Efficient Active Learning with Abstention

Yinglun Zhu
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706
yinglun@cs.wisc.edu

Robert Nowak
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706
rdnowak@wisc.edu

Abstract

The goal of active learning is to achieve the same accuracy achievable by passive learning, while using much fewer labels. Exponential savings in terms of label complexity have been proved in very special cases, but fundamental lower bounds show that such improvements are impossible in general. This suggests a need to explore alternative goals for active learning. Learning with abstention is one such alternative. In this setting, the active learning algorithm may abstain from prediction and incur an error that is marginally smaller than random guessing. We develop the first computationally efficient active learning algorithm with abstention. Our algorithm provably achieves $\mathsf{polylog}(\frac{1}{\varepsilon})$ label complexity, without any low noise conditions. Such performance guarantee reduces the label complexity by an exponential factor, relative to passive learning and active learning that is not allowed to abstain. Furthermore, our algorithm is guaranteed to only abstain on hard examples (where the true label distribution is close to a fair coin), a novel property we term *proper abstention* that also leads to a host of other desirable characteristics (e.g., recovering minimax guarantees in the standard setting, and avoiding the undesirable "noise-seeking" behavior often seen in active learning). We also provide novel extensions of our algorithm that achieve *constant* label complexity and deal with model misspecification.

1 Introduction

Active learning aims at learning an accurate classifier with a small number of labeled data points (Settles, 2009; Hanneke, 2014). Active learning has become increasingly important in modern application of machine learning, where unlabeled data points are abundant yet the labeling process requires expensive time and effort. Empirical successes of active learning have been observed in many areas (Tong and Koller, 2001; Gal et al., 2017; Sener and Savarese, 2018). In noise-free or certain low-noise cases (i.e., under Massart noise (Massart and Nédélec, 2006)), active learning algorithms with *provable* exponential savings over the passive counterpart have been developed (Balcan et al., 2007; Hanneke, 2007; Dasgupta et al., 2009; Hsu, 2010; Dekel et al., 2012; Hanneke, 2014; Zhang and Chaudhuri, 2014; Krishnamurthy et al., 2019; Katz-Samuels et al., 2021). On the other hand, however, not much can be said in the general case. In fact, Kääriäinen (2006) provides a $\Omega(\frac{1}{\varepsilon^2})$ lower bound by reducing active learning to a simple mean estimation problem: It takes $\Omega(\frac{1}{\varepsilon^2})$ samples to distinguish $\eta(x) = \frac{1}{2} + \varepsilon$ and $\eta(x) = \frac{1}{2} - \varepsilon$. Even with the relatively benign Tsybakov noise (Tsybakov, 2004), Castro and Nowak (2006, 2008) derive a $\Omega(\mathrm{poly}(\frac{1}{\varepsilon}))$ lower bound, again, indicating that exponential speedup over passive learning is not possible in general. These fundamental lower bounds lay out statistical barriers to active learning, and suggests considering a refinement of the label complexity goals in active learning (Kääriäinen, 2006).

Inspecting these lower bounds, one can see that active learning suffers from classifying hard examples that are close to the decision boundary. However, *do we really require a trained classifier to do well*

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

on those hard examples? In high-risk domains such as medical imaging, it makes more sense for the classifier to abstain from making the decision and leave the problem to a human expert. Such idea is formalized under Chow's error (Chow, 1970): Whenever the classifier chooses to abstain, a loss that is barely smaller than random guessing, i.e., $\frac{1}{2} - \gamma$, is incurred. The parameter γ should be thought as a small positive quantity, e.g., $\gamma = 0.01$. The inclusion of abstention is not only practically interesting, but also provides a statistical refinement of the label complexity goal of active learning: Achieving exponential improvement under Chow's excess error. When abstention is allowed as an action, Puchkin and Zhivotovskiy (2021) shows, for the first time, that exponential improvement in label complexity can be achieved by active learning in the general setting. However, the approach provided in Puchkin and Zhivotovskiy (2021) can not be efficiently implemented. Their algorithm follows the disagreement-based approach and requires maintaining a version space and checking whether or not an example lies in the region of disagreement. It is not clear how to generally implement these operations besides enumeration (Beygelzimer et al., 2010). Moreover, their algorithm relies on an Empirical Risk Minimization (ERM) oracle, which is known to be NP-Hard even for a simple linear hypothesis class (Guruswami and Raghavendra, 2009).

In this paper, we break the computational barrier and design an efficient active learning algorithm with exponential improvement in label complexity relative to conventional passive learning. The algorithm relies on weighted square loss regression oracle, which can be efficiently implemented in many cases (Krishnamurthy et al., 2017, 2019; Foster et al., 2018, 2020). The algorithm also abstains properly, i.e., abstain only when it is the optimal choice, which allows us to easily translate the guarantees to the *standard* excess error. Along the way, we propose new noise-seeking noise conditions and show that: "uncertainty-based" active learners can be easily trapped, yet our algorithm provably overcome these noise-seeking conditions. As an extension, we also provide the first algorithm that enjoys *constant* label complexity for a *general* set of regression functions.

1.1 Problem setting

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ denote the label space. We focus on the binary classification problem where $\mathcal{Y} = \{+1, -1\}$. The joint distribution over $\mathcal{X} \times \mathcal{Y}$ is denoted as $\mathcal{D}_{\mathcal{X}\mathcal{Y}}$. We use $\mathcal{D}_{\mathcal{X}}$ to denote the marginal distribution over the input space $\mathcal{X}$, and use $\mathcal{D}_{\mathcal{Y}|x}$ to denote the conditional distribution of $\mathcal{Y}$ with respect to any $x \in \mathcal{X}$. We define $\eta(x) := \mathbb{P}_{y \sim \mathcal{D}_{\mathcal{Y}|x}}(y = +1)$ as the conditional probability of taking a positive label. We consider the standard active learning setup where $(x, y) \sim \mathcal{D}_{\mathcal{X}\mathcal{Y}}$ but y is observed only after a label querying. We consider hypothesis class $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$. For any classifier $h \in \mathcal{H}$, its (standard) error is defined as $\mathrm{err}(h) := \mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathcal{X}\mathcal{Y}}}(h(x) \neq y)$.

Function approximation. We focus on the case where the hypothesis class $\mathcal{H}$ is induced from a set of regression functions $\mathcal{F} : \mathcal{X} \to [0, 1]$ that predicts the conditional probability $\eta(x)$. We write $\mathcal{H} = \mathcal{H}_{\mathcal{F}} := \{h_f : f \in \mathcal{F}\}$ where $h_f(x) := \mathrm{sign}(2f(x) - 1)$. The "size" of $\mathcal{F}$ is measured by the well-known complexity measure: the *Pseudo dimension* $\mathrm{Pdim}(\mathcal{F})$ (Pollard, 1984; Haussler, 1989, 1995). We assume $\mathrm{Pdim}(\mathcal{F}) < \infty$ throughout the paper.[1] Following existing works in active learning (Dekel et al., 2012; Krishnamurthy et al., 2017, 2019) and contextual bandits (Agarwal et al., 2012; Foster et al., 2018; Foster and Rakhlin, 2020; Simchi-Levi and Xu, 2020), we make the following *realizability* assumption.

Assumption 1 (Realizability)**.** *The learner is given a set of regressors $\mathcal{F} : \mathcal{X} \to [0, 1]$ such that there exists a $f^\star \in \mathcal{F}$ characterize the true conditional probability, i.e., $f^\star = \eta$.*

The realizability assumption allows *rich function approximation*, which strictly generalizes the setting with linear function approximation studied in active learning (e.g., in (Dekel et al., 2012)). We relax Assumption 1 in Section 4.2 to deal with model misspecification.

Regression oracle. We consider a regression oracle over $\mathcal{F}$, which is extensively studied in the literature in active learning and contextual bandits (Krishnamurthy et al., 2017, 2019; Foster et al., 2018, 2020). Given any set $\mathcal{S}$ of weighted examples $(w, x, y) \in \mathbb{R}_+ \times \mathcal{X} \times \mathcal{Y}$ as input, the regression

[1] See Appendix D for formal definition of the Pseudo dimension. Many function classes of practical interests have finite Pseudo dimension: (1) when $\mathcal{F}$ is finite, we have $\mathrm{Pdim}(\mathcal{F}) = O(\log|\mathcal{F}|)$; (2) when $\mathcal{F}$ is a set of linear functions/generalized linear function with non-decreasing link function, we have $\mathcal{F} = O(d)$; (3) when $\mathcal{F}$ is a set of degree-r polynomial in $\mathbb{R}^d$, we have $\mathrm{Pdim}(\mathcal{F}) = O(\binom{d+r}{r})$.

oracle outputs

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \sum_{(w,x,y) \in \mathcal{S}} w(f(x) - y)^2. \tag{1}$$

The regression oracle solves a convex optimization problem with respect to the regression function, and admits closed-form solutions in many cases, e.g., it is reduced to least squares when f is linear. We view the implementation of the regression oracle as an efficient operation and quantify the computational complexity in terms of the number of calls to the regression oracle.

Chow's excess error (Chow, 1970). Let $h^\star := h_{f^\star} \in \mathcal{H}$ denote the Bayes classifier. The *standard excess error* of classifier $h \in \mathcal{H}$ is defined as $\mathrm{err}(h) - \mathrm{err}(h^\star)$. Since achieving exponential improvement (of active over passive learning) with respect to the standard excess error is impossible in general (Kääriäinen, 2006), we introduce Chow's excess error next. We consider classifier of the form $\widehat{h} : \mathcal{X} \to \mathcal{Y} \cup \{\perp\}$ where $\perp$ denotes the action of abstention. For any fixed $0 < \gamma < \frac{1}{2}$, the Chow's error is defined as

$$\mathrm{err}_\gamma(\widehat{h}) := \mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathcal{X}\mathcal{Y}}}(\widehat{h}(x) \neq y, \widehat{h}(x) \neq \perp) + (1/2 - \gamma) \cdot \mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathcal{X}\mathcal{Y}}}(\widehat{h}(x) = \perp). \tag{2}$$

The parameter γ can be chosen as a small constant, e.g., $\gamma = 0.01$, to avoid excessive abstention: The price of abstention is only marginally smaller than random guess. The *Chow's excess error* is then defined as $\mathrm{err}_\gamma(\widehat{h}) - \mathrm{err}(h^\star)$ (Puchkin and Zhivotovskiy, 2021). For any fixed accuracy level $\varepsilon > 0$, we aim at constructing a classifier $\widehat{h} : \mathcal{X} \to \mathcal{Y} \cup \{\perp\}$ with ε Chow's excess error and $\mathrm{polylog}(\frac{1}{\varepsilon})$ label complexity. We also relate Chow's excess error to standard excess error in Section 3.

Remark 1. *Competing against the optimal Chow's error, i.e.,* $\mathrm{err}_\gamma(\widehat{h}) - \inf_{h:\mathcal{X}\to\{+1,-1,\perp\}} \mathrm{err}_\gamma(h)$*, will eliminate active learning gains. As in Kääriäinen (2006), it suffices to consider a simple problem with* $\mathcal{X} = \{x\}$*. In order to achieve* ε *excess error against the optimal Chow's classifier, we need to distinguish cases* $\eta(x) = \frac{1}{2} - \gamma - 2\varepsilon$ *and* $\eta(x) = \frac{1}{2} - \gamma + 2\varepsilon$*, which inevitably requires* $\Omega(\frac{1}{\varepsilon^2})$ *samples. We defer a detailed discussion (with pictorial explanations) of Chow's excess error in Appendix B.*

1.2 Contributions and paper organization

We provide informal statements of our main results in this section. Our results depend on complexity measures such as *value function* disagreement coefficient θ and eluder dimension $\mathfrak{e}$ (formally defined in Section 2 and Appendix C). These complexity measures are previously analyzed in contextual bandits (Russo and Van Roy, 2013; Foster et al., 2020) and we import them to the active learning setup. These complexity measures are well-bounded for many function classes of practical interests, e.g., we have $\theta, \mathfrak{e} = \widetilde{O}(d)$ for linear and generalized linear functions on $\mathbb{R}^d$.

Our first main contribution is that we design the first *computationally efficient* active learning algorithm (Algorithm 1) that achieves exponential labeling savings, *without any low noise assumptions*.

Theorem 1 (Informal)**.** *There exists an algorithm that constructs a classifier* $\widehat{h} : \mathcal{X} \to \{+1, -1, \perp\}$ *with Chow's excess error at most* ε *and label complexity* $\widetilde{O}(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\gamma^2} \cdot \mathrm{polylog}(\frac{1}{\varepsilon}))$*, without any low noise assumptions. The algorithm can be efficiently implemented via a regression oracle: It takes* $\widetilde{O}(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma^3})$ *oracle calls for general* $\mathcal{F}$*, and* $\widetilde{O}(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma})$ *oracle calls for convex* $\mathcal{F}$*.*

The formal statements are provided in Section 2. The *statistical* guarantees (i.e., label complexity) in Theorem 1 is similar to the one achieved in Puchkin and Zhivotovskiy (2021), with one critical difference: The label complexity provided in Puchkin and Zhivotovskiy (2021) is in terms of the *classifier-based* disagreement coefficient $\breve{\theta}$ (Hanneke, 2014). Even for a set of linear classifier, $\breve{\theta}$ is only known to be bounded in special cases, e.g., when $\mathcal{D}_\mathcal{X}$ is uniform over the unit sphere (Hanneke, 2007). On the other hand, we have $\theta \leq d$ for any $\mathcal{D}_\mathcal{X}$ (Foster et al., 2020).

We say that a classifier $\widehat{h} : \mathcal{X} \to \{+1, -1, \perp\}$ enjoys proper abstention if it abstains only if abstention is indeed the optimal choice (based on Eq. (2)). For any classifier that enjoys proper abstention, one can easily relate its *standard* excess error to the Chow's excess error, under commonly studied Massart/Tsybakov noises (Massart and Nédélec, 2006; Tsybakov, 2004). The classifier obtained in Theorem 1 enjoys proper abstention, and achieves the following guarantees (formally stated in Section 3.1).

Theorem 2 (Informal). *Under Massart/Tsybakov noise, with appropriate adjustments, the classifier learned in Theorem 1 achieves the minimax optimal label complexity under standard excess error.*

We also propose new noise conditions that *strictly* generalize the usual Massart/Tsybakov noises, which we call noise-seeking conditions. At a high-level, the noise-seeking conditions allow abundant data points with $\eta(x)$ equal/close to $\frac{1}{2}$. These points are somewhat "harmless" since it hardly matters what label is predicted at that point (in terms of excess error). These seemingly "harmless" data points can, however, cause troubles for any active learning algorithm that requests the label for any point that is uncertain, i.e., the algorithm cannot decide if $|\eta(x) - \frac{1}{2}|$ is strictly greater than 0. We call such algorithms "uncertainty-based" active learners. These algorithms could wastefully sample in these "harmless" regions, ignoring other regions where erring could be much more harmful. We derive the following proposition (formally stated in Section 3.2) under these noise-seeking conditions.

Proposition 1 (Informal). *For any labeling budget $B \gtrsim \frac{1}{\gamma^2} \cdot \mathrm{polylog}(\frac{1}{\varepsilon})$, there exists a learning problem such that (1) any uncertainty-based active learner suffers standard excess error $\Omega(B^{-1})$; yet (2) the classifier $\widehat{h}$ learned in Theorem 1 achieves standard excess error at most ε.*

The above result demonstrates the superiority of our algorithm over any "uncertainty-based" active learner. Moreover, we show that, under these strictly harder noise-seeking conditions, our algorithm still achieve guarantees similar to the ones stated in Theorem 2.

Before presenting our next main result, we first consider a simple active learning problem with $\mathcal{X} = \{x\}$. Under Massart noise, we have $|\eta(x) - \frac{1}{2}| \geq \tau_0$ for some constant $\tau_0 > 0$. Thus, it takes no more than $O(\tau_0^{-2} \log \frac{1}{\delta})$ labels to achieve ε standard excess error, no matter how small ε is. This example shows that, at least in simple cases, we can expect to achieve a *constant* label complexity for active learning, with no dependence on $\frac{1}{\varepsilon}$ at all. To the best of our knowledge, our next result provides the first generalization of such phenomenon to a *general* set of (finite) regression functions, as long as its eluder dimension $\mathfrak{e}$ is bounded.

Theorem 3 (Informal). *Under Massart noise with parameter τ_0 and a general (finite) set of regression function $\mathcal{F}$. There exists an algorithm that returns a classifier with standard excess error at most ε and label complexity $O(\frac{\mathfrak{e} \cdot \log(|\mathcal{F}|/\delta)}{\tau_0^2})$, which is independent of $\frac{1}{\varepsilon}$.*

A similar constant label complexity holds with Chow's excess error, without any low noise assumptions. We also provide discussion on why previous algorithms do not achieve such constant label complexity, even in the case with linear functions. We defer formal statements and discussion to Section 4.1. In Section 4.2, we relax Assumption 1 and propose an algorithm that can deal with model misspecification.

Paper organization. The rest of this paper is organized as follows. We present our main algorithm and its guarantees in Section 2. We further analyze our algorithm under standard excess error in Section 3, and discuss other important properties of the algorithm. Extensions of our algorithm, e.g., achieving *constant* label complexity and dealing with model misspecification, are provided in Section 4. We defer the discussion of additional related work and all proofs to the Appendix due to lack of space.

2 Efficient active learning with abstention

We provide our main algorithm (Algorithm 1) in this section. Algorithm 1 is an adaptation of the algorithm developed in Krishnamurthy et al. (2017, 2019), which studies active learning under the standard excess error (and Massart/Tsybakov noises). We additionally take the abstention option into consideration, and *manually construct* classifiers using the active set of (uneliminated) regression functions (which do not belong to the original hypothesis class). These new elements allow us to achieve ε Chow's excess error with $\mathrm{polylog}(\frac{1}{\varepsilon})$ label complexity, without any low noise assumptions.

Algorithm 1 Efficient Active Learning with Abstention

Input: Accuracy level $\varepsilon > 0$, abstention parameter $\gamma \in (0, 1/2)$ and confidence level $\delta \in (0, 1)$.

Define $T := \widetilde{O}(\frac{\theta \operatorname{Pdim}(\mathcal{F})}{\varepsilon\,\gamma})$, $M := \lceil \log_2 T \rceil$ and $C_\delta := O(\operatorname{Pdim}(\mathcal{F}) \cdot \log(T/\delta))$.
Define $\tau_m := 2^m$ for $m \geq 1$, $\tau_0 := 0$ and $\beta_m := (M - m + 1) \cdot C_\delta$.
for epoch $m = 1, 2, \ldots, M$ **do**
  Get $\widehat{f}_m := \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{\tau_{m-1}} Q_t (f(x_t) - y_t)^2$.
    // We use $Q_t \in \{0, 1\}$ to indicate whether the label of x_t is queried.
  (Implicitly) Construct active set of regression functions $\mathcal{F}_m \subseteq \mathcal{F}$ as

$$\mathcal{F}_m := \left\{ f \in \mathcal{F} : \sum_{t=1}^{\tau_{m-1}} Q_t (f(x_t) - y_t)^2 \leq \sum_{t=1}^{\tau_{m-1}} Q_t (\widehat{f}_m(x_t) - y_t)^2 + \beta_m \right\}.$$

  Construct classifier $\widehat{h}_m : \mathcal{X} \to \{+1, -1, \perp\}$ as

$$\widehat{h}_m(x) := \begin{cases} \perp, & \text{if } [\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)] \subseteq \left[\frac{1}{2} - \gamma, \frac{1}{2} + \gamma\right]; \\ \operatorname{sign}(2\widehat{f}_m(x) - 1), & \text{o.w.} \end{cases}$$

  and construct query function $g_m(x) := \mathbb{1}\big(\frac{1}{2} \in (\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m))\big) \cdot \mathbb{1}(\widehat{h}_m(x) \neq \perp)$.
  if epoch $m = M$ **then**
    Return classifier $\widehat{h}_M$.
  for time $t = \tau_{m-1} + 1, \ldots, \tau_m$ **do**
    Observe $x_t \sim \mathcal{D}_\mathcal{X}$. Set $Q_t := g_m(x_t)$.
    if $Q_t = 1$ **then**
      Query the label y_t of x_t.

