Toward a General Theory of Online Selective Sampling: Trading Off Mistakes and Queries

Steve Hanneke
Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Liu Yang
liu.yang0900@outlook.com

Abstract

While the literature on the theory of pool-based active learning has seen much progress in the past 15 years, and is now fairly mature, much less is known about its cousin problem: online selective sampling. In the stochastic online learning setting, there is a stream of iid data, and the learner is required to predict a label for each instance, and we are interested in the rate of growth of the number of mistakes the learner makes. In the selective sampling variant of this problem, after each prediction, the learner can optionally request to observe the true classification of the point. This introduces a trade-off between the number of these queries and the number of mistakes as a function of the number T of samples in the sequence. This work explores various properties of the optimal trade-off curve, both abstractly (for general VC classes), and more-concretely for several constructed examples that expose important properties of the trade-off.

1 Introduction

One common setting arising in practical machine learning is online prediction, where we are faced with a stream of test points X_t, and for each we are tasked with making a prediction $\hat{Y}_t$ for the value of some unobserved target variable Y_t. In such tasks, we are interested in achieving a small cumulative number of *mistakes*, where $\hat{Y}_t \neq Y_t$, as the total number T of rounds grows. To facilitate this, it is crucial that the learner is able to access some information about the Y_t values. At the extreme end of this, the traditional setting of *supervised* online learning supposes that, after every prediction, the learner is informed of the correct classification Y_t (Littlestone, 1988; Haussler, Littlestone, and Warmuth, 1994). However, there are a vast number of learning scenarios such that significant effort or resources would be required in order to determine the target label Y_t. For machine learning to be most useful in such scenarios, it is worthwhile to explore strategies that do not rely on access to Y_t after every prediction. In particular, in the present work, we consider *selective sampling* strategies.

After each prediction, a selective sampling strategy makes a decision about whether or not it wishes to observe the target label Y_t. A decision to observe Y_t is referred to as a "label query", following the active learning literature. The fact that the learner might refrain from querying some Y_t values introduces a trade-off between the number of queries and number of mistakes. Clearly a selective sampling algorithm cannot make fewer mistakes than a learner that observes every Y_t. But, for any larger number $\mathcal{M}_T$ of mistakes, we are interested in understanding how many queries $\mathcal{Q}_T$ are needed to guarantee at most that many mistakes, or vice versa.

The formal setting we consider in this work is the *stochastic* online setting, in which the sequence $X_1, \ldots, X_T$ is sampled iid from an unknown distribution $\mathcal{P}$ on a space $\mathcal{X}$. For simplicity, we also focus on the *realizable* case, wherein there is a fixed hypothesis class $\mathcal{H}$, and it is assumed that there is some (unknown) *target function* $f^\star \in \mathcal{H}$ such that $Y_t = f^\star(X_t)$ for all t. This problem has been studied in great detail in the traditional supervised learning setting (Vapnik and Chervonenkis, 1974; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Ehrenfeucht, Haussler, Kearns, and Valiant, 1989; Haussler, Littlestone, and Warmuth, 1994; Hanneke, 2016a). In particular, a very tight characterization of the minimax optimal expected number of mistakes was been established by Haussler, Littlestone, and Warmuth (1994), who showed that if $\mathcal{H}$ is infinite but has finite VC dimension (see definition

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

below), then the optimal expected number of mistakes in T rounds grows as $\Theta(\log(T))$.

1.1 Definitions

Before proceeding, we introduce some notation. There is an *instance space* $\mathcal{X}$ and a *label space* $\mathcal{Y} = \{-1, 1\}$, and we assume that $\mathcal{X}$ is equipped with a σ-algebra defining the measurable subsets. A *classifier* is a measurable function $\mathcal{X} \to \mathcal{Y}$. There is also a *hypothesis class* $\mathcal{H}$ of classifiers, and we denote by d the *VC dimension* of $\mathcal{H}$: that is, d is the largest n s.t. $\exists x_1, \ldots, x_n \in \mathcal{X}$ with all 2^n possible classifications realized by classifiers in $\mathcal{H}$. Throughout, we assume $d < \infty$.

In the learning problem, there is a probability measure $\mathcal{P}$ over $\mathcal{X}$ and an unknown *target concept* $f^\star \in \mathcal{H}$, and a *data sequence*: independent random variables $(X_1, Y_1), (X_2, Y_2), \ldots$ with $X_t \sim \mathcal{P}$ and $Y_t = f^\star(X_t)$; for notational simplicity, we leave the dependence of (X_t, Y_t) on the choice of $\mathcal{P}$ and $f^\star$ implicit, as the specific $\mathcal{P}$ and $f^\star$ will always be clear from context.

A selective sampling rule $\mathcal{A}$ is an algorithm that produces two sequences: $Q_t \in \{0, 1\}$ and $\hat{Y}_t \in \mathcal{Y}$. Both Q_t and $\hat{Y}_t$ may depend only on $X_1, \ldots, X_t$ and the subsequence $\{Y_{t'} : Q_{t'} = 1, t' < t\}$, and possibly additional random bits independent of the data sequence. We interpret Q_t as the indicator of whether the algorithm *queries* for the true label Y_t of X_t, and we interpret $\hat{Y}_t$ as the algorithm's *prediction* for the label of X_t.

We are interested in quantifying the total number of *mistakes* and *queries* for initial segements of the data sequence. Specifically, for the sequences $\hat{Y}_t$ and Q_t produced by an algorithm $\mathcal{A}$, define $\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P}) = \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[\hat{Y}_t \neq Y_t]\right]$ (the expected number of mistakes) and $\mathcal{Q}_T(\mathcal{A}; f^\star, \mathcal{P}) = \mathbb{E}\left[\sum_{t=1}^T Q_t\right]$ (the expected number of queries). Also define *worst-case* values of these: for distribution-free analysis, define $\mathcal{M}_T(\mathcal{A}) = \sup_{\mathcal{P}} \sup_{f^\star \in \mathcal{H}} \mathcal{M}_T(\mathcal{A}; \mathcal{P})$ and $\mathcal{Q}_T(\mathcal{A}) = \sup_{\mathcal{P}} \sup_{f^\star \in \mathcal{H}} \mathcal{Q}_T(\mathcal{A}; \mathcal{P})$; for distribution-dependent analysis, for any distribution $\mathcal{P}$, define $\mathcal{M}_T(\mathcal{A}; \mathcal{P}) = \sup_{f^\star \in \mathcal{H}} \mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P})$ and $\mathcal{Q}_T(\mathcal{A}; \mathcal{P}) = \sup_{f^\star \in \mathcal{H}} \mathcal{Q}_T(\mathcal{A}; f^\star, \mathcal{P})$.

For sequences $x_1, x_2, \ldots$, we let $x_{1:t} = \{x_1, \ldots, x_t\}$, and with a slight abuse we sometimes write $(x_{1:t}, y_{1:t})$ to denote $\{(x_1, y_1), \ldots, (x_t, y_t)\}$. Throughout, we write $a \lesssim b$ or $b \gtrsim a$ to indicate that there exists a numerical constant c such that $a \leq cb$. Also, we use big-O (and little-o) notation, interpreted strictly as indicating asymptotic dependence on the number of rounds (data points) T, considering any dependence on $\mathcal{H}$ (e.g., VC dimension) or other parameters of the relevant function as constant: for instance, in the statement "$\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P}) = O(\log(T))$," the big-$O$ may hide constants depending on $\mathcal{H}$, $\mathcal{A}$, $f^\star$, and $\mathcal{P}$.

1.2 Related Work

We briefly survey the related work on selective sampling. The reader is referred to (Hanneke, 2014) for a more-comprehensive summary of much of this literature.

Pool-Based Active Learning. In the pool-based active learning setting, the learner observes the full data set $X_1, \ldots, X_T$ at once, and may sequentially query for any of the labels Y_t without regard to their index order. Rather than number of mistakes, this setting is only concerned with the number of queries sufficient to learn a classifier $\hat{h}$ with low *error rate*: $\mathcal{P}(x : \hat{h}(x) \neq f^\star(x))$. The literature on the theory of pool-based active learning is vast, and we do not attempt a full summary here (see Hanneke, 2014). Instead, we briefly mention a few works particularly relevant to the present article. A classic approach to pool-based active learning is the *disagreement-based* strategy proposed by Cohn, Atlas, and Ladner (1994), known as CAL after these authors. The basic strategy is to query points for which there is some disagreement among classifiers in $\mathcal{H}$ consistent with previously queried labels. We discuss CAL in detail in Section 2.2 below. The number of queries sufficient for CAL to achieve a given error rate has been tightly characterized in terms of various related complexity measures such as the *disagreement coefficient*, *version space compression set size*, and *star number* (Hanneke, 2007, 2009, 2011, 2012, 2016b; Wiener, Hanneke, and El-Yaniv, 2015; Hanneke and Yang, 2015). Some refinements of the CAL algorithm have also been proposed, with a general approach of querying points in a well-chosen *subregion* of the region of disagreement (Dasgupta, Tauman Kalai, and Monteleoni, 2009; Balcan, Broder, and Zhang, 2007; Zhang and Chaudhuri, 2014). A more-sophisticated general technique, known as *splitting*, was proposed by Dasgupta (2005). One important advantage of splitting over the above methods is that it allows for a *trade-off* between the number of queries and the number of *unlabeled* samples. We discuss this technique in detail in Section 4, and we adapt the pool-based active learning strategy to be suitable for the online selective sampling setting. Another relevant thread of the pool-based active learning literature is on asymptotic improvements over passive learning in a distribution-dependent *and target-dependent* analysis. Specifically, Hanneke (2009, 2012); Balcan, Hanneke, and Vaughan (2010) proved that, if $d < \infty$, then for any passive learning algorithm (i.e., querying all T labels), there exists an active learning algorithm such that the number of queries sufficient to achieve error rate ϵ has a strictly slower rate of growth (as $\epsilon \to 0$)

compared to the number of samples the passive algorithm would require to achieve the same error rate ϵ. In particular, this implies there is an active learning algorithm that achieves ϵ error rate using a number of queries that is $o(1/\epsilon)$, something known to be impossible for passive learning algorithms Antos and Lugosi (1998); Schuurmans (1997). Section 5 discusses the natural corresponding question for the online selective sampling setting: namely, whether it is possible to use a number of queries growing *sublinearly* in T, while still making a number of mistakes competitive with a fully-supervised online prediction algorithm (namely, $O(\log(T))$ mistakes).

Stream-based Active Learning. A setting intermediate between pool-based active learning and online selective sampling is the *stream-based* active learning setting (e.g., Dasgupta, Tauman Kalai, and Monteleoni, 2009; Freund, Seung, Shamir, and Tishby, 1997; Sabato and Hess, 2018). In this case, the learning algorithm observes the X_t samples in sequence, and for each one must decide whether to query or not (as in online selective sampling). However, as in the pool-based setting, the objective in the end is to produce a classifier $\hat{h}$ of low error rate, and there is no requirement to make predictions for the X_t samples. Many of the above pool-based active learning strategies have also been expressed and studied in the stream-based setting. Generally, Sabato and Hess (2018) showed that stream-based active learning is essentially equivalent to pool-based active learning, in that any pool-based method can be converted to a stream-based method with roughly the same query complexity, though with a potential increase in the necessary number of unlabeled samples.

