_{\hat{m}_{\tilde{d}_f}} = V_{\hat{m}_{\tilde{d}_f}}^\star$. Recall that this implies that all $\hat{y}$ values obtained while $m \le \hat{m}_{\tilde{d}_f}$ are consistent with their respective

$f(X_m)$ values, so that every such m has $V_m = V_m^\star$ as well. In particular, $V_{\check{m}_{\tilde{d}_f}} = V^\star_{\check{m}_{\tilde{d}_f}}$. Also note that $n^{(iii)}(\delta,\lambda) \geq 24 \cdot \tau^{(iv)}(\gamma/16;\delta)$, so that $\tau^{(iv)}(\gamma/16;\delta) \leq \tilde{m}_n$, and recall we always have $\tilde{m}_n \leq \check{m}_{\tilde{d}_f}$. Thus, on $\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)} \cap \tilde{H}_n^{(iii)}(\delta,\lambda)$, (taking $\hat{\Delta}^{(k)}$ as in Meta-Algorithm 3)

$$\begin{aligned}
\hat{\Delta}^{(\tilde{d}_f)} &= \hat{\Delta}^{(\tilde{d}_f)}_{\check{m}_{\tilde{d}_f}}\left(W_1, W_2, V^\star_{\check{m}_{\tilde{d}_f}}\right) && \text{(Lemma 52)}\\
&\leq \mathcal{P}\left(x : p_x\left(\tilde{d}_f, \check{m}_{\tilde{d}_f}\right) \geq \gamma/8\right) + 4\check{m}_{\tilde{d}_f}^{-1} && \text{(Lemma 44)}\\
&\leq \frac{8\mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}\left(V^\star_{\check{m}_{\tilde{d}_f}}\right)\right)}{\gamma\mathcal{P}^{\tilde{d}_f-1}\left(\mathcal{S}^{\tilde{d}_f-1}\left(V^\star_{\check{m}_{\tilde{d}_f}}\right)\right)} + 4\check{m}_{\tilde{d}_f}^{-1} && \text{(Markov's ineq.)}\\
&\leq \left(8/\gamma\tilde{\delta}_f\right)\mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}\left(V^\star_{\check{m}_{\tilde{d}_f}}\right)\right) + 4\check{m}_{\tilde{d}_f}^{-1} && \text{(Lemma 35)}\\
&\leq \left(8/\gamma\tilde{\delta}_f\right)\mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}\left(V^\star_{\ell_{\hat{\imath}}}\right)\right) + 4\ell_{\hat{\imath}}^{-1} && \text{(defn of } \tilde{H}_n^{(iii)}(\delta,\lambda))\\
&\leq \left(8/\gamma\tilde{\delta}_f\right)\mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}\left(\mathrm{B}\left(f, \tilde{\phi}\left(\ell_{\hat{\imath}}, \delta, \lambda\right)\right)\right)\right) + 4\ell_{\hat{\imath}}^{-1} && \text{(Lemma 29)}\\
&\leq \left(8/\gamma\tilde{\delta}_f\right)\tilde{\theta}_f(d/\lambda)\tilde{\phi}\left(\ell_{\hat{\imath}}, \delta, \lambda\right) + 4\ell_{\hat{\imath}}^{-1} && \text{(defn of } \tilde{\theta}_f(d/\lambda))\\
&\leq \left(12/\gamma\tilde{\delta}_f\right)\tilde{\theta}_f(d/\lambda)\tilde{\phi}\left(\ell_{\hat{\imath}}, \delta, \lambda\right) && (\tilde{\phi}\left(\ell_{\hat{\imath}}, \delta, \lambda\right) \geq \ell_{\hat{\imath}}^{-1})\\
&= \frac{12\tilde{\theta}_f(d/\lambda)}{\gamma\tilde{\delta}_f}\max\left\{2\frac{d\ln\left(2e\max\left\{\ell_{\hat{\imath}}, d\right\}/d\right) + \ln\left(4\hat{\imath}/\delta\right)}{\ell_{\hat{\imath}}}, d/\lambda\right\}. && (73)
\end{aligned}$$

Plugging in the definition of $\hat{\imath}$ and $\ell_{\hat{\imath}}$,

$$\frac{d\ln\left(2e\max\left\{\ell_{\hat{\imath}}, d\right\}/d\right) + \ln\left(4\hat{\imath}/\delta\right)}{\ell_{\hat{\imath}}} \leq (d/\lambda)\gamma\tilde{\delta}_f 4^{-1-\tilde{d}_f}\ln\left(4^{1+\tilde{d}_f}\lambda/\delta\gamma\tilde{\delta}_f\right) \leq (d/\lambda)\ln\left(\lambda/\delta\right).$$

Therefore, (73) is at most $24\tilde{\theta}_f(d/\lambda)(d/\lambda)\ln\left(\lambda/\delta\right)/\gamma\tilde{\delta}_f$. Thus, since

$$n^{(iii)}(\delta,\lambda) \geq \max\left\{\tilde{c}_1^{(i)}\ln\left(\tilde{c}_2^{(i)}/\delta\right), \tilde{c}_1^{(ii)}\ln\left(\tilde{c}_2^{(ii)}/\delta\right)\right\},$$

Lemmas 51 and 52 imply that on $\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)} \cap \tilde{H}_n^{(iii)}(\delta,\lambda)$,

$$\begin{aligned}
\left|\mathcal{L}_{\tilde{d}_f}\right| = \left\lfloor T^\star_{\tilde{d}_f}/\left(3\hat{\Delta}^{(\tilde{d}_f)}\right)\right\rfloor &\geq \left\lfloor 4^{1-\tilde{d}_f}2n/\left(9\hat{\Delta}^{(\tilde{d}_f)}\right)\right\rfloor\\
&\geq \frac{4^{1-\tilde{d}_f}\gamma\tilde{\delta}_f n}{9\cdot 24\cdot\tilde{\theta}_f(d/\lambda)(d/\lambda)\ln(\lambda/\delta)} \geq \lambda\ln(\lambda/\delta) \geq \lambda.
\end{aligned}$$

Now we turn to bounding $\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)} \setminus \tilde{H}_n^{(iii)}(\delta,\lambda)\right)$. By a union bound, we have

$$1 - \mathbb{P}\left(\tilde{H}^{(3)}(\delta,\lambda)\right) \leq \sum_{i=1}^{\hat{\imath}}\left(1 - \mathbb{P}\left(H_{\ell_i}\left(\delta/2\hat{\imath}\right)\right)\right) \leq \delta/2. \tag{74}$$

Thus, it remains only to bound $\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)\cap\left\{\check{m}_{\tilde{d}_f}<\ell_{\hat{\imath}}\right\}\right)$.

For each $i\in\{0,1,\ldots,\hat{\imath}-1\}$, let $\check{Q}_i=\left|\left\{m\in(\ell_i,\ell_{i+1}]\cap\check{\mathcal{U}}_{\tilde{d}_f}:I^{\star}_{m\tilde{d}_f}=1\right\}\right|$. Now consider the set $\mathcal{I}$ of all $i\in\{0,1,\ldots,\hat{\imath}-1\}$ with $\ell_i\geq\tilde{m}_n$ and $(\ell_i,\ell_{i+1}]\cap\check{\mathcal{U}}_{\tilde{d}_f}\neq\emptyset$. Note that $n^{(iii)}(\delta,\lambda)\geq 48$, so that $\ell_0<\tilde{m}_n$. Fix any $i\in\mathcal{I}$. Since $n^{(iii)}(\lambda,\delta)\geq 24\cdot\tau(1/6;\delta)$, we have $\tilde{m}_n\geq\tau(1/6;\delta)$, so that Lemma 49 implies that on $\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)$, letting $\bar{Q}=2\cdot 4^{6+\tilde{d}_f}\left(d/\gamma^2\tilde{\delta}_f^2\right)\tilde{\theta}_f(d/\lambda)\ln(\lambda/\delta)$,

$$\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)\cap\left\{\check{Q}_i>\bar{Q}\right\}\middle|W_2,V^{\star}_{\ell_i}\right)$$
$$\leq\mathbb{P}\left(\left|\left\{m\in(\ell_i,\ell_{i+1}]\cap\mathbb{N}:\hat{\Delta}_m^{(\tilde{d}_f)}\left(X_m,W_2,V^{\star}_{\ell_i}\right)\geq 2\gamma/3\right\}\right|>\bar{Q}\middle|W_2,V^{\star}_{\ell_i}\right).\quad(75)$$

For $m>\ell_i$, the variables $\mathbb{1}_{[2\gamma/3,\infty)}\left(\hat{\Delta}_m^{(\tilde{d}_f)}\left(X_m,W_2,V^{\star}_{\ell_i}\right)\right)$ are conditionally (given $W_2,V^{\star}_{\ell_i}$) independent, each with respective conditional distribution Bernoulli with mean $\bar{p}_{2\gamma/3}\left(\tilde{d}_f,\ell_i,m\right)$. Since $n^{(iii)}(\delta,\lambda)\geq 24\cdot\tau(3/32;\delta)$, we have $\tilde{m}_n\geq\tau(3/32;\delta)$, so that Lemma 43 (with $\zeta=2\gamma/3$, $\alpha=3/4$, and $\beta=3/32$) implies that on $\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)$, each of these m values has

$$\begin{aligned}
\bar{p}_{2\gamma/3}\left(\tilde{d}_f,\ell_i,m\right)&\leq\mathcal{P}\left(x:p_x\left(\tilde{d}_f,\ell_i\right)\geq\gamma/2\right)+\exp\left\{-\tilde{M}(m)\gamma^2/256\right\}\\
&\leq\frac{2\mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}\left(V^{\star}_{\ell_i}\right)\right)}{\gamma\mathcal{P}^{\tilde{d}_f-1}\left(\mathcal{S}^{\tilde{d}_f-1}\left(V^{\star}_{\ell_i}\right)\right)}+\exp\left\{-\tilde{M}(\ell_i)\gamma^2/256\right\} && \text{(Markov's ineq.)}\\
&\leq\left(2/\gamma\tilde{\delta}_f\right)\mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}\left(V^{\star}_{\ell_i}\right)\right)+\exp\left\{-\tilde{M}(\ell_i)\gamma^2/256\right\} && \text{(Lemma 35)}\\
&\leq\left(2/\gamma\tilde{\delta}_f\right)\mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}\left(\mathrm{B}\left(f,\tilde{\phi}(\ell_i,\delta,\lambda)\right)\right)\right)+\exp\left\{-\tilde{M}(\ell_i)\gamma^2/256\right\} && \text{(Lemma 29)}\\
&\leq\left(2/\gamma\tilde{\delta}_f\right)\tilde{\theta}_f(d/\lambda)\tilde{\phi}(\ell_i,\delta,\lambda)+\exp\left\{-\tilde{M}(\ell_i)\gamma^2/256\right\} && \text{(defn of }\tilde{\theta}_f(d/\lambda)).
\end{aligned}$$

Denote the expression in this last line by p_i, and let $\mathbf{B}(\ell_i,p_i)$ be a Binomial(ℓ_i,p_i) random variable. Noting that $\ell_{i+1}-\ell_i=\ell_i$, we have that on $\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)$, (75) is at most $\mathbb{P}\left(\mathbf{B}(\ell_i,p_i)>\bar{Q}\right)$. Next, note that

$$\ell_i p_i=(2/\gamma\tilde{\delta}_f)\tilde{\theta}_f(d/\lambda)\ell_i\tilde{\phi}(\ell_i,\delta,\lambda)+\ell_i\cdot\exp\left\{-\ell_i^3\tilde{\delta}_f\gamma^2/512\right\}.$$

Since $u\cdot\exp\left\{-u^3\right\}\leq(3e)^{-1/3}$ for any u, letting $u=\ell_i\tilde{\delta}_f\gamma/8$ we have

$$\ell_i\cdot\exp\left\{-\ell_i^3\tilde{\delta}_f\gamma^2/512\right\}\leq\left(8/\gamma\tilde{\delta}_f\right)u\cdot\exp\left\{-u^3\right\}\leq 8/\left(\gamma\tilde{\delta}_f(3e)^{1/3}\right)\leq 4/\gamma\tilde{\delta}_f.$$