Algorithm 1 runs in epochs of geometrically increasing lengths. At the beginning of epoch $m \in [M]$, Algorithm 1 first computes the empirical best regression function $\widehat{f}_m$ that achieves the smallest cumulative square loss over previously labeled data points ($\widehat{f}_1$ can be selected arbitrarily); it then (implicitly) constructs an active set of regression functions $\mathcal{F}_m$, where the cumulative square loss of each $f \in \mathcal{F}_m$ is not too much larger than the cumulative square loss of empirical best regression function $\widehat{f}_m$. For any $x \in \mathcal{X}$, based on the active set of regression functions, Algorithm 1 constructs a lower bound $\mathsf{lcb}(x; \mathcal{F}_m) := \inf_{f \in \mathcal{F}_m} f(x)$ and an upper bound $\mathsf{ucb}(x; \mathcal{F}_m) := \sup_{f \in \mathcal{F}_m} f(x)$ for the true conditional probability $\eta(x)$. An empirical classifier $\widehat{h}_m : \mathcal{X} \to \{+1, -1, \perp\}$ and a query function $g_m : \mathcal{X} \to \{0, 1\}$ are then constructed based on these confidence ranges and the abstention parameter γ. For any time step t within epoch m, Algorithm 1 queries the label of the observed data point x_t if and only if $Q_t := g_m(x_t) = 1$. Algorithm 1 returns $\widehat{h}_M$ as the learned classifier.

We now discuss the empirical classifier $\widehat{h}_m$ and the query function g_m in more detail. Consider the event where $f^\star \in \mathcal{F}_m$ for all $m \in [M]$, which can be shown to hold with high probability. The constructed confidence intervals are valid under this event, i.e., $\eta(x) \in [\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)]$. First, let us examine the conditions that determine a label query. The label of x is *not* queried if

- **Case 1:** $\widehat{h}_m(x) = \perp$**.** We have $\eta(x) \in [\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)] \subseteq [\frac{1}{2} - \gamma, \frac{1}{2} + \gamma]$. Abstention leads to the smallest error (Herbei and Wegkamp, 2006), and no query is needed.
- **Case 2:** $\frac{1}{2} \notin (\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m))$**.** We have $\operatorname{sign}(2\widehat{f}_m(x) - 1) = \operatorname{sign}(2f^\star(x) - 1)$. Thus, no excess error is incurred and there is no need to query.

The only case when label query *is* issued, and thus when the classifier $\widehat{h}_m$ may suffer from excess error, is when

$$\frac{1}{2} \in (\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)) \quad \text{and} \quad [\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)] \nsubseteq \left[\frac{1}{2} - \gamma, \frac{1}{2} + \gamma\right] \tag{3}$$

hold simultaneously. Eq. (3) necessarily leads to the condition $w(x; \mathcal{F}_m) := \mathsf{ucb}(x; \mathcal{F}_m) - \mathsf{lcb}(x; \mathcal{F}_m) > \gamma$. Our theoretical analysis shows that the event must $\mathbb{1}(w(x; \mathcal{F}_m) > \gamma)$ happens in-

frequently, and its frequency is closely related to the so-called *value function disagreement coefficient* (Foster et al., 2020), which we introduce as follows.[2]

Definition 1 (Value function disagreement coefficient)**.** *For any $f^\star \in \mathcal{F}$ and $\gamma_0, \varepsilon_0 > 0$, the value function disagreement coefficient $\theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma_0, \varepsilon_0)$ is defined as*

$$\sup_{\mathcal{D}_\mathcal{X}} \sup_{\gamma > \gamma_0, \varepsilon > \varepsilon_0} \left\{ \frac{\gamma^2}{\varepsilon^2} \cdot \mathbb{P}_{\mathcal{D}_\mathcal{X}} \left(\exists f \in \mathcal{F} : |f(x) - f^\star(x)| > \gamma, \|f - f^\star\|_{\mathcal{D}_\mathcal{X}} \le \varepsilon \right) \right\} \vee 1,$$

where $\|f\|^2_{\mathcal{D}_\mathcal{X}} := \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[f^2(x)]$.

Combining the insights discussed above, we derive the following label complexity guarantee for Algorithm 1 (we use $\theta := \sup_{f^\star \in \mathcal{F}, \iota > 0} \theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma/2, \iota)$ and discuss its boundedness below). [3]

Theorem 4. *With probability at least $1 - 2\delta$, Algorithm 1 returns a classifier with Chow's excess error at most ε and label complexity $O(\frac{\theta \operatorname{Pdim}(\mathcal{F})}{\gamma^2} \cdot \log^2(\frac{\theta \operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma}) \cdot \log(\frac{\theta \operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma \delta}))$.*

Theorem 4 shows that Algorithm 1 achieves exponential label savings (i.e., $\operatorname{polylog}(\frac{1}{\varepsilon})$) without any low noise assumptions. We discuss the result in more detail next.

- **Boundedness of θ.** The value function disagreement coefficient is well-bounded for many function classes of practical interests. For instance, we have $\theta \le d$ for linear functions on $\mathbb{R}^d$ and $\theta \le C_{\mathsf{link}} \cdot d$ for generalized linear functions (where C_{link} is a quantity related to the link function). Moreover, θ is *always* upper bounded by complexity measures such as (squared) star number and eluder dimension (Foster et al., 2020). See Appendix C for the detailed definitions/bounds.
- **Comparison to Puchkin and Zhivotovskiy (2021).** The label complexity bound derived in Theorem 4 is similar to the one derived in Puchkin and Zhivotovskiy (2021), with one critical difference: The bound derived in Puchkin and Zhivotovskiy (2021) is in terms of *classifier-based* disagreement coefficient $\check{\theta}$ (Hanneke, 2014). Even in the case with linear classifiers, $\check{\theta}$ is only known to be bounded under additional assumptions, e.g., when $\mathcal{D}_\mathcal{X}$ is uniform over the unit sphere.

Computational efficiency. We discuss how to efficiently implement Algorithm 1 with the regression oracle defined in Eq. (1). [4] Our implementation relies on subroutines developed in Krishnamurthy et al. (2017); Foster et al. (2018), which allow us to approximate confidence bounds $\mathsf{ucb}(x; \mathcal{F}_m)$ and $\mathsf{lcb}(x; \mathcal{F}_m)$ up to α approximation error with $O(\frac{1}{\alpha^2} \log \frac{1}{\alpha})$ (or $O(\log \frac{1}{\alpha})$ when $\mathcal{F}$ is convex and closed under pointwise convergence) calls to the regression oracle. To achieve the same theoretical guarantees shown in Theorem 4 (up to changes in constant terms), we show that it suffices to (i) control the approximation error at level $O(\frac{\gamma}{\log T})$, (ii) construct the approximated confidence bounds $\widehat{\mathsf{lcb}}(x; \mathcal{F}_m)$ and $\widehat{\mathsf{ucb}}(x; \mathcal{F}_m)$ in a way such that the confidence region is non-increasing with respect to the epoch m, i.e., $(\widehat{\mathsf{lcb}}(x; \mathcal{F}_m), \widehat{\mathsf{ucb}}(x; \mathcal{F}_m)) \subseteq (\widehat{\mathsf{lcb}}(x; \mathcal{F}_{m-1}), \widehat{\mathsf{ucb}}(x; \mathcal{F}_{m-1}))$ (this ensures that the sampling region is non-increasing even with *approximated* confidence bounds, which is important to our theoretical analysis), and (iii) use the approximated confidence bounds $\widehat{\mathsf{lcb}}(x; \mathcal{F}_m)$ and $\widehat{\mathsf{ucb}}(x; \mathcal{F}_m)$ to construct the classifier $\widehat{h}_m$ and the query function g_m. We provide our guarantees as follows, and leave details to Appendix E (we redefine $\theta := \sup_{f^\star \in \mathcal{F}, \iota > 0} \theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma/4, \iota)$ in the Theorem 5 to account to approximation error).

Theorem 5. *Algorithm 1 can be efficiently implemented via the regression oracle and enjoys the same theoretical guarantees stated in Theorem 4. The number of oracle calls needed is $\widetilde{O}(\frac{\theta \operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma^3})$*

[2] Compared to the original definition studied in contextual bandits (Foster et al., 2020), our definition takes an additional "sup" over all possible marginal distributions $\mathcal{D}_\mathcal{X}$ to account for *distributional shifts* incurred by selective querying (which do not occur in contextual bandits). Nevertheless, as we show below, our disagreement coefficient is still well-bounded for many important function classes.

[3] It suffices to take $\theta := \theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma/2, \iota)$ with $\iota \propto \sqrt{\gamma \varepsilon}$ to derive a slightly different guarantee. See Appendix E.

[4] Recall that the implementation of the regression oracle should be viewed as an efficient operation since it solves a convex optimization problem with respect to the regression function, and it even admits closed-form solutions in many cases, e.g., it is reduced to least squares when f is linear. On the other hand, the ERM oracle used in Puchkin and Zhivotovskiy (2021) is NP-hard even for a set of linear classifiers (Guruswami and Raghavendra, 2009).

for a general set of regression functions $\mathcal{F}$, and $\widetilde{O}(\frac{\theta \operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma})$ when $\mathcal{F}$ is convex and closed under pointwise convergence. The per-example inference time of the learned $\widehat{h}_M$ is $\widetilde{O}(\frac{1}{\gamma^2} \log^2(\frac{\theta \operatorname{Pdim}(\mathcal{F})}{\varepsilon}))$ for general $\mathcal{F}$, and $\widetilde{O}(\log \frac{1}{\gamma})$ when $\mathcal{F}$ is convex and closed under pointwise convergence.

With Theorem 5, we provide the first computationally efficient active learning algorithm that achieves exponential label savings, without any low noise assumptions.

3 Guarantees under standard excess error

We provide guarantees for Algorithm 1 under *standard* excess error. In Section 3.1, we show that Algorithm 1 can be used to recover the usual minimax label complexity under Massart/Tsybakov noise; we also provide a new learning paradigm based on Algorithm 1 under limited budget. In Section 3.2, we show that Algorithm 1 provably avoid the undesired *noise-seeking* behavior often seen in active learning.

3.1 Recovering minimax optimal label complexity

One way to convert an abstaining classifier $\widehat{h} : \mathcal{X} \to \mathcal{Y} \cup \{\perp\}$ into a standard classifier $\check{h} : \mathcal{X} \to \mathcal{Y}$ is by randomizing the prediction in its abstention region, i.e., if $\widehat{h}(x) = \perp$, then its randomized version $\check{h}(x)$ predicts $+1/-1$ with equal probability (Puchkin and Zhivotovskiy, 2021). With such randomization, the *standard excess error* of $\check{h}$ can be characterized as

$$\operatorname{err}(\check{h}) - \operatorname{err}(h^\star) = \operatorname{err}_\gamma(\widehat{h}) - \operatorname{err}(h^\star) + \gamma \cdot \mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(\widehat{h}(x) = \perp). \tag{4}$$

The standard excess error depends on the (random) abstention region of $\widehat{h}$, which is difficult to quantify in general. To give a more practical characterization of the standard excess error, we introduce the concept of proper abstention in the following.

Definition 2 (Proper abstention)**.** *A classifier $\widehat{h} : \mathcal{X} \to \mathcal{Y} \cup \{\perp\}$ enjoys proper abstention if and only if it abstains in regions where abstention is indeed the optimal choice, i.e., $\{x \in \mathcal{X} : \widehat{h}(x) = \perp\} \subseteq \{x \in \mathcal{X} : \eta(x) \in [\frac{1}{2} - \gamma, \frac{1}{2} + \gamma]\} =: \mathcal{X}_\gamma$.*

Proposition 2. *The classifier $\widehat{h}$ returned by Algorithm 1 enjoys proper abstention. With randomization over the abstention region, we have the following upper bound on its standard excess error*

$$\operatorname{err}(\check{h}) - \operatorname{err}(h^\star) \le \operatorname{err}_\gamma(\widehat{h}) - \operatorname{err}(h^\star) + \gamma \cdot \mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(x \in \mathcal{X}_\gamma). \tag{5}$$

The proper abstention property of $\widehat{h}$ returned by Algorithm 1 is achieved via conservation: $\widehat{h}$ will avoid abstention unless it is absolutely sure that abstention is the optimal choice.[5] To characterize the standard excess error of classifier with proper abstention, we only need to upper bound the term $\mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(x \in \mathcal{X}_\gamma)$, which does *not* depends on the (random) classifier $\widehat{h}$. Instead, it only depends on the marginal distribution. We next introduce the common Massart/Tsybakov noise conditions.

Definition 3 (Massart noise, Massart and Nédélec (2006))**.** *A distribution $\mathcal{D}_{\mathcal{X}\mathcal{Y}}$ satisfies the Massart noise condition with parameter $\tau_0 > 0$ if $\mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(|\eta(x) - 1/2| \le \tau_0) = 0$.*

Definition 4 (Tsybakov noise, Tsybakov (2004))**.** *A distribution $\mathcal{D}_{\mathcal{X}\mathcal{Y}}$ satisfies the Tsybakov noise condition with parameter $\beta \ge 0$ and a universal constant $c > 0$ if $\mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(|\eta(x) - 1/2| \le \tau) \le c\,\tau^\beta$ for any $\tau > 0$.*

As in Balcan et al. (2007); Hanneke (2014), we assume knowledge of noise parameters (e.g., τ_0, β). Together with the active learning lower established in Castro and Nowak (2006, 2008), and focusing on the dependence of ε, our next theorem shows that Algorithm 1 can be used to recover the minimax label complexity in active learning, under the *standard* excess error.

[5] On the other hand, however, the algorithm provided in Puchkin and Zhivotovskiy (2021) is very unlikely to have such property. In fact, only a small but *nonzero* upper bound of abstention rate is provided (Proposition 3.6 therein) under the Massart noise with $\gamma \le \frac{\tau_0}{2}$; yet any classifier that enjoys proper abstention should have exactly zero abstention rate.

Theorem 6. *With an appropriate choice of the abstention parameter γ in Algorithm 1 and randomization over the abstention region, Algorithm 1 learns a classifier $\check{h}$ at the minimax optimal rates: To achieve ε standard excess error, it takes $\widetilde{\Theta}(\tau_0^{-2})$ labels under Massart noise and takes $\widetilde{\Theta}(\varepsilon^{-2/(1+\beta)})$ labels under Tsybakov noise.*

Remark 2. *In addition to recovering the minimax rates, the proper abstention property is desirable in practice: It guarantees that $\widehat{h}$ will not abstain on easy examples, i.e., it will not mistakenly flag easy examples as "hard-to-classify", thus eliminating unnecessary human labeling efforts.*

Algorithm 1 can also be used to provide new learning paradigms in the limited budget setting, which we introduce below. No prior knowledge of noise parameters are required in this setup.

New learning paradigm under limited budget. Given any labeling budget $B > 0$, we can then choose $\gamma \approx B^{-1/2}$ in Algorithm 1 to make sure the label complexity is never greater than B (with high probability). The learned classifier enjoys Chow's excess error (with parameter γ) at most ε; its standard excess error (with randomization over the abstention region) can be analyzed by relating the $\gamma \cdot \mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}(x \in \mathcal{X}_\gamma)$ term in Eq. (5) to the Massart/Tsybakov noise conditions, as discussed above.

3.2 Abstention to avoid noise-seeking

Active learning algorithms sometimes exhibit *noise-seeking* behaviors, i.e., oversampling in regions where $\eta(x)$ is close to the $\frac{1}{2}$ level. Such noise-seeking behavior is known to be a fundamental barrier to achieve low label complexity (under standard excess error), e.g., see Kääriäinen (2006). We show in this section that abstention naturally helps avoiding noise-seeking behaviors and speeding up active learning.

To better illustrate how properly abstaining classifiers avoid noise-seeking behavior, we first propose new noise conditions below, which strictly generalize the usual Massart/Tsybakov noises.

Definition 5 (Noise-seeking Massart noise)**.** *A distribution $\mathcal{D}_{\mathcal{X}\mathcal{Y}}$ satisfies the noise-seeking Massart noise condition with parameters $0 \le \zeta_0 < \tau_0 \le 1/2$ if $\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}(\zeta_0 < |\eta(x) - 1/2| \le \tau_0) = 0$.*

Definition 6 (Noise-seeking Tsybakov noise)**.** *A distribution $\mathcal{D}_{\mathcal{X}\mathcal{Y}}$ satisfies the noise-seeking Tsybakov noise condition with parameters $0 \le \zeta_0 < 1/2$, $\beta \ge 0$ and a universal constant $c > 0$ if $\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}(\zeta_0 < |\eta(x) - 1/2| \le \tau) \le c\,\tau^\beta$ for any $\tau > \zeta_0$.*

Compared to the standard Massart/Tsybakov noises, these newly proposed noise-seeking conditions allow arbitrary probability mass of data points whose conditional probability $\eta(x)$ is equal/close to $1/2$. As a result, they can trick standard active learning algorithms into exhibiting the noise-seeking bahaviors (and hence their names). We also mention that the parameter ζ_0 should be considered as an *extremely small quantity* (e.g., $\zeta_0 \ll \varepsilon$), with the extreme case corresponding to $\zeta_0 = 0$ (which still allow arbitrary probability for region $\{x \in \mathcal{X} : \eta(x) = 1/2\}$).

Ideally, any active learning algorithm should not be heavily affected by these noise conditions since it hardly matters (in terms of excess error) what label is predicted over region $\{x \in \mathcal{X} : |\eta(x) - 1/2| \le \zeta_0\}$. However, these seemingly benign noise-seeking conditions can cause troubles for any "uncertainty-based" active learner, i.e., any active learning algorithm that requests the label for any point that is uncertain (see Definition 10 in Appendix F for formal definition). In particular, under limited budget, we derive the following result.

Proposition 3. *Fix $\varepsilon, \delta, \gamma > 0$. For any labeling budget $B \gtrsim \frac{1}{\gamma^2} \cdot \log^2(\frac{1}{\varepsilon\,\gamma}) \cdot \log(\frac{1}{\varepsilon\,\gamma\,\delta})$, there exists a learning problem (with a set of linear regression functions) satisfying Definition 5/Definition 6 such that (1) any "uncertainty-based" active learner suffers expected standard excess error $\Omega(B^{-1})$; yet (2) with probability at least $1 - \delta$, Algorithm 1 returns a classifier with standard excess error at most ε.*

The above result demonstrates the superiority of our Algorithm 1 over any "uncertainty-based" active learner. Moreover, we show that Algorithm 1 achieves similar guarantees as in Theorem 6 under the strictly harder noise-seeking conditions. Specifically, we have the following guarantees.

Theorem 7. *With an appropriate choice of the abstention parameter γ in Algorithm 1 and randomization over the abstention region, Algorithm 1 learns a classifier $\check{h}$ with $\varepsilon + \zeta_0$ standard excess error after querying $\widetilde{\Theta}(\tau_0^{-2})$ labels under Definition 5 or querying $\widetilde{\Theta}(\varepsilon^{-2/(1+\beta)})$ labels under Definition 6.*

The special case of the noise-seeking condition with $\zeta_0 = 0$ is recently studied in (Kpotufe et al., 2021), where the authors conclude that no active learners can outperform the passive counterparts in the *nonparametric* regime. Theorem 7 shows that, in the *parametric* setting (with function approximation), Algorithm 1 provably overcomes these noise-seeking conditions.

4 Extensions

We provide two adaptations of our main algorithm (Algorithm 1) that can (1) achieve constant label complexity for a general set of regression functions (Section 4.1); and (2) adapt to model misspecification (Section 4.2). These two adaptations can also be efficiently implemented via regression oracle and enjoy similar guarantees stated in Theorem 5. We defer computational analysis to Appendix G and Appendix H.

4.1 Constant label complexity

We start by considering a simple problem instance with $\mathcal{X} = \{x\}$, where active learning is reduced to mean estimation of $\eta(x)$. Consider the Massart noise case where $\eta(x) \notin [\frac{1}{2} - \tau_0, \frac{1}{2} + \tau_0]$. No matter how small the desired accuracy level $\varepsilon > 0$ is, the learner should not spend more than $O(\frac{\log(1/\delta)}{\tau_0^2})$ labels to correctly classify x with probability at least $1 - \delta$, which ensures 0 excess error. In the general setting, but with Chow's excess error, a similar result follows: It takes at most $O(\frac{\log(1/\delta)}{\gamma^2})$ samples to verify if $\eta(x)$ is contained in $[\frac{1}{2} - \gamma, \frac{1}{2} + \gamma]$ or not. Taking the optimal action within $\{+1, -1, \perp\}$ (based on Eq. (2)) then leads to 0 Chow's excess error. This reasoning shows that, at least in simple cases, one should be able to achieve *constant* label complexity no matter how small ε is. One natural question to ask is as follows.