Online Selective Sampling. In contrast to pool-based and stream-based active learning, the existing literature on online selective sampling is relatively sparse. In some cases, analyses of the stream-based expressions of certain active learning algorithms (such as CAL) reveal bounds on the error rate one can guarantee if the algorithm were stopped at any sample size T. Such guarantees trivially lead to results for these methods in the online selective sampling setting, expressing an expected number of queries (based on the analysis from the stream-based setting) and an expected number of mistakes (based on summing the error rate guarantees over the T rounds); see e.g., (Yang, 2011) for an explicit expression of such a result. There have been several works proposing various online selective sampling algorithms, each accompanied by specific bounds on the number of mistakes and queries under various specialized contexts, conditions, and assumptions (e.g., Cesa-Bianchi, Gentile, and Zaniboni, 2006; Dekel, Gentile, and Sridharan, 2012; Yang, 2011; Hanneke, Kanade, and Yang, 2015). These results are all interesting and valuable, and taken altogether we may get a partial picture of the mistakes-vs-queries feasible region. In the present work we are interested in initiating the direct study of the optimal trade-off curve itself, with the aim of leading toward a general theory applicable to any hypothesis class $\mathcal{H}$. Our main approach at this stage is to examine general principles, in combination with specially constructed examples that demonstrate that certain important behaviors are sometimes possible. We also provide some more-speculative stabs toward a general theory, proposing one approach to the design of abstract selective sampling strategies, and offering a conjecture about achievable asymptotic behaviors: both of these presented with the aim of stirring future work in this direction.

1.3 Summary of Main Results

This paper represents a first step toward a general theory of online selective sampling, exploring the trade-off between number of mistakes vs number of queries. We specifically study three levels of dependence in the analysis: distribution-free, distribution-dependent, and distribution- and target-dependent. In these contexts, we make the following contributions:

- We identify a family of algorithms, termed the *trivial modifications of* CAL, which serves as a key baseline for comparison throughout.
- In distribution-free analysis, we prove that the optimal points $(\mathcal{M}_T(\mathcal{A}), \mathcal{Q}_T(\mathcal{A}))$ are always (nearly) achieved by trivial modifications of CAL.
- In contrast, in a distribution-dependent analysis, we show that there exist spaces $\mathcal{H}$ and distributions $\mathcal{P}$ where there are feasible points $(\mathcal{M}_T(\mathcal{A};\mathcal{P}), \mathcal{Q}_T(\mathcal{A};\mathcal{P}))$ achievable by *some* $\mathcal{A}$, yet no $\mathcal{A}'$ trivial modification of CAL can achieve $(\mathcal{M}_T(\mathcal{A}';\mathcal{P}), \mathcal{Q}_T(\mathcal{A}';\mathcal{P}))$ close to it.
- In a distribution-dependent and target-dependent analysis, we pose the question of whether there always exists a selective sampling algorithm $\mathcal{A}$ with $\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P}) = O(\log(T))$ so that it is asymptotically competitive with fully-supervised learning, yet $\mathcal{Q}_T(\mathcal{A}; f^\star, \mathcal{P}) = o(T)$ so that asymptotically it very rarely needs to query for a label.
- We define a general flexible online selective sampling strategy applicable to any hypothesis class $\mathcal{H}$, a variant of the *splitting* active learning strategy.

Altogether, these results reveal that online selective sampling is an interesting and rich subject to be explored, which is not fully addressed (even implicitly)

by the existing general theories of active learning, especially when the analysis depends on the distribution $\mathcal{P}$ or target function $f^\star$.

2 Distribution-Free Analysis

As a starting point, we first present a distribution-free analysis. It turns out that the optimal trade-off between distribution-free guarantees on mistakes and queries is always nearly achieved by certain trivial modifications of the CAL active learning algorithm. We note that this fact is not particularly surprising, since CAL is known to be nearly optimal in the pool-based setting as well (for distribution-free analysis), and yet guarantees a number of mistakes comparable to a fully-supervised learning algorithm. However, this family of modifications of CAL will serve as the baseline for comparison in later sections where we discuss distribution-dependent analysis (in which we will find more-interesting non-trivial trade-offs are possible).

2.1 Trivial Modifications

For any given selective sampling algorithm, one can trivially generate an entire spectrum of different behaviors for the number of mistakes and queries, simply by arbitrarily preventing the algorithm from querying at certain times. As a simple illustrative example of this, consider the *fully-supervised* learning method of *empirical risk minimization* (ERM), which simply queries *every* label and chooses any $\hat{h}_t \in \mathcal{H}$ with $h(X_{t'}) = Y_{t'}$ for all $t' < t$, and predicts $\hat{Y}_t = \hat{h}_t(X_t)$. We can trivially modify this algorithm to reduce the number of queries from T to $T/2$ simply by only querying every *second* label, and predicting each $\hat{Y}_t$ based on choosing any $\hat{h}_t \in \mathcal{H}$ with $h(X_{t'}) = Y_{t'}$ for every $t' < t$ for which $Y_{t'}$ *was* queried. In this way, though the number of queries is reduced by a factor of 2, the expected number of mistakes would correspondingly *increase* by a factor of 2. We refer to this as a *trivial modification* of the algorithm.

More generally, for any selective sampling rule $\mathcal{A}$, we define the *trivial* modifications of $\mathcal{A}$, denoted by $\mathbb{TM}(\mathcal{A})$, as a family of algorithms $\mathcal{A}'$ given by defining any deterministic set $\mathcal{I} \subseteq \mathbb{N}$, and running $\mathcal{A}$ with the *subsequence* $\{(X_i, Y_i) : i \in \mathcal{I}\}$: that is, for each t, values of Q_t and $\hat{Y}_t$ are produced as if the data sequence so far is $\{(X_i, Y_i) : i < t, i \in \mathcal{I}\}$; in particular, the values of Q_t and $\hat{Y}_t$ will only depend on $\{X_i : i < t, i \in \mathcal{I}\}$, X_t, and the values $\{Y_i : Q_i = 1, i < t, i \in \mathcal{I}\}$. We will also only query Y_t if $t \in I$ (and $Q_t = 1$), so that we generally force $Q_t = 0$ if $t \notin \mathcal{I}$.

The point here is that there is a screening process that is independent of the X_t example, which determines whether we will even feed the example X_t into the base selective sampling algorithm $\mathcal{A}$ (besides for the purpose of generating a prediction $\hat{Y}_t$). These *trivial modifications* allow us to vary the trade-off between the number of queries and number of mistakes via the specification of $\mathcal{I}$, without actually significantly changing the general strategy of the algorithm.

2.2 Baseline: Trivial Modifications of CAL

In the theoretical active learning literature, one of the earliest-proposed general methods for realizable-case active learning is a disagreement-based technique proposed by Cohn, Atlas, and Ladner (1994), usually referred to as CAL after its authors. The algorithm is particularly interesting for the fact that it loses *no* information compared to passive learning: that is, it only elects not to query Y_t if the value of Y_t can be perfectly *inferred* based on the examples observed so far. Formally, the algorithm is defined as follows. Let $\mathcal{A}_p : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \mathcal{Y}$ be a (passive) supervised learning algorithm: that is, $\mathcal{A}_p$ takes a labeled training set, and a test point x, and returns a prediction y for the label of x. For our present purposes, we specifically take $\mathcal{A}_p$ as any minimax-optimal passive learner (i.e., guaranteeing $\mathbb{P}(\mathcal{A}_p(X_{1:t}, Y_{1:t}, X_{t+1}) \neq Y_{t+1}) \lesssim \frac{d}{t}$), such as the one-inclusion graph predictor of Haussler, Littlestone, and Warmuth (1994) or the optimal PAC learner of Hanneke (2016a).

Algorithm: CAL
0. For $t = 1, 2, \ldots$
1. Predict $\hat{Y}_t = \mathcal{A}_p(X_{1:(t-1)}, \tilde{Y}_{1:(t-1)}, X_t)$ for Y_t
2. Let $V_{t-1} = \{h \in \mathcal{H} : h(X_{1:(t-1)}) = \tilde{Y}_{1:(t-1)}\}$
3. If $\exists h, h' \in V_{t-1}$ with $h(X_t) \neq h'(X_t)$
4. Set $Q_t = 1$ (query for Y_t) and define $\tilde{Y}_t = Y_t$
5. Else set $Q_t = 0$ and define $\tilde{Y}_t = h(X_t)$ agreed by all $h \in V_{t-1}$

There is a significant body of work on the analysis of CAL and related methods (see Hanneke, 2009, 2011, 2012, 2014, 2016b; Wiener, Hanneke, and El-Yaniv, 2015; Hanneke and Yang, 2015), and CAL is known to be nearly minimax optimal in its distribution-free label complexity guarantees (Hanneke and Yang, 2015; Hanneke, 2016b). Our aim in this section is to argue that this optimality also extends to distribution-free analysis of the number of mistakes and queries. However, in this case, since there can be an entire spectrum of "optimal" $\mathcal{Q}_T$ values, depending on the constraint on $\mathcal{M}_T$ (or vice versa), we cannot simply analyze the single algorithm CAL. Instead, we argue that every point on the optimal $\mathcal{Q}_T$ vs $\mathcal{M}_T$ trade-off is nearly matched by one of the trivial modifications of CAL. Specifically, we establish the following result.

Theorem 1. *For any selective sampling algorithm $\mathcal{A}$ and any T, there exists $\mathcal{A}' \in \mathbb{TM}(\text{CAL})$ such that $\mathcal{M}_T(\mathcal{A}') \lesssim d\mathcal{M}_T(\mathcal{A}) + d\log(T)$ and $\mathcal{Q}_T(\mathcal{A}') \lesssim \mathcal{Q}_T(\mathcal{A})\log(T)$.*

Our proof of this result is comprised of two parts, presented in Theorems 3 and 4 below. The proofs of these results take their roots in the work of Hanneke and Yang (2015) characterizing the optimal distribution-free label complexity of pool-based active learning in terms of the *star number* (defined below), together with upper bounds for CAL in terms of the star number from Hanneke (2016b). Our contribution on top of these results is to extend those arguments to the entire spectrum of $\mathcal{Q}_T$ vs $\mathcal{M}_T$ trade-offs. We rely on the following definition.

Definition 2. *(Hanneke and Yang, 2015) The* star number*, denoted by $\mathfrak{s}$, is the largest $s \in \mathbb{N}$ such that there exist $x_1, \ldots, x_s \in \mathcal{X}$ and $h_0, h_1, \ldots, h_s \in \mathcal{H}$ such that, $\forall i$, $\{j : h_i(x_j) \neq h_0(x_j)\} = \{i\}$; we say $\{x_1, \ldots, x_s\}$ is a* star set *witnessed by $h_0, \ldots, h_s$. If no such largest s exists, define $\mathfrak{s} = \infty$.*

We have the following theorems, which together imply Theorem 1; the proofs are deferred to Appendix A, along with the proof of Theorem 1 based on them.