Therefore, since $\tilde{\phi}(\ell_i,\delta,\lambda) \geq \ell_i^{-1}$, we have that $\ell_i p_i$ is at most

$$\begin{aligned}
\frac{6}{\gamma\tilde{\delta}_f}\tilde{\theta}_f(d/\lambda)\ell_i\tilde{\phi}(\ell_i,\delta,\lambda) &\leq \frac{6}{\gamma\tilde{\delta}_f}\tilde{\theta}_f(d/\lambda)\max\left\{2d\ln(2e\ell_{\hat{\imath}})+2\ln\left(\frac{4\hat{\imath}}{\delta}\right),\ell_{\hat{\imath}}d/\lambda\right\}\\
&\leq \frac{6}{\gamma\tilde{\delta}_f}\tilde{\theta}_f(d/\lambda)\max\left\{2d\ln\left(\frac{4^{3+\tilde{d}_f}e\lambda}{\gamma\tilde{\delta}_f}\right)+2\ln\left(\frac{4^{3+\tilde{d}_f}2\lambda}{\gamma\tilde{\delta}_f\delta}\right),\frac{d4^{3+\tilde{d}_f}}{\gamma\tilde{\delta}_f}\right\}\\
&\leq \frac{6}{\gamma\tilde{\delta}_f}\tilde{\theta}_f(d/\lambda)\max\left\{4d\ln\left(\frac{4^{3+\tilde{d}_f}\lambda}{\gamma\tilde{\delta}_f\delta}\right),\frac{d4^{3+\tilde{d}_f}}{\gamma\tilde{\delta}_f}\right\}\\
&\leq \frac{6}{\gamma\tilde{\delta}_f}\tilde{\theta}_f(d/\lambda)\cdot\frac{d4^{4+\tilde{d}_f}}{\gamma\tilde{\delta}_f}\ln\left(\frac{\lambda}{\delta}\right) \leq \frac{4^{6+\tilde{d}_f}d}{\gamma^2\tilde{\delta}_f^2}\tilde{\theta}_f(d/\lambda)\ln\left(\frac{\lambda}{\delta}\right) = \bar{Q}/2.
\end{aligned}$$

Therefore, a Chernoff bound implies $\mathbb{P}\left(\mathbf{B}(\ell_i,p_i) > \bar{Q}\right) \leq \exp\left\{-\bar{Q}/6\right\} \leq \delta/2\hat{\imath}$, so that on $\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)$, (75) is at most $\delta/2\hat{\imath}$. The law of total probability implies there exists an event $\tilde{H}_n^{(4)}(i,\delta,\lambda)$ with $\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)\setminus\tilde{H}_n^{(4)}(i,\delta,\lambda)\right) \leq \delta/2\hat{\imath}$ such that, on $\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)\cap\tilde{H}_n^{(4)}(i,\delta,\lambda)$, $\check{Q}_i \leq \bar{Q}$.

Note that

$$\begin{aligned}
\hat{\imath}\bar{Q} &\leq \log_2\left(4^{2+\tilde{d}_f}\lambda/\gamma\tilde{\delta}_f\right)\cdot 4^{7+\tilde{d}_f}\left(d/\gamma^2\tilde{\delta}_f^2\right)\tilde{\theta}_f(d/\lambda)\ln(\lambda/\delta)\\
&\leq \left(\tilde{d}_f 4^{9+\tilde{d}_f}/\gamma^3\tilde{\delta}_f^3\right)d\tilde{\theta}_f(d/\lambda)\ln^2(\lambda/\delta) \leq 4^{1-\tilde{d}_f}n/12. \qquad (76)
\end{aligned}$$

Since $\sum_{m\leq 2\tilde{m}_n} I^\star_{m\tilde{d}_f} \leq n/12$, if $\tilde{d}_f = 1$ then (76) implies that on the event $\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)\cap\bigcap_{i\in\mathcal{I}}\tilde{H}_n^{(4)}(i,\delta,\lambda)$, $\sum_{m\leq\ell_{\hat{\imath}}} I^\star_{m1} \leq n/12+\sum_{i\in\mathcal{I}}\check{Q}_i \leq n/12+\hat{\imath}\bar{Q} \leq n/6 \leq \lceil T_1^\star/4\rceil$, so that $\check{m}_1 \geq \ell_{\hat{\imath}}$. Otherwise, if $\tilde{d}_f > 1$, then every $m\in\check{\mathcal{U}}_{\tilde{d}_f}$ has $m > 2\tilde{m}_n$, so that $\sum_{i\leq\hat{\imath}}\check{Q}_i = \sum_{i\in\mathcal{I}}\check{Q}_i$; thus, on $\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)\cap\bigcap_{i\in\mathcal{I}}\tilde{H}_n^{(4)}(i,\delta,\lambda)$, $\sum_{i\in\mathcal{I}}\check{Q}_i \leq \hat{\imath}\bar{Q} \leq 4^{1-\tilde{d}_f}n/12$; Lemma 50 implies $4^{1-\tilde{d}_f}n/12 \leq \left\lceil T^\star_{\tilde{d}_f}/4\right\rceil$, so that again we have $\check{m}_{\tilde{d}_f} \geq \ell_{\hat{\imath}}$. Combined with a union bound, this implies

$$\begin{aligned}
&\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)\cap\left\{\check{m}_{\tilde{d}_f} < \ell_{\hat{\imath}}\right\}\right)\\
&\quad\leq \mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)\setminus\bigcap_{i\in\mathcal{I}}\tilde{H}_n^{(4)}(i,\delta,\lambda)\right)\\
&\quad\leq \sum_{i\in\mathcal{I}}\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\cap\tilde{H}^{(3)}(\delta,\lambda)\setminus\tilde{H}_n^{(4)}(i,\delta,\lambda)\right) \leq \delta/2. \qquad (77)
\end{aligned}$$

Therefore, $\mathbb{P}\left(\tilde{H}_n^{(i)}(\delta)\cap\tilde{H}_n^{(ii)}\setminus\tilde{H}_n^{(iii)}(\delta,\lambda)\right) \leq \delta$, obtained by summing (77) and (74). ■

Proof [Theorem 16] If $\Lambda_p(\varepsilon/4,f,\mathcal{P}) = \infty$ then the result trivially holds. Otherwise, suppose $\varepsilon\in(0,10e^{-3})$, let $\delta = \varepsilon/10$, $\lambda = \Lambda_p(\varepsilon/4,f,\mathcal{P})$, $\tilde{c}_2 = \max\left\{10\tilde{c}_2^{(i)},10\tilde{c}_2^{(ii)},10\tilde{c}_2^{(iii)},10e(d+1)\right\}$, and $\tilde{c}_1 = \max\left\{\tilde{c}_1^{(i)},\tilde{c}_1^{(ii)},\tilde{c}_1^{(iii)},2\cdot 6^3(d+1)\tilde{d}\ln(e(d+1))\right\}$, and consider running Meta-Algorithm

3 with passive algorithm $\mathcal{A}_p$ and budget $n \geq \tilde{c}_1 \tilde{\theta}_f(d/\lambda)\ln^2(\tilde{c}_2\lambda/\varepsilon)$, while f is the target function and $\mathcal{P}$ is the data distribution. On the event $\tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)} \cap \tilde{H}_n^{(iii)}(\delta,\lambda)$, Lemma 53 implies $\left|\mathcal{L}_{\tilde{d}_f}\right| \geq \lambda$, while Lemma 52 implies $V^{(\tilde{d}_f)} = V^\star_{\hat{m}_{\tilde{d}_f}}$; recalling that Lemma 35 implies that $V^\star_{\hat{m}_{\tilde{d}_f}} \neq \emptyset$ on this event, we must have $\mathrm{er}_{\mathcal{L}_{\tilde{d}_f}}(f) = 0$. Furthermore, if $\hat{h}$ is the classifier returned by Meta-Algorithm 3, then Lemma 34 implies that $\mathrm{er}(\hat{h})$ is at most $2\,\mathrm{er}(\mathcal{A}_p(\mathcal{L}_{\tilde{d}_f}))$, on a high probability event (call it $\hat{E}_2$ in this context). Letting $\hat{E}_3(\delta) = \hat{E}_2 \cap \tilde{H}_n^{(i)}(\delta) \cap \tilde{H}_n^{(ii)} \cap \tilde{H}_n^{(iii)}(\delta,\lambda)$, a union bound implies the total failure probability $1-\mathbb{P}(\hat{E}_3(\delta))$ from all of these events is at most $4\delta + e(d+1)\cdot\exp\left\{-\lfloor n/3\rfloor/\left(72\tilde{d}_f(d+1)\ln(e(d+1))\right)\right\} \leq 5\delta = \varepsilon/2$. Since, for $\ell \in \mathbb{N}$ with $\mathbb{P}\left(\left|\mathcal{L}_{\tilde{d}_f}\right| = \ell\right) > 0$, the sequence of X_m values appearing in $\mathcal{L}_{\tilde{d}_f}$ are conditionally distributed as $\mathcal{P}^\ell$ given $|\mathcal{L}_{\tilde{d}_f}| = \ell$, and this is the same as the (unconditional) distribution of $\{X_1, X_2, \ldots, X_\ell\}$, we have that

$$\begin{aligned}\mathbb{E}\left[\mathrm{er}(\hat{h})\right] &\leq \mathbb{E}\left[2\,\mathrm{er}\Big(\mathcal{A}_p\left(\mathcal{L}_{\tilde{d}_f}\right)\Big)\mathbb{1}_{\hat{E}_3(\delta)}\right] + \varepsilon/2 = \mathbb{E}\left[\mathbb{E}\left[2\,\mathrm{er}\Big(\mathcal{A}_p\left(\mathcal{L}_{\tilde{d}_f}\right)\Big)\mathbb{1}_{\hat{E}_3(\delta)}\Big| |\mathcal{L}_{\tilde{d}_f}|\right]\right] + \varepsilon/2 \\ &\leq 2 \sup_{\ell \geq \Lambda_p(\varepsilon/4, f, \mathcal{P})} \mathbb{E}\left[\mathrm{er}(\mathcal{A}_p(\mathcal{Z}_\ell))\right] + \varepsilon/2 \leq \varepsilon.\end{aligned}$$

To specialize to the specific variant of Meta-Algorithm 3 stated in Section 5.2, take $\gamma = 1/2$. ■

Appendix E. Proofs Related to Section 6: Agnostic Learning

This appendix contains the proofs of our results on learning with noise. Specifically, Appendix E.1 provides the proof of the counterexample from Theorem 22, demonstrating that there is no activizer for the $\check{\mathcal{A}}_p$ passive learning algorithm described in Section 6.2 in the agnostic case. Appendix E.2 presents the proof of Lemma 26 from Section 6.7, bounding the label complexity of Algorithm 5 under Condition 1. Finally, Appendix E.3 presents a proof of Theorem 28, demonstrating that any active learning algorithm can be modified to trivialize the misspecified model case. The notation used throughout Appendix E is taken from Section 6.

E.1 Proof of Theorem 22: Negative Result for Agnostic Activized Learning

It suffices to show that $\check{\mathcal{A}}_p$ achieves a label complexity Λ_p such that, for any label complexity Λ_a achieved by any active learning algorithm $\mathcal{A}_a$, there exists a distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1,+1\}$ such that $\mathcal{P}_{XY} \in \mathrm{Nontrivial}(\Lambda_p;\mathbb{C})$ and yet $\Lambda_a(\nu + c\varepsilon, \mathcal{P}_{XY}) \neq o\left(\Lambda_p(\nu+\varepsilon, \mathcal{P}_{XY})\right)$ for every constant $c \in (0,\infty)$. Specifically, we will show that there is a distribution $\mathcal{P}_{XY}$ for which $\Lambda_p(\nu+\varepsilon,\mathcal{P}_{XY}) = \Theta(1/\varepsilon)$ and $\Lambda_a(\nu+\varepsilon,\mathcal{P}_{XY}) \neq o(1/\varepsilon)$.