Is it possible to achieve constant label complexity in the general case of active learning?

We provide the first affirmative answer to the above question with a *general* set of regression function $\mathcal{F}$ (finite), and under *general* action space $\mathcal{X}$ and marginal distribution $\mathcal{D}_{\mathcal{X}}$. The positive result is achieved by Algorithm 2 (deferred to Appendix G.2), which differs from Algorithm 1 in two aspects: (1) we drop the epoch scheduling, and (2) apply a tighter elimination step derived from an optimal stopping theorem. Another change comes from the analysis of the algorithm: Instead of analyzing with respect to the disagreement coefficient, we work with the *eluder dimension* $\mathfrak{e} := \sup_{f^\star \in \mathcal{F}} \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2)$.[6] To do that, we analyze active learning from the perspective of *regret minimization with selective querying* (Dekel et al., 2012), which allows us to incorporate techniques developed in the field of contextual bandits (Russo and Van Roy, 2013; Foster et al., 2020). We defer a detailed discussion to Appendix G.1 and provide the following guarantees.

Theorem 8. *With probability at least $1 - 2\delta$, Algorithm 2 returns a classifier with expected Chow's excess error at most ε and label complexity $O(\frac{\mathfrak{e} \cdot \log(|\mathcal{F}|/\delta)}{\gamma^2})$, which is independent of $\frac{1}{\varepsilon}$.*

Based on discussion in Section 3, we can immediately translate the above results into *standard* excess error guarantees under the Massart noise (with γ replaced by τ_0). We next discuss why existing algorithms/analyses do not guarantee constant label complexity, even in the linear case.

1. **Epoch scheduling.** Many algorithms proceed in epochs and aim at *halving* the excess error after each epoch (Balcan et al., 2007; Zhang and Chaudhuri, 2014; Puchkin and Zhivotovskiy, 2021). One inevitably needs $\log \frac{1}{\varepsilon}$ epochs to achieve ε excess error.
2. **Relating to disagreement coefficient.** The algorithm presented in Krishnamurthy et al. (2019) does not use epoch scheduling. However, their label complexity are analyzed with disagreement coefficient, which incurs a $\sum_{t=1}^{1/\varepsilon} \frac{1}{t} = O(\log \frac{1}{\varepsilon})$ term in the label complexity.

Remark 3. *Algorithm 2 also provides guarantees when x is selected by an adaptive adversary (instead of i.i.d. sampled $x \sim \mathcal{D}_{\mathcal{X}}$). In that case, we simultaneously upper bound the regret and the label complexity (see Theorem 10 in Appendix G.2). Our results can be viewed as a generalization of the results developed in the linear case (Dekel et al., 2012).*

[6] We formally define eluder dimension in Appendix C. As examples, we have $\mathfrak{e} = O(d \cdot \log \frac{1}{\gamma})$ for linear functions in $\mathbb{R}^d$, and $\mathfrak{e} = O(C_{\mathsf{link}} \cdot d \log \frac{1}{\gamma})$ for generalized linear functions (where C_{link} is a quantity related to the link function).

4.2 Dealing with model misspecification

Our main results are developed under realizability (Assumption 1), which assumes that there exists a $f^\star \in \mathcal{F}$ such that $f^\star = \eta$. In this section, we relax that assumption and allow model misspecification. We assume the learner is given a set of regression function $\mathcal{F} : \mathcal{X} \to [0, 1]$ that may only *approximates* the conditional probability η. More specifically, we make the following assumption.

Assumption 2 (Model misspecification)**.** *There exists a $\bar{f} \in \mathcal{F}$ such that $\bar{f}$ approximate η up to $\kappa > 0$ accuracy, i.e.,* $\sup_{x \in \mathcal{X}} \left| \bar{f}(x) - \eta(x) \right| \leq \kappa$.

We use a variation of Algorithm 1 to adapt to model misspecification (Algorithm 3, deferred to Appendix H.1). Compared to Algorithm 1, the main change in Algorithm 3 is to apply a more conservative step in determining the active set $\mathcal{F}_m$ at each epoch: We maintain a larger active set of regression function to ensure that $\bar{f}$ is not eliminated throughout all epochs. Our algorithm proceeds *without* knowing the misspecification level κ. However, the excess error bound presented next holds under the condition that $\kappa \leq \varepsilon$ (i.e., it requires that the misspecification is no larger than the desired accuracy). Abbreviate $\bar{\theta} := \sup_{\iota > 0} \theta^{\mathrm{val}}_{\bar{f}}(\mathcal{F}, \gamma/2, \iota)$, we achieve the following guarantees.

Theorem 9. *Suppose $\kappa \leq \varepsilon$. With probability at least $1 - 2\delta$, Algorithm 3 returns a classifier with Chow's excess error $O(\varepsilon \cdot \bar{\theta} \cdot \log(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma\,\delta}))$ and label complexity $O(\frac{\bar{\theta}\,\mathrm{Pdim}(\mathcal{F})}{\gamma^2} \cdot \log^2(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma}) \cdot \log(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma\,\delta}))$.*

We only provide guarantee when $\kappa \leq \varepsilon$, since the learned classifier suffers from an additive κ term in the excess error (see Appendix H.2 for more discussion). On the other hand, the (inefficient) algorithm provided in Puchkin and Zhivotovskiy (2021) works without any assumption on the approximation error. An interesting future direction is to study the relation between computational efficiency and learning with *general* approximation error.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their helpful comments. This work is partially supported by NSF grant 1934612 and AFOSR grant FA9550-18-1-0166.

References

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

Martin Anthony. Uniform glivenko-cantelli theorems and concentration of measure in the mathematical modelling of learning. *Research Report LSE-CDAM-2002–07*, 2002.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

Alina Beygelzimer, Daniel J Hsu, John Langford, and Tong Zhang. Agnostic active learning without constraints. *Advances in neural information processing systems*, 23, 2010.

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

Rui M Castro and Robert D Nowak. Upper and lower error bounds for active learning. In *The 44th Annual Allerton Conference on Communication, Control and Computing*, volume 2, page 1, 2006.

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

CK Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Sanjoy Dasgupta, Adam Tauman Kalai, and Adam Tauman. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10(2), 2009.

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research*, 13(1):2655–2697, 2012.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th international conference on Machine learning*, pages 353–360, 2007.

Steve Hanneke. Theory of active learning. *Foundations and Trends in Machine Learning*, 7(2-3), 2014.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. 1989.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

Radu Herbei and Marten H Wegkamp. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 709–721, 2006.

Daniel Joseph Hsu. *Algorithms for active learning*. PhD thesis, UC San Diego, 2010.

Tzu-Kuo Huang, Alekh Agarwal, Daniel J Hsu, John Langford, and Robert E Schapire. Efficient and parsimonious agnostic active learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Matti Kääriäinen. Active learning in the non-realizable case. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2006.

Julian Katz-Samuels, Jifan Zhang, Lalit Jain, and Kevin Jamieson. Improved algorithms for agnostic pool-based active classification. In *International Conference on Machine Learning*, pages 5334–5344. PMLR, 2021.

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

Samory Kpotufe, Gan Yuan, and Yunfan Zhao. Nuances in margin conditions determine gains in active learning. *arXiv preprint arXiv:2110.08418*, 2021.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924. PMLR, 2017.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. *Journal of Machine Learning Research*, 20:1–50, 2019.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

Andrea Locatelli, Alexandra Carpentier, and Samory Kpotufe. Adaptivity to noise parameters in nonparametric active learning. In *Proceedings of the 2017 Conference on Learning Theory, PMLR*, 2017.

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

Stanislav Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13 (1), 2012.

D Pollard. *Convergence of Stochastic Processes*. David Pollard, 1984.

Nikita Puchkin and Nikita Zhivotovskiy. Exponential savings in agnostic active learning through abstention. *arXiv preprint arXiv:2102.00451*, 2021.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pages 2256–2264. Citeseer, 2013.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Burr Settles. Active learning literature survey. 2009.

Shubhanshu Shekhar, Mohammad Ghavamzadeh, and Tara Javidi. Active learning for classification with abstention. *IEEE Journal on Selected Areas in Information Theory*, 2(2):705–719, 2021.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Available at SSRN 3562765*, 2020.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

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

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Our results in the misspecification setting (i.e., in Section 4.2) are derived under an assumption. We provide discussion on such assumption in Appendix H.2, and leave a comprehensive study of the problem for future work.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our paper is theoretical in nature, and there is no negative societal impact of our work in the foreseeable future.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Assumptions are clearly stated in the statement of each theorem.
   (b) Did you include complete proofs of all theoretical results? [Yes] Complete proofs are provided in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional related work

Learning with Chow's excess error is closely related to learning under Massart noise (Massart and Nédélec, 2006), which assumes that no data point has conditional expectation close to the decision boundary, i.e., $\mathbb{P}(|\eta(x) - 1/2| \leq \tau_0) = 0$ for some constant $\tau_0 > 0$. Learning with Massart noise is commonly studied in active learning (Balcan et al., 2007; Hanneke, 2014; Zhang and Chaudhuri, 2014; Krishnamurthy et al., 2019), where $\widetilde{O}(\tau_0^{-2})$ type of guarantees are achieved. Instead of making explicit assumptions on the underlying distribution, learning with Chow's excess error empowers the learner with the ability to abstain: There is no need to make predictions on hard data points that are close to the decision boundary, i.e., $\{x : |\eta(x) - 1/2| \leq \gamma\}$. Learning with Chow's excess error thus works on more general settings and still enjoys the $\widetilde{O}(\gamma^{-2})$ type of guarantee as learning under Massart noise (Puchkin and Zhivotovskiy, 2021).[7] We show in Section 3 that statistical guarantees achieved under Chow's excess error can be directly translated to guarantees under (usual and more challenging versions of) Massart/Tsybakov noise (Massart and Nédélec, 2006; Tsybakov, 2004).

Active learning at aim competing the best in-class classifier with few labels. A long line of work directly works with the set of classifiers (Balcan et al., 2007; Hanneke, 2007, 2014; Huang et al., 2015; Puchkin and Zhivotovskiy, 2021), where the algorithms are developed with (in general) hard-to-implement ERM oracles (Guruswami and Raghavendra, 2009) and the the guarantees dependence on the so-called disagreement coefficient (Hanneke, 2014). More recently, learning with function approximation have been studied inactive learning and contextual bandits (Dekel et al., 2012; Agarwal et al., 2012; Foster et al., 2018; Krishnamurthy et al., 2019). The function approximation scheme permits efficient regression oracles, which solve convex optimization problems with respect to regression functions (Krishnamurthy et al., 2017, 2019; Foster et al., 2018). It can also be analyzed with the scale-sensitive version of disagreement coefficient, which is usually tighter than the original one (Foster et al., 2020; Russo and Van Roy, 2013). Our algorithms are inspired Krishnamurthy et al. (2019), where the authors study active learning under the standard excess error. The main deviation from Krishnamurthy et al. (2019) is that we need to *manually* construct a classifier $\widehat{h}$ with an abstention option and $\widehat{h} \notin \mathcal{H}$, which leads to differences in the analysis of excess error and label complexity. We borrow techniques developed in contextual bandits Russo and Van Roy (2013); Foster et al. (2020) to analyze our algorithm.

Although one can also apply our algorithms in the nonparametric regime with proper pre-processing schemes such discretizations, our algorithm primarily works in the parametric setting with finite pseudo dimension (Haussler, 1995) and finite (value function) disagreement coefficient (Foster et al., 2020). Active learning has also been studied in the nonparametric regime (Castro and Nowak, 2008; Koltchinskii, 2010; Minsker, 2012; Locatelli et al., 2017). Notably, Shekhar et al. (2021) studies Chow's excess error with margin-type of assumptions. Their setting is different to ours and $\mathrm{poly}(\frac{1}{\varepsilon})$ label complexities are achieved. If abundant amounts of data points are allowed to be exactly at the decision boundary, i.e., $\eta(x) = \frac{1}{2}$, Kpotufe et al. (2021) recently shows that, in the nonparametric regime, no active learner can outperform the passive counterpart.

B Why Chow's excess error helps?

For illustration purpose, we focus on the simple case where $\mathcal{X} = \{x\}$ in this section. The active learning problem is then reduced to mean estimation of the conditional probability $\eta(x) \in [0, 1]$.

Learning with standard excess error. Fix any $\varepsilon > 0$. With respect to the conditional probability $\eta(x)$, we define the positive region $\mathcal{S}_{+,\varepsilon} := \left[\frac{1-\varepsilon}{2}, 1\right]$ and the negative region $\mathcal{S}_{-,\varepsilon} := [0, \frac{1+\varepsilon}{2}]$. These regions are interpreted as follows: If $\eta(x) \in \mathcal{S}_{+,\varepsilon}$ (resp. $\eta(x) \in \mathcal{S}_{-,\varepsilon}$), then labeling x as positive (resp. negative) incurs no more than ε excess error. Under standard excess error, we define the flexible region as $\mathcal{S}^{\text{standard}}_{\text{flexible},\varepsilon} := \mathcal{S}_{+,\varepsilon} \cap \mathcal{S}_{-,\varepsilon} = [\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}]$ (the grey area in the top plot in Fig. 1). Two important implications of the flexible region are as follows: (1) if $\eta(x) \in \mathcal{S}^{\text{standard}}_{\text{flexible},\varepsilon}$, labeling x as either positive or negative would lead to excess error at most ε; and (2) if $\eta(x) \notin \mathcal{S}^{\text{standard}}_{\text{flexible},\varepsilon}$, then a classifier must correctly label x as either positive or negative to guarantee ε excess error. Since the

[7] However, passive learning with abstention only achieves error rate $\frac{1}{n\gamma}$ with n samples (Bousquet and Zhivotovskiy, 2021).

flexible region is of length ε under standard excess error, distinguishing two points at the edge of the flexible region, e.g., $\eta(x) = \frac{1}{2} - \varepsilon$ and $\eta(x) = \frac{1}{2} + \varepsilon$, leads to label complexity lower bound $\Omega(\frac{1}{\varepsilon^2})$.

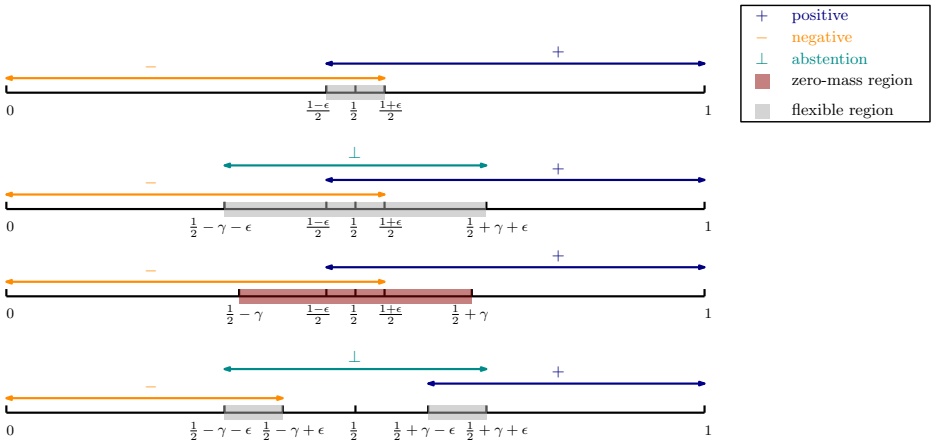

Figure 1: Illustration of regions $\mathcal{S}_{+,\varepsilon}$, $\mathcal{S}_{-,\varepsilon}$ and $\mathcal{S}_{\perp,\varepsilon}$. (1) top plot: learning with standard excess error $\mathrm{err}(\widehat{h}) - \mathrm{err}(h^\star)$; (2) second plot: learning with standard excess error $\mathrm{err}(\widehat{h}) - \mathrm{err}(h^\star)$ and Massart noise with parameter γ; (3) third plot: learning with Chow's excess error $\mathrm{err}_\gamma(\widehat{h}) - \mathrm{err}(h^\star)$; and (4) bottom plot: learning against the optimal Chow's excess error, i.e., $\mathrm{err}_\gamma(\widehat{h}) - \inf_{h:\mathcal{X}\to\{+1,-1,\perp\}} \mathrm{err}_\gamma(h)$.

Learning with Chow's excess error. Now we turn our attention to the Chow's error. We consider $\mathcal{S}_{+,\varepsilon}$ and $\mathcal{S}_{-,\varepsilon}$ as before, and define a third abstention region $\mathcal{S}_{\perp,\varepsilon} := [\frac{1}{2} - \gamma - \varepsilon, \frac{1}{2} + \gamma + \varepsilon]$ If $\eta(x) \in \mathcal{S}_{\perp,\varepsilon}$, then abstaining on x leads to Chow's excess error at most ε. We define the positive flexible region as $\mathcal{S}^{\mathrm{Chow}}_{\mathrm{flexible},+,\varepsilon} := \mathcal{S}_{\perp,\varepsilon} \cap \mathcal{S}_{\perp,\varepsilon} = [\frac{1-\varepsilon}{2}, \frac{1}{2} + \gamma + \varepsilon]$, which is of length $\gamma + \frac{3\varepsilon}{2}$. The negative flexible region is defined similarly (see the second plot in Fig. 1). Since both positive and negative flexible regions are longer than γ, it takes at most $\widetilde{O}(\frac{1}{\gamma^2})$ samples to identify a labeling action with at most ε Chow's excess error. Without loss of generality, we assume $\eta(x) < \frac{1}{2}$. It takes $\widetilde{O}(\frac{1}{\gamma^2})$ samples to construct a confidence interval $[\mathsf{lcb}(x), \mathsf{ucb}(x)]$ (of $\eta(x)$) of length at most $\frac{\gamma}{2}$. If $\eta(x) < \frac{1-\gamma}{2}$, we observe $\mathsf{ucb}(x) \le \frac{1}{2}$ and know that labeling x negative ensures at most ε excess error. If $\eta(x) \in [\frac{1-\gamma}{2}, \frac{1}{2}]$, we observe $[\mathsf{lcb}(x), \mathsf{ucb}(x)] \subseteq [\frac{1}{2} - \gamma, \frac{1}{2} + \gamma] \subseteq \mathcal{S}_{\perp,\varepsilon}$. Thus, choosing to abstain on x again ensures at most ε excess error. To summarize, learning with Chow's excess error in the general case resembles the behavior of learning under Massart noise: Hard examples that are close to the decision boundary are not selected for labeling (see the third plot in Fig. 1), and learning is possible with $\widetilde{O}(\frac{1}{\gamma^2})$ labels.

Why not compete against the optimal Chow's error. As shown in the last plot in Fig. 1, the flexible regions become narrow (of length $O(\varepsilon)$) again when competing against the optimal Chow's error. Abstention becomes the only action that guarantees at most ε excess error over region $(\frac{1}{2} - \gamma - \varepsilon, \frac{1}{2} + \gamma + \varepsilon)$. One then needs to distinguish cases $\eta(x) = \frac{1}{2} + \gamma - 2\varepsilon$ and $\eta(x) = \frac{1}{2} + \gamma + 2\varepsilon$, which requires $\Omega(\frac{1}{\varepsilon^2})$ samples. It is also unreasonable to compete against the optimal Chow's error: When $\eta(x) = \frac{1}{2} + \gamma - 2\varepsilon$, $\Omega(\frac{1}{\varepsilon^2})$ samples are required to decide whether take action $+1$ or action $\perp$; however, with only $\widetilde{O}(\frac{1}{\gamma^2})$ samples, one can already guarantee $\eta(x) > \frac{1}{2}$ (and take the positive label in the usual case).

C Disagreement coefficient, star number and eluder dimension

We provide formal definitions/guarantees of value function disagreement coefficient, eluder dimension and star number in this section. These results are developed in Foster et al. (2020); Russo and Van Roy (2013). Since our guarantees are developed in terms of these complexity measures, any future developments on these complexity measures (e.g., with respect to richer function classes) directly lead to broader applications of our algorithms.

We first state known upper bound on value function disagreement coefficient with respect to nice sets of regression functions.