Theorem 3. *For any $q_T \leq T$, there exists $\mathcal{A} \in \mathbb{TM}(\text{CAL})$ such that $\mathcal{Q}_T(\mathcal{A}) \leq q_T$ and*

$$\mathcal{M}_T(\mathcal{A}) \leq d\log(T) + \frac{dT}{q_T}\mathbb{1}[q_T < \mathfrak{s}\ln(eT)]\,.$$

The following minimax lower bound reveals that the upper bound for the trivial modifications of CAL in Theorem 3 are essentially the best achievable by any selective sampling algorithm, up to log factors and dependence on d.

Theorem 4. *For any selective sampling algorithm $\mathcal{A}$,*

$$\mathcal{M}_T(\mathcal{A}) \gtrsim \min\{d, T\} + \frac{T}{\mathcal{Q}_T(\mathcal{A})}\mathbb{1}[\mathcal{Q}_T(\mathcal{A}) < \mathfrak{s}/16]\,.$$

3 Distribution-Dependent Analysis

Above, we found that in distribution-free analysis, *no* significant improvements over the trivial modifications of CAL are possible. As noted by Dasgupta (2005), allowing distribution-dependence in the analysis can sometimes be important for revealing advantages of active learning over passive learning. The present section explores whether distribution-dependent analysis may also be able to reveal advantages over the trivial modifications of CAL. In contrast to the distribution-free analysis, we find that in distribution-dependent analysis significant improvements over the trivial modifications of CAL *are* possible.

The main demonstration of the possibility of such improvements is via a carefully-constructed example, witnessing the improvements. We also discuss a more abstract approach to designing distribution-specific selective sampling algorithms, based on the *splitting* approach of Dasgupta (2005).

3.1 An Example: Significant Improvements over the Trivial Modifications of CAL

The purpose of this subsection is to describe a construction of a learning problem that verifies the following proposition, which claims that there exist distribution-dependent scenarios where some algorithms are capable of achieving a number of mistakes and queries that no trivial modification of CAL can come close to simultaneously achieving.

Proposition 5. *There exists a space $\mathcal{X}$, a hypothesis class $\mathcal{H}$ with VC dimension $d = 1$, and a distribution $\mathcal{P}$ such that, there is a selective sampling algorithm $\mathcal{A}$ that, for infinitely many T, satisfies $\mathcal{Q}_T(\mathcal{A};\mathcal{P}) \lesssim \log^2(T)$ and $\mathcal{M}_T(\mathcal{A};\mathcal{P}) \lesssim (T\log(T))^{1/2}$, yet for these same times T, $\forall \mathcal{A}' \in \mathbb{TM}(\text{CAL})$, if $\mathcal{Q}_T(\mathcal{A}';\mathcal{P}) < T^{1/17}$, then $\mathcal{M}_T(\mathcal{A}';\mathcal{P}) > T^{7/8}$.*

The construction of this space $\mathcal{H}$ and distribution $\mathcal{P}$ is rather technical, largely due to the fact that it should be a *single* $\mathcal{H}$ and $\mathcal{P}$ for *all* T. However, it is based on a recursive application of a familiar construction of Hanneke (2014) exhibiting a case where CAL is suboptimal for pool-based active learning. Specifically, the basic idea is to have two regions, one with high probability mass and all of the points in disagreement, but very few nontrivially-informative points, and another region with very low probability mass, but where it is easy to identify highly-informative points.

We apply this idea recursively, so that a single distribution can remain fixed as $T \to \infty$ in Proposition 5. Such recursive constructions of difficult active learning problems have roots in the work of Balcan, Hanneke, and Vaughan (2010). Our construction below represents a coupling of these two types of constructions. There are a number of substantial technical challenges addressed in the proofs in order to successfully couple these constructions, and particularly to adapt the techniques to the online selective sampling setting.

We now proceed to describe the construction. Let $g : (0,1) \to \mathbb{N}$ be a nonincreasing function such that $g(\varepsilon) \to \infty$ and $\varepsilon g(\varepsilon) \to 0$ as $\varepsilon \to 0$. Let $\{\ell_i\}_{i=1}^{\infty}$ be any sequence of strictly positive values with $\sum_{i=1}^{\infty} \ell_i = 1$. Let $p_1 \in (0, 1/2)$ satisfy $g(p_1) \geq 4$ and $p_1 g(p_1) \leq \ell_1$, and for each integer $i \geq 2$, inductively define p_i as any value in $(0, 1/2)$ with $p_i \prod_{j=1}^{i} g(p_j) \leq \min\{p_{i-1}, \ell_i\}$, $g(p_i) \geq \prod_{j=1}^{i-1} g(p_j)$, $p_i g(p_i) \leq p_{i-1}/2$, and $g(p_i)^2 \geq$

$g(p_{i-1})^2/p_{i-1}$. Also define $p_0 = 1-\sum_{i=1}^{\infty} p_i \prod_{j=1}^{i} g(p_j)$, which is nonnegative since every $p_i \prod_{j=1}^{i} g(p_j) \leq \ell_i$. In particular, Proposition 5 will be established by choosing $g(\epsilon) = \lfloor \epsilon^{-1/2} \rfloor$, $\ell_i = 2^{-i}$, and $p_i = 2^{-2^{2i+1}}$.

Now consider two infinite trees (constructed in parallel) defined as follows. The set of nodes can be described as distinct points $x_{\mathbf{z}}$ (in the first tree) and $y_{\mathbf{z}}$ (in the second tree), defined for every $\mathbf{z} = (z_1, \ldots, z_k)$, $k \in \mathbb{N}$, with $\forall i \leq k$, $z_i \in \{1, \ldots, g(p_i)\}$, and also defined for $\mathbf{z} = ()$. Define $x_{()}$ as the "root" node in the first tree, and $y_{()}$ as the "root" node in the second tree, and we define $\text{Children}(x_{()}) = \{x_{(1)}, \ldots, x_{(g(p_1))}\}$ and $\text{Children}(y_{()}) = \{y_{(1)}, \ldots, y_{(g(p_1))}\}$. For any $k \in \mathbb{N}$ and $\mathbf{z} = (z_1, \ldots, z_k)$ s.t. $\forall i \leq k$, $z_i \in \{1, \ldots, g(p_i)\}$, define $\text{Children}(x_{\mathbf{z}}) = \{x_{(z_1,\ldots,z_k,j)} : j \in \{1, \ldots, g(p_{k+1})\}\}$ and $\text{Children}(y_{\mathbf{z}}) = \{y_{(z_1,\ldots,z_k,j)} : j \in \{1, \ldots, g(p_{k+1})\}\}$. This defines the structure of the trees, and we let $\mathcal{X}$ be simply the set of all such nodes $x_{\mathbf{z}}$ and $y_{\mathbf{z}}$. For any node x defined in either of these trees, we denote by $\text{Subtree}(x)$ the set of all nodes in the subtree rooted at x: that is, we inductively define $\text{Subtree}(x) = \{x\} \cup \bigcup_{x' \in \text{Children}(x)} \text{Subtree}(x')$.

To specify the probability measure $\mathcal{P}$, define $\mathcal{P}(\{x_{()}\}) = (1/2)\mathcal{P}(\{y_{()}\}) = (1/2)p_0$, and for any $k \in \mathbb{N}$, letting $\alpha_k = 1/g(p_k)^2$, for any $\mathbf{z} = (z_1, \ldots, z_k)$ with $\forall i \leq k$, $z_i \in \{1, \ldots, g(p_i)\}$, define $\mathcal{P}(\{x_{\mathbf{z}}\}) = (1-\alpha_k)p_k$ and $\mathcal{P}(\{y_{\mathbf{z}}\}) = \alpha_k p_k$. This uniquely defines a probability measure on $\mathcal{X}$.

Next, we specify the concept space $\mathcal{H}$. Let $Z = \{\{z_i\}_{i=1}^{\infty} : \forall i \in \mathbb{N}, z_i \in \{1, \ldots, g(p_i)\}\}$, and for every $\mathbf{z} = \{z_i\}_{i=1}^{\infty} \in Z$, define a classifier $h_{\mathbf{z}}$ such that any $x \in \mathcal{X}$ has $h_{\mathbf{z}}(x) = +1$ if and only if $x \in \{x_{()}, y_{()}\} \cup \{x_{(z_1,\ldots,z_k)} : k \in \mathbb{N}\} \cup \{y_{(z_1,\ldots,z_{k-1},j)} : k \in \mathbb{N}, j \in \{z_k, \ldots, g(p_k)\}\}$: that is, $h_{\mathbf{z}}$ is positive on a single infinite path in the first tree, starting from the root, and in the second tree it labels the corresponding path as positive but also labels as positive any sibling nodes with larger index. It is an easy exercise to verify that the VC dimension of $\mathcal{H}$ is 1.

For $G \subseteq \mathcal{H}$, let $\text{DIS}(G) = \{x : \exists h, h' \in G \text{ s.t. } h(x) \neq h'(x)\}$. Consider the following algorithm.

Algorithm: PickyActive
0. Let $V_0 = \mathcal{H}$ and let $\hat{h}_0$ be any classifier in V_0
1. For $t = 1, 2, \ldots$
2. Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$ as the prediction for Y_t
3. If $X_t \in \text{DIS}(V_{t-1}) \cap \text{Subtree}(y_{()})$
4. Set $Q_t = 1$ (Query for Y_t)
5. Let $V_t = \{h \in V_{t-1} : h(X_t) = Y_t\}$
6. Else set $Q_t = 0$ and let $V_t = V_{t-1}$
7. Let $\hat{h}_t$ be any classifier in V_t

The following theorem provides guarantees for the PickyActive algorithm.

Theorem 6. *For $\mathcal{H}$, $\mathcal{P}$ above, for $\mathcal{A}$ the* PickyActive *algorithm, for any $T \in \mathbb{N}$, denoting by k_T^* the smallest $k \in \mathbb{N}$ with $T < \frac{\ln(2T)}{\alpha_k p_k}$, we have $\mathcal{Q}_T(\mathcal{A}; \mathcal{P}) \lesssim \log^2(T)$ and $\mathcal{M}_T(\mathcal{A}; \mathcal{P}) \lesssim g(p_{k_T^*})^2 \log(T)$.*

This guarantee becomes most interesting when contrasted with the following lower bound for the trivial modifications of CAL for this same scenario.

Theorem 7. *For any $T \in \mathbb{N}$, letting k_T^* be as in Theorem 6, if T is large enough that $g(p_{k_T^*-1}) \geq 800$, $\forall \mathcal{A}' \in \mathbb{TM}(\text{CAL})$, if $\mathcal{Q}_T(\mathcal{A}'; \mathcal{P}) \leq (1/800)g(p_{k_T^*-1})$, then $\mathcal{M}_T(\mathcal{A}'; \mathcal{P}) \geq e^{-1} p_{k_T^*-1} T$.*

Proposition 5 follows from these two theorems for well-chosen values of g, ℓ_i, and p_i (described above). Proofs of Theorems 6 and 7 and Proposition 5 are provided in Appendix B.