Let $\mathcal{P}(\{0\}) = 1/2$, and for any measurable $A \subseteq (0,1]$, $\mathcal{P}(A) = \lambda(A)/2$, where λ is Lebesgue measure. Let $\mathbb{D}$ be the family of distributions $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1,+1\}$ characterized by the properties that the marginal distribution on $\mathcal{X}$ is $\mathcal{P}$, $\eta(0;\mathcal{P}_{XY}) \in (1/8, 3/8)$, and $\forall x \in (0,1]$,

$$\eta(x;\mathcal{P}_{XY}) = \eta(0;\mathcal{P}_{XY}) + (x/2)\cdot(1-\eta(0;\mathcal{P}_{XY})).$$

Thus, $\eta(x;\mathcal{P}_{XY})$ is a linear function. For any $\mathcal{P}_{XY} \in \mathbb{D}$, since the point $z^* = \frac{1-2\eta(0;\mathcal{P}_{XY})}{1-\eta(0;\mathcal{P}_{XY})}$ has $\eta(z^*;\mathcal{P}_{XY}) = 1/2$, we see that $f = h_{z^*}$ is a Bayes optimal classifier. Furthermore, for any $\eta_0 \in$

$[1/8,3/8]$,

$$\left|\frac{1-2\eta_0}{1-\eta_0} - \frac{1-2\eta(0;\mathcal{P}_{XY})}{1-\eta(0;\mathcal{P}_{XY})}\right| = \frac{|\eta(0;\mathcal{P}_{XY})-\eta_0|}{(1-\eta_0)(1-\eta(0;\mathcal{P}_{XY}))},$$

and since $(1-\eta_0)(1-\eta(0;\mathcal{P}_{XY})) \in (25/64, 49/64) \subset (1/3,1)$, the value $z = \frac{1-2\eta_0}{1-\eta_0}$ satisfies

$$|\eta_0 - \eta(0;\mathcal{P}_{XY})| \leq |z - z^*| \leq 3|\eta_0 - \eta(0;\mathcal{P}_{XY})|. \tag{78}$$

Also note that under $\mathcal{P}_{XY}$, since $(1-2\eta(0;\mathcal{P}_{XY})) = (1-\eta(0;\mathcal{P}_{XY}))z^*$, any $z \in (0,1)$ has

$$\begin{aligned}\mathrm{er}(h_z) - \mathrm{er}(h_{z^*}) &= \int_z^{z^*} \big(1-2\eta(x;\mathcal{P}_{XY})\big)\mathrm{d}x = \int_z^{z^*} \big(1-2\eta(0;\mathcal{P}_{XY}) - x(1-\eta(0;\mathcal{P}_{XY}))\big)\mathrm{d}x \\ &= (1-\eta(0;\mathcal{P}_{XY}))\int_z^{z^*}(z^*-x)\,\mathrm{d}x = \frac{(1-\eta(0;\mathcal{P}_{XY}))}{2}(z^*-z)^2,\end{aligned}$$

so that

$$\frac{5}{16}(z-z^*)^2 \leq \mathrm{er}(h_z) - \mathrm{er}(h_{z^*}) \leq \frac{7}{16}(z-z^*)^2. \tag{79}$$

Finally, note that any $x,x' \in (0,1]$ with $|x-z^*| < |x'-z^*|$ has

$$|1-2\eta(x;\mathcal{P}_{XY})| = |x-z^*|(1-\eta(0;\mathcal{P}_{XY})) < |x'-z^*|(1-\eta(0;\mathcal{P}_{XY})) = |1-2\eta(x';\mathcal{P}_{XY})|.$$

Thus, for any $q \in (0,1/2]$, there exists $z'_q \in [0,1]$ such that $z^* \in [z'_q, z'_q+2q] \subseteq [0,1]$, and the classifier $h'_q(x) = h_{z^*}(x)\cdot\left(1-2\mathbb{1}_{(z'_q,z'_q+2q]}(x)\right)$ has $\mathrm{er}(h) \geq \mathrm{er}(h'_q)$ for every classifier h with $h(0) = -1$ and $\mathcal{P}(x: h(x) \neq h_{z^*}(x)) = q$. Noting that $\mathrm{er}(h'_q) - \mathrm{er}(h_{z^*}) = \left(\lim_{z\downarrow z'_q}\mathrm{er}(h_z) - \mathrm{er}(h_{z^*})\right) + \left(\mathrm{er}(h_{z'_q+2q}) - \mathrm{er}(h_{z^*})\right)$, (79) implies that $\mathrm{er}(h'_q) - \mathrm{er}(h_{z^*}) \geq \frac{5}{16}\left(\left(z'_q - z^*\right)^2 + \left(z'_q+2q-z^*\right)^2\right)$, and since $\max\{z^*-z'_q, z'_q+2q-z^*\} \geq q$, this is at least $\frac{5}{16}q^2$. In general, any h with $h(0) = +1$ has $\mathrm{er}(h) - \mathrm{er}(h_{z^*}) \geq 1/2 - \eta(0;\mathcal{P}_{XY}) > 1/8 \geq (1/8)\mathcal{P}(x: h(x) \neq h_{z^*}(x))^2$. Combining these facts, we see that any classifier h has

$$\mathrm{er}(h) - \mathrm{er}(h_{z^*}) \geq (1/8)\mathcal{P}\left(x: h(x) \neq h_{z^*}(x)\right)^2. \tag{80}$$

Lemma 54 *The passive learning algorithm $\check{\mathcal{A}}_p$ achieves a label complexity Λ_p such that, for every $\mathcal{P}_{XY} \in \mathbb{D}$, $\Lambda_p(\nu+\varepsilon, \mathcal{P}_{XY}) = \Theta(1/\varepsilon)$.*

Proof Consider the values $\hat{\eta}_0$ and $\hat{z}$ from $\check{\mathcal{A}}_p(\mathcal{Z}_n)$ for some $n \in \mathbb{N}$. Combining (78) and (79), we have $\mathrm{er}(h_{\hat{z}}) - \mathrm{er}(h_{z^*}) \leq \frac{7}{16}(\hat{z}-z^*)^2 \leq \frac{63}{16}(\hat{\eta}_0 - \eta(0;\mathcal{P}_{XY}))^2 \leq 4(\hat{\eta}_0 - \eta(0;\mathcal{P}_{XY}))^2$. Let $N_n = |\{i \in \{1,\ldots,n\} : X_i = 0\}|$, and $\bar{\eta}_0 = N_n^{-1}|\{i \in \{1,\ldots,n\} : X_i = 0, Y_i = +1\}|$ if $N_n > 0$, or $\bar{\eta}_0 = 0$ if $N_n = 0$. Note that $\hat{\eta}_0 = \left(\bar{\eta}_0 \vee \frac{1}{8}\right) \wedge \frac{3}{8}$, and since $\eta(0;\mathcal{P}_{XY}) \in (1/8,3/8)$, we have $|\hat{\eta}_0 - \eta(0;\mathcal{P}_{XY})| \leq |\bar{\eta}_0 - \eta(0;\mathcal{P}_{XY})|$. Therefore, for any $\mathcal{P}_{XY} \in \mathbb{D}$,

$$\begin{aligned}\mathbb{E}\left[\mathrm{er}(h_{\hat{z}}) - \mathrm{er}(h_{z^*})\right] &\leq 4\mathbb{E}\left[(\hat{\eta}_0 - \eta(0;\mathcal{P}_{XY}))^2\right] \leq 4\mathbb{E}\left[(\bar{\eta}_0 - \eta(0;\mathcal{P}_{XY}))^2\right] \\ &\leq 4\mathbb{E}\left[\mathbb{E}\left[(\bar{\eta}_0 - \eta(0;\mathcal{P}_{XY}))^2 \middle| N_n\right]\mathbb{1}_{[n/4,n]}(N_n)\right] + 4\mathbb{P}(N_n < n/4).\end{aligned} \tag{81}$$

By a Chernoff bound, $\mathbb{P}(N_n < n/4) \leq \exp\{-n/16\}$, and since the conditional distribution of $N_n\bar{\eta}_0$ given N_n is $\mathrm{Binomial}(N_n, \eta(0;\mathcal{P}_{XY}))$, (81) is at most

$$4\mathbb{E}\left[\frac{1}{N_n \vee n/4}\eta(0;\mathcal{P}_{XY})(1-\eta(0;\mathcal{P}_{XY}))\right] + 4\cdot\exp\{-n/16\} \leq 4\cdot\frac{4}{n}\cdot\frac{15}{64} + 4\cdot\frac{16}{n} < \frac{68}{n}.$$

For any $n \geq \lceil 68/\varepsilon \rceil$, this is at most ε. Therefore, $\check{\mathcal{A}}_p$ achieves a label complexity Λ_p such that, for any $\mathcal{P}_{XY} \in \mathbb{D}$, $\Lambda_p(\nu + \varepsilon, \mathcal{P}_{XY}) = \lceil 68/\varepsilon \rceil = \Theta(1/\varepsilon)$. ■

Next we establish a corresponding lower bound for any active learning algorithm. Note that this requires more than a simple minimax lower bound, since we must have an asymptotic lower bound for a *fixed* $\mathcal{P}_{XY}$, rather than selecting a different $\mathcal{P}_{XY}$ for each ε value; this is akin to the *strong* minimax lower bounds proven by Antos and Lugosi (1998) for passive learning in the realizable case. For this, we proceed by reduction from the task of estimating a binomial mean; toward this end, the following lemma will be useful.

Lemma 55 *For any nonempty $(a,b) \subset [0,1]$, and any sequence of estimators $\hat{p}_n : \{0,1\}^n \to [0,1]$, there exists $p \in (a,b)$ such that, if $B_1, B_2, \ldots$ are independent* Bernoulli(p) *random variables, also independent from every $\hat{p}_n$, then $\mathbb{E}\left[(\hat{p}_n(B_1,\ldots,B_n) - p)^2\right] \neq o(1/n)$.*

Proof We first establish the claim when $a = 0$ and $b = 1$. For any $p \in [0,1]$, let $B_1(p), B_2(p), \ldots$ be i.i.d. Bernoulli(p) random variables, independent from any internal randomness of the $\hat{p}_n$ estimators. We proceed by reduction from hypothesis testing, for which there are known lower bounds. Specifically, it is known (e.g., Wald, 1945; Bar-Yossef, 2003) that for any $p, q \in (0,1)$, $\delta \in (0, e^{-1})$, any (possibly randomized) $\hat{q} : \{0,1\}^n \to \{p,q\}$, and any $n \in \mathbb{N}$,

$$n < \frac{(1-8\delta)\ln(1/8\delta)}{8\mathrm{KL}(p\|q)} \implies \max_{p^* \in \{p,q\}} \mathbb{P}\left(\hat{q}(B_1(p^*),\ldots,B_n(p^*)) \neq p^*\right) > \delta,$$

where $\mathrm{KL}(p\|q) = p\ln(p/q) + (1-p)\ln((1-p)/(1-q))$. It is also known (e.g., Poland and Hutter, 2006) that for $p, q \in [1/4, 3/4]$, $\mathrm{KL}(p\|q) \leq (8/3)(p-q)^2$. Combining this with the above fact, we have that for $p, q \in [1/4, 3/4]$,

$$\max_{p^* \in \{p,q\}} \mathbb{P}\left(\hat{q}(B_1(p^*),\ldots,B_n(p^*)) \neq p^*\right) \geq (1/16) \cdot \exp\left\{-128(p-q)^2 n/3\right\}. \tag{82}$$

Given the estimator $\hat{p}_n$ from the lemma statement, we construct a sequence of hypothesis tests as follows. For $i \in \mathbb{N}$, let $\alpha_i = \exp\left\{-2^i\right\}$ and $n_i = \left\lfloor 1/\alpha_i^2 \right\rfloor$. Define $p_0^* = 1/4$, and for $i \in \mathbb{N}$, inductively define $\hat{q}_i(b_1,\ldots,b_{n_i}) = \operatorname{argmin}_{p \in \{p_{i-1}^*, p_{i-1}^* + \alpha_i\}} |\hat{p}_{n_i}(b_1,\ldots,b_{n_i}) - p|$ for $b_1,\ldots,b_{n_i} \in \{0,1\}$, and $p_i^* = \operatorname{argmax}_{p \in \{p_{i-1}^*, p_{i-1}^* + \alpha_i\}} \mathbb{P}\left(\hat{q}_i(B_1(p),\ldots,B_{n_i}(p)) \neq p\right)$. Finally, define $p^* = \lim_{i\to\infty} p_i^*$. Note that $\forall i \in \mathbb{N}$, $p_i^* < 1/2$, $p_{i-1}^*, p_{i-1}^* + \alpha_i \in [1/4, 3/4]$, and $0 \leq p^* - p_i^* \leq \sum_{j=i+1}^{\infty} \alpha_j < 2\alpha_{i+1} = 2\alpha_i^2$. We generally have

$$\begin{aligned}\mathbb{E}\left[(\hat{p}_{n_i}(B_1(p^*),\ldots,B_{n_i}(p^*)) - p^*)^2\right] &\geq \frac{1}{3}\mathbb{E}\left[(\hat{p}_{n_i}(B_1(p^*),\ldots,B_{n_i}(p^*)) - p_i^*)^2\right] - (p^* - p_i^*)^2 \\ &\geq \frac{1}{3}\mathbb{E}\left[(\hat{p}_{n_i}(B_1(p^*),\ldots,B_{n_i}(p^*)) - p_i^*)^2\right] - 4\alpha_i^4.\end{aligned}$$