Proposition 4 (Foster et al. (2020)). *For any $f^\star \in \mathcal{F}$ and $\gamma, \varepsilon > 0$, let $\theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma, \varepsilon)$ be the value function disagreement coefficient defined in Definition 1. Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be a fixed feature mapping and $\mathcal{W} \subseteq \mathbb{R}^d$ be a fixed set. The following upper bounds hold true.*

- *Suppose $\mathcal{F} := \{x \mapsto \langle \phi(x), w\rangle : w \in \mathcal{W}\}$ is a set of linear functions. We then have $\sup_{f\in\mathcal{F},\gamma>0,\varepsilon>0} \theta^{\mathrm{val}}_f(\mathcal{F}, \gamma, \varepsilon) \le d$.*
- *Suppose $\mathcal{F} := \{x \mapsto \sigma(\langle \phi(x), w\rangle) : w \in \mathcal{W}\}$ is a set of generalized linear functions with any fixed link function $\sigma : \mathbb{R} \to \mathbb{R}$ such that $0 < c_l < \sigma' \le c_u$. We then have $\sup_{f\in\mathcal{F},\gamma>0,\varepsilon>0} \theta^{\mathrm{val}}_f(\mathcal{F}, \gamma, \varepsilon) \le \frac{c_u}{c_l}^2 \cdot d$.*

We next provide the formal definition of value function eluder dimension and star number (Foster et al., 2020; Russo and Van Roy, 2013).

Definition 7 (Value function eluder dimension). *For any $f^\star \in \mathcal{F}$ and $\gamma > 0$, let $\check{\mathfrak{e}}_{f^\star}(\mathcal{F}, \gamma)$ be the length of the longest sequence of data points $x^1, \ldots, x^m$ such that for all i, there exists $f^i \in \mathcal{F}$ such that*

$$|f^i(x^i) - f^\star(x^i)| > \gamma, \quad \text{and} \quad \sum_{j<i} (f^i(x^j) - f^\star(x^j))^2 \le \gamma^2.$$

The value function eluder dimension is defined as $\mathfrak{e}_{f^\star}(\mathcal{F}, \gamma_0) := \sup_{\gamma\ge\gamma_0} \check{\mathfrak{e}}_{f^\star}(\mathcal{F}, \gamma)$.

Definition 8 (Value function star number). *For any $f^\star \in \mathcal{F}$ and $\gamma > 0$, let $\check{\mathfrak{s}}_{f^\star}(\mathcal{F}, \gamma)$ be the length of the longest sequence of data points $x^1, \ldots, x^m$ such that for all i, there exists $f^i \in \mathcal{F}$ such that*

$$|f^i(x^i) - f^\star(x^i)| > \gamma, \quad \text{and} \quad \sum_{j\neq i} (f^i(x^j) - f^\star(x^j))^2 \le \gamma^2.$$

The value function eluder dimension is defined as $\mathfrak{s}_{f^\star}(\mathcal{F}, \gamma_0) := \sup_{\gamma\ge\gamma_0} \check{\mathfrak{s}}_{f^\star}(\mathcal{F}, \gamma)$.

Since the second constrain in the definition of star number is more stringent than the counterpart in the definition of eluder dimension, one immediately have that $\mathfrak{s}_{f^\star}(\mathcal{F}, \gamma) \le \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma)$. We provide known upper bounds for eluder dimension next.

Proposition 5 (Russo and Van Roy (2013)). *Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be a fixed feature mapping and $\mathcal{W} \subseteq \mathbb{R}^d$ be a fixed set. Suppose $\sup_{x\in\mathcal{X}} \|\phi(x)\|_2 \le 1$ and $\sup_{w\in\mathcal{W}} \|w\|_2 \le 1$. The following upper bounds hold true.*

- *Suppose $\mathcal{F} := \{x \mapsto \langle \phi(x), w\rangle : w \in \mathcal{W}\}$ is a set of linear functions. We then have $\sup_{f^\star\in\mathcal{F}} \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma) = O(d \log \frac{1}{\gamma})$.*
- *Suppose $\mathcal{F} := \{x \mapsto \sigma(\langle \phi(x), w\rangle) : w \in \mathcal{W}\}$ is a set of generalized linear functions with any fixed link function $\sigma : \mathbb{R} \to \mathbb{R}$ such that $0 < c_l < \sigma' \le c_u$. We then have $\sup_{f^\star\in\mathcal{F}} \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma) = O\big(\big(\frac{c_u}{c_l}\big)^2 d \log\big(\frac{c_u}{\gamma}\big)\big)$.*

The next result shows that the disagreement coefficient (with our Definition 1) can be always upper bounded by (squared) star number and eluder dimension.

Proposition 6 (Foster et al. (2020)). *Suppose $\mathcal{F}$ is a uniform Glivenko-Cantelli class.*

For any $f^\star : \mathcal{X} \to [0, 1]$ and $\gamma, \varepsilon > 0$, we have $\theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma, \varepsilon) \le 4(\mathfrak{s}_{f^\star}(\mathcal{F}, \gamma))^2$, and $\theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma, \varepsilon) \le 4\,\mathfrak{e}_{f^\star}(\mathcal{F}, \gamma)$.

The requirement that $\mathcal{F}$ is a uniform Glivenko-Cantelli class is rather weak: It is satisfied as long as $\mathcal{F}$ has finite Pseudo dimension (Anthony, 2002).

In our analysis, we sometimes work with sub probability measure (due to selective sampling). Our next result shows that defining the disagreement coefficient over all (sub) probability measures will not affect its value. More specifically, denote $\widetilde{\theta}^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma, \varepsilon)$ be the disagreement coefficient defined in Definition 1, but with sup taking over all probability and sub probability measures. We then have the following equivalence.

Proposition 7. *Fix any $\gamma_0, \varepsilon_0 \geq 0$. We have $\widetilde{\theta}^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma_0, \varepsilon_0) = \theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma_0, \varepsilon_0)$.*

Proof. We clearly have $\widetilde{\theta}^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma_0, \varepsilon_0) \geq \theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma_0, \varepsilon_0)$ by additionally considering sub probability measures. We next show the opposite direction.

Fix any sub probability measure $\widetilde{\mathcal{D}}_{\mathcal{X}}$ that is non-zero (otherwise we have $\mathbb{P}_{x\sim\widetilde{\mathcal{D}}_{\mathcal{X}}}(\cdot) = 0$). Suppose $\mathbb{E}_{x\sim\widetilde{\mathcal{D}}_{\mathcal{X}}}[1] = \kappa < 1$. We can now consider its normalized probability measure $\overline{\mathcal{D}}_{\mathcal{X}}$ such that $\overline{\mathcal{D}}_{\mathcal{X}}(\omega) = \frac{\widetilde{\mathcal{D}}_{\mathcal{X}}(\omega)}{\kappa}$ (for any ω in the sigma algebra). Now fix any $\gamma > \gamma_0$ and $\varepsilon > \varepsilon_0$. We have

$$
\begin{aligned}
&\frac{\gamma^2}{\varepsilon^2} \cdot \mathbb{P}_{\widetilde{\mathcal{D}}_{\mathcal{X}}}\left(\exists f \in \mathcal{F} : |f(x) - f^\star(x)| > \gamma, \|f - f^\star\|^2_{\widetilde{\mathcal{D}}_{\mathcal{X}}} \leq \varepsilon^2\right) \\
&= \frac{\gamma^2}{\varepsilon^2/\kappa} \cdot \mathbb{P}_{\overline{\mathcal{D}}_{\mathcal{X}}}\left(\exists f \in \mathcal{F} : |f(x) - f^\star(x)| > \gamma, \|f - f^\star\|^2_{\overline{\mathcal{D}}_{\mathcal{X}}} \leq \varepsilon^2/\kappa\right) \\
&= \frac{\gamma^2}{\bar{\varepsilon}^2} \cdot \mathbb{P}_{\overline{\mathcal{D}}_{\mathcal{X}}}\left(\exists f \in \mathcal{F} : |f(x) - f^\star(x)| > \gamma, \|f - f^\star\|^2_{\overline{\mathcal{D}}_{\mathcal{X}}} \leq \bar{\varepsilon}^2\right) \\
&\leq \theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma_0, \varepsilon_0),
\end{aligned}
$$

where we denote $\bar{\varepsilon} := \frac{\varepsilon}{\sqrt{\kappa}} > \varepsilon$, and the last follows from the fact that $\overline{\mathcal{D}}_{\mathcal{X}}$ is a probability measure. We then have $\widetilde{\theta}^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma_0, \varepsilon_0) \leq \theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma_0, \varepsilon_0)$, and thus the desired result. □

D Concentration results

The Freedman's inequality is quite common in the field of active learning and contextual bandits, e.g., (Freedman, 1975; Agarwal et al., 2014; Krishnamurthy et al., 2019; Foster et al., 2020). We thus state the result without proof.

Lemma 1 (Freedman's inequality)**.** *Let $(Z_t)_{t\leq T}$ be a real-valued martingale difference sequence adapted to a filtration $\mathfrak{F}_t$, and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathfrak{F}_{t-1}]$. If $|Z_t| \leq B$ almost surely, then for any $\eta \in (0, 1/B)$ it holds with probability at least $1-\delta$,*

$$
\sum_{t=1}^{T} Z_t \leq \eta \sum_{t=1}^{T} \mathbb{E}_t[Z_t^2] + \frac{\log \delta^{-1}}{\eta}.
$$

Lemma 2. *Let $(Z_t)_{t\leq T}$ to be real-valued sequence of random variables adapted to a filtration $\mathfrak{F}_t$. If $|Z_t| \leq B$ almost surely, then with probability at least $1-\delta$,*

$$
\sum_{t=1}^{T} Z_t \leq \frac{3}{2} \sum_{t=1}^{T} \mathbb{E}_t[Z_t] + 4B \log(2\delta^{-1}),
$$

and

$$
\sum_{t=1}^{T} \mathbb{E}_t[Z_t] \leq 2 \sum_{t=1}^{T} Z_t + 8B \log(2\delta^{-1}).
$$

Proof. This is a direct consequence of Lemma 1. □

We define/recall some notations first. Fix any epoch $m \in [M]$ and any time step t within epoch m. For any $f \in \mathcal{F}$, we denote $M_t(f) := Q_t((f(x_t) - y_t)^2 - (f^\star(x_t) - y_t)^2)$, and

$\widehat{R}_m(f) := \sum_{t=1}^{\tau_{m-1}} Q_t(f(x_t) - y_t)^2$. Recall that we have $Q_t = g_m(x_t)$. We define filtration $\mathfrak{F}_t := \sigma((x_1, y_1), \ldots, (x_t, y_t))$,[8] and denote $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathfrak{F}_{t-1}]$.

We first provide a simple concentration result with respect to a finite $\mathcal{F}$.

Lemma 3. *Suppose $\mathcal{F}$ is finite. Fix any $\delta \in (0,1)$. For any $\tau, \tau' \in [T]$ such that $\tau < \tau'$, with probability at least $1-\delta$, we have*

$$\sum_{t=\tau}^{\tau'} M_t(f) \le \sum_{t=\tau}^{\tau'} \frac{3}{2}\mathbb{E}_t[M_t(f)] + C_\delta(\mathcal{F}),$$

and

$$\sum_{t=\tau}^{\tau'} \mathbb{E}_t[M_t(f)] \le 2\sum_{t=\tau}^{\tau'} M_t(f) + C_\delta(\mathcal{F}),$$

where $C_\delta(\mathcal{F}) = 8\log\Big(\frac{|\mathcal{F}|\cdot T^2}{\delta}\Big)$.

Proof. We first notice that $M_t(f)$ adapts to filtration $\mathfrak{F}_t$, and satisfies $|M_t(f)| \le 1$. The results follow by taking Lemma 2 together with a union bound over $f \in \mathcal{F}$ and $\tau, \tau' \in [T]$. □

Although one can not directly apply a union bound as in Lemma 3 in the case when the set of regression function $\mathcal{F}$ is infinite (but has finite Pseudo dimension by assumption), it turns out that similar guarantees as in Lemma 3 can be derived. We first recall the formal definition of the Pseudo dimension of $\mathcal{F}$.

Definition 9 (Pseudo Dimension, Pollard (1984); Haussler (1989, 1995))**.** *Consider a set of real-valued function $\mathcal{F} : \mathcal{X} \to \mathbb{R}$. The pseudo-dimension $\mathrm{Pdim}(\mathcal{F})$ of $\mathcal{F}$ is defined as the VC dimension of the set of threshold functions $\{(x, \zeta) \mapsto \mathbb{1}(f(x) > \zeta) : f \in \mathcal{F}\}$.*

Lemma 4 (Krishnamurthy et al. (2019))**.** *Suppose $\mathrm{Pdim}(\mathcal{F}) < \infty$. Fix any $\delta \in (0,1)$. For any $\tau, \tau' \in [T]$ such that $\tau < \tau'$, with probability at least $1-\delta$, we have*

$$\sum_{t=\tau}^{\tau'} M_t(f) \le \sum_{t=\tau}^{\tau'} \frac{3}{2}\mathbb{E}_t[M_t(f)] + C_\delta(\mathcal{F}),$$

and

$$\sum_{t=\tau}^{\tau'} \mathbb{E}_t[M_t(f)] \le 2\sum_{t=\tau}^{\tau'} M_t(f) + C_\delta(\mathcal{F}),$$

where $C_\delta(\mathcal{F}) = C \cdot \Big(\mathrm{Pdim}(\mathcal{F}) \cdot \log T + \log\Big(\frac{\mathrm{Pdim}(\mathcal{F})\cdot T}{\delta}\Big)\Big) \le C' \cdot \big(\mathrm{Pdim}(\mathcal{F}) \cdot \log\big(\frac{T}{\delta}\big)\big)$, where $C, C' > 0$ are universal constants.

We will be primarily using Lemma 4 in the following. However, one can replace Lemma 4 with Lemma 3 to derive results with respect to a finite set of regressions $\mathcal{F}$.

E Proofs of results in Section 2

We give the proof of Theorem 4 and Theorem 5. Supporting lemmas used in the proofs are deferred to Appendix E.1.

Fix any classifier $\widehat{h} : \mathcal{X} \to \{+1, -1, \perp\}$. For any $x \in \mathcal{X}$, we introduce the notion

$$\begin{aligned}
&\mathsf{excess}_\gamma(\widehat{h}; x) := \\
&\mathbb{P}_{y|x}\big(y \ne \mathrm{sign}(\widehat{h}(x))\big) \cdot \mathbb{1}\big(\widehat{h}(x) \ne \perp\big) + \big(1/2 - \gamma\big) \cdot \mathbb{1}\big(\widehat{h}(x) = \perp\big) - \mathbb{P}_{y|x}\big(y \ne \mathrm{sign}(h^\star(x))\big) \\
&= \mathbb{1}\big(\widehat{h}(x) \ne \perp\big) \cdot \big(\mathbb{P}_{y|x}\big(y \ne \mathrm{sign}(\widehat{h}(x))\big) - \mathbb{P}_{y|x}\big(y \ne \mathrm{sign}(h^\star(x))\big)\big) \\
&\quad + \mathbb{1}\big(\widehat{h}(x) = \perp\big) \cdot \big(\big(1/2 - \gamma\big) - \mathbb{P}_{y|x}\big(y \ne \mathrm{sign}(h^\star(x))\big)\big)
\end{aligned} \tag{6}$$

[8] y_t is not observed (and thus not included in the filtration) when $Q_t = 0$. Note that Q_t is measurable with respect to $\sigma((\mathfrak{F}_{t-1}, x_t))$.

to represent the excess error of $\widehat{h}$ at point $x \in \mathcal{X}$. Excess error of classifier $\widehat{h}$ can be then written as $\mathsf{excess}_\gamma(\widehat{h}) := \mathrm{err}_\gamma(\widehat{h}) - \mathrm{err}(h^\star) = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{excess}_\gamma(\widehat{h}; x)]$.

Theorem 4. *With probability at least $1-2\delta$, Algorithm 1 returns a classifier with Chow's excess error at most ε and label complexity $O(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\gamma^2}\cdot\log^2(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma})\cdot\log(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma\,\delta}))$.*

Proof. We analyze under the good event $\mathcal{E}$ defined in Lemma 4, which holds with probability at least $1-\delta$. Note that all supporting lemmas stated in Appendix E.1 hold true under this event.

We analyze the Chow's excess error of $\widehat{h}_m$, which is measurable with respect to $\mathfrak{F}_{\tau_{m-1}}$. For any $x \in \mathcal{X}$, if $g_m(x) = 0$, Lemma 9 implies that $\mathsf{excess}_\gamma(\widehat{h}_m; x) \le 0$. If $g_m(x) = 1$, we know that $\widehat{h}_m(x) \neq \perp$ and $\frac{1}{2} \in (\mathsf{lcb}(x;\mathcal{F}_m), \mathsf{ucb}(x;\mathcal{F}_m))$. Note that $\widehat{h}_m(x) \neq h^\star(x)$ only if $\mathrm{sign}(2f^\star(x)-1)\cdot\mathrm{sign}(2\widehat{f}_m(x)-1) \le 0$. Since $f^\star, \widehat{f}_m \in \mathcal{F}_m$ by Lemma 5. The error incurred in this case can be upper bounded by $2|f^\star(x)-1/2| \le 2w(x;\mathcal{F}_m)$, which results in $\mathsf{excess}_\gamma(\widehat{h}_m; x) \le 2w(x;\mathcal{F}_m)$. Combining these two cases together, we have

$$\mathsf{excess}_\gamma(\widehat{h}_m) \le 2\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x)=1)\cdot w(x;\mathcal{F}_m)].$$

Take $m = M$ and apply Lemma 8, with notation $\rho_m := 2\beta_m + C_\delta$, leads to the following guarantee.

$$\begin{aligned}\mathsf{excess}_\gamma(\widehat{h}_M) &\le \frac{8\rho_M}{\tau_{M-1}\gamma}\cdot\theta^{\mathrm{val}}_{f^\star}\Big(\mathcal{F},\gamma/2,\sqrt{\rho_M/2\tau_{M-1}}\Big)\\ &= O\bigg(\frac{\mathrm{Pdim}(\mathcal{F})\cdot\log(T/\delta)}{T\,\gamma}\cdot\theta^{\mathrm{val}}_{f^\star}\Big(\mathcal{F},\gamma/2,\sqrt{C_\delta/T}\Big)\bigg),\end{aligned}$$

where we use the fact that $\frac{T}{2} \le \tau_{M-1} \le T$ and definitions of β_m and C_δ. Simply considering $\theta := \sup_{f^\star\in\mathcal{F},\iota>0}\theta^{\mathrm{val}}_{f^\star}(\mathcal{F},\gamma/2,\iota)$ as an upper bound of $\theta^{\mathrm{val}}_{f^\star}(\mathcal{F},\gamma/2,\sqrt{C_\delta/T})$ and taking

$$T = O\bigg(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma}\cdot\log\Big(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma\,\delta}\Big)\bigg)$$

ensures that $\mathsf{excess}_\gamma(\widehat{h}_M) \le \varepsilon$.

We now analyze the label complexity (note that the sampling process of Algorithm 1 stops at time $t = \tau_{M-1}$). Note that $\mathbb{E}[\mathbb{1}(Q_t=1)\mid\mathfrak{F}_{t-1}] = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x)=1)]$ for any epoch $m \ge 2$ and time step t within epoch m. Combine Lemma 2 with Lemma 7 leads to

$$\begin{aligned}\sum_{t=1}^{\tau_{M-1}}\mathbb{1}(Q_t=1) &\le \frac{3}{2}\sum_{t=1}^{\tau_{M-1}}\mathbb{E}[\mathbb{1}(Q_t=1)\mid\mathfrak{F}_{t-1}] + 4\log\delta^{-1}\\ &\le 3 + \frac{3}{2}\sum_{m=2}^{M-1}\frac{(\tau_m-\tau_{m-1})\cdot 4\rho_m}{\tau_{m-1}\gamma^2}\cdot\theta^{\mathrm{val}}_{f^\star}\Big(\mathcal{F},\gamma/2,\sqrt{\rho_m/2\tau_{m-1}}\Big) + 4\log\delta^{-1}\\ &\le 3 + 6\sum_{m=2}^{M-1}\frac{\rho_m}{\gamma^2}\cdot\theta^{\mathrm{val}}_{f^\star}\Big(\mathcal{F},\gamma/2,\sqrt{\rho_m/2\tau_{m-1}}\Big) + 4\log\delta^{-1}\\ &\le 3 + 4\log\delta^{-1} + \frac{18\log T\cdot M\cdot C_\delta}{\gamma^2}\cdot\theta^{\mathrm{val}}_{f^\star}\Big(\mathcal{F},\gamma/2,\sqrt{C_\delta/T}\Big)\\ &= O\bigg(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\gamma^2}\cdot\Big(\log\Big(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma}\Big)\Big)^2\cdot\log\Big(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma\,\delta}\Big)\bigg),\end{aligned}$$

with probability at least $1-2\delta$ (due to an additional application of Lemma 2); where we plug the above choice of T and upper bound other terms as before. □

A slightly different guarantee for Algorithm 1. The stated Algorithm 1 takes $\theta := \sup_{f^\star\in\mathcal{F},\iota>0}\theta^{\mathrm{val}}_{f^\star}(\mathcal{F},\gamma/2,\iota)$ as an input (the value of θ can be upper bounded for many function class $\mathcal{F}$, as discussed in Appendix C). However, we don't necessarily need to take θ as an input to the algorithm. Indeed, we can simply run a modified version of Algorithm 1 with $T = \frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma}$.