4 A General $\mathcal{P}$-Dependent Algorithm

In this subsection, we present an attempt at generalizing the principles underlying the PickyActive algorithm, by providing a flexible distribution-dependent selective sampling strategy applicable to *any* hypothesis class $\mathcal{H}$, with the ability to control the mistakes-vs-queries trade-off beyond mere trivial modifications. We specifically base this strategy on an adaptation of the pool-based active learning strategy known as *splitting* (Dasgupta, 2005). This is reasonable, since the splitting technique has a built-in mechanism for trading off number of queries with number of *unlabeled examples* (in the pool-based setting), the latter of which is (weakly) related to the number of mistakes in the online selective sampling setting. However, one challenge in presenting this strategy is to formulate concise characterizations of the number of mistakes and queries; we provide only a coarse description, leaving a more-refined characterization for future work.

The splitting strategy seems a natural starting place for investigating the trade-off between number of queries and number of mistakes in selective sampling, since there is a rough analogue between number of mistakes in selective sampling and number of unlabeled samples in pool-based active learning. However, there are a number of challenges in adapting the algorithm to this setting. One issue is that the original technique relied upon a given desired error rate for the classifier it should produce, whereas in selective sampling we need the error rate to shrink to zero as t grows (to avoid having $\Omega(T)$ mistakes). Another challenge is that we wish the algorithm to be flexible enough to express the trade-off of $\mathcal{M}_T$ and $\mathcal{Q}_T$ as $T \to \infty$, so that it may be necessary to vary the querying frequency or desired informativeness of queried points, as T grows.

To address these issues, we suppose the algorithm is parameterized by a sequence $\{T_i\}_{i=1}^{\infty}$ of values in $\mathbb{N}$, which crucially affect the behavior of the algorithm and provide control over the trade-off between queries and mistakes: smaller T_i values lead to more queries and fewer mistakes, and larger T_i values lead to fewer queries and more mistakes.

For any $x \in \mathcal{X}$ and finite set $E \subseteq \mathcal{H}^2$, define $\text{Split}(E,x) = |E| - \max_{y \in \mathcal{Y}} |\{(h,h') \in E : h(x) = h'(x) = y\}|$.

Let $T_0 = 0$ and let $T_1, T_2, \ldots$ be any elements in $\mathbb{N}$ with each $T_i \geq 2e$. Let $t_0 = 0$ and for each $i \in \mathbb{N}$ define $t_i = \sum_{j=1}^{i} T_j$. Let $j_0 = \bar{t}_0 = 0$ and for each $k \in \mathbb{N}$ inductively define $j_k = \min\{i > j_{k-1} : t_i \geq 2t_{j_{k-1}}\}$ and $\tilde{t}_k = \sum_{k'=1}^{k} t_{j_{k'}}$. For each $k \in \mathbb{N} \cup \{0\}$, let $\varepsilon_k = 1/t_{j_{k+1}}^2$. Consider the following $\mathcal{P}$-dependent algorithm.

Algorithm: PickySplitting
0. Let V_0 be a minimal ε_0-cover of $\mathcal{H}$, let $\hat{h}_0 \in V_0$, $\Delta_0 = 1$, $E_0 = \{\}$, $k = i = 1$
1. Repeat
2. If $i = j_k$
3. Set $\hat{h}_0 = \hat{h}_{i-1}$, V_0 as a minimal ε_k-cover of $\mathcal{H}$, set $\Delta_0 = 1$, $E_0 = \{\}$, $k = k+1$, $i = 1$
4. If $E_{i-1} = \{\}$,
5. Set $\Delta_i = \Delta_{i-1}/2$ and $E_{i-1} = \{(h,h') \in V_{i-1}^2 : \mathcal{P}(x : h(x) \neq h'(x)) \geq \Delta_i\}$
6. Else set $\Delta_i = \Delta_{i-1}$
7. For $t = \tilde{t}_{k-1} + t_{i-1} + 1, \ldots, \tilde{t}_{k-1} + t_i$
8. Predict $\hat{Y}_t = \hat{h}_{i-1}(X_t)$ as the prediction for Y_t
9. If $t > \tilde{t}_{k-1} + t_{i-1} + T_i/e$, $Q_{t'} = 0$ $\forall t' \in \{\tilde{t}_{k-1} + t_{i-1} + 1, \ldots, t-1\}$, and $\text{Split}(E_{i-1}, X_t) \geq \max\{\text{Split}(E_{i-1}, X_{t'}) : 1 \leq t' - t_{i-1} - \tilde{t}_{k-1} \leq T_i/e\}$
10. Set $Q_t = 1$ (Query for Y_t), $V_i = \{h \in V_{i-1} : h(X_t) = Y_t\}$, $E_i = E_{i-1} \cap V_i^2$
11. Else set $Q_t = 0$, $V_i = V_{i-1}$, $E_i = E_{i-1}$
12. Let $\hat{h}_i$ be any element of V_i; set $i = i + 1$

The essential strategy is to take the most-informative query within consecutive batches of sizes T_i, where informativeness is measured by the splitting criterion. However, since we cannot anticipate what the most informative point will be in advance of seeing the points X_t in the batch, we apply a solution to the well-known *secretary problem*, wherein we use an initial fraction of each batch to set a target for roughly how large we can expect $\text{Split}(E_{i-1}, X_t)$ to be at the t in the batch that maximizes this, and then we query the next point in the batch that meets this target (if there is one). As is well known, this strategy is guaranteed to find the t of largest $\text{Split}(E_{i-1}, X_t)$ in the batch, with at least a constant probability. Thus, with high probability, this algorithm will succeed in picking the highest-splitting t, in at least a constant fraction of the batches.

The algorithm must be slightly more complicated than this, due to the fact that the $\text{Split}(E,x)$ measure of informativeness requires E to be a *finite* set. Following the strategy of Dasgupta (2005), we resolve this issue by replacing $\mathcal{H}$ with an ε-cover of $\mathcal{H}$. However, unlike the setting considered by Dasgupta (2005), in our setting we are interested in running the algorithm on an infinite stream of data, so that no fixed value of ε suffices. For this reason, we also update the value of ε periodically, and effectively reset the learning process (in Step 3) after each such update. For simplicity, we have taken V_i as an ε_k-cover of $\mathcal{H}$ in Step 3; however, it would also be reasonable (though would not affect the results we establish) to instead set V_i as an ε_k-cover of $\{h \in \mathcal{H} : h(X_t) = Y_t \text{ for all previous } t \text{ s.t. } Q_t = 1\}$.

Following Dasgupta (2005), define the *splitting index* as follows. For any $\rho, \Delta, \tau \in (0,1)$, we say a set $V \subseteq \mathcal{H}$ is (ρ, Δ, τ)-*splittable* if, for every finite $E \subseteq \{(h,h') : h, h' \in V, \mathcal{P}(x : h(x) \neq h'(x)) \geq \Delta\}$, it holds that $\mathcal{P}(x : \text{Split}(E,x) \geq \rho|E|) \geq \tau$. For any $r > 0$, define $\text{B}(f^\star, r) = \{h \in \mathcal{H} : \mathcal{P}(x : h(x) \neq f^\star(x)) \leq r\}$. Then for any $\varepsilon, \tau \in [0,1]$, define the *splitting index* (of $(f^\star, \mathcal{P})$)

$$\rho(\varepsilon;\tau) = \sup\{\rho \in [0,1] : \forall \Delta \geq \varepsilon, \text{B}(f^\star, 4\Delta) \text{ is } (\rho,\Delta,\tau)\text{-splittable}\}.$$

It is possible to express bounds on the expected numbers of queries and mistakes by the PickySplitting algorithm, in terms of the values of the splitting index $\rho(\tilde{\varepsilon}_i; \tilde{\tau}_i)$ at appropriate values of $\tilde{\varepsilon}_i$ and $\tilde{\tau}_i$. However, as an abstract expression, it is rather complex, and provides little additional insight into the behavior of the algorithm beyond the literal description of the algorithm itself above. As such, we do not discuss the details of this abstract analysis, and we leave for future work the issue of expressing concise general bounds. However, we do find that, at least in special cases in which the T_i sequence is chosen carefully so that the relevant ρ values are *bounded away from zero*, concise bounds are possible. Specifically, we have the following theorem. A proof sketch is provided in Appendix B.

Theorem 8. *Suppose* $\exists \tilde{\rho} > 0$ *such that,* $\forall i$, $\rho(2^{-i}; 1/T_i) \geq \tilde{\rho}$. *Then for* $\mathcal{A} =$ PickySplitting, $\mathcal{Q}_T(\mathcal{A}; f^\star, \mathcal{P}) \lesssim \max\{i : t_{i-1} < T\} \log(T)$, *and*

$$\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P}) \lesssim \sum_{i : t_{i-1} < T} T_i \left(2^{-ci\tilde{\rho}/(d \log(T))} + t_i^{-1}\right) \log(T)$$

for a numerical constant $c > 0$.

Note that this result already contains within it a trade-off between the sizes of T_i values and number of queries: larger T_i values lead to fewer queries and more mistakes. Furthermore, it is clear from the algorithm that in some scenarios these trade-offs would be non-trivial (in that they are not equivalent to trivial modifications of the

algorithm), at least to the extent that small values of T_i might not admit ρ values bounded below, while in some cases larger T_i values would. However, we suspect the specific result in Theorem 8 should be improvable in a number of ways, such as in the arguments to $\rho(\cdot,\cdot)$. Moreover, it seems the PickySplitting algorithm itself can exhibit many favorable behaviors not captured by Theorem 8. Indeed, to truly understand this algorithm, it is important to also characterize these trade-offs *beyond* merely enabling a constant ρ value: for instance, setting the T_i sequence in order to control the rate of decrease of $\rho(\varepsilon_i, 1/T_i)$ toward zero (for appropriate ε_i), which would lead to a more-involved expression of the bounds. We leave for future work the question of whether relatively simple expressions for such interesting trade-offs are possible.

5 Target-Dependent Analysis: Improvements in Asymptotic Rates

In the previous sections, we discussed the trade-off between mistakes and queries in a distribution-free analysis (Section 2.2) and a distribution-dependent analysis (Section 3). In this section, we follow this line to its extreme with an analysis that is both distribution-dependent *and target-dependent*. We are specifically interested in the asymptotic guarantees achievable by the sequences $\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P})$ and $\mathcal{Q}_T(\mathcal{A}; f^\star, \mathcal{P})$ for all $f^\star \in \mathcal{H}$ and all $\mathcal{P}$.

In the pool-based active learning setting, the analogous subject was explored by Hanneke (2009, 2012); Balcan, Hanneke, and Vaughan (2010). These works found that the label complexity of pool-based active learning can always have a dependence on the desired error rate ϵ that is $o(1/\epsilon)$, where the constants in this bound may depend on $\mathcal{H}$, $f^\star$, and $\mathcal{P}$. This result is significant since it is known that there are classes $\mathcal{H}$ where this is definitely *not* achievable by passive learning algorithms.