Furthermore, note that for any $m \in \{0,\ldots,n_i\}$,

$$\begin{aligned}\frac{(p^*)^m(1-p^*)^{n_i-m}}{(p_i^*)^m(1-p_i^*)^{n_i-m}} &\geq \left(\frac{1-p^*}{1-p_i^*}\right)^{n_i} \geq \left(\frac{1-p_i^*-2\alpha_i^2}{1-p_i^*}\right)^{n_i} \\ &\geq \left(1-4\alpha_i^2\right)^{n_i} \geq \exp\left\{-8\alpha_i^2 n_i\right\} \geq e^{-8},\end{aligned}$$

so that the probability mass function of $(B_1(p^*),\dots,B_{n_i}(p^*))$ is never smaller than e^{-8} times that of $(B_1(p_i^*),\dots,B_{n_i}(p_i^*))$, which implies (by the law of the unconscious statistician)

$$\mathbb{E}\left[(\hat{p}_{n_i}(B_1(p^*),\dots,B_{n_i}(p^*)) - p_i^*)^2\right] \geq e^{-8}\mathbb{E}\left[(\hat{p}_{n_i}(B_1(p_i^*),\dots,B_{n_i}(p_i^*)) - p_i^*)^2\right].$$

By a triangle inequality, we have

$$\mathbb{E}\left[(\hat{p}_{n_i}(B_1(p_i^*),\dots,B_{n_i}(p_i^*)) - p_i^*)^2\right] \geq \frac{\alpha_i^2}{4}\mathbb{P}\left(\hat{q}_i(B_1(p_i^*),\dots,B_{n_i}(p_i^*)) \neq p_i^*\right).$$

By (82), this is at least

$$\frac{\alpha_i^2}{4}(1/16)\cdot\exp\left\{-128\alpha_i^2 n_i/3\right\} \geq 2^{-6}e^{-43}\alpha_i^2.$$

Combining the above, we have

$$\mathbb{E}\left[(\hat{p}_{n_i}(B_1(p^*),\dots,B_{n_i}(p^*)) - p^*)^2\right] \geq 3^{-1}2^{-6}e^{-51}\alpha_i^2 - 4\alpha_i^4 \geq 2^{-9}e^{-51}n_i^{-1} - 4n_i^{-2}.$$

For $i \geq 5$, this is larger than $2^{-11}e^{-51}n_i^{-1}$. Since n_i diverges as $i \to \infty$, we have that

$$\mathbb{E}\left[(\hat{p}_{n_i}(B_1(p^*),\dots,B_{n_i}(p^*)) - p^*)^2\right] \neq o(1/n),$$

which establishes the result for $a = 0$ and $b = 1$.

To extend this result to general nonempty ranges (a,b), we proceed by reduction from the above problem. Specifically, suppose $p' \in (0,1)$, and consider the following independent random variables (also independent from the $B_i(p')$ variables and $\hat{p}_n$ estimators). For each $i \in \mathbb{N}$, $C_{i1} \sim$ Bernoulli(a), $C_{i2} \sim$ Bernoulli$((b-a)/(1-a))$. Then for $b_i \in \{0,1\}$, define $B_i'(b_i) = \max\{C_{i1}, C_{i2}\cdot b_i\}$. For any given $p' \in (0,1)$, the random variables $B_i'(B_i(p'))$ are i.i.d. Bernoulli(p), with $p = a + (b-a)p' \in (a,b)$ (which forms a bijection between $(0,1)$ and (a,b)). Defining $\hat{p}_n'(b_1,\dots,b_n) = (\hat{p}_n(B_1'(b_1),\dots,B_n'(b_n)) - a)/(b-a)$, we have

$$\mathbb{E}\left[(\hat{p}_n(B_1(p),\dots,B_n(p)) - p)^2\right] = (b-a)^2\cdot\mathbb{E}\left[\left(\hat{p}_n'(B_1(p'),\dots,B_n(p')) - p'\right)^2\right]. \tag{83}$$

We have already shown there exists a value of $p' \in (0,1)$ such that the right side of (83) is not $o(1/n)$. Therefore, the corresponding value of $p = a + (b-a)p' \in (a,b)$ has the left side of (83) not $o(1/n)$, which establishes the result. ∎

We are now ready for the lower bound result for our setting.

Lemma 56 *For any label complexity Λ_a achieved by any active learning algorithm $\mathcal{A}_a$, there exists a $\mathcal{P}_{XY} \in \mathbb{D}$ such that $\Lambda_a(\nu + \varepsilon, \mathcal{P}_{XY}) \neq o(1/\varepsilon)$.*

Proof The idea here is to reduce from the task of estimating the mean of iid Bernoulli trials, corresponding to the Y_i values. Specifically, consider any active learning algorithm $\mathcal{A}_a$; we use $\mathcal{A}_a$ to construct an estimator for the mean of iid Bernoulli trials as follows. Suppose we have $B_1, B_2, \dots, B_n$ i.i.d. Bernoulli(p), for some $p \in (1/8, 3/8)$ and $n \in \mathbb{N}$. We take the sequence of $X_1, X_2, \dots$ random variables i.i.d. with distribution $\mathcal{P}$ defined above (independent from the B_j variables). For each i, we additionally have a random variable C_i with conditional distribution Bernoulli$(X_i/2)$ given X_i, where the C_i are conditionally independent given the X_i sequence, and independent from the B_i sequence as well.

We run $\mathcal{A}_a$ with this sequence of X_i values. For the t^{th} label request made by the algorithm, say for the Y_i value corresponding to some X_i, if it has previously requested this Y_i already, then we simply repeat the same answer for Y_i again, and otherwise we return to the algorithm the value $2\max\{B_t, C_i\} - 1$ for Y_i. Note that in the latter case, the conditional distribution of $\max\{B_t, C_i\}$ is Bernoulli$(p + (1-p)X_i/2)$, given the X_i that $\mathcal{A}_a$ requests the label of; thus, the Y_i response has the same conditional distribution given X_i as it would have for the $\mathcal{P}_{XY} \in \mathbb{D}$ with $\eta(0;\mathcal{P}_{XY}) = p$ (i.e., $\eta(X_i;\mathcal{P}_{XY}) = p + (1-p)X_i/2$). Since this Y_i value is conditionally (given X_i) independent from the previously returned labels and X_j sequence, this is distributionally equivalent to running $\mathcal{A}_a$ under the $\mathcal{P}_{XY} \in \mathbb{D}$ with $\eta(0;\mathcal{P}_{XY}) = p$.

Let $\hat{h}_n$ be the classifier returned by $\mathcal{A}_a(n)$ in the above context, and let $\hat{z}_n$ denote the value of $z \in [2/5, 6/7]$ with minimum $\mathcal{P}(x : h_z(x) \neq \hat{h}_n(x))$. Then define $\hat{p}_n = \frac{1-\hat{z}_n}{2-\hat{z}_n} \in [1/8, 3/8]$ and $z^* = \frac{1-2p}{1-p} \in (2/5, 6/7)$. By a triangle inequality, we have $|\hat{z}_n - z^*| = 2\mathcal{P}(x : h_{\hat{z}_n}(x) \neq h_{z^*}(x)) \leq 4\mathcal{P}(x : \hat{h}_n(x) \neq h_{z^*}(x))$. Combining this with (80) and (78) implies that

$$\operatorname{er}(\hat{h}_n) - \operatorname{er}(h_{z^*}) \geq \frac{1}{8}\mathcal{P}\left(x : \hat{h}_n(x) \neq h_{z^*}(x)\right)^2 \geq \frac{1}{128}\left(\hat{z}_n - z^*\right)^2 \geq \frac{1}{128}\left(\hat{p}_n - p\right)^2. \tag{84}$$

In particular, by Lemma 55, we can choose $p \in (1/8, 3/8)$ so that $\mathbb{E}\left[(\hat{p}_n - p)^2\right] \neq o(1/n)$, which, by (84), implies $\mathbb{E}\left[\operatorname{er}(\hat{h}_n)\right] - \nu \neq o(1/n)$. This means there is an increasing infinite sequence of values $n_k \in \mathbb{N}$, and a constant $c \in (0, \infty)$ such that $\forall k \in \mathbb{N}$, $\mathbb{E}\left[\operatorname{er}(\hat{h}_{n_k})\right] - \nu \geq c/n_k$. Supposing $\mathcal{A}_a$ achieves label complexity Λ_a, and taking the values $\varepsilon_k = c/(2n_k)$, we have $\Lambda_a(\nu + \varepsilon_k, \mathcal{P}_{XY}) > n_k = c/(2\varepsilon_k)$. Since $\varepsilon_k > 0$ and approaches 0 as $k \to \infty$, we have $\Lambda_a(\nu + \varepsilon, \mathcal{P}_{XY}) \neq o(1/\varepsilon)$. ■

Proof [of Theorem 22] The result follows from Lemmas 54 and 56. ■

E.2 Proof of Lemma 26: Label Complexity of Algorithm 5

The proof of Lemma 26 essentially runs parallel to that of Theorem 16, with variants of each lemma from that proof adapted to the noise-robust Algorithm 5.

As before, in this section we will fix a particular joint distribution $\mathcal{P}_{XY}$ on $\mathcal{X} \times \{-1, +1\}$ with marginal $\mathcal{P}$ on $\mathcal{X}$, and then analyze the label complexity achieved by Algorithm 5 for that particular distribution. For our purposes, we will suppose $\mathcal{P}_{XY}$ satisfies Condition 1 for some finite parameters μ and κ. We also fix any $f \in \bigcap_{\varepsilon>0} \operatorname{cl}(\mathbb{C}(\varepsilon))$. Furthermore, we will continue using the notation of Appendix B, such as $\mathcal{S}^k(\mathcal{H})$, etc., and in particular we continue to denote $V_m^\star = \{h \in \mathbb{C} : \forall \ell \leq m, h(X_\ell) = f(X_\ell)\}$ (though note that in this case, we may sometimes have $f(X_\ell) \neq Y_\ell$, so that $V_m^\star \neq \mathbb{C}[\mathcal{Z}_m]$). As in the above proofs, we will prove a slightly more general result in which the "1/2" threshold in Step 5 can be replaced by an arbitrary constant $\gamma \in (0, 1)$.

For the estimators $\hat{P}_{4m}$ used in the algorithm, we take the same definitions as in Appendix B.1. To be clear, we assume the sequences W_1 and W_2 mentioned there are independent from the entire

$(X_1,Y_1),(X_2,Y_2),\ldots$ sequence of data points; this is consistent with the earlier discussion of how these W_1 and W_2 sequences can be constructed in a preprocessing step.

We will consider running Algorithm 5 with label budget $n \in \mathbb{N}$ and confidence parameter $\delta \in (0,e^{-3})$, and analyze properties of the internal sets V_i. We will denote by $\hat{V}_i$, $\hat{\mathcal{L}}_i$, and $\hat{i}_k$, the *final* values of V_i, $\mathcal{L}_i$, and i_k, respectively, for each i and k in Algorithm 5. We also denote by $\hat{m}^{(k)}$ and $\hat{V}^{(k)}$ the final values of m and V_{i_k+1}, respectively, obtained while k has the specified value in Algorithm 5; $\hat{V}^{(k)}$ may be smaller than $\hat{V}_{\hat{i}_k}$ when $\hat{m}^{(k)}$ is not a power of 2. Additionally, define $\mathcal{L}_i^\star = \{(X_m,Y_m)\}_{m=2^{i-1}+1}^{2^i}$. After establishing a few results concerning these, we will show that for n satisfying the condition in Lemma 26, the conclusion of the lemma holds. First, we have a few auxiliary definitions. For $\mathcal{H} \subseteq \mathbb{C}$, and any $i \in \mathbb{N}$, define

$$\phi_i(\mathcal{H}) = \mathbb{E} \sup_{h_1,h_2 \in \mathcal{H}} \left| \left(\mathrm{er}(h_1) - \mathrm{er}_{\mathcal{L}_i^\star}(h_1)\right) - \left(\mathrm{er}(h_2) - \mathrm{er}_{\mathcal{L}_i^\star}(h_2)\right) \right|$$
$$\text{and} \quad \tilde{U}_i(\mathcal{H},\delta) = \min\left\{ \tilde{K}\left(\phi_i(\mathcal{H}) + \sqrt{\mathrm{diam}(\mathcal{H}) \frac{\ln(32i^2/\delta)}{2^{i-1}}} + \frac{\ln(32i^2/\delta)}{2^{i-1}} \right), 1 \right\},$$

where for our purposes we can take $\tilde{K} = 8272$. It is known (see, e.g., Massart and Nédélec, 2006; Giné and Koltchinskii, 2006) that for some universal constant $c' \in [2,\infty)$,

$$\phi_{i+1}(\mathcal{H}) \le c' \max\left\{ \sqrt{\mathrm{diam}(\mathcal{H}) 2^{-i} d \log_2 \frac{2}{\mathrm{diam}(\mathcal{H})}}, 2^{-i} di \right\}. \tag{85}$$

We also generally have $\phi_i(\mathcal{H}) \le 2$ for every $i \in \mathbb{N}$. The next lemma is taken from the work of Koltchinskii (2006) on data-dependent Rademacher complexity bounds on the excess risk.