Following similar analyses in proof of Theorem 4, set $\iota := \sqrt{C_\delta/T} \propto \sqrt{\gamma\varepsilon}$, the modified version achieves excess error

$$\mathsf{excess}_\gamma(\widehat{h}_M) = O\left(\varepsilon \cdot \theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma/2, \iota) \cdot \log\left(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\delta\,\gamma}\right)\right)$$

with label complexity

$$O\left(\frac{\theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma/2, \iota) \cdot \mathrm{Pdim}(\mathcal{F})}{\gamma^2} \cdot \left(\log\left(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma}\right)\right)^2 \cdot \log\left(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma\,\delta}\right)\right).$$

We now discuss the efficient implementation of Algorithm 1 and its computational complexity. We first state some known results in computing the confidence intervals with respect to a set of regression functions $\mathcal{F}$.

Proposition 8 (Krishnamurthy et al. (2017); Foster et al. (2018, 2020))**.** *Consider the setting studied in Algorithm 1. Fix any epoch $m \in [M]$ and denote $\mathcal{B}_m := \{(x_t, Q_t, y_t)\}_{t=1}^{\tau_{m-1}}$. Fix any $\alpha > 0$. For any data point $x \in \mathcal{X}$, there exists algorithms $\mathbf{Alg}_{\mathsf{lcb}}$ and $\mathbf{Alg}_{\mathsf{ucb}}$ that certify*

$$\mathsf{lcb}(x; \mathcal{F}_m) - \alpha \le \mathbf{Alg}_{\mathsf{lcb}}(x; \mathcal{B}_m, \beta_m, \alpha) \le \mathsf{lcb}(x; \mathcal{F}_m) \quad \text{and}$$
$$\mathsf{ucb}(x; \mathcal{F}_m) \le \mathbf{Alg}_{\mathsf{ucb}}(x; \mathcal{B}_m, \beta_m, \alpha) \le \mathsf{ucb}(x; \mathcal{F}_m) + \alpha.$$

The algorithms take $O(\frac{1}{\alpha^2}\log\frac{1}{\alpha})$ calls of the regression oracle for general $\mathcal{F}$ and take $O(\log\frac{1}{\alpha})$ calls of the regression oracle if $\mathcal{F}$ is convex and closed under pointwise convergence.

Proof. See Algorithm 2 in Krishnamurthy et al. (2017) for the general case; and Algorithm 3 in Foster et al. (2018) for the case when $\mathcal{F}$ is convex and closed under pointwise convergence. □

We next discuss the computational efficiency of Algorithm 1. Recall that we redefine $\theta := \sup_{f^\star \in \mathcal{F}, \iota > 0} \theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma/4, \iota)$ in the Theorem 5 to account to approximation error.

Theorem 5. *Algorithm 1 can be efficiently implemented via the regression oracle and enjoys the same theoretical guarantees stated in Theorem 4. The number of oracle calls needed is $\widetilde{O}(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma^3})$ for a general set of regression functions $\mathcal{F}$, and $\widetilde{O}(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma})$ when $\mathcal{F}$ is convex and closed under pointwise convergence. The per-example inference time of the learned $\widehat{h}_M$ is $\widetilde{O}(\frac{1}{\gamma^2}\log^2(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon}))$ for general $\mathcal{F}$, and $\widetilde{O}(\log\frac{1}{\gamma})$ when $\mathcal{F}$ is convex and closed under pointwise convergence.*

Proof. Fix any epoch $m \in [M]$. Denote $\bar{\alpha} := \frac{\gamma}{4M}$ and $\alpha_m := \frac{(M-m)\gamma}{4M}$. With any observed $x \in \mathcal{X}$, we construct the approximated confidence intervals $\widehat{\mathsf{lcb}}(x; \mathcal{F}_m)$ and $\widehat{\mathsf{ucb}}(x; \mathcal{F}_m)$ as follows.

$$\widehat{\mathsf{lcb}}(x; \mathcal{F}_m) := \mathbf{Alg}_{\mathsf{lcb}}(x; \mathcal{B}_m, \beta_m, \bar{\alpha}) - \alpha_m \quad \text{and} \quad \widehat{\mathsf{ucb}}(x; \mathcal{F}_m) := \mathbf{Alg}_{\mathsf{ucb}}(x; \mathcal{B}_m, \beta_m, \bar{\alpha}) + \alpha_m.$$

For efficient implementation of Algorithm 1, we replace $\mathsf{lcb}(x; \mathcal{F}_m)$ and $\mathsf{ucb}(x; \mathcal{F}_m)$ with $\widehat{\mathsf{lcb}}(x; \mathcal{F}_m)$ and $\widehat{\mathsf{ucb}}(x; \mathcal{F}_m)$ in the construction of $\widehat{h}_m$ and g_m.

Based on Proposition 8, we know that

$$\mathsf{lcb}(x; \mathcal{F}_m) - \alpha_m - \bar{\alpha} \le \widehat{\mathsf{lcb}}(x; \mathcal{F}_m) \le \mathsf{lcb}(x; \mathcal{F}_m) - \alpha_m \quad \text{and}$$
$$\mathsf{ucb}(x; \mathcal{F}_m) + \alpha_m \le \widehat{\mathsf{ucb}}(x; \mathcal{F}_m) \le \mathsf{ucb}(x; \mathcal{F}_m) + \alpha_m + \bar{\alpha}.$$

Since $\alpha_m + \bar{\alpha} \le \frac{\gamma}{4}$ for any $m \in [M]$, the guarantee in Lemma 6 can be modified as $g_m(x) = 1 \implies w(x; \mathcal{F}_m) \ge \frac{\gamma}{2}$.

Fix any $m \ge 2$. Since $\mathcal{F}_m \subseteq \mathcal{F}_{m-1}$ by Lemma 5, we have

$$\widehat{\mathsf{lcb}}(x; \mathcal{F}_m) \ge \mathsf{lcb}(x; \mathcal{F}_m) - \alpha_m - \bar{\alpha} \ge \mathsf{lcb}(x; \mathcal{F}_{m-1}) - \alpha_{m-1} \ge \widehat{\mathsf{lcb}}(x; \mathcal{F}_{m-1}) \quad \text{and}$$
$$\widehat{\mathsf{ucb}}(x; \mathcal{F}_m) \le \mathsf{ucb}(x; \mathcal{F}_m) + \alpha_m + \bar{\alpha} \le \mathsf{ucb}(x; \mathcal{F}_{m-1}) + \alpha_{m-1} \le \widehat{\mathsf{ucb}}(x; \mathcal{F}_{m-1}).$$

These ensure $\mathbb{1}(g_m(x) = 1) \le \mathbb{1}(g_{m-1}(x) = 1)$. Thus, the guarantees stated in Lemma 7 and Lemma 8 still hold (with $\frac{\gamma}{2}$ replaced by $\frac{\gamma}{4}$ due to modification of Lemma 6). The guarantee stated in

Lemma 9 also hold since $\widehat{\mathsf{lcb}}(x;\mathcal{F}_m) \le \mathsf{lcb}(x;\mathcal{F}_m)$ and $\widehat{\mathsf{ucb}}(x;\mathcal{F}_m) \ge \mathsf{ucb}(x;\mathcal{F}_m)$ by construction. As a result, the guarantees stated in Theorem 4 hold true with changes only in constant terms.

We now discuss the computational complexity of the efficient implementation. At the beginning of each epoch m. We use one oracle call to compute $\widehat{f}_m = \arg\min_{f\in\mathcal{F}} \sum_{t=1}^{\tau_{m-1}} Q_t(f(x_t)-y_t)^2$. The main computational cost comes from computing $\widehat{\mathsf{lcb}}$ and $\widehat{\mathsf{ucb}}$ at each time step. We take $\alpha = \overline{\alpha} := \frac{\gamma}{4M}$ into Proposition 8, which leads to $O(\frac{(\log T)^2}{\gamma^2}\cdot\log(\frac{\log T}{\gamma}))$ calls of the regression oracle for general $\mathcal{F}$ and $O(\log(\frac{\log T}{\gamma}))$ calls of the regression oracle for any convex $\mathcal{F}$ that is closed under pointwise convergence. This also serves as the per-example inference time for $\widehat{h}_M$. The total computational cost of Algorithm 1 is then derived by multiplying the per-round cost by T and plugging $T = \widetilde{O}(\frac{\theta\,\mathrm{Pdim}(\mathcal{F})}{\varepsilon\gamma})$ into the bound (for any parameter, we only keep poly factors in the total computational cost and keep poly or polylog dependence in the per-example computational cost). □

E.1 Supporting lemmas

We use $\mathcal{E}$ to denote the good event considered in Lemma 4, and analyze under this event in this section. We abbreviate $C_\delta := C_\delta(\mathcal{F})$ in the following analysis.

Lemma 5. *The followings hold true:*

1. $f^\star \in \mathcal{F}_m$ *for any* $m \in [M]$.
2. $\sum_{t=1}^{\tau_{m-1}} \mathbb{E}_t[M_t(f)] \le 2\beta_m + C_\delta$ *for any* $f \in \mathcal{F}_m$.
3. $\mathcal{F}_{m+1} \subseteq \mathcal{F}_m$ *for any* $m \in [M-1]$.

Proof. 1. Fix any epoch $m \in [M]$ and time step t within epoch m. Since $\mathbb{E}[y_t] = f^\star(x_t)$, we have $\mathbb{E}_t[M_t(f)] = \mathbb{E}[Q_t(f(x)-f^\star(x))^2] = \mathbb{E}[g_m(x)(f(x)-f^\star(x))^2] \ge 0$ for any $f \in \mathcal{F}$. By Lemma 4, we then have $\widehat{R}_m(f^\star) \le \widehat{R}_m(f) + C_\delta/2 \le \widehat{R}_m(f) + \beta_m$ for any $f \in \mathcal{F}$. The elimination rule in Algorithm 2 then implies that $f^\star \in \mathcal{F}_m$ for any $m \in [M]$.

2. Fix any $f \in \mathcal{F}_m$. With Lemma 4, we have

$$\begin{aligned}\sum_{t=1}^{\tau_{m-1}} \mathbb{E}_t[M_t(f)] &\le 2\sum_{t=1}^{\tau_{m-1}} M_t(f) + C_\delta \\ &= 2\widehat{R}_m(f) - 2\widehat{R}_m(f^\star) + C_\delta \\ &\le 2\widehat{R}_m(f) - 2\widehat{R}_m(\widehat{f}_m) + C_\delta \\ &\le 2\beta_m + C_\delta,\end{aligned}$$

where the third line comes from the fact that $\widehat{f}_m$ is the minimizer of $\widehat{R}_m(\cdot)$; and the last line comes from the fact that $f \in \mathcal{F}_m$.

3. Fix any $f \in \mathcal{F}_{m+1}$. We have

$$\begin{aligned}\widehat{R}_m(f) - \widehat{R}_m(\widehat{f}_m) &\le \widehat{R}_m(f) - \widehat{R}_m(f^\star) + \frac{C_\delta}{2} \\ &= \widehat{R}_{m+1}(f) - \widehat{R}_{m+1}(f^\star) - \sum_{t=\tau_{m-1}+1}^{\tau_m} M_t(f) + \frac{C_\delta}{2} \\ &\le \widehat{R}_{m+1}(f) - \widehat{R}_{m+1}(\widehat{f}_{m+1}) - \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{E}_t[M_t(f)]/2 + C_\delta \\ &\le \beta_{m+1} + C_\delta \\ &= \beta_m,\end{aligned}$$

where the first line comes from Lemma 4; the third line comes from the fact that $\widehat{f}_{m+1}$ is the minimizer with respect to $\widehat{R}_{m+1}$ and Lemma 4; the last line comes from the definition of β_m.

□

Lemma 6. *For any $m \in [M]$, we have $g_m(x) = 1 \implies w(x; \mathcal{F}_m) > \gamma$.*

Proof. We only need to show that $\mathsf{ucb}(x; \mathcal{F}_m) - \mathsf{lcb}(x; \mathcal{F}_m) \le \gamma \implies g_m(x) = 0$. Suppose otherwise $g_m(x) = 1$, which implies that both

$$\frac{1}{2} \in (\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)) \quad \text{and} \quad [\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)] \not\subseteq \left[\frac{1}{2} - \gamma, \frac{1}{2} + \gamma\right]. \tag{7}$$

If $\frac{1}{2} \in (\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m))$ and $\mathsf{ucb}(x; \mathcal{F}_m) - \mathsf{lcb}(x; \mathcal{F}_m) \le \gamma$, we must have $\mathsf{lcb}(x; \mathcal{F}_m) \ge \frac{1}{2} - \gamma$ and $\mathsf{ucb}(x; \mathcal{F}_m) \le \frac{1}{2} + \gamma$, which contradicts with Eq. (7). □

We introduce more notations. Fix any $m \in [M]$. We use $n_m := \tau_m - \tau_{m-1}$ to denote the length of epoch m, and use abbreviation $\rho_m := 2\beta_m + C_\delta$. Denote $(\mathcal{X}, \Sigma, \mathcal{D}_\mathcal{X})$ as the (marginal) probability space, and denote $\overline{\mathcal{X}}_m := \{x \in \mathcal{X} : g_m(x) = 1\} \in \Sigma$ be the region where query *is* requested within epoch m. Since we have $\mathcal{F}_{m+1} \subseteq \mathcal{F}_m$ by Lemma 5, we clearly have $\overline{\mathcal{X}}_{m+1} \subseteq \overline{\mathcal{X}}_m$. We now define a sub probability measure $\bar{\mu}_m := (\mathcal{D}_\mathcal{X})_{|\overline{\mathcal{X}}_m}$ such that $\bar{\mu}_m(\omega) = \mathcal{D}_\mathcal{X}(\omega \cap \overline{\mathcal{X}}_m)$ for any $\omega \in \Sigma$. Fix any time step t within epoch m and any $\overline{m} \le m$. Consider any measurable function F (that is $\mathcal{D}_\mathcal{X}$ integrable), we have

$$\begin{aligned}
\mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1) \cdot F(x)] &= \int_{x \in \overline{\mathcal{X}}_m} F(x)\, d\mathcal{D}_\mathcal{X}(x) \\
&\le \int_{x \in \overline{\mathcal{X}}_{\overline{m}}} F(x)\, d\mathcal{D}_\mathcal{X}(x) \\
&= \int_{x \in \mathcal{X}} F(x)\, d\bar{\mu}_{\overline{m}}(x) \\
&=: \mathbb{E}_{x \sim \bar{\mu}_{\overline{m}}}[F(x)],
\end{aligned} \tag{8}$$

where, by a slightly abuse of notations, we use $\mathbb{E}_{x \sim \mu}[\cdot]$ to denote the integration with any sub probability measure μ. In particular, Eq. (8) holds with equality when $\overline{m} = m$.

Lemma 7. *Fix any epoch $m \ge 2$. We have*

$$\mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1)] \le \frac{4\rho_m}{\tau_{m-1}\gamma^2} \cdot \theta^{\mathrm{val}}_{f^\star}\left(\mathcal{F}, \gamma/2, \sqrt{\rho_m / 2\tau_{m-1}}\right).$$

Proof. We know that $\mathbb{1}(g_m(x) = 1) = \mathbb{1}(g_m(x) = 1) \cdot \mathbb{1}(w(x; \mathcal{F}_m) > \gamma)$ from Lemma 6. Thus, for any $\overline{m} \le m$, we have

$$\begin{aligned}
\mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1)] &= \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1) \cdot \mathbb{1}(w(x; \mathcal{F}_m) > \gamma)] \\
&\le \mathbb{E}_{x \sim \bar{\mu}_{\overline{m}}}[\mathbb{1}(w(x; \mathcal{F}_m) > \gamma)] \\
&\le \mathbb{E}_{x \sim \bar{\mu}_{\overline{m}}}\Big(\mathbb{1}\big(\exists f \in \mathcal{F}_m, |f(x) - f^\star(x)| > \gamma/2\big)\Big),
\end{aligned} \tag{9}$$

where the second line uses Eq. (8) and the last line comes from the facts that $f^\star \in \mathcal{F}_m$ and $w(x; \mathcal{F}_m) > \gamma \implies \exists f \in \mathcal{F}_m, |f(x) - f^\star(x)| > \gamma/2$.

For any time step t, let $m(t)$ denote the epoch where t belongs to. From Lemma 5, we know that, $\forall f \in \mathcal{F}_m$,

$$\begin{aligned}
\rho_m &\ge \sum_{t=1}^{\tau_{m-1}} \mathbb{E}_t\Big[Q_t\big(f(x_t) - f^\star(x_t)\big)^2\Big] \\
&= \sum_{t=1}^{\tau_{m-1}} \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}\Big[\mathbb{1}(g_{m(t)}(x) = 1) \cdot \big(f(x) - f^\star(x)\big)^2\Big] \\
&= \sum_{\overline{m}=1}^{m-1} n_{\overline{m}} \cdot \mathbb{E}_{x \sim \bar{\mu}_{\overline{m}}}\Big[\big(f(x) - f^\star(x)\big)^2\Big] \\
&= \tau_{m-1} \mathbb{E}_{x \sim \bar{\nu}_m}\Big[\big(f(x) - f^\star(x)\big)^2\Big],
\end{aligned} \tag{10}$$

where we use $Q_t = g_{m(t)}(x_t) = \mathbb{1}(g_{m(t)}(x) = 1)$ and Eq. (8) on the second line, and define a new sub probability measure

$$\bar{\nu}_m := \frac{1}{\tau_{m-1}} \sum_{\overline{m}=1}^{m-1} n_{\overline{m}} \cdot \bar{\mu}_{\overline{m}}$$

on the third line.