Here we are interested in the analogous question for selective sampling. We propose the following question:

Open Problem: Is it true that, for every $\mathcal{H}$ of finite VC dimension, there exists a selective sampling algorithm $\mathcal{A}$ such that, for every $\mathcal{P}$ and every $f^\star \in \mathcal{H}$, $\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P}) = O(\log(T))$ and $\mathcal{Q}_T(\mathcal{A}; f^\star, \mathcal{P}) = o(T)$?

Here the big-O and little-o hide constant factors that may depend on $\mathcal{H}$, $\mathcal{A}$, $f^\star$, and $\mathcal{P}$, all of which are considered constant (the algorithm itself may only depend on $\mathcal{H}$). Since passive learning would typically make a number of mistakes $\Theta(\log(T))$, the above problem is essentially asking whether there is an active learning algorithm making roughly the same number of mistakes as passive learning, but requesting only a *sublinear* number of labels. We note that the aforementioned pool-based techniques achieving $o(1/\varepsilon)$ sample complexity *cannot* directly resolve this problem, since they require an unbounded number of unlabeled examples, which here translates into worse than $\log(T)$ mistakes.

The above question remains open at this time. However, in this section we present an example illustrating that, at least in certain interesting and nontrivial cases, it is indeed possible to obtain $O(\log(T))$ mistakes, even with $O(\log(T))$ queries, even when most general active learning algorithms, such as CAL (or its trivial modifications), fail to achieve this guarantee. This example is indeed *nontrivial* since these strong improvements (or indeed, *any* improvements) over passive sampling would *not* be observed in an analysis that is distribution-dependent but *not* target-dependent.

Perhaps the simplest example we could describe, exhibiting this type of strong distinction between target-dependent and target-independent analysis, is the class of *interval* classifiers. However, to consider a more expressive class, we instead study *unions of k intervals*.

Specifically, let $\mathcal{X} = [0,1]$, and (defining $\mathbb{1}_A^{\pm}(x) = 2\mathbb{1}[x \in A] - 1$ for any $A \subseteq \mathcal{X}$) define $\mathcal{H}_k = \{\mathbb{1}^{\pm}_{\bigcup_{i=1}^{k}[a_i,b_i]} : \forall i \leq k, a_i, b_i \in [0,1]\}$ for each $k \in \mathbb{N}$, and $\mathcal{H}_0 = \{\mathbb{1}^{\pm}_{\emptyset}\}$.

Algorithm: UIntActive$_k$
0. Let $S_0 = \{\}$, $k_0 = 0$, and let $\hat{h}_0 = \mathbb{1}^{\pm}_{\emptyset}$
1. For $t = 1, 2, \ldots$
2. Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$ as the prediction for Y_t
3. If $X_t \in \mathrm{DIS}(\{h \in \mathcal{H}_{k_{t-1}} : \forall (x,y) \in S_{t-1}, h(x) = y\})$ or $\log_2(2t) \in \mathbb{N}$
4. Set $Q_t = 1$ (query for Y_t), $S_t = S_{t-1} \cup \{(X_t, Y_t)\}$
5. Else Set $Q_t = 0$ and let $S_t = S_{t-1}$
6. If $\nexists h \in \mathcal{H}_{k_{t-1}}$ s.t. $\forall (x,y) \in S_t$, $h(x) = y$
7. Let $k_t = k_{t-1} + 1$
8. Else let $k_t = k_{t-1}$
9. Let $\hat{h}_t$ be any $h \in \mathcal{H}_{k_t}$ s.t. $\forall (x,y) \in S_t, h(x) = y$

We have the following theorem for this algorithm. The proof is presented in Appendix C.

Theorem 9. *Let $k \in \mathbb{N}$ and $\mathcal{H} = \mathcal{H}_k$. For $\mathcal{A} =$ UIntActive$_k$, for any $\mathcal{P}$ and any $f^\star \in \mathcal{H}$, $\mathcal{Q}_T(\mathcal{A}; f^\star, \mathcal{P}) = O(\log(T))$ and $\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P}) = O(\log(T))$.*

6 Conclusions

This work represents mere first steps toward a general theory of the trade-off between mistakes and queries in online selective sampling. In addition to identifying the trivial modifications of CAL as the fundamental baseline for comparison, we believe the main contributions of this work are in providing examples illustrating

possible behaviors, far more interesting than exhibited by trivially-modified algorithms. Such examples are an important part of the development of a general theory. We have also made speculative advances toward such a theory, providing a general flexible distribution-dependent selective sampling strategy (in Section 4), and proposing an open question regarding the ability to always achieve strong asymptotic improvements over passive sampling. It is our hope that these discussions will stir future work in this direction.

References

A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Machine Learning*, 30(1):31–56, 1998.

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

N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7(7): 1205–1230, 2006.

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. Tauman Kalai, and C. Monteleoni. Analysis of Perceptron-based active learning. *Journal of Machine Learning Research*, 10(11):281–299, 2009.

O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, To Appear, 2012.

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

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

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

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

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

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

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

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

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

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

S. Hanneke, V. Kanade, and L. Yang. Learning with a drifting target concept. In *Proceedings of the* 26th *International Conference on Algorithmic Learning Theory*, 2015.

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

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

S. Sabato and T. Hess. Interactive algorithms: Pool, stream and precognitive stream. *Journal of Machine Learning Research*, 18(229):1–39, 2018.

D. Schuurmans. Characterizing rational versus exponential learning curves. *Journal of Computer and System Sciences*, 55(1):140–160, 1997.

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

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

L. Yang. Active learning with a drifting distribution. In *Advances in Neural Information Processing Systems 24*, 2011.

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

A Proofs for the Distribution-Free Analysis

Proof of Theorem 3. For any set $G \subseteq \mathcal{H}$, define $\mathrm{DIS}(G) = \{x : \exists h, h' \in G \text{ s.t. } h(x) \neq h'(x)\}$. Hanneke (2016b) proves that

$$\mathbb{E}[\mathcal{P}(\mathrm{DIS}(V_{t-1}))] \leq \frac{\mathfrak{s}}{t}. \tag{1}$$

The particular trivial modification of CAL we will consider here takes $\mathcal{I} = \{1, \ldots, i_T\}$, where we define $i_T = \max\{i \leq T : q_T \geq \min\{i, \mathfrak{s}\ln(ei)\}\}$. Let $\mathcal{A}$ denote this (trivially-modified) algorithm. In particular, this algorithm $\mathcal{A}$ satisfies the constraint on $\mathcal{Q}_T(\mathcal{A})$ since (1) and the definition of i_T imply

$$\begin{aligned}\mathcal{Q}_T(\mathcal{A}) &\leq \sum_{i=1}^{i_T} \min\Big\{1, \frac{\mathfrak{s}}{i}\Big\} \\ &\leq \min\{i_T, \mathfrak{s}\ln(ei_T)\} \leq q_{i_T} \leq q_T.\end{aligned}$$

Now we analyze $\mathcal{M}_T(\mathcal{A})$. Since $f^\star$ is in $\mathcal{H}$, the definition of CAL ensures every $\tilde{Y}_t$ is equal $f^\star(X_t) = Y_t$. Thus, in the modification $\mathcal{A}$, the predictions $\hat{Y}_t$ are each equal $\mathcal{A}_p(\{X_i, Y_i\}_{i=1}^{\min\{i_T, t-1\}}, X_t)$. Thus, by the assumed property of $\mathcal{A}_p$ (namely, that its probability of making a mistake on a new point is bounded by $\frac{cd}{m}$ when its training set has size m), we obtain that $\mathbb{P}(\hat{Y}_t \neq Y_t) \lesssim \frac{d}{\min\{i_T, t-1\}}$. Therefore, by linearity of the expectation, $\mathcal{M}_T(\mathcal{A}) \lesssim d\sum_{t=1}^{T} \frac{1}{\min\{i_T, t-1\}} \leq d\log(T) + \frac{dT}{i_T}$. If $q_T < \mathfrak{s}\ln(eT)$, then since the definition of i_T implies we always have $i_T \geq q_T$, we have $\mathcal{M}_T(\mathcal{A}) \lesssim d\log(T) + \frac{dT}{q_T}$. On the other hand, if $q_T \geq \mathfrak{s}\ln(eT)$, then $i_T = T$, so that $\mathcal{M}_T(\mathcal{A}) \lesssim d\log(T)$. □

Proof of Theorem 4. The $\min\{d, T\}$ lower bound follows from existing results for fully-supervised (passive) learning (see Haussler, Littlestone, and Warmuth, 1994).

For the rest, let $k \in \mathbb{N}$ and suppose $x_1, \ldots, x_k$ are a star set witnessed by $h_0, h_1, \ldots, h_k \in \mathcal{H}$. Let $\mathcal{P}$ be uniform on $\{x_1, \ldots, x_k\}$, and choose $f^\star$ uniformly at random among $h_1, \ldots, h_k$. Consider the sequence of queries Q_t and predictions $\hat{Y}_t$ the algorithm would make if the target were h_0. With probability at least 3/4, $\sum_{t=1}^{T} Q_t \leq 4\mathcal{Q}_T(\mathcal{A})$. Furthermore, on this event, if $f^\star$ is instead chosen uniformly among $h_1, \ldots, h_k$, and $4\mathcal{Q}_T(\mathcal{A}) \leq k/4$, then with probability at least 3/4 the points (X_t, Y_t) with $Q_t = 1$ will all have $Y_t = h_0(X_t)$. Furthermore, given these observations Y_t, the posterior distribution of $f^\star$ remains uniform on the h_i functions consistent with these observations, of which there are at least two. Thus, on each round t, the probability $\hat{Y}_t \neq Y_t$ is at least $\frac{1}{4k}$. In particular, this implies that for any given $\mathcal{A}$ there exists a deterministic choice of $f^\star$ for which $\mathbb{P}(\hat{Y}_t \neq Y_t) \geq \frac{1}{4k}$ for every $t \leq T$.

To complete the proof, we note that if $\mathcal{Q}_T(\mathcal{A}) < \mathfrak{s}/16$, then we can find a star set with size $k = 16\mathcal{Q}_T(\mathcal{A})$, and we have $\mathcal{M}_T(\mathcal{A}) \geq \frac{T}{64\mathcal{Q}_T(\mathcal{A})}$. □

From these results, we derive the proof of Theorem 1, regarding the optimal trade-off for selective sampling.

Proof of Theorem 1. Let $q_T = 16\mathcal{Q}_T(\mathcal{A})\ln(eT)$ and let $\mathcal{A}'$ be the trivial modification of CAL for this q_T implied by Theorem 3. In particular, note that $q_T < \mathfrak{s}\ln(eT)$ iff $\mathcal{Q}_T(\mathcal{A}) < \mathfrak{s}/16$. The result then follows immediately from the bounds in Theorems 3 and 4. □

B Proofs for the Distribution-Dependent Analysis

Proof of Theorem 6. Fix any $f^\star \in \mathcal{H}$, and let $\mathbf{z}^\star = \{z_i^\star\}_{i=1}^{\infty}$ be the element of Z such that $f^\star = h_{\mathbf{z}^\star}$. First note that the update in Step 5 always guarantees $f^\star \in V_t$. Also, due to the criterion in Step 3, for *every* t with X_t in $\mathrm{Subtree}(y_{()})$, every $h \in V_t$ has $h(X_t) = Y_t$.