Lemma 57 *For any $\delta \in (0,e^{-3})$, any $\mathcal{H} \subseteq \mathbb{C}$ with $f \in \mathrm{cl}(\mathcal{H})$, and any $i \in \mathbb{N}$, on an event K_i with $\mathbb{P}(K_i) \ge 1 - \delta/4i^2$, $\forall h \in \mathcal{H}$,*

$$\mathrm{er}_{\mathcal{L}_i^\star}(h) - \min_{h' \in \mathcal{H}} \mathrm{er}_{\mathcal{L}_i^\star}(h') \le \mathrm{er}(h) - \mathrm{er}(f) + \hat{U}_i(\mathcal{H},\delta)$$
$$\mathrm{er}(h) - \mathrm{er}(f) \le \mathrm{er}_{\mathcal{L}_i^\star}(h) - \mathrm{er}_{\mathcal{L}_i^\star}(f) + \hat{U}_i(\mathcal{H},\delta)$$
$$\min\left\{\hat{U}_i(\mathcal{H},\delta), 1\right\} \le \tilde{U}_i(\mathcal{H},\delta).$$

Lemma 57 essentially follows from a version of Talagrand's inequality. The details of the proof may be extracted from the proofs of Koltchinskii (2006), and related derivations have previously been presented by Hanneke (2011) and Koltchinskii (2010). The only minor twist here is that f need only be in $\mathrm{cl}(\mathcal{H})$, rather than in $\mathcal{H}$ itself, which easily follows from Koltchinskii's original results, since the Borel-Cantelli lemma implies that with probability one, every $\varepsilon > 0$ has some $g \in \mathcal{H}(\varepsilon)$ (very close to f) with $\mathrm{er}_{\mathcal{L}_i^\star}(g) = \mathrm{er}_{\mathcal{L}_i^\star}(f)$.

For our purposes, the important implications of Lemma 57 are summarized by the following lemma.

Lemma 58 *For any $\delta \in (0,e^{-3})$ and any $n \in \mathbb{N}$, when running Algorithm 5 with label budget n and confidence parameter δ, on an event $J_n(\delta)$ with $\mathbb{P}(J_n(\delta)) \ge 1-\delta/2$, $\forall i \in \{0,1,\ldots,\hat{i}_{d+1}\}$, if $V_{2^i}^\star \subseteq \hat{V}_i$*

then $\forall h \in \hat{V}_i$,

$$\mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h) - \min_{h' \in \hat{V}_i} \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h') \leq \mathrm{er}(h) - \mathrm{er}(f) + \hat{U}_{i+1}(\hat{V}_i, \delta)$$
$$\mathrm{er}(h) - \mathrm{er}(f) \leq \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h) - \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(f) + \hat{U}_{i+1}(\hat{V}_i, \delta)$$
$$\min\left\{\hat{U}_{i+1}(\hat{V}_i, \delta), 1\right\} \leq \tilde{U}_{i+1}(\hat{V}_i, \delta).$$

Proof For each i, consider applying Lemma 57 under the conditional distribution given $\hat{V}_i$. The set $\mathcal{L}^{\star}_{i+1}$ is independent from $\hat{V}_i$, as are the Rademacher variables in the definition of $\hat{R}_{i+1}(\hat{V}_i)$. Furthermore, by Lemma 35, on H', $f \in \mathrm{cl}\left(V^{\star}_{2^i}\right)$, so that the conditions of Lemma 57 hold. The law of total probability then implies the existence of an event J_i of probability $\mathbb{P}(J_i) \geq 1 - \delta/4(i+1)^2$, on which the claimed inequalities hold for that value of i if $i \leq \hat{i}_{d+1}$. A union bound over values of i then implies the existence of an event $J_n(\delta) = \bigcap_i J_i$ with probability $\mathbb{P}(J_n(\delta)) \geq 1 - \sum_i \delta/4(i+1)^2 \geq 1 - \delta/2$ on which the claimed inequalities hold for all $i \leq \hat{i}_{d+1}$. ∎

Lemma 59 *For some* $(\mathbb{C}, \mathcal{P}_{XY}, \gamma)$*-dependent constants* $c, c^* \in [1, \infty)$*, for any* $\delta \in (0, e^{-3})$ *and integer* $n \geq c^* \ln(1/\delta)$*, when running Algorithm 5 with label budget* n *and confidence parameter* δ*, on event* $J_n(\delta) \cap H_n^{(i)} \cap H_n^{(ii)}$*, every* $i \in \{0, 1, \ldots, \hat{i}_{\tilde{d}_f}\}$ *satisfies*

$$V^{\star}_{2^i} \subseteq \hat{V}_i \subseteq \mathbb{C}\left(c\left(\frac{di + \ln(1/\delta)}{2^i}\right)^{\frac{\kappa}{2\kappa-1}}\right),$$

and furthermore $V^{\star}_{\hat{m}^{(\tilde{d}_f)}} \subseteq \hat{V}^{(\tilde{d}_f)}$.

Proof Define $c = \left(24\tilde{K}c'\sqrt{\mu}\right)^{\frac{2\kappa}{2\kappa-1}}$, $c^* = \max\left\{\tau^*, 8d\left(\frac{\mu c^{1/\kappa}}{r_{(1-\gamma)/6}}\right)^{\frac{1}{2\kappa-1}} \log_2\left(\frac{4\mu c^{1/\kappa}}{r_{(1-\gamma)/6}}\right)\right\}$, and suppose $n \geq c^* \ln(1/\delta)$. We now proceed by induction. As the right side equals $\mathbb{C}$ for $i = 0$, the claimed inclusions are certainly true for $\hat{V}_0 = \mathbb{C}$, which serves as our base case. Now suppose some $i \in \{0, 1, \ldots, \hat{i}_{\tilde{d}_f}\}$ satisfies

$$V^{\star}_{2^i} \subseteq \hat{V}_i \subseteq \mathbb{C}\left(c\left(\frac{di + \ln(1/\delta)}{2^i}\right)^{\frac{\kappa}{2\kappa-1}}\right). \tag{86}$$

In particular, Condition 1 implies

$$\mathrm{diam}(\hat{V}_i) \leq \mathrm{diam}\left(\mathbb{C}\left(c\left(\frac{di + \ln(1/\delta)}{2^i}\right)^{\frac{\kappa}{2\kappa-1}}\right)\right) \leq \mu c^{\frac{1}{\kappa}}\left(\frac{di + \ln(1/\delta)}{2^i}\right)^{\frac{1}{2\kappa-1}}. \tag{87}$$

If $i < \hat{i}_{\tilde{d}_f}$, then let k be the integer for which $\hat{i}_{k-1} \leq i < \hat{i}_k$, and otherwise let $k = \tilde{d}_f$. Note that we certainly have $\hat{i}_1 \geq \lfloor \log_2(n/2) \rfloor$, since $m = \lfloor n/2 \rfloor \geq 2^{\lfloor \log_2(n/2) \rfloor}$ is obtained while $k = 1$. Therefore, if $k > 1$,

$$\frac{di + \ln(1/\delta)}{2^i} \leq \frac{4d \log_2(n) + 4\ln(1/\delta)}{n},$$

so that (87) implies

$$\operatorname{diam}\left(\hat{V}_i\right) \leq \mu c^{\frac{1}{\kappa}} \left(\frac{4d\log_2(n)+4\ln(1/\delta)}{n}\right)^{\frac{1}{2\kappa-1}}.$$

By our choice of c^*, the right side is at most $r_{(1-\gamma)/6}$. Therefore, since Lemma 35 implies $f \in \operatorname{cl}\left(V^\star_{2^i}\right)$ on $H_n^{(i)}$, we have $\hat{V}_i \subseteq \mathrm{B}\left(f, r_{(1-\gamma)/6}\right)$ when $k > 1$. Combined with (86), we have that $V^\star_{2^i} \subseteq \hat{V}_i$, and either $k=1$, or $\hat{V}_i \subseteq \mathrm{B}(f, r_{(1-\gamma)/6})$ and $4m > 4\lfloor n/2 \rfloor \geq n$. Now consider any m with $2^i+1 \leq m \leq \min\left\{2^{i+1}, \hat{m}^{(\tilde{d}_f)}\right\}$, and for the purpose of induction suppose $V^\star_{m-1} \subseteq V_{i+1}$ upon reaching Step 5 for that value of m in Algorithm 5. Since $V_{i+1} \subseteq \hat{V}_i$ and $n \geq \tau^*$, Lemma 41 (with $\ell = m-1$) implies that on $H_n^{(i)} \cap H_n^{(ii)}$,

$$\hat{\Delta}^{(k)}_{4m}(X_m, W_2, V_{i+1}) < \gamma \implies \hat{\Gamma}^{(k)}_{4m}(X_m, -f(X_m), W_2, V_{i+1}) < \hat{\Gamma}^{(k)}_{4m}(X_m, f(X_m), W_2, V_{i+1}),$$

so that after Step 8 we have $V^\star_m \subseteq V_{i+1}$. Since (86) implies that the $V^\star_{m-1} \subseteq V_{i+1}$ condition holds if Algorithm 5 reaches Step 5 with $m = 2^i+1$ (at which time $V_{i+1} = \hat{V}_i$), we have by induction that on $H_n^{(i)} \cap H_n^{(ii)}$, $V^\star_m \subseteq V_{i+1}$ upon reaching Step 9 with $m = \min\left\{2^{i+1}, \hat{m}^{(\tilde{d}_f)}\right\}$. This establishes the final claim of the lemma, given that the first claim holds. For the remainder of this inductive proof, suppose $i < \hat{i}_{\tilde{d}_f}$. Since Step 8 enforces that, upon reaching Step 9 with $m = 2^{i+1}$, every $h_1, h_2 \in V_{i+1}$ have $\operatorname{er}_{\hat{\mathcal{L}}_{i+1}}(h_1) - \operatorname{er}_{\hat{\mathcal{L}}_{i+1}}(h_2) = \operatorname{er}_{\mathcal{L}^\star_{i+1}}(h_1) - \operatorname{er}_{\mathcal{L}^\star_{i+1}}(h_2)$, on $J_n(\delta) \cap H_n^{(i)} \cap H_n^{(ii)}$ we have

$$\begin{aligned}
\hat{V}_{i+1} &\subseteq \left\{h \in \hat{V}_i : \operatorname{er}_{\mathcal{L}^\star_{i+1}}(h) - \min_{h' \in V^\star_{2^{i+1}}} \operatorname{er}_{\mathcal{L}^\star_{i+1}}(h') \leq \hat{U}_{i+1}\left(\hat{V}_i, \delta\right)\right\} \\
&\subseteq \left\{h \in \hat{V}_i : \operatorname{er}_{\mathcal{L}^\star_{i+1}}(h) - \operatorname{er}_{\mathcal{L}^\star_{i+1}}(f) \leq \hat{U}_{i+1}\left(\hat{V}_i, \delta\right)\right\} \\
&\subseteq \hat{V}_i \cap \mathbb{C}\left(2\hat{U}_{i+1}\left(\hat{V}_i, \delta\right)\right) \subseteq \mathbb{C}\left(2\tilde{U}_{i+1}\left(\hat{V}_i, \delta\right)\right),
\end{aligned} \tag{88}$$

where the second line follows from Lemma 35 and the last two inclusions follow from Lemma 58. Focusing on (88), combining (87) with (85) (and the fact that $\phi_{i+1}(\hat{V}_i) \leq 2$), we can bound the value of $\tilde{U}_{i+1}\left(\hat{V}_i, \delta\right)$ as follows.