Plugging Eq. (10) into Eq. (9) leads to the bound

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1)] \\
&\leq \mathbb{E}_{x\sim\bar{\nu}_m}\left[\mathbb{1}\Big(\exists f \in \mathcal{F}, \big|f(x) - f^\star(x)\big| > \gamma/2, \mathbb{E}_{x\sim\bar{\nu}_m}\Big[\big(f(x) - f^\star(x)\big)^2\Big] \leq \frac{\rho_m}{\tau_{m-1}}\Big)\right],
\end{aligned}$$

where we use the definition of $\bar{\nu}_m$ again (note that Eq. (9) works with any $\overline{m} \leq m$). Combining the above result with the discussion around Proposition 7 and Definition 1, we then have

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1)] \leq \frac{4\rho_m}{\tau_{m-1}\,\gamma^2} \cdot \theta^{\mathrm{val}}_{f^\star}\Big(\mathcal{F}, \gamma/2, \sqrt{\rho_m/2\tau_{m-1}}\Big).$$

□

Lemma 8. *Fix any epoch $m \geq 2$. We have*

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1) \cdot w(x; \mathcal{F}_m)] \leq \frac{4\rho_m}{\tau_{m-1}\,\gamma} \cdot \theta^{\mathrm{val}}_{f^\star}\Big(\mathcal{F}, \gamma/2, \sqrt{\rho_m/2\tau_{m-1}}\Big).$$

Proof. Similar to the proof of Lemma 7, we have

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1) \cdot w(x; \mathcal{F}_m)] &= \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1) \cdot \mathbb{1}(w(x; \mathcal{F}_m) > \gamma) \cdot w(x; \mathcal{F}_m)] \\
&\leq \mathbb{E}_{x\sim\bar{\mu}_{\overline{m}}}[\mathbb{1}(w(x; \mathcal{F}_m) > \gamma) \cdot w(x; \mathcal{F}_m)]
\end{aligned}$$

for any $\overline{m} \leq m$. With $\bar{\nu}_m = \frac{1}{\tau_{m-1}} \sum_{\overline{m}=1}^{m-1} n_{\overline{m}} \cdot \bar{\mu}_{\overline{m}}$, we then have

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x) = 1) \cdot w(x; \mathcal{F}_m)] \\
&\leq \mathbb{E}_{x\sim\bar{\nu}_m}[\mathbb{1}(w(x; \mathcal{F}_m) > \gamma) \cdot w(x; \mathcal{F}_m)] \\
&\leq \mathbb{E}_{x\sim\bar{\nu}_m}\left[\mathbb{1}(\exists f \in \mathcal{F}_m, \big|f(x) - f^\star(x)\big| > \gamma/2) \cdot \left(\sup_{f,f'\in\mathcal{F}_m} |f(x) - f'(x)|\right)\right] \\
&\leq 2\mathbb{E}_{x\sim\bar{\nu}_m}\left[\mathbb{1}(\exists f \in \mathcal{F}_m, \big|f(x) - f^\star(x)\big| > \gamma/2) \cdot \left(\sup_{f\in\mathcal{F}_m} |f(x) - f^\star(x)|\right)\right] \\
&\leq 2\int_{\gamma/2}^1 \mathbb{E}_{x\sim\bar{\nu}_m}\left[\mathbb{1}\left(\sup_{f\in\mathcal{F}_m} \big|f(x) - f^\star(x)\big| \geq \omega\right)\right] d\,\omega \\
&\leq 2\int_{\gamma/2}^1 \frac{1}{\omega^2}\, d\,\omega \cdot \left(\frac{\rho_m}{\tau_{m-1}} \cdot \theta^{\mathrm{val}}_{f^\star}\Big(\mathcal{F}, \gamma/2, \sqrt{\rho_m/2\tau_{m-1}}\Big)\right) \\
&\leq \frac{4\rho_m}{\tau_{m-1}\,\gamma} \cdot \theta^{\mathrm{val}}_{f^\star}\Big(\mathcal{F}, \gamma/2, \sqrt{\rho_m/2\tau_{m-1}}\Big),
\end{aligned}$$

where we use similar steps as in the proof of Lemma 7. □

Lemma 9. *Fix any $m \in [M]$. We have* $\mathsf{excess}_\gamma(\widehat{h}_m; x) \leq 0$ *if* $g_m(x) = 0$.

Proof. Recall that

$$\begin{aligned}
\mathsf{excess}_\gamma(\widehat{h}; x) = \mathbb{1}\big(\widehat{h}(x) \neq \perp\big) \cdot \big(\mathbb{P}_y\big(y \neq \mathrm{sign}(\widehat{h}(x))\big) - \mathbb{P}_y\big(y \neq \mathrm{sign}(h^\star(x))\big)\big) \\
+ \mathbb{1}\big(\widehat{h}(x) = \perp\big) \cdot \big((1/2 - \gamma) - \mathbb{P}_y\big(y \neq \mathrm{sign}(h^\star(x))\big)\big).
\end{aligned}$$

We now analyze the event $\{g_m(x) = 0\}$ in two cases.

Case 1: $\widehat{h}_m(x) = \perp$**.**

Since $\eta(x) = f^\star(x) \in [\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)]$, we know that $\eta(x) \in [\frac{1}{2} - \gamma, \frac{1}{2} + \gamma]$ and thus $\mathbb{P}_y\big(y \neq \mathrm{sign}(h^\star(x))\big) \geq \frac{1}{2} - \gamma$. As a result, we have $\mathsf{excess}_\gamma(\widehat{h}_m; x) \leq 0$.

Case 2: $\widehat{h}_m(x) \neq \perp$ but $\frac{1}{2} \notin (\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m))$.

In this case, we know that $\mathrm{sign}(\widehat{h}_m(x)) = \mathrm{sign}(h^\star(x))$ whenever $\eta(x) \in [\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)]$. As a result, we have $\mathsf{excess}_\gamma(\widehat{h}_m; x) \leq 0$ as well. □

F Proofs of results in Section 3

Proposition 2. *The classifier $\widehat{h}$ returned by Algorithm 1 enjoys proper abstention. With randomization over the abstention region, we have the following upper bound on its standard excess error*

$$\mathrm{err}(\check{h}) - \mathrm{err}(h^\star) \leq \mathrm{err}_\gamma(\widehat{h}) - \mathrm{err}(h^\star) + \gamma \cdot \mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(x \in \mathcal{X}_\gamma). \tag{5}$$

Proof. The proper abstention property of $\widehat{h}$ returned by Algorithm 1 is achieved via conservation: $\widehat{h}$ will avoid abstention unless it is absolutely sure that abstention is the optimal choice. The proper abstention property implies that $\mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(\widehat{h}(x) = \perp) \leq \mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(x \in \mathcal{X}_\gamma)$. The desired result follows by combining this inequality with Eq. (4). □

Theorem 6. *With an appropriate choice of the abstention parameter γ in Algorithm 1 and randomization over the abstention region, Algorithm 1 learns a classifier $\check{h}$ at the minimax optimal rates: To achieve ε standard excess error, it takes $\widetilde{\Theta}(\tau_0^{-2})$ labels under Massart noise and takes $\widetilde{\Theta}(\varepsilon^{-2/(1+\beta)})$ labels under Tsybakov noise.*

Proof. The results follow by taking the corresponding γ in Algorithm 1 and then apply Proposition 2. In the case with Massart noise, we have $\mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(x \in \mathcal{X}_\gamma) = 0$ when $\gamma = \tau_0$; and the corresponding label complexity scales as $\widetilde{O}(\tau_0^{-2})$. In the case with Tsybakov noise, we have $\mathbb{P}_{x \sim \mathcal{D}_\mathcal{X}}(x \in \mathcal{X}_\gamma) = \frac{\varepsilon}{2}$ when $\gamma = (\frac{\varepsilon}{2c})^{1/(1+\beta)}$. Applying Algorithm 1 to achieve $\frac{\varepsilon}{2}$ Chow's excess error thus leads to $\frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ standard excess error. The corresponding label complexity scales as $\widetilde{O}(\varepsilon^{-2/(1+\beta)})$. □

Theorem 7. *With an appropriate choice of the abstention parameter γ in Algorithm 1 and randomization over the abstention region, Algorithm 1 learns a classifier $\check{h}$ with $\varepsilon + \zeta_0$ standard excess error after querying $\widetilde{\Theta}(\tau_0^{-2})$ labels under Definition 5 or querying $\widetilde{\Theta}(\varepsilon^{-2/(1+\beta)})$ labels under Definition 6.*

Proof. For any abstention parameter $\gamma > 0$, we denote $\mathcal{X}_{\zeta_0,\gamma} := \{x \in \mathcal{X} : \eta(x) \in [\frac{1}{2} - \gamma, \frac{1}{2} + \gamma], |\eta(x) - 1/2| > \zeta_0\}$ as the intersection of the region controlled by noise-seeking conditions and the (possible) abstention region. Let $\widehat{h}$ be the classifier returned by Algorithm 1 and $\check{h}$ be its randomized version (over the abstention region). We denote $\mathcal{S} := \{x \in \mathcal{X} : \widehat{h}(x) = \perp\}$ be the abstention region of $\widehat{h}$. Since $\widehat{h}$ abstains properly, we have $\mathcal{S} \subseteq \{x \in \mathcal{X} : |\eta(x) - 1/2| \leq \gamma\} =: \mathcal{X}_\gamma$. We write $\mathcal{S}_0 := \mathcal{S} \cap \mathcal{X}_{\zeta_0,\gamma}$, $\mathcal{S}_1 := \mathcal{S} \setminus \mathcal{S}_0$ and $\mathcal{S}_2 := \mathcal{X} \setminus \mathcal{S}$. For any $h : \mathcal{X} \to \mathcal{Y}$, we define the notation $\mathsf{excess}(h; x) := (\mathbb{P}_{y|x}\big(y \neq \mathrm{sign}(h(x))\big) - \mathbb{P}_{y|x}\big(y \neq \mathrm{sign}(h^\star(x))\big))$, and have $\mathsf{excess}(h) = \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathsf{excess}(h; x)]$. We then have

$$\begin{aligned}
&\mathsf{excess}(\check{h}) \\
&= \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}\left[\mathsf{excess}(\check{h}; x) \cdot \mathbb{1}(x \in \mathcal{S}_0)\right] + \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}\left[\mathsf{excess}(\check{h}; x) \cdot \mathbb{1}(x \in \mathcal{S}_1)\right] + \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}\left[\mathsf{excess}(\check{h}; x) \cdot \mathbb{1}(x \in \mathcal{S}_2)\right] \\
&\leq \gamma \cdot \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathbb{1}(x \in \mathcal{S}_0)] + \zeta_0 \cdot \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathbb{1}(x \in \mathcal{S}_1)] + \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathsf{excess}_\gamma(\widehat{h}; x) \cdot \mathbb{1}(x \in \mathcal{S}_2)] \\
&\leq \gamma \cdot \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathbb{1}(x \in \mathcal{X}_{\zeta_0,\gamma})] + \zeta_0 + \varepsilon/2,
\end{aligned}$$

where the bound on the second term comes from the fact that $\mathcal{S} \subseteq \mathcal{X}_\gamma$ and the bound on the third term comes from the same analysis that appears in the proof of Theorem 4 (with $\varepsilon/2$ accuracy). One can then tune γ in ways discussed in the proof of Theorem 6 to bound the first term by $\varepsilon/2$, i.e., $\gamma \cdot \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathbb{1}(x \in \mathcal{X}_{\zeta_0,\gamma})] \leq \varepsilon/2$, with similar label complexity. □

Proposition 3. *Fix $\varepsilon, \delta, \gamma > 0$. For any labeling budget $B \gtrsim \frac{1}{\gamma^2} \cdot \log^2(\frac{1}{\varepsilon\,\gamma}) \cdot \log(\frac{1}{\varepsilon\,\gamma\,\delta})$, there exists a learning problem (with a set of linear regression functions) satisfying Definition 5/Definition 6 such that (1) any "uncertainty-based" active learner suffers expected standard excess error $\Omega(B^{-1})$; yet (2) with probability at least $1-\delta$, Algorithm 1 returns a classifier with standard excess error at most ε.*

Before proving Proposition 3, we first construct a simple problem with linear regression function and give the formal definition of "uncertainty-based" active learner.

Example 1. *We consider the case where $\mathcal{X} = [0,1]$ and $\mathcal{D}_{\mathcal{X}} = \mathrm{unif}(\mathcal{X})$. We consider feature embedding $\phi : \mathcal{X} \to \mathbb{R}^2$, i.e., $\phi(x) = [\phi_1(x), \phi_2(x)]^\top$. We take $\phi_1(x) := 1$ for any $x \in \mathcal{X}$, and define $\phi_2(x)$ as*

$$\phi_2(x) := \begin{cases} 0, & x \in \mathcal{X}_{\mathrm{hard}}, \\ 1, & x \in \mathcal{X}_{\mathrm{easy}}, \end{cases}$$

where $\mathcal{X}_{\mathrm{easy}} \subseteq \mathcal{X}$ is any subset such that $\mathcal{D}_{\mathcal{X}}(\mathcal{X}_{\mathrm{easy}}) = p$, for some constant $p \in (0,1)$, and $\mathcal{X}_{\mathrm{hard}} = \mathcal{X} \setminus \mathcal{X}_{\mathrm{easy}}$. We consider a set of linear regression function $\mathcal{F} := \{f_\theta : f_\theta(x) = \langle \phi(x), \theta \rangle, \|\theta\|_2 \le 1\}$. We set $f^\star = f_{\theta^\star}$, where $\theta^\star = [\theta_1^\star, \theta_2^\star]^\top$ is selected such that $\theta_1^\star = \frac{1}{2}$ and $\theta_2^\star = \mathrm{unif}(\{\pm\frac{1}{2}\})$.

Definition 10. *We say an algorithm is a "uncertainty-based" active learner if, for any $x \in \mathcal{X}$, the learner*

- *constructs an open confidence interval $(\mathsf{lcb}(x), \mathsf{ucb}(x))$ such that $\eta(x) \in (\mathsf{lcb}(x), \mathsf{ucb}(x))$;*[9]
- *queries the label of $x \in \mathcal{X}$ if $\frac{1}{2} \in (\mathsf{lcb}(x), \mathsf{ucb}(x))$.*

Proof. With any given labeling budget B, we consider the problem instance described in Example 1 with $p = B^{-1}/2$. We can easily see that this problem instance satisfy Definition 5 and Definition 6.

We first consider any "uncertainty-based" active learner. Let Z denote the number of data points lie in $\mathcal{X}_{\mathrm{easy}}$ among the first B random draw of examples. We see that $Z \sim \mathcal{B}(B, B^{-1}/2)$ follows a binomial distribution with B trials and $B^{-1}/2$ success rate. By Markov inequality, we have

$$\mathbb{P}\left(Z \ge \frac{3}{2}\mathbb{E}[Z]\right) = \mathbb{P}\left(Z \ge \frac{3}{4}\right) \le \frac{2}{3}.$$

That being said, with probability at least $1/3$, there will be $Z = 0$ data point that randomly drawn from the easy region $\mathcal{X}_{\mathrm{easy}}$. We denote that event as $\mathcal{E}$. Since $\eta(x) = f^\star(x) = \frac{1}{2}$ for any $x \in \mathcal{X}_{\mathrm{hard}}$, any "uncertainty-based" active learner will query the label of any data point $x \in \mathcal{X}_{\mathrm{hard}}$. As a result, under event $\mathcal{E}$, the active learner will use up all the labeling budget in the first B rounds and observe zero label for any data point $x \in \mathcal{X}_{\mathrm{easy}}$. Since the easy region $\mathcal{X}_{\mathrm{easy}}$ has measure $B^{-1}/2$ and $\theta_2^\star = \mathrm{unif}(\{\pm\frac{1}{2}\})$, any classification rule over the easy region would results in expected excess error lower bounded by $B^{-1}/4$. To summarize, with probability at least $\frac{1}{3}$, any "uncertainty-based" active learner without abstention suffers expected excess error $\Omega(B^{-1})$.

We now consider the classifier returned by Algorithm 1.[10] For the linear function considered in Example 1, we have $\mathrm{Pdim}(\mathcal{F}) \le 2$ (Haussler, 1989) and $\theta^{\mathrm{val}}_{f^\star}(\mathcal{F}, \gamma/2, \varepsilon) \le 2$ for any $\varepsilon \ge 0$ (see Appendix C). Thus, by setting $T = O(\frac{1}{\varepsilon\,\gamma} \cdot \log(\frac{1}{\varepsilon\,\gamma\,\delta}))$, with probability at least $1-\delta$, Algorithm 1 return a classifier $\widehat{h}$ with Chow's excess error at most ε and label complexity $O(\frac{1}{\gamma^2} \cdot \log^2(\frac{1}{\varepsilon\,\gamma}) \cdot \log(\frac{1}{\varepsilon\,\gamma\,\delta})) = \mathrm{poly}(\frac{1}{\gamma}, \log(\frac{1}{\varepsilon\,\gamma\,\delta}))$. Since $\widehat{h}$ enjoys proper abstention, it never abstains for $x \in \mathcal{X}_{\mathrm{easy}}$. Note that we have $\eta(x) = \frac{1}{2}$ for any $x \in \mathcal{X}_{\mathrm{hard}}$. By randomizing the decision of $\widehat{h}$ over the abstention region, we obtain a classifier with standard excess error at most ε. □

[9] By restricting to learners that construct an open confidence interval containing $\eta(x)$, we do not consider the corner cases when $\mathsf{lcb}(x) = \frac{1}{2}$ or $\mathsf{ucb}(x) = \frac{1}{2}$ and the confidence interval close.

[10] The version that works with an infinite set of regression functions using concentration results presented in Lemma 4. Or, one can first discretie the set of regression function and then use the version presented in Algorithm 1.

G Omitted details for Section 4.1

We introduce a new perspective for designing and analyzing active learning algorithms in Appendix G.1 (with new notations introduced). Based on this new perspective, we present our algorithm and its theoretical guarantees in Appendix G.2. Supporting lemmas are deferred to Appendix G.3.

G.1 The perspective: Regret minimization with selective sampling

We view active learning as a decision making problem: at each round, the learner selects an action, suffers a loss (that may not be observable), and decides to query the label or not. At a high level, the learner aims at *simultaneously* minimizing the regret and the number of queries; and will randomly return a classifier/decision rule at the end of the learning process.

The perspective is inspired by the seminal results derived in Dekel et al. (2012), where the authors study active learning with linear regression functions and focus on standard excess error guarantees. With this regret minimization perspective, we can also take advantage of fruitful results developed in the field of contextual bandits (Russo and Van Roy, 2013; Foster et al., 2020).

Decision making for regret minimization. To formulate the regret minimization problem, we consider the action set $\mathcal{A} = \{+1, -1, \perp\}$, where the action $+1$ (resp. -1) represents labeling any data point $x \in \mathcal{X}$ as positive (resp. negative); and the action $\perp$ represents abstention. At each round $t \in [T]$, the learner observes a data point $x_t \in \mathcal{X}$ (which can be chosen by an adaptive adversary), takes an action $a_t \in \mathcal{A}$, and then suffers a loss, which is defined as

$$\ell_t(a_t) = \mathbb{1}(\mathrm{sign}(y_t) \neq a_t, a_t \neq \perp) + \left(\frac{1}{2} - \gamma\right) \cdot \mathbb{1}(a_t = \perp).$$

We use $a_t^\star := \mathrm{sign}(2f^\star(x_t) - 1) = \mathrm{sign}(2\eta(x_t) - 1)$ to denote the action taken by the Bayes optimal classifier $h^\star \in \mathcal{H}$. Denote filtration $\mathfrak{F}_t := \sigma((x_i, y_i)_{i=1}^t)$. We define the (conditionally) expected regret at time step $t \in [T]$ as

$$\mathbf{Regret}_t = \mathbb{E}[\ell_t(a_t) - \ell_t(a_t^\star) \mid \mathfrak{F}_{t-1}].$$

The (conditionally) expected cumulative regret across T rounds is defined as

$$\mathbf{Regret}(T) = \sum_{t=1}^{T} \mathbf{Regret}_t,$$

which is the target that the learner aims at minimizing.

Selective querying for label efficiency. Besides choosing an action $a_t \in \mathcal{A}$ at each time step, our algorithm also determines *whether or not* to query the label y_t with respect to x_t. Note that such selective querying protocol makes our problem different from contextual bandits (Russo and Van Roy, 2013; Foster et al., 2020): The loss $\ell_t(a_t)$ of an chosen a_t may not be even observed.

We use Q_t to indicate the query status at round t, i.e.,

$$Q_t = \mathbb{1}(\text{label } y_t \text{ of } x_t \text{ is queried}).$$

The learner also aims at minimizing the total number of queries across T rounds, i.e., $\sum_{t=1}^T Q_t$.

Connection to active learning. We consider the following learner for the above mentioned decision making problem. At each round, the learner constructs a classifier $\widehat{h}_t : \mathcal{X} \to \{+1, -1, \perp\}$ and a query function $g_t : \mathcal{X} \to \{0, 1\}$; the learner then takes action $a_t = \widehat{h}_t(x_t)$ and decides the query status as $Q_t = g_t(x_t)$.