For any $t \in \mathbb{N} \cup \{0\}$, let k_t denote the smallest $k \in \mathbb{N}$ such that

$$\mathrm{DIS}(V_t) \cap \{y_{(z_1,\ldots,z_k)} : \forall i \leq k, z_i \in \{1, \ldots, g(p_i)\}\} \neq \emptyset.$$

Now we establish the result by verifying three claims:

Claim 1: $\forall t, \mathcal{P}(x : \hat{h}_t(x) \neq f^\star(x)) \leq 12p_{k_t}$.
Proof: $\hat{h}_t$ is some $h_{\mathbf{z}}$ with $\mathbf{z} = \{z_i\}_{i=1}^{\infty} \in Z$, satisfying $z_i = z_i^\star$ for every $i < k_t$. It therefore satisfies

$$\begin{aligned}&\mathrm{DIS}(\{\hat{h}_t, f^\star\}) \\ &\subseteq \bigcup_{i \geq k_t} \Big(\{x_{(z_1,\ldots,z_i)}, x_{(z_1^\star,\ldots,z_i^\star)}\} \cup \mathrm{Children}(y_{(z_1,\ldots,z_{i-1})}) \\ &\qquad\qquad \cup \mathrm{Children}(y_{(z_1^\star,\ldots,z_{i-1}^\star)}) \Big).\end{aligned}$$

Note that

$$\begin{aligned}&\mathcal{P}\left(\bigcup_{i \geq k_t} \{x_{(z_1,\ldots,z_i)}, x_{(z_1^\star,\ldots,z_i^\star)}\} \right) \\ &\leq \sum_{i \geq k_t} 2(1-\alpha_i)p_i \leq \sum_{i \geq k_t} 4p_i \leq 8p_{k_t}, \end{aligned} \tag{2}$$

where the last inequality is due to the fact that $p_i \leq p_{i-1}/2$. Also note that, due to the fact that $p_i \leq p_{i-1}/2$,

we have that

$$\mathcal{P}\left(\bigcup_{i\geq k_t}\text{Children}(y_{(z_1,\ldots,z_{i-1})})\cup\text{Children}(y_{(z_1^\star,\ldots,z_{i-1}^\star)})\right)$$
$$\leq \sum_{i\geq k_t} 2\alpha_i p_i g(p_i) \leq \sum_{i\geq k_t} 2p_i \leq 4p_{k_t}. \quad (3)$$

Claim 1 follows from (2) and (3) by the union bound.

Claim 2: For any $\delta\in(0,1)$ and $t\in\mathbb{N}$, let $\bar{k}_t(\delta)$ be a maximum integer such that $t\geq \frac{1}{\alpha_{\bar{k}_t(\delta)}p_{\bar{k}_t(\delta)}}\ln\left(\frac{2}{\delta}\right)$, or $\bar{k}_t(\delta)=0$ if $t<\frac{1}{\alpha_1 p_1}\ln\left(\frac{2}{\delta}\right)$. Then with probability at least $1-\delta$, we have $k_t>\bar{k}_t(\delta)$.
Proof: Fix any $k\in\mathbb{N}$, and note that we will have $k_t>k$ if the data $X_1,\ldots,X_t$ contain $y_{(z_1^\star,\ldots,z_k^\star)}$ and (in the case that $z_k^\star>1$) $y_{(z_1^\star,\ldots,z_k^\star-1)}$. Each of these two (or one, if $z_k^\star=1$) points has probability mass $\alpha_k p_k$, and thus, they occur in the sample with probability at least $1-2(1-\alpha_k p_k)^t\geq 1-2\exp\{-\alpha_k p_k t\}$, which is at least $1-\delta$ for any $t\geq\frac{1}{\alpha_k p_k}\ln\left(\frac{2}{\delta}\right)$. Claim 2 follows.

Claim 3: For any $t\in\mathbb{N}$, $\mathbb{E}[Q_t]\leq\frac{5\ln(2t)}{t}$.
Proof: We follow a *leave-one-out* argument of Hanneke (2016b) for bounding the probability that CAL queries label Y_t, but modified to apply to the PickyActive algorithm instead. Specifically, for each $i\leq t$, let Q_t^i denote the hypothetical value of Q_t if the samples (X_i,Y_i) and (X_t,Y_t) were swapped in the data sequence; in particular, Q_t^t denotes the value of Q_t for the unmodified (original) order of the sequence. Since the samples are i.i.d., we have that

$$\mathbb{E}[Q_t]=\frac{1}{t}\sum_{i=1}^t\mathbb{E}[Q_t^i]=\frac{1}{t}\mathbb{E}\left[\sum_{i=1}^t Q_t^i\right].$$

Now consider which points X_i would have $Q_t^i=1$. For any k, we refer to points $y_{(z_1,\ldots,z_k)}$ as being "at level k". If there exists a $j\leq t$ with X_j at some level $k\geq k_t$ and with $Y_j=1$, then all of the points (X_i,Y_i) with X_i at levels $<k$ have $Q_t^i=0$; in the special case that k_t is the largest k for which such an X_j exists, there can be at most two points X_i, $X_{i'}$ at level k_t with $Q_t^i=1$ and $Q_t^{i'}=1$: namely, the (X_i,Y_i) with $X_i=y_{(z_1^*,\ldots,z_{k_t-1}^*,n)}$ of largest n such that $Y_i=-1$ (if one exists), and the $(X_{i'},Y_{i'})$ with $X_{i'}=y_{(z_1^*,\ldots,z_{k_t-1}^*,n)}$ of smallest n. On the other hand, if every X_j $(j\leq t)$ at levels $k\geq k_t$ has $Y_j=-1$, then (by minimality of k_t) it must be that $y_{(z_1^*,\ldots,z_{k_t-1}^*)}$ and (if $z_{k_t-1}^*>1$) $y_{(z_1^*,\ldots,z_{k_t-1}^*-1)}$ are present in $X_1,\ldots,X_t$, so that the corresponding X_i and $X_{i'}$ are the only points at level k_t-1 with $Q_t^i=1$ and $Q_t^{i'}=1$; in this case, it also holds that there is at most one X_j at level k_t with $Q_t^j=1$: namely, if there are any points in $\text{Children}(y_{(z_1^*,\ldots,z_{k_t-1}^*)})$ among $X_1,\ldots,X_t$, then the $y_{(z_1^*,\ldots,z_{k_t-1}^*,n)}$ among $X_1,\ldots,X_t$ with largest n. Thus, in either case, there are at most three points X_i at levels $\leq k_t$ with $Q_t^i=1$. All of the other values i with $Q_t^i=1$ have X_i at levels $>k_t$. Altogether, we have that $\sum_{i\leq t}Q_t^i\leq 3+\sum_{i\leq t}\mathbb{1}[X_i\text{ at level }>k_t]$, so that

$$\mathbb{E}[Q_t]\leq\frac{3}{t}+\frac{1}{t}\sum_{i\leq t}\mathbb{P}(X_i\text{ at level }>k_t)$$
$$\leq\frac{3}{t}+\mathbb{P}\big(k_t\leq\bar{k}_t(\delta)\big)+\frac{1}{t}\sum_{i\leq t}\mathbb{P}\big(X_i\text{ at level}\geq 2+\bar{k}_t(\delta)\big).$$

We then note that

$$\mathbb{P}\big(X_i\text{ at level }\geq 2+\bar{k}_t(\delta)\big)$$
$$\leq\sum_{k\geq 2+\bar{k}_t(\delta)}\alpha_k\ell_k\leq\alpha_{2+\bar{k}_t(\delta)}\leq\alpha_{1+\bar{k}_t(\delta)}p_{1+\bar{k}_t(\delta)},$$

where the final inequality uses the assumed property that $g(p_i)^2\geq g(p_{i-1})^2/p_{i-1}$. Then note that, by maximality of $\bar{k}_t(\delta)$, we have $t<\frac{1}{\alpha_{1+\bar{k}_t(\delta)}p_{1+\bar{k}_t(\delta)}}\ln\left(\frac{2}{\delta}\right)$, which implies $\alpha_{1+\bar{k}_t(\delta)}p_{1+\bar{k}_t(\delta)}<\frac{1}{t}\ln\left(\frac{2}{\delta}\right)$. Altogether, we have that

$$\mathbb{E}[Q_t]\leq\frac{3}{t}+\delta+\frac{1}{t}\ln\left(\frac{2}{\delta}\right).$$

Claim 3 immediately follows by setting $\delta=\frac{1}{t}$.

To complete the proof of the theorem, note that Claim 3 implies

$$\mathcal{Q}_T(\mathcal{A};\mathcal{P})\leq\sum_{t=1}^T\frac{5\ln(2t)}{t}\lesssim\log^2(T).$$

Furthermore, Claim 1 implies

$$\mathcal{M}_T(\mathcal{A};\mathcal{P})\leq 12\sum_{t=1}^T\mathbb{E}[p_{k_t}]$$
$$\leq 12\sum_{k=0}^{\bar{k}_T(1/T)}|\{t\leq T:\bar{k}_t(1/T)=k\}|p_{k+1}$$
$$\leq 12\sum_{k=0}^{\bar{k}_T(1/T)}\frac{1}{\alpha_{k+1}}\ln(2T),$$

where the last inequality is due to maximality of $\bar{k}_t(1/T)$ from its definition. The fact that $g(p_{k+1})^2\geq g(p_k)^2/p_k$ implies $g(p_{k+1})^2\geq 2g(p_k)^2$. Therefore,

$$\sum_{k=0}^{\bar{k}_T(1/T)}\frac{1}{\alpha_{k+1}}\leq 2g(p_{1+\bar{k}_T(1/T)})^2=2g(p_{k_T^*})^2,$$

and the theorem follows immediately. □

Proof of Theorem 7. We proceed by the probabilistic method. Choose $f^\star = h_{\mathbf{z}^*}$ where $\mathbf{z}^* = \{z_i^*\}_{i=1}^\infty$ is such that, independently, each z_i^* is uniform random from $\{1, \ldots, g(p_i)\}$: that is, the target function is a random path in the tree.

Fix any T large enough that $g(p_{k_T^*-1}) \geq 800$ and consider any $\mathcal{A}' \in \mathbb{TM}(\mathrm{CAL})$ guaranteeing $\mathcal{Q}_T(\mathcal{A}'; \mathcal{P}) \leq (1/800) g(p_{k_T^*-1})$. Let $\mathcal{I}$ be the set of indices defining the modification $\mathcal{A}'$. Then for each $t \leq T$, $\mathcal{A}'$ makes its prediction $\hat{Y}_t$ by $\hat{Y}_t = \hat{h}_t(X_t) = \mathcal{A}_p(S_{t-1}, X_t)$, where $S_{t-1} = \{(X_i, Y_i) : i < t, i \in \mathcal{I}\}$ is a (conditionally, given $f^\star$) i.i.d. sample of size $n_{t-1} := |\{i \in \mathcal{I} : i < t\}|$; this is due to the nature of the CAL active learner, wherein $\tilde{Y}_{t'} = Y_{t'}$ for all $t' \in \mathcal{I}$ processed in the algorithm (including those not queried). Let Q_t be the query indicators for $\mathcal{A}'$.