$$\begin{aligned}
\sqrt{\operatorname{diam}(\hat{V}_i)\frac{\ln(32(i+1)^2/\delta)}{2^i}} &\leq \sqrt{\mu} c^{\frac{1}{2\kappa}} \left(\frac{di+\ln(1/\delta)}{2^i}\right)^{\frac{1}{4\kappa-2}} \left(\frac{\ln(32(i+1)^2/\delta)}{2^i}\right)^{\frac{1}{2}} \\
&\leq \sqrt{\mu} c^{\frac{1}{2\kappa}} \left(\frac{2di+2\ln(1/\delta)}{2^{i+1}}\right)^{\frac{1}{4\kappa-2}} \left(\frac{8(i+1)+2\ln(1/\delta)}{2^{i+1}}\right)^{\frac{1}{2}} \\
&\leq 4\sqrt{\mu} c^{\frac{1}{2\kappa}} \left(\frac{d(i+1)+\ln(1/\delta)}{2^{i+1}}\right)^{\frac{\kappa}{2\kappa-1}},
\end{aligned}$$

$$\begin{aligned}
\phi_{i+1}(\hat{V}_i) &\leq c'\sqrt{\mu} c^{\frac{1}{2\kappa}} \left(\frac{di+\ln(1/\delta)}{2^i}\right)^{\frac{1}{4\kappa-2}} \left(\frac{d(i+2)}{2^i}\right)^{\frac{1}{2}} \\
&\leq 4c'\sqrt{\mu} c^{\frac{1}{2\kappa}} \left(\frac{d(i+1)+\ln(1/\delta)}{2^{i+1}}\right)^{\frac{\kappa}{2\kappa-1}},
\end{aligned}$$

and thus

$$\tilde{U}_{i+1}(\hat{V}_i,\delta) \leq \min\left\{8\tilde{K}c'\sqrt{\mu}c^{\frac{1}{2\kappa}}\left(\frac{d(i+1)+\ln(1/\delta)}{2^{i+1}}\right)^{\frac{\kappa}{2\kappa-1}} + \tilde{K}\frac{\ln(32(i+1)^2/\delta)}{2^i}, 1\right\}$$
$$\leq 12\tilde{K}c'\sqrt{\mu}c^{\frac{1}{2\kappa}}\left(\frac{d(i+1)+\ln(1/\delta)}{2^{i+1}}\right)^{\frac{\kappa}{2\kappa-1}} = (c/2)\left(\frac{d(i+1)+\ln(1/\delta)}{2^{i+1}}\right)^{\frac{\kappa}{2\kappa-1}}.$$

Combining this with (88) now implies

$$\hat{V}_{i+1} \subseteq \mathbb{C}\left(c\left(\frac{d(i+1)+\ln(1/\delta)}{2^{i+1}}\right)^{\frac{\kappa}{2\kappa-1}}\right).$$

To complete the inductive proof, it remains only to show $V^{\star}_{2^{i+1}} \subseteq \hat{V}_{i+1}$. Toward this end, recall we have shown above that on $H_n^{(i)} \cap H_n^{(ii)}$, $V^{\star}_{2^{i+1}} \subseteq V_{i+1}$ upon reaching Step 9 with $m = 2^{i+1}$, and that every $h_1, h_2 \in V_{i+1}$ at this point have $\mathrm{er}_{\hat{\mathcal{L}}_{i+1}}(h_1) - \mathrm{er}_{\hat{\mathcal{L}}_{i+1}}(h_2) = \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h_1) - \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h_2)$. Consider any $h \in V^{\star}_{2^{i+1}}$, and note that any other $g \in V^{\star}_{2^{i+1}}$ has $\mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(g) = \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h)$. Thus, on $H_n^{(i)} \cap H_n^{(ii)}$,

$$\mathrm{er}_{\hat{\mathcal{L}}_{i+1}}(h) - \min_{h' \in V_{i+1}} \mathrm{er}_{\hat{\mathcal{L}}_{i+1}}(h') = \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h) - \min_{h' \in V_{i+1}} \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h')$$
$$\leq \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h) - \min_{h' \in \hat{V}_i} \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h') = \inf_{g \in V^{\star}_{2^{i+1}}} \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(g) - \min_{h' \in \hat{V}_i} \mathrm{er}_{\mathcal{L}^{\star}_{i+1}}(h'). \quad (89)$$

Lemma 58 and (86) imply that on $J_n(\delta) \cap H_n^{(i)} \cap H_n^{(ii)}$, the last expression in (89) is not larger than $\inf_{g \in V^{\star}_{2^{i+1}}} \mathrm{er}(g) - \mathrm{er}(f) + \hat{U}_{i+1}(\hat{V}_i,\delta)$, and Lemma 35 implies $f \in \mathrm{cl}\left(V^{\star}_{2^{i+1}}\right)$ on $H_n^{(i)}$, so that $\inf_{g \in V^{\star}_{2^{i+1}}} \mathrm{er}(g) = \mathrm{er}(f)$. We therefore have

$$\mathrm{er}_{\hat{\mathcal{L}}_{i+1}}(h) - \min_{h' \in V_{i+1}} \mathrm{er}_{\hat{\mathcal{L}}_{i+1}}(h') \leq \hat{U}_{i+1}(\hat{V}_i,\delta),$$

so that $h \in \hat{V}_{i+1}$ as well. Since this holds for any $h \in V^{\star}_{2^{i+1}}$, we have $V^{\star}_{2^{i+1}} \subseteq \hat{V}_{i+1}$. The lemma now follows by the principle of induction. ■

Lemma 60 *There exist $(\mathbb{C}, \mathcal{P}_{XY}, \gamma)$-dependent constants $c_1^*, c_2^* \in [1,\infty)$ such that, for any $\varepsilon, \delta \in (0, e^{-3})$ and integer*

$$n \geq c_1^* + c_2^* \tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\varepsilon^{\frac{2}{\kappa}-2}\log_2^2\left(\frac{1}{\varepsilon\delta}\right),$$

when running Algorithm 5 with label budget n and confidence parameter δ, on an event $J_n^*(\varepsilon,\delta)$ with $\mathbb{P}(J_n^*(\varepsilon,\delta)) \geq 1-\delta$, we have $\hat{V}_{\hat{i}_{\tilde{d}_f}} \subseteq \mathbb{C}(\varepsilon)$.

Proof Define

$$c_1^* = \max\left\{2^{\tilde{d}_f+5}\left(\frac{\mu c^{1/\kappa}}{r_{(1-\gamma)/6}}\right)^{2\kappa-1} d\log_2\frac{d\mu c^{1/\kappa}}{r_{(1-\gamma)/6}}, \frac{2}{\tilde{\delta}_f^{1/3}}\ln\left(8c^{(i)}\right), \frac{120}{\tilde{\delta}_f^{1/3}}\ln\left(8c^{(ii)}\right)\right\}$$

and

$$c_2^* = \max\left\{c^*, 2^{\tilde{d}_f+5}\cdot\left(\frac{\mu c^{1/\kappa}}{r_{(1-\gamma)/6}}\right)^{2\kappa-1}, 2^{\tilde{d}_f+15}\cdot\frac{\mu c^2 d}{\gamma\tilde{\delta}_f}\log_2^2(4dc)\right\}.$$

Fix any $\varepsilon,\delta \in (0,e^{-3})$ and integer $n \geq c_1^* + c_2^*\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\varepsilon^{\frac{2}{\kappa}-2}\log_2^2\left(\frac{1}{\varepsilon\delta}\right)$.

For each $i \in \{0,1,\ldots\}$, let $\tilde{r}_i = \mu c^{\frac{1}{\kappa}}\left(\frac{di+\ln(1/\delta)}{2^i}\right)^{\frac{1}{2\kappa-1}}$. Also define

$$\tilde{i} = \left\lceil\left(2-\frac{1}{\kappa}\right)\log_2\frac{c}{\varepsilon} + \log_2\left[8d\log_2\frac{2dc}{\varepsilon\delta}\right]\right\rceil.$$

and let $\check{i} = \min\left\{i \in \mathbb{N} : \sup_{j\geq i}\tilde{r}_j < r_{(1-\gamma)/6}\right\}$. For any $i \in \left\{\check{i},\ldots,\hat{i}_{\tilde{d}_f}\right\}$, let

$$\mathcal{Q}_{i+1} = \left\{m \in \left\{2^i+1,\ldots,2^{i+1}\right\} : \hat{\Delta}_{4m}^{(\tilde{d}_f)}\left(X_m, W_2, \mathrm{B}\left(f,\tilde{r}_i\right)\right) \geq 2\gamma/3\right\}.$$

Also define

$$\tilde{\mathcal{Q}} = \frac{96}{\gamma\tilde{\delta}_f}\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\cdot 2\mu c^2\cdot\left(8d\log_2\frac{2dc}{\varepsilon\delta}\right)\cdot\varepsilon^{\frac{2}{\kappa}-2}.$$

By Lemma 59 and Condition 1, on $J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}$, if $i \leq \hat{i}_{\tilde{d}_f}$,

$$\hat{V}_i \subseteq \mathbb{C}\left(c\left(\frac{di+\ln(1/\delta)}{2^i}\right)^{\frac{\kappa}{2\kappa-1}}\right) \subseteq \mathrm{B}\left(f,\tilde{r}_i\right). \tag{90}$$

Lemma 59 also implies that, on $J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}$, for i with $\hat{i}_{\tilde{d}_f-1} \leq i \leq \hat{i}_{\tilde{d}_f}$, all of the sets V_{i+1} obtained in Algorithm 5 while $k=\tilde{d}_f$ and $m \in \left\{2^i+1,\ldots,2^{i+1}\right\}$ satisfy $V_{2^{i+1}}^{\star} \subseteq V_{i+1} \subseteq \hat{V}_i$. Recall that $\hat{i}_1 \geq \lfloor\log_2(n/2)\rfloor$, so that we have either $\tilde{d}_f = 1$ or else every $m \in \left\{2^i+1,\ldots,2^{i+1}\right\}$ has $4m > n$. Also recall that Lemma 49 implies that when the above conditions are satisfied, and $i \geq \check{i}$, on $H'\cap G_n^{(i)}$, $\hat{\Delta}_{4m}^{(\tilde{d}_f)}\left(X_m,W_2,V_{i+1}\right) \leq (3/2)\hat{\Delta}_{4m}^{(\tilde{d}_f)}\left(X_m,W_2,\mathrm{B}\left(f,\tilde{r}_i\right)\right)$, so that $|\mathcal{Q}_{i+1}|$ upper bounds the number of $m \in \left\{2^i+1,\ldots,2^{i+1}\right\}$ for which Algorithm 5 requests the label Y_m in Step 6 of the $k=\tilde{d}_f$ round. Thus, on $J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}$, $2^{\check{i}} + \sum_{i=\max\left\{\check{i},\hat{i}_{\tilde{d}_f-1}\right\}}^{\hat{i}_{\tilde{d}_f}}|\mathcal{Q}_{i+1}|$ upper bounds the total number of label requests by Algorithm 5 while $k=\tilde{d}_f$; therefore, by the constraint in Step 3, we know that either this quantity is at least as big as $\left\lfloor 2^{-\tilde{d}_f}n\right\rfloor$, or else we have $2^{\hat{i}_{\tilde{d}_f}+1} > \tilde{d}_f\cdot 2^n$. In particular, on this event, if we can show that

$$2^{\check{i}} + \sum_{i=\max\left\{\check{i},\hat{i}_{\tilde{d}_f-1}\right\}}^{\min\left\{\hat{i}_{\tilde{d}_f},\tilde{i}\right\}}|\mathcal{Q}_{i+1}| < \left\lfloor 2^{-\tilde{d}_f}n\right\rfloor \text{ and } 2^{\tilde{i}+1} \leq \tilde{d}_f\cdot 2^n, \tag{91}$$

then it must be true that $\tilde{i} < \hat{i}_{\tilde{d}_f}$. Next, we will focus on establishing this fact.