Conditioned on $\mathfrak{F}_{t-1}$, taking expectation over $\ell_t(a_t)$ leads to the following equivalence.

$$\begin{aligned}
\mathbb{E}[\ell_t(a_t) \mid \mathfrak{F}_{t-1}] &= \mathbb{E}\left[\mathbb{1}(\operatorname{sign}(y_t) \neq a_t, a_t \neq \perp) + \left(\frac{1}{2} - \gamma\right) \cdot \mathbb{1}(a_t = \perp) \mid \mathfrak{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mathbb{1}\big(\operatorname{sign}(y_t) \neq \widehat{h}(x_t), \widehat{h}(x_t) \neq \perp\big) + \left(\frac{1}{2} - \gamma\right) \cdot \mathbb{1}\big(\widehat{h}(x_t) = \perp\big) \mid \mathfrak{F}_{t-1}\right] \\
&= \mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathcal{X}\mathcal{Y}}}\big(\operatorname{sign}(y) \neq \widehat{h}(x), \widehat{h}(x) \neq \perp\big) + \left(\frac{1}{2} - \gamma\right) \cdot \mathbb{P}(\widehat{h}(x) = \perp) \\
&= \operatorname{err}_\gamma(\widehat{h}_t).
\end{aligned}$$

This shows that the (conditionally) expected instantaneous loss precisely captures the Chow's error of classifier $\widehat{h}_t$. Similarly, we have

$$\mathbb{E}[\ell_t(a_t^\star) \mid \mathfrak{F}_{t-1}] = \mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathcal{X}\mathcal{Y}}}(\mathbb{1}(y \neq \operatorname{sign}(2\eta(x) - 1))) = \operatorname{err}(h^\star).$$

Combining the above two results, we notice that the (conditionally) expected instantaneous *regret* exactly captures the Chow's excess error of classifier $\widehat{h}_t$, i.e.,

$$\mathbf{Regret}_t = \operatorname{err}_\gamma(\widehat{h}_t) - \operatorname{err}(h^\star).$$

Let $\widehat{h} \sim \operatorname{unif}(\{\widehat{h}_t\}_{t=1}^T)$ be a classifier randomly selected from all the constructed classifiers. Taking expectation with respect to this random selection procedure, we then have

$$\mathbb{E}_{\widehat{h}\sim\operatorname{unif}(\{\widehat{h}_t\}_{t=1}^T)}[\operatorname{err}_\gamma(\widehat{h}) - \operatorname{err}(h^\star)] = \sum_{t=1}^T (\operatorname{err}_\gamma(\widehat{h}_t) - \operatorname{err}(h^\star))/T = \mathbf{Regret}(T)/T. \tag{11}$$

That being said, the expected Chow's excess error of $\widehat{h}$ can be sublinear in T. If the total number of queries is logarithmic in T, this immediately implies learning a classifier with exponential savings in label complexity.

G.2 Algorithm and main results

We present an algorithm that achieves constant label complexity next (Algorithm 2). Compared to Algorithm 1, Algorithm 2 drops the epoch scheduling, uses a sharper elimination rule for the active set (note that β doesn't depend on T, due to applying optimal stopping theorem in Lemma 10), and is analyzed with respect to eluder dimension (Definition 7) instead of disagreement coefficient. As a result, we shave all three sources of $\log \frac{1}{\varepsilon}$, and achieve constant label complexity for general $\mathcal{F}$ (as long as it's finite and has finite eluder dimension). We abbreviate $\mathfrak{e} := \sup_{f^\star\in\mathcal{F}} \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2)$.

Algorithm 2 Efficient Active Learning with Abstention (Constant Label Complexity)

Input: Time horizon $T \in \mathbb{N}$, abstention parameter $\gamma \in (0, 1/2)$ and confidence level $\delta \in (0, 1)$.

Initialize $\widehat{\mathcal{H}} := \emptyset$. Set $T := O(\frac{\mathfrak{e}}{\varepsilon\gamma} \cdot \log(\frac{|\mathcal{F}|}{\delta}))$ and $\beta := \frac{1}{2}\log(\frac{|\mathcal{F}|}{\delta})$.

for $t = 1, 2, \ldots, T$ **do**

  Get $\widehat{f}_t := \arg\min_{f \in \mathcal{F}} \sum_{i<t} Q_i(f(x_i) - y_i)^2$.

  // We use $Q_t \in \{0, 1\}$ to indicate whether the label of x_t is queried.

  (Implicitly) Construct active set of regression function $\mathcal{F}_t \subseteq \mathcal{F}$ as

$$\mathcal{F}_t := \left\{ f \in \mathcal{F} : \sum_{i=1}^{t-1} Q_i(f(x_i) - y_i)^2 \le \sum_{i=1}^{t-1} Q_i(\widehat{f}_t(x_i) - y_i)^2 + \beta \right\}.$$

  Construct classifier $\widehat{h}_t : \mathcal{X} \to \{+1, -1, \perp\}$ as

$$\widehat{h}_t(x) := \begin{cases} \perp, & \text{if } [\mathsf{lcb}(x; \mathcal{F}_t), \mathsf{ucb}(x; \mathcal{F}_t)] \subseteq \left[\frac{1}{2} - \gamma, \frac{1}{2} + \gamma\right]; \\ \operatorname{sign}(2\widehat{f}_t(x) - 1), & \text{o.w.} \end{cases}$$

  Update $\widehat{\mathcal{H}} = \widehat{\mathcal{H}} \cup \{\widehat{h}_t\}$. Construct query function $g_m : \mathcal{X} \to \{0, 1\}$ as

$$g_t(x) := \mathbb{1}\left(\frac{1}{2} \in (\mathsf{lcb}(x; \mathcal{F}_t), \mathsf{ucb}(x; \mathcal{F}_t))\right) \cdot \mathbb{1}(\widehat{h}_t(x) \neq \perp).$$

  Observe $x_t \sim \mathcal{D}_{\mathcal{X}}$. Take action $a_t := \widehat{h}_t(x_t)$. Set $Q_t := g_t(x_t)$.

  if $Q_t = 1$ **then**

    Query the label y_t of x_t.

Return $\widehat{h} := \mathrm{unif}(\widehat{\mathcal{H}})$.

Before proving Theorem 8. We define some notations that are specialized to Appendix G.

We define filtrations $\mathfrak{F}_{t-1} := \sigma(x_1, y_1, \ldots, x_{t-1}, y_{t-1})$ and $\overline{\mathfrak{F}}_{t-1} := \sigma(x_1, y_1, \ldots, x_t)$. Note that we additionally include the data point x_t in the filtration $\overline{\mathfrak{F}}_{t-1}$ at time step $t-1$. We denote $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \overline{\mathfrak{F}}_{t-1}]$. For any $t \in [T]$, we denote $M_t(f) := Q_t((f(x_t) - y_t)^2 - (f^\star(x_t) - y_t)^2)$. We have $\sum_{i=1}^{\tau} \mathbb{E}_t[M_t(f)] = \sum_{t=1}^{\tau} Q_t(f(x_t) - f^\star(x_t))^2$. For any given data point $x_t \in \mathcal{X}$, we use abbreviations

$$\mathsf{ucb}_t := \mathsf{ucb}(x_t; \mathcal{F}_t) = \sup_{f \in \mathcal{F}_t} f(x_t) \quad \text{and} \quad \mathsf{lcb}_t := \mathsf{lcb}(x_t; \mathcal{F}_t) = \inf_{f \in \mathcal{F}_t} f(x_t)$$

to denote the upper and lower confidence bounds of $\eta(x_t) = f^\star(x_t)$. We also denote

$$w_t := \mathsf{ucb}_t - \mathsf{lcb}_t = \sup_{f, f' \in \mathcal{F}_t} |f(x_t) - f'(x_t)|$$

as the width of confidence interval.

Theorem 8. *With probability at least $1 - 2\delta$, Algorithm 2 returns a classifier with expected Chow's excess error at most ε and label complexity $O(\frac{\mathfrak{e} \cdot \log(|\mathcal{F}|/\delta)}{\gamma^2})$, which is independent of $\frac{1}{\varepsilon}$.*

Proof. We first analyze the label complexity of Algorithm 2. Note that Algorithm 2 constructs $\widehat{h}_t$ and g_t in forms similar to the ones constructed in Algorithm 1, and Lemma 6 holds for Algorithm 2 as well. Based on Lemma 6, we have $Q_t = g_t(x_t) = 1 \implies w_t > \gamma$. Thus, taking $\zeta = \gamma$ in Lemma 13 leads to

$$\sum_{t=1}^{T} \mathbb{1}(Q_t = 1) < \frac{17 \log(2|\mathcal{F}|/\delta)}{2\gamma^2} \cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2),$$

with probability one. The label complexity of Algorithm 2 is then upper bounded by a constant as long as $\mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2)$ is upper bounded by a constant (which has no dependence on T or $\frac{1}{\varepsilon}$).

We next analyze the excess error of $\widehat{h}$. We consider the good event $\mathcal{E}$ defined in Lemma 12, which holds true with probability at least $1-\delta$. Under event $\mathcal{E}$, Lemma 15 shows that

$$\sum_{t=1}^{T}\mathbb{E}[\ell_t(a_t)-\ell_t(a_t^\star)\mid\overline{\mathfrak{F}}_{t-1}]\le\frac{17\sqrt{2}\beta}{\gamma}\cdot\mathfrak{e}_{f^\star}(\mathcal{F},\gamma/2).$$

Since

$$\mathbb{E}\Big[\mathbb{E}[\ell_t(a_t)-\ell_t(a_t^\star)\mid\overline{\mathfrak{F}}_{t-1}]\mid\mathfrak{F}_{t-1}\Big]=\mathbb{E}[\ell_t(a_t)-\ell_t(a_t^\star)\mid\mathfrak{F}_{t-1}],$$

and $|\mathbb{E}[\ell_t(a_t)-\ell_t(a_t^\star)\mid\overline{\mathfrak{F}}_{t-1}]|\le 1$ by construction, applying Lemma 2 with respect to $\mathbb{E}[\ell_t(a_t)-\ell_t(a_t^\star)\mid\overline{\mathfrak{F}}_{t-1}]$ further leads to

$$\mathbf{Regret}(T)=\sum_{t=1}^{T}\mathbb{E}[\ell_t(a_t)-\ell_t(a_t^\star)\mid\mathfrak{F}_{t-1}]\le\frac{34\sqrt{2}\beta}{\gamma}\cdot\mathfrak{e}_{f^\star}(\mathcal{F},\gamma/2)+8\log(2\delta^{-1}),$$

with probability at least $1-2\delta$ (due to the additional application of Lemma 2). Since $\widehat{h}\sim\mathrm{unif}(\widehat{\mathcal{H}})$, based on Eq. (11), we thus know that

$$\begin{aligned}\mathbb{E}_{\widehat{h}\sim\mathrm{unif}(\widehat{\mathcal{H}})}[\mathrm{err}_\gamma(\widehat{h})-\mathrm{err}(h^\star)]&=\sum_{t=1}^{T}\big(\mathrm{err}_\gamma(\widehat{h}_t)-\mathrm{err}(h^\star)\big)/T\\&\le\left(\frac{34\sqrt{2}\beta}{\gamma}\cdot\mathfrak{e}_{f^\star}(\mathcal{F},\gamma/2)+8\log\big(2\delta^{-1}\big)\right)/T\end{aligned}$$

Since $T:=O(\frac{\mathfrak{e}}{\varepsilon\gamma}\cdot\log(\frac{|\mathcal{F}|}{\delta}))$, we then know that the expected Chow's excess error is at most ε. □

Theorem 10. *Consider the setting where the data points $\{x_t\}_{t=1}^T$ are chosen by an adaptive adversary with $y_t\sim\mathcal{D}_{\mathcal{Y}|x_t}$. With probability at least $1-\delta$, Algorithm 2 simultaneously guarantees*

$$\sum_{t=1}^{T}\mathbb{E}[\ell_t(a_t)-\ell_t(a_t^\star)\mid\overline{\mathfrak{F}}_{t-1}]\le\frac{34\sqrt{2}\beta}{\gamma}\cdot\mathfrak{e}_{f^\star}(\mathcal{F},\gamma/2),$$

and

$$\sum_{t=1}^{T}\mathbb{1}(Q_t=1)<\frac{17\log(2|\mathcal{F}|/\delta)}{2\gamma^2}\cdot\mathfrak{e}_{f^\star}(\mathcal{F},\gamma/2).$$

Proof. The proof follows the same analysis as in the first part of the proof of Theorem 8 (simply stopped at the step with conditioning on $\overline{\mathfrak{F}}_{t-1}$). □

We redefine $\mathfrak{e}:=\sup_{f^\star\in\mathcal{F}}\mathfrak{e}_{f^\star}(\mathcal{F},\gamma/4)$ in the following Theorem 11 to account for the induced approximation error in efficient implementation.

Theorem 11. *Algorithm 2 can be efficiently implemented via the regression oracle and enjoys the same theoretical guarantees stated in Theorem 8 or Theorem 10. The number of oracle calls needed is $O(\frac{\mathfrak{e}}{\varepsilon\gamma^3}\cdot\log(\frac{|\mathcal{F}|}{\delta})\cdot\log(\frac{1}{\gamma}))$ for a general set of regression functions $\mathcal{F}$, and $O(\frac{\mathfrak{e}}{\varepsilon\gamma}\cdot\log(\frac{|\mathcal{F}|}{\delta})\cdot\log(\frac{1}{\gamma}))$ when $\mathcal{F}$ is convex and closed under pointwise convergence. The per-example inference time of the learned $\widehat{h}_M$ is $O(\frac{1}{\gamma^2}\log\frac{1}{\gamma})$ for general $\mathcal{F}$, and $O(\log\frac{1}{\gamma})$ when $\mathcal{F}$ is convex and closed under pointwise convergence.*

Proof. Denote $\mathcal{B}_t:=\{(x_i,Q_i,y_i)\}_{i=1}^{\tau_{t-1}}$ At any time step $t\in[T]$ of Algorithm 2, we construct classifier $\widehat{h}_t$ and query function g_t with approximated confidence bounds, i.e.,

$$\widehat{\mathsf{lcb}}(x;\mathcal{F}_t):=\mathbf{Alg}_{\mathsf{lcb}}(x;\mathcal{B}_t,\beta_t,\alpha)\quad\text{and}\qquad\widehat{\mathsf{ucb}}(x;\mathcal{F}_t):=\mathbf{Alg}_{\mathsf{ucb}}(x;\mathcal{B}_t,\beta_t,\alpha),$$

where $\mathbf{Alg}_{\mathsf{lcb}}$ and $\mathbf{Alg}_{\mathsf{ucb}}$ are subroutines discussed in Proposition 8 and $\alpha:=\frac{\gamma}{4}$.

Since the theoretical analysis of Theorem 8 and Theorem 10 do not require an non-increasing (with respect to time step t) sampling region, i.e., $\{x\in\mathcal{X}:g_t(x)=1\}$, we only need to approximate the confidence intervals at $\frac{\gamma}{4}$ level. This slightly save the computational complexity compared to Theorem 5, which approximates the confidence interval at $\frac{\gamma}{4\lceil\log T\rceil}$ level. The rest of the analysis of computational complexity follows similar steps in the proof of Theorem 5. □

G.3 Supporting lemmas

Consider a sequence of random variables $(Z_t)_{t\in\mathbb{N}}$ adapted to filtration $\overline{\mathfrak{F}}_t$. We assume that $\mathbb{E}[\exp(\lambda Z_t)] < \infty$ for any λ and $\mu_t := \mathbb{E}\big[Z_t \mid \overline{\mathfrak{F}}_{t-1}\big]$. We also denote

$$\psi_t(\lambda) := \log \mathbb{E}\big[\exp(\lambda \cdot (Z_t - \mu_t)) \mid \overline{\mathfrak{F}}_{t-1}\big].$$

Lemma 10 (Russo and Van Roy (2013))**.** *With notations defined above. For any $\lambda \geq 0$ and $\delta > 0$, we have*

$$\mathbb{P}\left(\forall \tau \in \mathbb{N}, \sum_{t=1}^{\tau} \lambda Z_t \leq \sum_{t=1}^{\tau} (\lambda \mu_t + \psi_t(\lambda)) + \log\left(\frac{1}{\delta}\right)\right) \geq 1 - \delta. \tag{12}$$

Lemma 11. *Fix any $\delta \in (0,1)$. For any $\tau \in [T]$, with probability at least $1-\delta$, we have*

$$\sum_{t=1}^{\tau} M_t(f) \leq \sum_{t=1}^{\tau} \frac{3}{2}\mathbb{E}_t[M_t(f)] + C_\delta,$$

and

$$\sum_{t=1}^{\tau} \mathbb{E}_t[M_t(f)] \leq 2\sum_{t=1}^{\tau} M_t(f) + C_\delta,$$

where $C_\delta := \log\left(\frac{2|\mathcal{F}|}{\delta}\right)$.

Proof. Fix any $f \in \mathcal{F}$. We take $Z_t = M_t(f)$ in Lemma 10. We can rewrite

$$Z_t = Q_t\big((f(x_t) - f^\star(x_t))^2 + 2(f(x_t) - f^\star(x_t))\varepsilon_t\big),$$

where we use the notation $\varepsilon_t := f^\star(x_t) - y_t$. Since $\mathbb{E}_t[\varepsilon_t] = 0$ and $\mathbb{E}_t[\exp(\lambda\varepsilon_t) \mid \overline{\mathfrak{F}}_{t-1}] \leq \exp(\frac{\lambda^2}{8})$ a.s. by assumption, we have

$$\mu_t = \mathbb{E}_t[Z_t] = Q_t(f(x_t) - f^\star(x_t))^2,$$

and

$$\begin{aligned}
\psi_t(\lambda) &= \log \mathbb{E}\big[\exp(\lambda \cdot (Z_t - \mu_t)) \mid \overline{\mathfrak{F}}_{t-1}\big] \\
&= \log \mathbb{E}_t[\exp(2\lambda Q_t(f(x_t) - f^\star(x_t) \cdot \varepsilon_t))] \\
&\leq \frac{\lambda^2 (Q_t(f(x_t) - f^\star(x_t))^2}{2} \\
&= \frac{\lambda^2 \mu_t}{2},
\end{aligned}$$

where the last line comes from the fact that $Q_t \in \{0,1\}$. We can similarly upper bound $\mathbb{E}[\exp(\lambda Z_t)] = \mathbb{E}[\mathbb{E}_t[\exp(\lambda Z_t)]] \leq \exp(\lambda + \frac{\lambda^2}{2})$ by noticing the range fact that $\mu_t \leq 1$.

Plugging the above results into Lemma 10 with $\lambda = 1$ leads to

$$\sum_{t=1}^{\tau} M_t(f) \leq \sum_{t=1}^{\tau} \frac{3}{2}\mathbb{E}_t[M_t(f)] + \log \delta^{-1}.$$

Following the same procedures above with $Z_t = -M_t(f)$ and $\lambda = 1$ leads to

$$\sum_{t=1}^{\tau} \frac{3}{2}\mathbb{E}_t[M_t(f)] \leq 2\sum_{t=1}^{\tau} M_t(f) + \log \delta^{-1}.$$

The final guarantees comes from taking a union abound over $f \in \mathcal{F}$ and splitting the probability for both directions. □

We use $\mathcal{E}$ to denote the good event considered in Lemma 11, we use it through out the rest of this section.

Lemma 12. *With probability at least $1-\delta$, the followings hold true:*

1. *$f^\star \in \mathcal{F}_t$ for any $t \in [T]$.*
2. *$\sum_{t=1}^{\tau-1} \mathbb{E}_t[M_t(f)] \le 2C_\delta$ for any $f \in \mathcal{F}_\tau$.*

Proof. The first statement immediately follows from Lemma 11 (the second inequality) and the fact that $\beta := C_\delta/2$ in Algorithm 2.

For any $f \in \mathcal{F}_\tau$, we have

$$\begin{aligned}\sum_{t=1}^{\tau-1} \mathbb{E}_t[M_t(f)] &\le 2\sum_{t=1}^{\tau-1} Q_t\big((f(x_t)-y_t)^2 - (f^\star(x_t)-y_t)^2\big) + C_\delta \\ &\le 2\sum_{t=1}^{\tau-1} Q_t\Big((f(x_t)-y_t)^2 - (\widehat{f}_\tau(x_t)-y_t)^2\Big) + C_\delta \\ &\le 2C_\delta, \end{aligned} \tag{13}$$

where the first line comes from Lemma 11, the second line comes from the fact that $\widehat{f}_\tau$ is the minimize among $\mathcal{F}_\tau$, and the third line comes from the fact that $f \in \mathcal{F}_\tau$ and $2\beta = C_\delta$. □

Lemma 13. *For any $\zeta > 0$, with probability 1, we have*

$$\sum_{t=1}^{T} \mathbb{1}(Q_t = 1)\cdot \mathbb{1}(w_t > \zeta) < \left(\frac{16\beta}{\zeta^2}+1\right)\cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \zeta/2).$$

Remark 4. *Similar upper bound has been established in the contextual bandit settings for $\sum_{t=1}^{T} \mathbb{1}(w_t > \zeta)$ (Russo and Van Roy, 2013; Foster et al., 2020). Our results is established with an additional $\mathbb{1}(Q_t = 1)$ term due to selective querying in active learning.*

Proof. We give some definitions first. We say that x is ζ-independent of a sequence $x_1, \dots, x_\tau$ if there exists a $f \in \mathcal{F}$ such that $|f(x) - f^\star(x)| > \zeta$ and $\sum_{i\le\tau}(f(x_i) - f^\star(x_i))^2 \le \zeta^2$. We say that x is ζ-dependent of $x_1, \dots, x_\tau$ if we have $|f(x) - f^\star(x)| \le \zeta$ for all $f \in \mathcal{F}$ such that $\sum_{i\le\tau}(f(x_i) - f^\star(x_i))^2 \le \zeta^2$. The eluder dimension $\check{\mathfrak{e}}_{f^\star}(\mathcal{F}, \zeta)$ can be equivalently defined as the length of the longest sequence $x_1, \dots, x_\tau$ such that each x_i is ζ-independent of all its predecessors.