Enumerate $\mathcal{I} \cap \{1, \ldots, T\}$ as $i_1, \ldots, i_{n_T}$, and let $i_0 = 0$ and $i_{n_T+1} = T$. For any $n \leq n_T$, with probability at least $1 - n\mathcal{P}\Big(\mathrm{Subtree}(y_{(z_1^*, \ldots, z_{k_T^*-2}^*)})\Big) \geq 1 - n\sum_{k \geq k_T^*-1} \alpha_k p_k \prod_{i=k_T^*-1}^k g(p_i) \geq 1 - n\sum_{k \geq k_T^*-1} \alpha_k p_k g(p_k)^2 \geq 1 - 2np_{k_T^*-1}$ (using $g(p_i) \geq \prod_{j=1}^{i-1} g(p_j)$ and $p_i \leq p_{i-1}/2$), it holds that $\{X_{i_1}, \ldots, X_{i_n}\} \cap \mathrm{Subtree}(y_{(z_1^*, \ldots, z_{k_T^*-2}^*)}) = \emptyset$. Furthermore, with probability at least $1 - n\mathcal{P}\Big(\mathrm{Subtree}(x_{(z_1^*, \ldots, z_{k_T^*-2}^*)}) \setminus \mathrm{Children}(x_{(z_1^*, \ldots, z_{k_T^*-2}^*)})\Big) \geq 1 - n\sum_{k \geq k_T^*} p_k \prod_{i=k_T^*-1}^k g(p_i) \geq 1 - n\sum_{k \geq k_T^*} p_{k-1} \geq 1 - 2np_{k_T^*-1}$ (using the facts that $p_k \prod_{i=1}^k g(p_i) \leq p_{k-1}$ and $p_i \leq p_{i-1}/2$), it holds that $\{X_{i_1}, \ldots, X_{i_n}\} \cap \mathrm{Subtree}(x_{(z_1^*, \ldots, z_{k_T^*-2}^*)}) \setminus \mathrm{Children}(x_{(z_1^*, \ldots, z_{k_T^*-2}^*)}) = \emptyset$. Also, with probability at least $1 - np_{k_T^*-1}$, $x_{(z_1^*, \ldots, z_{k_T^*-1}^*)} \notin \{X_{i_1}, \ldots, X_{i_n}\}$. In particular, if all three of these events occur, we may note that $\sum_{n' \leq n} Q_{i_{n'}}$ is at least as large as the number of distinct elements of $\{X_{i_1}, \ldots, X_{i_n}\} \cap \mathrm{Children}(x_{(z_1^*, \ldots, z_{k_T^*-2}^*)})$. If $n \geq (1/25) p_{k_T^*-1}^{-1}$, then with probability at least $1 - e^{-1}$, the number of such distinct elements is at least $\min\big\{(1/8) np_{k_T^*-1} g(p_{k_T^*-1}), (1/8) g(p_{k_T^*-1})\big\}$. Altogether, if $(1/25) p_{k_T^*-1}^{-1} \leq n \leq (1/20) p_{k_T^*-1}^{-1}$, then with probability at least $1 - e^{-1} - 5np_{k_T^*-1} > 1/4$, we have $\sum_{n' \leq n} Q_{i_{n'}} \geq (1/200) g(p_{k_T^*-1})$. Thus, since $n \leq n_T$, we have $\mathcal{Q}_T(\mathcal{A}'; \mathcal{P}) > (1/800) g(p_{k_T^*-1})$, which contradicts the stated assumption on $\mathcal{A}'$. Therefore, no such n exists with $n \leq n_T$, so that we conclude that $n_T \leq (1/25) p_{k_T^*-1}^{-1}$.

Next we prove the lower bound on $\mathcal{M}_T(\mathcal{A}'; \mathcal{P})$. Now fix any $n \in \{0, \ldots, n_T\}$ and if $n > 0$ then suppose that the first three events above occur for this n, which happens with probability at least $1 - 5n_T p_{k_T^*-1} \geq 4/5$. Furthermore, with probability at least $1 - e^{-1}$, there are at most $(1/12) g(p_{k_T^*-1})$ points $X_{i_{n'}}$ with $n' \leq n$ and $X_{i_{n'}} \in \mathrm{Children}(x_{(z_1^*, \ldots, z_{k_T^*-2}^*)})$. In particular, when these events occur together, the posterior distribution of $x_{(z_1^*, \ldots, z_{k_T^*-1}^*)}$ after round i_n is uniform on a set of at least $(11/12) g(p_{k_T^*-1})$ points; this latter fact is also true for $n = 0$: i.e., before seeing any examples. Consider any $t \in \{i_n + 1, \ldots, i_{n+1}\}$, and note that $\hat{h}_t = \hat{h}_{i_n+1}$. If $\hat{h}_t$ does not predict $\hat{h}_t(x_{(z_1, \ldots, z_{k_T^*-1})}) = 1$ for any $z_1, \ldots, z_{k_T^*-1}$, then certainly $\mathcal{P}(x : \hat{h}_t(x) \neq f^\star(x)) \geq p_{k_T^*-1}$. Otherwise, if either $n = 0$, or $n > 0$ and the above events hold, if there exists $x_{(z_1, \ldots, z_{k_T^*-1})}$ on which $\hat{h}_t(x_{(z_1, \ldots, z_{k_T^*-1})}) = 1$, then there is posterior probability at least $11/12$ that $(z_1, \ldots, z_{k_T^*-1}) \neq (z_1^*, \ldots, z_{k_T^*-1}^*)$ and hence also in this case $\mathcal{P}(x : \hat{h}_t(x) \neq f^\star(x)) \geq p_{k_T^*-1}$. Altogether we have that $\mathbb{E}[\mathcal{P}(x : \hat{h}_t(x) \neq f^\star(x))] \geq (11/12)((4/5) - e^{-1}) p_{k_T^*-1} \geq e^{-1} p_{k_T^*-1}$. We therefore have that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[\hat{Y}_t \neq Y_t]\right] &= \sum_{n=0}^{n_T} \sum_{t=i_n+1}^{i_{n+1}} \mathbb{E}\Big[\mathbb{1}[\hat{h}_t(X_t) \neq Y_t]\Big] \\
&= \sum_{n=0}^{n_T} \sum_{t=i_n+1}^{i_{n+1}} \mathbb{E}\Big[\mathcal{P}(x : \hat{h}_t(x) \neq f^\star(x))\Big] \\
&\geq e^{-1} p_{k_T^*-1} T.
\end{aligned}$$

In particular, by the law of total expectation, this also implies there exists a deterministic choice of $f^\star \in \mathcal{H}$ such that $\mathcal{M}_T(\mathcal{A}'; f^\star, \mathcal{P}) \geq e^{-1} p_{k_T^*-1} T$, which completes the proof. □

Proof of Proposition 5. For this concrete result, we specify $\ell_i = 2^{-i}$, $p_i = 2^{-2^{2i+1}}$, and $g(\epsilon) = \lfloor \epsilon^{-1/2} \rfloor$. One can easily check that these specifications satisfy the requirements stated above. In particular, note that this implies $\alpha_k = g(p_k)^{-2} = p_k$. Let $T \in \mathbb{N}$ satisfy $g(p_{k_T^*-1}) \geq 800$, and $T + 1 \geq \frac{\ln(2T)}{\alpha_{k_T^*} p_{k_T^*}}$. There are an infinite number of such values T. In particular, this implies $p_{k_T^*} \geq \left(\frac{\ln(2T)}{T+1}\right)^{1/2}$. We use the spaces $\mathcal{X}$ and $\mathcal{H}$ and the distribution $\mathcal{P}$, all as described above. For $\mathcal{A}$ the PickyActive algorithm, we have from Theorem 6 that $\mathcal{Q}_T(\mathcal{A}; \mathcal{P}) \lesssim \log^2(T)$. Furthermore, $\mathcal{M}_T(\mathcal{A}; \mathcal{P}) \lesssim g(p_{k_T^*})^2 \log(T) \leq \left(\frac{T+1}{\ln(2T)}\right)^{1/2} \log(T) \lesssim (T \log(T))^{1/2}$. On the other hand, for any $\mathcal{A}' \in \mathbb{TM}(\mathrm{CAL})$, Theorem 7 implies that if $\mathcal{Q}_T(\mathcal{A}'; \mathcal{P}) \leq (1/800) g(p_{k_T^*-1})$ then necessarily $\mathcal{M}_T(\mathcal{A}'; \mathcal{P}) \geq e^{-1} p_{k_T^*-1} T$. Since $g(p_{k_T^*-1}) = p_{k_T^*-1}^{-1/2} = p_{k_T^*}^{-1/8} > \left(\frac{T}{\ln(2T)}\right)^{1/16}$, if T is sufficiently large, satisfying $\mathcal{Q}_T(\mathcal{A}'; \mathcal{P}) < T^{1/17}$ would imply $\mathcal{Q}_T(\mathcal{A}'; \mathcal{P}) \leq (1/800) g(p_{k_T^*-1})$. Furthermore, since $p_{k_T^*-1} = p_{k_T^*}^{1/4} \geq \left(\frac{\ln(2T)}{T+1}\right)^{1/8}$, if T is sufficiently

large, and $\mathcal{M}_T(\mathcal{A}';\mathcal{P}) \geq e^{-1}p_{k^*_T-1}T$, then it holds that $\mathcal{M}_T(\mathcal{A}';\mathcal{P}) > T^{7/8}$. □

Proof Sketch of Theorem 8. The value of $\mathcal{Q}_T(\mathcal{A}; f^\star, \mathcal{P})$ is trivially at most $\max\{i : t_{i-1} < T\}\log(T)$ since the algorithm queries at most one Y_t value in each batch of T_i points, and since t_{j_k} grows at least exponentially, each i is encountered at most $\log(T)$ times. It only remains to bound $\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P})$.

First we address the issue of using the sequence of covers V_0 instead of $\mathcal{H}$. For each k encountered in the algorithm, within the corresponding V_0, there exists h^*_k with $\mathcal{P}(x : h^*_k(x) \neq f^\star(x)) \leq \varepsilon_k$. In particular, on each round i that this h^*_k remains in V_{i-1}, we have $V_{i-1} \subseteq \mathrm{B}(h^*_k, 2\Delta_i) \subseteq \mathrm{B}(f^\star, 2\Delta_i+\varepsilon_k)$. In particular, the probability that some X_t has $h^*_k(X_t) \neq f^\star(X_t)$ after reaching $i = j_k$ for the first time, but before reaching $i = j_{k+1}$ for the first time, is at most $t_{j_{k+1}}\varepsilon_k \leq 1/t_{j_{k+1}}$. Thus, at any given time t, for the values of k and i occurring with that t in the algorithm (during Step 8), the probability that h^*_k no longer remains in V_{i-1} is at most $1/t$. Let E_t denote the event that h^*_k remains in V_{i-1} at time t. We then have that $\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P}) \leq \sum_{t=1}^T(1 - \mathbb{P}(E_t)) + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[\hat{Y}_t \neq Y_t]\mathbb{1}_{E_t}\right] \lesssim \log(T) + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[\hat{Y}_t \neq Y_t]\mathbb{1}_{E_t}\right]$. The effect of this is that we may essentially disregard the possibility of h^*_k being eliminated from V_{i-1} at any point it the algorithm. For simplicity, for the remainder of the proof we will omit the $\mathbb{1}_{E_t}$ indicators, and simply work with the presumption that h^*_k is never eliminated.