Consider any $i \in \left\{\max\left\{\check{i},\hat{i}_{\tilde{d}_f-1}\right\},\ldots,\min\left\{\hat{i}_{\tilde{d}_f},\tilde{i}\right\}\right\}$ and any $m \in \left\{2^i+1,\ldots,2^{i+1}\right\}$. If $\tilde{d}_f = 1$, then

$$\mathbb{P}\left(\hat{\Delta}_{4m}^{(\tilde{d}_f)}\left(X_m,W_2,\mathrm{B}\left(f,\tilde{r}_i\right)\right) \geq 2\gamma/3 \middle| W_2\right) = \mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}\left(\mathrm{B}\left(f,\tilde{r}_i\right)\right)\right).$$

Otherwise, if $\tilde{d}_f > 1$, then by Markov's inequality and the definition of $\hat{\Delta}_{4m}^{(\tilde{d}_f)}(\cdot,\cdot,\cdot)$ from (15),

$$\mathbb{P}\left(\hat{\Delta}_{4m}^{(\tilde{d}_f)}(X_m, W_2, \mathrm{B}(f,\tilde{r}_i)) \geq 2\gamma/3 \middle| W_2\right) \leq \frac{3}{2\gamma}\mathbb{E}\left[\hat{\Delta}_{4m}^{(\tilde{d}_f)}(X_m, W_2, \mathrm{B}(f,\tilde{r}_i)) \middle| W_2\right]$$
$$= \frac{3}{2\gamma}\frac{1}{M_{4m}^{(\tilde{d}_f)}(\mathrm{B}(f,\tilde{r}_i))}\sum_{s=1}^{(4m)^3}\mathbb{P}\left(S_s^{(\tilde{d}_f)} \cup \{X_m\} \in \mathcal{S}^{\tilde{d}_f}(\mathrm{B}(f,\tilde{r}_i)) \middle| S_s^{(\tilde{d}_f)}\right).$$

By Lemma 39, Lemma 59, and (90), on $J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}$, this is at most

$$\frac{3}{\tilde{\delta}_f\gamma}\frac{1}{(4m)^3}\sum_{s=1}^{(4m)^3}\mathbb{P}\left(S_s^{(\tilde{d}_f)} \cup \{X_m\} \in \mathcal{S}^{\tilde{d}_f}(\mathrm{B}(f,\tilde{r}_i)) \middle| S_s^{(\tilde{d}_f)}\right)$$
$$\leq \frac{24}{\tilde{\delta}_f\gamma}\frac{1}{4^3 2^{3i+3}}\sum_{s=1}^{4^3 2^{3i+3}}\mathbb{P}\left(S_s^{(\tilde{d}_f)} \cup \{X_m\} \in \mathcal{S}^{\tilde{d}_f}(\mathrm{B}(f,\tilde{r}_i)) \middle| S_s^{(\tilde{d}_f)}\right).$$

Note that this value is invariant to the choice of $m \in \left\{2^i+1,\ldots,2^{i+1}\right\}$. By Hoeffding's inequality, on an event $J_n^*(i)$ of probability $\mathbb{P}(J_n^*(i)) \geq 1-\delta/(16i^2)$, this is at most

$$\frac{24}{\tilde{\delta}_f\gamma}\left(\sqrt{\frac{\ln(4i/\delta)}{4^3 2^{3i+3}}} + \mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}(\mathrm{B}(f,\tilde{r}_i))\right)\right). \tag{92}$$

Since $i \geq \hat{i}_1 > \log_2(n/4)$ and $n \geq \ln(1/\delta)$, we have

$$\sqrt{\frac{\ln(4i/\delta)}{4^3 2^{3i+3}}} \leq 2^{-i}\sqrt{\frac{\ln(4\log_2(n/4)/\delta)}{128n}} \leq 2^{-i}\sqrt{\frac{\ln(n/\delta)}{128n}} \leq 2^{-i}.$$

Thus, (92) is at most

$$\frac{24}{\tilde{\delta}_f\gamma}\left(2^{-i} + \mathcal{P}^{\tilde{d}_f}\left(\mathcal{S}^{\tilde{d}_f}(\mathrm{B}(f,\tilde{r}_i))\right)\right).$$

In either case ($\tilde{d}_f = 1$ or $\tilde{d}_f > 1$), by definition of $\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)$, on $J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}\cap J_n^*(i)$, $\forall m \in \left\{2^i+1,\ldots,2^{i+1}\right\}$ we have

$$\mathbb{P}\left(\hat{\Delta}_{4m}^{(\tilde{d}_f)}(X_m, W_2, \mathrm{B}(f,\tilde{r}_i)) \geq 2\gamma/3 \middle| W_2\right) \leq \frac{24}{\tilde{\delta}_f\gamma}\left(2^{-i} + \tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\cdot\max\left\{\tilde{r}_i, \varepsilon^{\frac{1}{\kappa}}\right\}\right). \tag{93}$$

Furthermore, the $\mathbb{1}_{[2\gamma/3,\infty)}\left(\hat{\Delta}_{4m}^{(\tilde{d}_f)}(X_m, W_2, \mathrm{B}(f,\tilde{r}_i))\right)$ indicators are conditionally independent given W_2, so that we may bound $\mathbb{P}\left(|Q_{i+1}| > \tilde{Q} \middle| W_2\right)$ via a Chernoff bound. Toward this end, note that on $J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}\cap J_n^*(i)$, (93) implies

$$\mathbb{E}\left[|Q_{i+1}| \middle| W_2\right] = \sum_{m=2^i+1}^{2^{i+1}}\mathbb{P}\left(\hat{\Delta}_{4m}^{(\tilde{d}_f)}(X_m, W_2, \mathrm{B}(f,\tilde{r}_i)) \geq 2\gamma/3 \middle| W_2\right)$$
$$\leq 2^i\cdot\frac{24}{\tilde{\delta}_f\gamma}\left(2^{-i} + \tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\cdot\max\left\{\tilde{r}_i, \varepsilon^{\frac{1}{\kappa}}\right\}\right) \leq \frac{24}{\tilde{\delta}_f\gamma}\left(1 + \tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\cdot\max\left\{2^i\tilde{r}_i, 2^{\tilde{i}}\varepsilon^{\frac{1}{\kappa}}\right\}\right). \tag{94}$$

Note that

$$
\begin{aligned}
2^i \tilde{r}_i &= \mu c^{\frac{1}{\kappa}} \left(di + \ln(1/\delta)\right)^{\frac{1}{2\kappa-1}} \cdot 2^{i\left(1-\frac{1}{2\kappa-1}\right)} \\
&\leq \mu c^{\frac{1}{\kappa}} \left(d\tilde{i} + \ln(1/\delta)\right)^{\frac{1}{2\kappa-1}} \cdot 2^{\tilde{i}\left(1-\frac{1}{2\kappa-1}\right)} \leq \mu c^{\frac{1}{\kappa}} \left(8d\log_2 \frac{2dc}{\varepsilon\delta}\right)^{\frac{1}{2\kappa-1}} \cdot 2^{\tilde{i}\left(1-\frac{1}{2\kappa-1}\right)}.
\end{aligned}
$$

Then since $2^{-\tilde{i}\frac{1}{2\kappa-1}} \leq \left(\frac{\varepsilon}{c}\right)^{\frac{1}{\kappa}} \cdot \left(8d\log_2 \frac{2dc}{\varepsilon\delta}\right)^{-\frac{1}{2\kappa-1}}$, we have that the rightmost expression in (94) is at most

$$
\frac{24}{\gamma\tilde{\delta}_f}\left(1+\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\cdot\mu\cdot 2^{\tilde{i}}\varepsilon^{\frac{1}{\kappa}}\right) \leq \frac{24}{\gamma\tilde{\delta}_f}\left(1+\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\cdot 2\mu c^2\cdot\left(8d\log_2\frac{2dc}{\varepsilon\delta}\right)\cdot\varepsilon^{\frac{2}{\kappa}-2}\right) \leq \tilde{\mathcal{Q}}/2.
$$

Therefore, a Chernoff bound implies that on $J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}\cap J_n^*(i)$, we have

$$
\begin{aligned}
\mathbb{P}\left(|\mathcal{Q}_{i+1}| > \tilde{\mathcal{Q}}\middle|W_2\right) &\leq \exp\left\{-\tilde{\mathcal{Q}}/6\right\} \leq \exp\left\{-8\log_2\left(\frac{2dc}{\varepsilon\delta}\right)\right\} \\
&\leq \exp\left\{-\log_2\left(\frac{48\log_2(2dc/\varepsilon\delta)}{\delta}\right)\right\} \leq \delta/(8\tilde{i}).
\end{aligned}
$$

Combined with the law of total probability and a union bound over i values, this implies there exists an event $J_n^*(\varepsilon,\delta) \subseteq J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}$ with

$$
\mathbb{P}\left(J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}\setminus J_n^*(\varepsilon,\delta)\right) \leq \sum_{i=\check{i}}^{\tilde{i}}\left(\delta/(16i^2)+\delta/(8\tilde{i})\right) \leq \delta/4,
$$

on which every $i \in \left\{\max\left\{\check{i},\hat{i}_{\tilde{d}_f-1}\right\},\ldots,\min\left\{\hat{i}_{\tilde{d}_f},\tilde{i}\right\}\right\}$ has $|\mathcal{Q}_{i+1}| \leq \tilde{\mathcal{Q}}$.

We have chosen c_1^* and c_2^* large enough that $2^{\tilde{i}+1} < \tilde{d}_f \cdot 2^n$ and $2^{\check{i}} < 2^{-\tilde{d}_f-2}n$. In particular, this means that on $J_n^*(\varepsilon,\delta)$,

$$
2^{\check{i}} + \sum_{i=\max\left\{\check{i},\hat{i}_{\tilde{d}_f-1}\right\}}^{\min\left\{\tilde{i},\hat{i}_{\tilde{d}_f}\right\}} |\mathcal{Q}_{i+1}| < 2^{-\tilde{d}_f-2}n + \tilde{i}\tilde{\mathcal{Q}}.
$$

Furthermore, since $\tilde{i} \leq 3\log_2\frac{4dc}{\varepsilon\delta}$, we have

$$
\begin{aligned}
\tilde{i}\tilde{\mathcal{Q}} &\leq \frac{2^{13}\mu c^2 d}{\gamma\tilde{\delta}_f}\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\cdot\varepsilon^{\frac{2}{\kappa}-2}\cdot\log_2^2\frac{4dc}{\varepsilon\delta} \\
&\leq \frac{2^{13}\mu c^2 d\log_2^2(4dc)}{\gamma\tilde{\delta}_f}\tilde{\theta}_f\left(\varepsilon^{\frac{1}{\kappa}}\right)\cdot\varepsilon^{\frac{2}{\kappa}-2}\cdot\log_2^2\frac{1}{\varepsilon\delta} \leq 2^{-\tilde{d}_f-2}n.
\end{aligned}
$$

Combining the above, we have that (91) is satisfied on $J_n^*(\varepsilon,\delta)$, so that $\hat{i}_{\tilde{d}_f} > \tilde{i}$. Combined with Lemma 59, this implies that on $J_n^*(\varepsilon,\delta)$,

$$
\hat{V}_{\hat{i}_{\tilde{d}_f}} \subseteq \hat{V}_{\tilde{i}} \subseteq \mathbb{C}\left(c\left(\frac{d\tilde{i}+\ln(1/\delta)}{2^{\tilde{i}}}\right)^{\frac{\kappa}{2\kappa-1}}\right),
$$

and by definition of $\tilde{i}$ we have

$$c\left(\frac{d\tilde{i}+\ln(1/\delta)}{2^{\tilde{i}}}\right)^{\frac{\kappa}{2\kappa-1}} \leq c\left(8d\log_2\frac{2dc}{\varepsilon\delta}\right)^{\frac{\kappa}{2\kappa-1}}\cdot 2^{-\tilde{i}\frac{\kappa}{2\kappa-1}}$$
$$\leq c\left(8d\log_2\frac{2dc}{\varepsilon\delta}\right)^{\frac{\kappa}{2\kappa-1}}\cdot(\varepsilon/c)\cdot\left(8d\log_2\frac{2dc}{\varepsilon\delta}\right)^{-\frac{\kappa}{2\kappa-1}} = \varepsilon,$$

so that $\hat{V}_{\hat{i}_{\tilde{d}_f}} \subseteq \mathbb{C}(\varepsilon)$.

Finally, to prove the stated bound on $\mathbb{P}(J_n^*(\varepsilon,\delta))$, by a union bound we have

$$1-\mathbb{P}\left(J_n^*(\varepsilon,\delta)\right) \leq \left(1-\mathbb{P}(J_n(\delta))\right)+\left(1-\mathbb{P}\left(H_n^{(i)}\right)\right)+\mathbb{P}\left(H_n^{(i)}\setminus H_n^{(ii)}\right)$$
$$+\mathbb{P}\left(J_n(\delta)\cap H_n^{(i)}\cap H_n^{(ii)}\setminus J_n^*(\varepsilon,\delta)\right)$$
$$\leq 3\delta/4 + c^{(i)}\cdot\exp\left\{-n^3\tilde{\delta}_f/8\right\}+c^{(ii)}\cdot\exp\left\{-n\tilde{\delta}_f^{1/3}/120\right\}\leq\delta.$$

■

We are now ready for the proof of Lemma 26.