Next we address the issue of the "reset" occurring when k is incremented. With the above remark about h^*_k remaining in V_{i-1} in mind, note that since j_k grows exponentially, we can upper bound $\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P})$ by analyzing the algorithm that would simply omit these resets and instead initializes V_0 to the *union* of the various V_0 covers for values of k encountered in the algorithm, and omits Steps 2 and 3; this V_0 has size $\lesssim \max_k(c/\varepsilon_k)^d\log(T)$ for a numerical constant c, where k ranges over those values of k encountered in the algorithm (specifically, those values which hold during some execution of Step 12 before reaching $t = T$, since this is the step defining the predictor $\hat{h}_i$ used for predictions in the next batch). In particular, note that among these values of k, $\varepsilon_k \geq 1/T^2$, so that $\log(|V_0|) \lesssim d\log(T)$. We can analyze this alternate algorithm $\mathcal{A}'$, and it follows that $\mathcal{M}_T(\mathcal{A}; f^\star, \mathcal{P}) \leq \mathcal{M}_T(\mathcal{A}'; f^\star, \mathcal{P})\log(T)$.

Finally, we analyze this $\mathcal{A}'$ algorithm. The result will be proven if we can argue that $\mathbb{E}[\Delta_i] \lesssim 2^{-c\tilde{\rho}/d\log(T)}$. Note that each time the condition $E_{i-1} = \{\}$ holds in Step 4, we reduce Δ_i by a factor of 2. Furthermore, the algorithm maintains the invariant that $V_{i-1} \subseteq \mathrm{B}(f^\star, 2\Delta_i+\varepsilon_k)$, which (wlog) is contained in $\mathrm{B}(f^\star, 4\Delta_i)$. Thus, by definition of $\tilde{\rho}$, on any round i that $\Delta_i \geq 2^{-i}$, there is at least $1/T_i$ probability that a random point x will have $\mathrm{Split}(E_{i-1}, x) \geq \tilde{\rho}|E_{i-1}|$, and hence probability at least $(1 - 1/T_i)^{T_i} \geq 1/4$ that the batch will contain a point X_t with $\mathrm{Split}(E_{i-1}, X_t) \geq \tilde{\rho}|E_{i-1}|$. Moreover, by the standard analysis of the secretary problem, if such an X_t exists, there is (independent) probability at least a constant $c > 0$ that the algorithm will query this Y_t (or one of at least as large of a Split value). Thus, on each round that $\Delta_i \geq 2^{-i}$, there is some constant probability $p > 0$ that the algorithm will query some X_t with $\mathrm{Split}(E_{i-1}, X_t) \geq \tilde{\rho}|E_{i-1}|$. Therefore, with probability at least $1 - 1/T^2$, the number of rounds i having $\Delta_i \geq 2^{-i}$ before $E_{i-1} = \{\}$ is $\leq (1/c\tilde{\rho})d\log(T)$ for a constant c. If we suppose, for induction, that some particular time i satisfies $\Delta_i \lesssim 2^{-\lfloor ic\tilde{\rho}/d\log(T)\rfloor}$. Then the above analysis indicates that, with probability at least $1 - 1/T^2$, by some round i' with $i' - i = (1/c\tilde{\rho})d\log(T)$, we will either have $\Delta_{i'} \leq 2^{-i'}$, or we will have $\Delta_{i'} \leq (1/2)\Delta_i \leq 2^{-\lfloor i'c\tilde{\rho}/d\log(T)\rfloor}$. In either case, we obtain $\Delta_{i'} \leq 2^{-\lfloor i'c\tilde{\rho}/d\log(T)\rfloor}$, and (since $i' - i = (1/c\tilde{\rho})d\log(T)$), every i'' strictly between i and i' has $\Delta_{i''} \leq \Delta_i \leq 2^{-\lfloor ic\tilde{\rho}/d\log(T)\rfloor} = 2^{-\lfloor i''c\tilde{\rho}/d\log(T)\rfloor}$ Altogether, with probability at least $1-\sum_{i:t_{i-1}<T} 1/T^2 \geq 1 - 1/T$, every i has $\Delta_i \leq 2^{-\lfloor ic\tilde{\rho}/d\log(T)\rfloor}$. Therefore, $\mathcal{M}_T(\mathcal{A}'; f^\star, \mathcal{P}) \leq 1 + \sum_{i:t_{i-1}<T} T_i 4 \cdot 2^{-\lfloor ic\tilde{\rho}/d\log(T)\rfloor}$, which completes the proof. □

C Proofs for the Target-Dependent Analysis

Proof of Theorem 9. We first note that $\hat{h}_t$ is always guaranteed to exist, since (by induction) we always start each round with at least one $h \in \mathcal{H}_{k_{t-1}}$ correct on S_{t-1}, and we need only increment k_t by one (Step 7) to preserve this property for the next round (when needed, due to the condition in Step 6).

Fix any $\mathcal{P}$ and any $f^\star \in \mathcal{H}$. Let $k^\star \in \mathbb{N} \cup \{0\}$ be minimal such that $\inf_{h\in\mathcal{H}_{k^\star}} \mathcal{P}(x : h(x) \neq f^\star(x)) = 0$, and note that (since $f^\star \in \mathcal{H}_k$), if $k^\star > 0$ this implies $\exists a^\star_1, \ldots, a^\star_{k^\star}, b^\star_1, \ldots, b^\star_{k^\star} \in [0,1]$ such that $\forall i \leq k^\star$, $a^\star_i \leq b^\star_i$ and $\mathcal{P}([a^\star_i, b^\star_i]) > 0$, and $\forall i < k^\star$, $b^\star_i < a^\star_{i+1}$ and $\mathcal{P}((b^\star_i, a^\star_{i+1})) > 0$, and defining $h^* = \mathbb{1}^{\pm}_{\bigcup_{i\leq k^\star}[a^\star_i, b^\star_i]} \in \mathcal{H}_{k^\star}$, we have $\mathcal{P}(x : h^*(x) \neq f^\star(x)) = 0$; this follows from minimality of $k^\star$. Let

$$w_{\min} = \min\left(\{\mathcal{P}([a^\star_i, b^\star_i]) : i \leq k^\star\} \cup \{\mathcal{P}((b^\star_i, a^\star_{i+1})) : i < k^\star\}\right).$$

In the case $k^\star = 0$, define $h^* = \mathbb{1}^{\pm}_{\emptyset}$.

Let E denote the event, of probability one, on which $h^*(X_t) = f^\star(X_t)$ for every $t \in \mathbb{N}$. First note that, if $k^\star = 0$, then on the event E, we have $k_t = 0$ for every $t \in \mathbb{N} \cup \{0\}$, and furthermore $\hat{h}_t = h^*$ for every

$t \in \mathbb{N} \cup \{0\}$, so that the algorithm makes no mistakes, and (since $\mathrm{DIS}(\mathcal{H}_0) = \emptyset$) only makes queries when $\log_2(2t) \in \mathbb{N}$, and hence only makes $O(\log(T))$ queries. Now for the nontrivial case, suppose $k^\star > 0$. Then note that for any $t \in \mathbb{N}$, on event E, if S_t contains elements x in each $[a_i^\star, b_i^\star]$, $i \leq k^\star$, and contains elements x in each $(b_i^\star, a_{i+1}^\star)$, $i < k^\star$, then $k_t = k^\star$. Furthermore, note that on E, we never have $k_t > k^\star$. By the well-known "coupon collector" calculation, the expected number of random samples needed to obtain samples in all of these regions is at most $(c/w_{\min}) \log(2k^\star)$ for a numerical constant c. Thus, since the algorithm queries all of the iid samples $\{X_t : \log_2(2t) \in \mathbb{N}\}$, defining $T^\star = \min\{t : k_t = k^\star\}$, the expected value of $T^\star$ is at most $2^{(c/w_{\min}) \log(2k^\star)}$, which is a finite constant (dependent on $\mathcal{P}$ and $f^\star$); denote this constant by $C^\star$.

Now for any $t > T^\star$, the standard analysis of CAL applies. Specifically, define $V_t^\star = \{h \in \mathcal{H}_{k^\star} : \forall (x,y) \in S_t, h(x) = y\}$. Denote by $\mathfrak{s}^\star$ the star number of $V_{T^\star}^\star$ (recalling Definition 2). Note that, due to the definition of $T^\star$, it holds that $\mathfrak{s}^\star \leq 4k^\star$ (since no $h \in \mathcal{H}_{k^\star - 1}$ is consistent with S_{T^*}, it follows that star sets can only be constructed with one point on either side of each decision boundary). Also, on E, from the criterion in Step 3, we observe that $V_t^\star \subseteq \{h \in \mathcal{H}_{k^\star} : \forall t' \in \{T^\star + 1, \ldots, t\}, h(X_t) = Y_t\}$. Thus, (1) implies any $t > T^\star$ has

$$\mathbb{E}\big[\mathbb{P}\big(\big\{X_t \in \mathrm{DIS}(V_{t-1}^\star)\big\} \cap E \big| T^\star\big)\big] \leq \frac{4k^\star}{t - T^\star}.$$

Furthermore, since (on E) both of h^* and $\hat{h}_{t-1}$ are in $V_{t-1}^\star$, we have $\{x : \hat{h}_{t-1}(x) \neq h^*(x)\} \subseteq \mathrm{DIS}(V_{t-1}^\star)$, so that it also holds that

$$\mathbb{E}\Big[\mathbb{P}\Big(\Big\{\hat{h}_{t-1}(X_t) \neq Y_t\Big\} \cap E \Big| T^\star\Big)\Big] \leq \frac{4k^\star}{t - T^\star}.$$

Thus, the expected number of mistakes up to a time T is at most

$$\mathbb{E}\left[T^\star + \sum_{t=T^\star+1}^{T} \frac{4k^\star}{t - T^\star}\right] \leq C^\star + 4k^\star \ln(eT) = O(\log(T)),$$

and the expected number of queries up to a time T is at most

$$\mathbb{E}\left[T^\star + \sum_{t=T^\star+1}^{T} \left(\frac{4k^\star}{t - T^\star} + \mathbb{1}[\log_2(2t) \in \mathbb{N}]\right)\right] \leq C^\star + 4k^\star \ln(eT) + \log_2(2T) = O(\log(T)).$$

□