Proof [Lemma 26] First, note that because we break ties in the argmax of Step 7 in favor of a $\hat{y}$ value with $V_{i_k+1}[(X_m,\hat{y})] \neq \emptyset$, if $V_{i_k+1} \neq \emptyset$ before Step 8, then this remains true after Step 8. Furthermore, the $\hat{U}_{i_k+1}$ estimator is nonnegative, and thus the update in Step 10 never removes from V_{i_k+1} the minimizer of $\mathrm{er}_{\hat{\mathcal{L}}_{i_k+1}}(h)$ among $h \in V_{i_k+1}$. Therefore, by induction we have $V_{i_k} \neq \emptyset$ at all times in Algorithm 5. In particular, $\hat{V}_{\hat{i}_{d+1}+1} \neq \emptyset$ so that the return classifier $\hat{h}$ exists. Also, by Lemma 60, for n as in Lemma 60, on $J_n^*(\varepsilon,\delta)$, running Algorithm 5 with label budget n and confidence parameter δ results in $\hat{V}_{\hat{i}_{\tilde{d}_f}} \subseteq \mathbb{C}(\varepsilon)$. Combining these two facts implies that for such a value of n, on $J_n^*(\varepsilon,\delta)$, $\hat{h} \in \hat{V}_{\hat{i}_{d+1}+1} \subseteq \hat{V}_{\hat{i}_{\tilde{d}_f}} \subseteq \mathbb{C}(\varepsilon)$, so that $\mathrm{er}\left(\hat{h}\right) \leq \nu+\varepsilon$. ■

E.3 The Misspecified Model Case

Here we present a proof of Theorem 28, including a specification of the method $\mathcal{A}'_a$ from the theorem statement.

Proof [Theorem 28] Consider a weakly universally consistent passive learning algorithm $\mathcal{A}_u$ (Devroye, Györfi, and Lugosi, 1996). Such a method must exist in our setting; for instance, Hoeffding's inequality and a union bound imply that it suffices to take $\mathcal{A}_u(\mathcal{L}) = \mathrm{argmin}_{\mathbb{1}^{\pm}_{B_i}} \mathrm{er}_{\mathcal{L}}(\mathbb{1}^{\pm}_{B_i}) + \sqrt{\frac{\ln(4i^2|\mathcal{L}|)}{2|\mathcal{L}|}}$, where $\{B_1, B_2, \ldots\}$ is a countable algebra that generates $\mathcal{F}_{\mathcal{X}}$.

Then $\mathcal{A}_u$ achieves a label complexity Λ_u such that for any distribution $\mathcal{P}_{XY}$ on $\mathcal{X}\times\{-1,+1\}$, $\forall\varepsilon\in(0,1)$, $\Lambda_u(\varepsilon+\nu^*(\mathcal{P}_{XY}),\mathcal{P}_{XY}) < \infty$. In particular, if $\nu^*(\mathcal{P}_{XY}) < \nu(\mathbb{C};\mathcal{P}_{XY})$, then we have $\Lambda_u((\nu^*(\mathcal{P}_{XY})+\nu(\mathbb{C};\mathcal{P}_{XY}))/2,\mathcal{P}_{XY}) < \infty$.

Fix any $n \in \mathbb{N}$ and describe the execution of $\mathcal{A}'_a(n)$ as follows. In a preprocessing step, withhold the first $m_{un} = n - \lfloor n/2\rfloor - \lfloor n/3\rfloor \geq n/6$ examples $\{X_1,\ldots,X_{m_{un}}\}$ and request their labels $\{Y_1,\ldots,Y_{m_{un}}\}$. Run $\mathcal{A}_a(\lfloor n/2\rfloor)$ on the remainder of the sequence $\{X_{m_{un}+1}, X_{m_{un}+2},\ldots\}$ (i.e., shift any index references in the algorithm by m_{un}), and let h_a denote the classifier it returns. Also request the labels $Y_{m_{un}+1}, \ldots Y_{m_{un}+\lfloor n/3 \rfloor}$, and let

$$h_u = \mathcal{A}_u\left(\left\{(X_{m_{un}+1}, Y_{m_{un}+1}), \ldots, (X_{m_{un}+\lfloor n/3 \rfloor}, Y_{m_{un}+\lfloor n/3 \rfloor})\right\}\right).$$

If $\mathrm{er}_{m_{un}}(h_a) - \mathrm{er}_{m_{un}}(h_u) > n^{-1/3}$, return $\hat{h} = h_u$; otherwise, return $\hat{h} = h_a$. This method achieves the stated result, for the following reasons.

First, let us examine the final step of this algorithm. By Hoeffding's inequality, with probability at least $1 - 2 \cdot \exp\left\{-n^{1/3}/12\right\}$,

$$|(\mathrm{er}_{m_{un}}(h_a) - \mathrm{er}_{m_{un}}(h_u)) - (\mathrm{er}(h_a) - \mathrm{er}(h_u))| \leq n^{-1/3}.$$

When this is the case, a triangle inequality implies $\mathrm{er}(\hat{h}) \leq \min\{\mathrm{er}(h_a), \mathrm{er}(h_u) + 2n^{-1/3}\}$.

If $\mathcal{P}_{XY}$ satisfies the benign noise case, then for any

$$n \geq 2\Lambda_a(\varepsilon/2 + \nu(\mathbb{C}; \mathcal{P}_{XY}), \mathcal{P}_{XY}),$$

we have $\mathbb{E}[\mathrm{er}(h_a)] \leq \nu(\mathbb{C}; \mathcal{P}_{XY}) + \varepsilon/2$, so $\mathbb{E}[\mathrm{er}(\hat{h})] \leq \nu(\mathbb{C}; \mathcal{P}_{XY}) + \varepsilon/2 + 2 \cdot \exp\{-n^{1/3}/12\}$, which is at most $\nu(\mathbb{C}; \mathcal{P}_{XY}) + \varepsilon$ if $n \geq 12^3 \ln^3(4/\varepsilon)$. So in this case, we can take $\lambda(\varepsilon) = \left\lceil 12^3 \ln^3(4/\varepsilon) \right\rceil$.

On the other hand, if $\mathcal{P}_{XY}$ is not in the benign noise case (i.e., the misspecified model case), then for any $n \geq 3\Lambda_u((\nu^*(\mathcal{P}_{XY}) + \nu(\mathbb{C}; \mathcal{P}_{XY}))/2, \mathcal{P}_{XY})$, $\mathbb{E}\left[\mathrm{er}(h_u)\right] \leq (\nu^*(\mathcal{P}_{XY}) + \nu(\mathbb{C}; \mathcal{P}_{XY}))/2$, so that

$$\begin{aligned}
\mathbb{E}[\mathrm{er}(\hat{h})] &\leq \mathbb{E}[\mathrm{er}(h_u)] + 2n^{-1/3} + 2 \cdot \exp\{-n^{1/3}/12\} \\
&\leq (\nu^*(\mathcal{P}_{XY}) + \nu(\mathbb{C}; \mathcal{P}_{XY}))/2 + 2n^{-1/3} + 2 \cdot \exp\{-n^{1/3}/12\}.
\end{aligned}$$

Again, this is at most $\nu(\mathbb{C}; \mathcal{P}_{XY}) + \varepsilon$ if $n \geq \max\left\{12^3 \ln^3 \frac{2}{\varepsilon}, 64(\nu(\mathbb{C}; \mathcal{P}_{XY}) - \nu^*(\mathcal{P}_{XY}))^{-3}\right\}$. So in this case, we can take

$$\lambda(\varepsilon) = \left\lceil \max\left\{12^3 \ln^3 \frac{2}{\varepsilon}, 3\Lambda_u\left(\frac{\nu^*(\mathcal{P}_{XY}) + \nu(\mathbb{C}; \mathcal{P}_{XY})}{2}, \mathcal{P}_{XY}\right), \frac{64}{(\nu(\mathbb{C}; \mathcal{P}_{XY}) - \nu^*(\mathcal{P}_{XY}))^3}\right\} \right\rceil.$$

In either case, we have $\lambda(\varepsilon) \in \mathrm{Polylog}(1/\varepsilon)$. ■

References

N. Abe and H. Mamitsuka. Query learning strategies using boosting and bagging. In *Proceedings of the* 15th *International Conference on Machine Learning*, 1998.

M. Alekhnovich, M. Braverman, V. Feldman, A. Klivans, and T. Pitassi. Learnability and automatizability. In *Proceedings of the* 45th *Foundations of Computer Science*, 2004.

K. Alexander. Probability inequalities for empirical processes and a law of the iterated logarithm. *The Annals of Probability*, 4:1041–1067, 1984.

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

A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Machine Learning*, 30:31–56, 1998.

R. B. Ash and C. A. Doléans-Dade. *Probability & Measure Theory*. Academic Press, 2000.

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

M.-F. Balcan, A. Blum, and S. Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning Journal*, 65(1):79–94, 2006b.

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the* 20th *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.

J. Baldridge and A. Palmer. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2009.

Z. Bar-Yossef. Sampling lower bounds via information theory. In *Proceedings of the* 35th *Annual ACM Symposium on the Theory of Computing*, 2003.

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

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 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.

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

F. Bunea, A. B. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2009.

C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *Proceedings of the* 17th *International Conference on Machine Learning*, 2000.

R. Castro and R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

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

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

S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the* 18th *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.

S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10:281–299, 2009.

O. Dekel, C. Gentile, and K. Sridharan. Robust selective sampling from single and multiple teachers. In *Proceedings of the* 23rd *Conference on Learning Theory*, 2010.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, Inc., 1996.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

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* 22nd *Conference on Learning Theory*, 2009.

R. Gangadharaiah, R. D. Brown, and J. Carbonell. Active learning in example-based machine translation. In *Proceedings of the* 17th *Nordic Conference on Computational Linguistics*, 2009.

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

S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.

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

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

S. Hanneke. Adaptive rates of convergence in active learning. In *Proceedings of the* 22nd *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. Har-Peled, D. Roth, and D. Zimak. Maximum margin coresets for active and noise tolerant learning. In *Proceedings of the* 20th *International Joint Conference on Artificial Intelligence*, 2007.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

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

T. Hegedüs. Generalized teaching dimension 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.

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

S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the* 23rd *International Conference on Machine Learning*, 2006.

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

N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4:373–395, 1984.

M. J. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1994.

M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

L. G. Khachiyan. A polynomial algorithm in linear programming. *Soviet Mathematics Doklady*, 20:191–194, 1979.

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*, 11:2457–2485, 2010.

V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems. In *École d'Été de Probabilités de Saint-Flour XXXVIII-2008*. Lecture Notes in Mathematics, 2033, Springer, 2011.

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

M. Lindenbaum, S. Markovitch, and D. Rusakov. Selective sampling for nearest neighbor classifiers. *Machine Learning*, 54:125–152, 2004.

T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, 6: 589–613, 2005.

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: 1808–1829, 1999.

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

A. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *Proceedings of the* 15th *International Conference on Machine Learning*, 1998.

P. Mitra, C. A. Murthy, and S. K. Pal. A probabilistic active support vector learning algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):413–418, 2004.

J. R. Munkres. *Topology*. Prentice Hall, Inc., 2nd edition, 2000.

I. Muslea, S. Minton, and C. A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the* 19th *International Conference on Machine Learning*, 2002.

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

L. Pitt and L. G. Valiant. Computational limitations on learning from examples. *Journal of the Association for Computing Machinery*, 35(4):965–984, 1988.

J. Poland and M. Hutter. MDL convergence speed for Bernoulli sequences. *Statistics and Computing*, 16:161–175, 2006.

G. V. Rocha, X. Wang, and B. Yu. Asymptotic distribution and sparsistency for l1-penalized parametric M-estimators with applications to linear SVM and logistic regression. *arXiv:0908.1940v1*, 2009.

D. Roth and K. Small. Margin-based active learning for structured output spaces. In *European Conference on Machine Learning*, 2006.

N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the* 18th *International Conference on Machine Learning*, 2001.

A. I. Schein and L. H. Ungar. Active learning for logistic regression: An evaluation. *Machine Learning*, 68(3):235–265, 2007.

G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *Proceedings of the* 17th *International Conference on Machine Learning*, 2000.

B. Settles. Active learning literature survey. *http://active-learning.net*, 2010.

S. M. Srivastava. *A Course on Borel Sets*. Springer-Verlag, 1998.

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2, 2001.

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

L. G. Valiant. A theory of the learnable. *Communications of the Association for Computing Machinery*, 27(11):1134–1142, 1984.

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

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

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

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945.

L. Wang. Sufficient conditions for agnostic active learnable. In *Advances in Neural Information Processing Systems 22*, 2009.

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

L. Wang and X. Shen. On L1-norm multiclass support vector machines. *Journal of the American Statistical Association*, 102(478):583–594, 2007.

L. Yang, S. Hanneke, and J. Carbonell. The sample complexity of self-verifying Bayesian active learning. In *Proceedings of the* 14^{th} *International Conference on Artificial Intelligence and Statistics*, 2011.