For any $t \in [T]$, and we denote $\mathcal{S}_t = \{x_i : Q_i = g_i(x_i) = 1, i \in [t]\}$ as the *queried* data points up to time step t. We assume that $|\mathcal{S}_t| = \tau$ and denote $\mathcal{S}_t = (x_{g(1)}, \dots, x_{g(\tau)})$, where $g(i)$ represents the time step where the i-th *queried* data point is queried.

Claim 1. For any $j \in [\tau]$, $x_{g(j)}$ is $\frac{\zeta}{2}$-dependent on at most $\frac{16\beta}{\zeta^2}$ disjoint subsequences of $x_{g(1)}, \dots, x_{g(j-1)}$.

For any $x_{g(j)} \in \mathcal{S}_t$, recall that

$$w_{g(j)} = \mathsf{ucb}_{g(j)} - \mathsf{lcb}_{g(j)} = \max_{f, f' \in \mathcal{F}_{g(j)}} |f(x_t) - f'(x_t)|.$$

If $m_{g(j)} > \zeta$, there must exists a $f \in \mathcal{F}_{g(j)}$ such that $\left|f(x_{g(j)}) - f^\star(x_{g(j)})\right| > \frac{\zeta}{2}$. Focus on this specific $f \in \mathcal{F}_{g(j)} \subseteq \mathcal{F}$. If $x_{g(j)}$ is $\frac{\zeta}{2}$-dependent on a subsequence $x_{g(i_1)}, \dots, x_{g(i_m)}$ (of $x_{g(1)}, \dots, x_{g(j-1)}$), we must have

$$\sum_{k\le m}(f(x_{g(i_k)}) - f^\star(x_{g(i_k)}))^2 > \frac{\zeta^2}{4}.$$

Suppose $x_{g(j)}$ is $\frac{\zeta}{2}$-dependent on K *disjoint* subsequences of $x_{g(1)}, \dots, x_{g(j-1)}$, according to Lemma 12, we must have

$$K\cdot\frac{\zeta^2}{4} < \sum_{i<j}(f(x_{g(i)}) - f^\star(x_{g(i)}))^2 = \sum_{k<g(j)} Q_k(f(x_k) - f^\star(x_k))^2 \le 4\beta,$$

which implies that $K < \frac{16\beta}{\zeta^2}$.

Claim 2. Denote $d := \check{\mathfrak{e}}_{f^\star}(\mathcal{F}, \zeta/2) \geq 1$ and $K = \left\lfloor \frac{\tau-1}{d} \right\rfloor$. There must exists a $j \in [\tau]$ such that $x_{g(j)}$ is $\frac{\zeta}{2}$-dependent on at least K disjoint subsequences of $x_{g(1)}, \ldots, x_{g(j-1)}$.

We initialize K subsequences $\mathcal{C}_i = \{x_{g(i)}\}$. If $x_{g(K+1)}$ is $\frac{\zeta}{2}$-dependent on each $\mathcal{C}_i$, we are done. If not, select a subsequence $\mathcal{C}_i$ such that $x_{g(K+1)}$ is $\frac{\zeta}{2}$-independent of and add $x_{g(K+1)}$ into this subsequence. Repeat this procedure with $j > K+1$ until $x_{g(j)}$ is $\frac{\zeta}{2}$-dependent of all $\mathcal{C}_i$ or $j = \tau$. In the later case, we have $\sum_{i \leq K} |\mathcal{C}_i| = \tau - 1 \geq Kd$. Since $|\mathcal{C}_i| \leq d$ by definition, we must have $|\mathcal{C}_i| = d$ for all $i \in [K]$. As a result, $x_{g(\tau)}$ must be $\frac{\zeta}{2}$-dependent of all $\mathcal{C}_i$.

It's easy to check that $\left\lfloor \frac{\tau-1}{d} \right\rfloor \geq \frac{\tau}{d} - 1$. Combining Claim 1 and 2, we have

$$\frac{\tau}{d} - 1 \leq \left\lfloor \frac{\tau - 1}{d} \right\rfloor \leq K < \frac{16\beta}{\zeta^2}.$$

Rearranging leads to the desired result. □

The following Lemma 14 is a restatement of Lemma 9 in the regret minimization setting.

Lemma 14. *If $Q_t = 0$, we have $\mathbb{E}\big[\ell_t(a_t) - \ell_t(a_t^\star) \mid \overline{\mathfrak{F}}_{t-1}\big] \leq 0$.*

Proof. Recall we have $a_t = \widehat{h}_t(x_t)$. We then have

$$\begin{aligned}
&\mathbb{E}\big[\ell_t(a_t) - \ell_t(a_t^\star) \mid \overline{\mathfrak{F}}_{t-1}\big] \\
&= \mathbb{P}_{y_t|x_t}\big(y_t \neq \operatorname{sign}(\widehat{h}_t(x_t))\big) \cdot \mathbb{1}\big(\widehat{h}_t(x_t) \neq \perp\big) + \big(1/2 - \gamma\big) \cdot \mathbb{1}\big(\widehat{h}_t(x_t) = \perp\big) - \mathbb{P}_{y_t|x_t}\big(y_t \neq \operatorname{sign}(h^\star(x_t))\big) \\
&= \mathbb{1}\big(\widehat{h}_t(x_t) \neq \perp\big) \cdot \big(\mathbb{P}_{y_t|x_t}\big(y_t \neq \operatorname{sign}(\widehat{h}_t(x_t))\big) - \mathbb{P}_{y_t|x_t}\big(y_t \neq \operatorname{sign}(h^\star(x_t))\big)\big) \\
&\quad + \mathbb{1}\big(\widehat{h}_t(x_t) = \perp\big) \cdot \big(\big(1/2 - \gamma\big) - \mathbb{P}_{y_t|x_t}\big(y_t \neq \operatorname{sign}(h^\star(x_t))\big)\big).
\end{aligned}$$

We now analyze the event $\{Q_t = 0\}$ in two cases.

Case 1: $\widehat{h}_t(x_t) = \perp$.

Since $\eta(x_t) = f^\star(x_t) \in [\mathsf{lcb}_t, \mathsf{ucb}_t]$, we further know that $\eta(x_t) \in [\frac{1}{2} - \gamma, \frac{1}{2} + \gamma]$ and thus $\mathbb{P}_{y_t|x_t}\big(y_t \neq \operatorname{sign}(h^\star(x_t))\big) \geq \frac{1}{2} - \gamma$. As a result, we have $\mathbb{E}\big[\ell_t(a_t) - \ell_t(a_t^\star) \mid \overline{\mathfrak{F}}_{t-1}\big] \leq 0$.

Case 2: $\widehat{h}_t(x_t) \neq \perp$ but $\frac{1}{2} \notin (\mathsf{lcb}_t, \mathsf{ucb}_t)$.

In this case, we know that $\operatorname{sign}(\widehat{h}_t(x_t)) = \operatorname{sign}(h^\star(x_t))$ whenever $\eta(x_t) \in [\mathsf{lcb}_t, \mathsf{ucb}_t]$. As a result, we have $\mathbb{E}\big[\ell_t(a_t) - \ell_t(a_t^\star) \mid \overline{\mathfrak{F}}_{t-1}\big] = 0$. □

Lemma 15. *Assume $\mu(x_t) \in [\mathsf{lcb}_t, \mathsf{ucb}_t]$ and $f^\star$ is not eliminated across all $t \in [T]$. We have*

$$\sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t) - \ell_t(a_t^\star) \mid \overline{\mathfrak{F}}_{t-1}] \leq \frac{17\sqrt{2}\beta}{\gamma} \cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2). \tag{14}$$

Proof. Lemma 14 shows that non-positive conditional regret is incurred at whenever $Q_t = 0$, we then have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\ell_t(a_t) - \ell_t(a_t^\star) \mid \overline{\mathfrak{F}}_{t-1}] &\leq \sum_{t=1}^{T} \mathbb{1}(Q_t = 1) \mathbb{E}\big[\ell_t(a_t) - \ell_t(a_t^\star) \mid \overline{\mathfrak{F}}_{t-1}\big] \\
&= \sum_{t=1}^{T} \mathbb{1}(Q_t = 1) \cdot \mathbb{1}(w_t > \gamma) \cdot |2f^\star(x_t) - 1| \\
&\leq \sum_{t=1}^{T} \mathbb{1}(Q_t = 1) \cdot \mathbb{1}(w_t > \gamma) \cdot 2w_t,
\end{aligned}$$

where we use Lemma 6 and Lemma 14 on the second line; and the last line comes from the fact that $|f^\star - \frac{1}{2}| \le w_t$ whenever $f^\star$ is not eliminated and a query is issued. We can directly apply $w_t \le 1$ and Lemma 13 to bound the above terms by $\widetilde{O}(\frac{\mathfrak{e}_{f^\star}(\mathcal{F},\gamma/2)}{\gamma^2})$, which has slightly worse dependence on γ. Following Foster et al. (2020), we take a slightly tighter analysis below.

Let $\mathcal{S}_T := \{x_i : Q_i = 1, i \in [T]\}$ denote the set of queried data points. Suppose $|\mathcal{S}_T| = \tau$. Let $i_1, \dots, i_\tau$ be a reordering of indices within $\mathcal{S}_T$ such that $w_{i_1}(x_{i_1}) \ge w_{i_2}(x_{i_2}) \ge \dots \ge w_{i_\tau}(x_{i_\tau})$. Consider any index $t \in [\tau]$ such that $w_{i_t}(x_{i_t}) \ge \gamma$. For any $\zeta \ge \gamma$, Lemma 13 implies that

$$t \le \sum_{t=1}^{T} \mathbb{1}(Q_t = 1) \cdot \mathbb{1}(w_t(x_t) > \zeta) \le \frac{17\beta}{\zeta^2} \cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \zeta/2) \le \frac{17\beta}{\zeta^2} \cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2). \tag{15}$$

Taking $\zeta = w_{i_t}(x_{i_t})$ in Eq. (15) leads to the fact that

$$w_{i_t}(x_{i_t}) \le \sqrt{\frac{17\beta \cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2)}{t}}.$$

Taking $\zeta = \gamma$ in Eq. (15) leads to the fact that

$$\tau \le \frac{17\beta}{\gamma^2} \cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2).$$

We now have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}(Q_t = 1) \cdot \mathbb{1}(w_t > \gamma) \cdot 2w_t &= \sum_{t=1}^{\tau} \mathbb{1}(w_{i_t} > \gamma) \cdot 2w_{i_t}(x_{i_t}) \\
&\le 2 \sum_{t=1}^{\tau} \sqrt{\frac{17\beta \cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2)}{t}} \\
&\le \sqrt{34\beta \cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2) \cdot \tau} \\
&\le \frac{17\sqrt{2}\beta}{\gamma} \cdot \mathfrak{e}_{f^\star}(\mathcal{F}, \gamma/2).
\end{aligned}$$

□

H Omitted details for Section 4.2

H.1 Algorithm and main results

Algorithm 3 achieves the guarantees stated in Theorem 9. Theorem 9 is proved based on supporting lemmas derived in Appendix H.3. Note that, under the condition $\kappa \le \varepsilon$, we still compete against the Bayes classifier $h^\star = h_{f^\star}$ in the analysis of Chow's excess error Eq. (2).

Theorem 9. *Suppose $\kappa \le \varepsilon$. With probability at least $1 - 2\delta$, Algorithm 3 returns a classifier with Chow's excess error $O(\varepsilon \cdot \bar{\theta} \cdot \log(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma\,\delta}))$ and label complexity $O(\frac{\bar{\theta}\,\mathrm{Pdim}(\mathcal{F})}{\gamma^2} \cdot \log^2(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma}) \cdot \log(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma\,\delta}))$.*

Proof. We analyze under the good event $\mathcal{E}$ defined in Lemma 4, which holds with probability at least $1 - \delta$. Note that all supporting lemmas stated in Appendix H.3 hold true under this event.

We analyze the Chow's excess error of $\widehat{h}_m$, which is measurable with respect to $\mathfrak{F}_{\tau_{m-1}}$. For any $x \in \mathcal{X}$, if $g_m(x) = 0$, Lemma 20 implies that $\mathsf{excess}_\gamma(\widehat{h}_m; x) \le 2\kappa$. If $g_m(x) = 1$, we know that $\widehat{h}_m(x) \ne \perp$ and $\frac{1}{2} \in (\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m))$. Since $\bar{f} \in \mathcal{F}_m$ by Lemma 17 and $\sup_{x \in \mathcal{X}} |\bar{f}(x) - f^\star(x)| \le \kappa$ by assumption. The error incurred in this case is upper bounded by

$$\begin{aligned}
\mathsf{excess}_\gamma(\widehat{h}_m; x) &\le 2|f^\star(x) - 1/2| \\
&\le 2\kappa + 2|\bar{f}(x) - 1/2| \\
&\le 2\kappa + 2w(x; \mathcal{F}_m).
\end{aligned}$$

Algorithm 3 Efficient Active Learning with Abstention under Misspecification

Input: Accuracy level $\varepsilon > 0$, abstention parameter $\gamma \in (\varepsilon, 1/2)$ and confidence level $\delta \in (0,1)$.

Define $T := \frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma}$, $M := \lceil \log_2 T \rceil$ and $C_\delta := O(\mathrm{Pdim}(\mathcal{F}) \cdot \log(T/\delta))$.

Define $\tau_m := 2^m$ for $m \geq 1$, $\tau_0 = 0$ and $\beta_m := (M - m + 1) \cdot \big(2\varepsilon^2 \tau_{M-1} + 2C_\delta\big)$.

for epoch $m = 1, 2, \ldots, M$ **do**

  Get $\widehat{f}_m := \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{\tau_{m-1}} Q_t (f(x_t) - y_t)^2$.

  // We use $Q_t \in \{0,1\}$ to indicate whether the label of x_t is queried.

  (Implicitly) Construct active set of regression function $\mathcal{F}_m \subseteq \mathcal{F}$ as

$$\mathcal{F}_m := \left\{ f \in \mathcal{F} : \sum_{t=1}^{\tau_{m-1}} Q_t (f(x_t) - y_t)^2 \leq \sum_{t=1}^{\tau_{m-1}} Q_t (\widehat{f}_m(x_t) - y_t)^2 + \beta_m \right\}.$$

  Construct classifier $\widehat{h}_m : \mathcal{X} \to \{+1, -1, \perp\}$ as

$$\widehat{h}_m(x) := \begin{cases} \perp, & \text{if } [\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m)] \subseteq \left[\tfrac{1}{2} - \gamma, \tfrac{1}{2} + \gamma\right]; \\ \mathrm{sign}(2\widehat{f}_m(x) - 1), & \text{o.w.} \end{cases}$$

  and query function $g_m : \mathcal{X} \to \{0,1\}$ as

$$g_m(x) := \mathbb{1}\left(\frac{1}{2} \in (\mathsf{lcb}(x; \mathcal{F}_m), \mathsf{ucb}(x; \mathcal{F}_m))\right) \cdot \mathbb{1}(\widehat{h}_m(x) \neq \perp).$$

  if epoch $m = M$ **then**

    Return classifier $\widehat{h}_M$.

  for time $t = \tau_{m-1} + 1, \ldots, \tau_m$ **do**

    Observe $x_t \sim \mathcal{D}_{\mathcal{X}}$. Set $Q_t := g_m(x_t)$.

    if $Q_t = 1$ **then**

      Query the label y_t of x_t.

Combining these two cases together, we have

$$\mathsf{excess}_\gamma(\widehat{h}_m) \leq 2\kappa + 2\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathbb{1}(g_m(x) = 1) \cdot w(x; \mathcal{F}_m)].$$

Take $m = M$ and apply Lemma 19 leads to the following guarantee.

$$\begin{aligned} \mathsf{excess}_\gamma(\widehat{h}_M) &\leq 2\kappa + \frac{72\beta_M}{\tau_{M-1}\gamma} \cdot \theta^{\mathrm{val}}_{\bar{f}}\left(\mathcal{F}, \gamma/2, \sqrt{\beta_M/\tau_{M-1}}\right) \\ &\leq 2\kappa + O\left(\frac{\varepsilon^2}{\gamma} + \frac{\mathrm{Pdim}(\mathcal{F}) \cdot \log(T/\delta)}{T\,\gamma}\right) \cdot \theta^{\mathrm{val}}_{\bar{f}}\left(\mathcal{F}, \gamma/2, \sqrt{C_\delta/T}\right) \\ &= O\left(\varepsilon \cdot \bar{\theta} \cdot \log\left(\frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma\,\delta}\right)\right), \end{aligned}$$

where we take $\bar{\theta} := \sup_{\iota > 0} \theta^{\mathrm{val}}_{\bar{f}}(\mathcal{F}, \gamma/2, \iota)$ as an upper bound of $\theta^{\mathrm{val}}_{\bar{f}}(\mathcal{F}, \gamma/2, \sqrt{C_\delta/T})$, and use the fact that $T = \frac{\mathrm{Pdim}(\mathcal{F})}{\varepsilon\,\gamma}$ and the assumptions that $\kappa \leq \varepsilon < \gamma$.

We now analyze the label complexity (note that the sampling process of Algorithm 3 stops at time $t = \tau_{M-1}$). Note that $\mathbb{E}[\mathbb{1}(Q_t = 1) \mid \mathfrak{F}_{t-1}] = \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathbb{1}(g_m(x) = 1)]$ for any epoch $m \geq 2$ and

time step t within epoch m. Combine Lemma 2 with Lemma 18 leads to

$$\begin{aligned}
\sum_{t=1}^{\tau_{M-1}} \mathbb{1}(Q_t = 1) &\le \frac{3}{2} \sum_{t=1}^{\tau_{M-1}} \mathbb{E}[\mathbb{1}(Q_t = 1) \mid \mathfrak{F}_{t-1}] + 4 \log \delta^{-1} \\
&\le 3 + \frac{3}{2} \sum_{m=2}^{M-1} \frac{(\tau_m - \tau_{m-1}) \cdot 36 \beta_m}{\tau_{m-1} \gamma^2} \cdot \theta_{\bar{f}}^{\mathrm{val}}\left(\mathcal{F}, \gamma/2, \sqrt{\beta_m / \tau_{m-1}}\right) + 4 \log \delta^{-1} \\
&\le 3 + 48 \sum_{m=2}^{M-1} \frac{\beta_m}{\gamma^2} \cdot \theta_{\bar{f}}^{\mathrm{val}}\left(\mathcal{F}, \gamma/2, \sqrt{\beta_m / \tau_{m-1}}\right) + 4 \log \delta^{-1} \\
&\le 3 + 4 \log \delta^{-1} + O\left(\frac{M^2 \cdot \varepsilon^2 \cdot T}{\gamma^2} + \frac{M^2 \cdot C_\delta}{\gamma^2}\right) \cdot \theta_{\bar{f}}^{\mathrm{val}}\left(\mathcal{F}, \gamma/2, \sqrt{C_\delta / T}\right) \\
&= O\left(\frac{\bar{\theta} \operatorname{Pdim}(\mathcal{F})}{\gamma^2} \cdot \left(\log\left(\frac{\operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma}\right)\right)^2 \cdot \log\left(\frac{\operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma \delta}\right)\right)
\end{aligned}$$

with probability at least $1 - 2\delta$ (due to an additional application of Lemma 2); where we use the fact that $T = \frac{\operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma}$ and the assumptions that $\kappa \le \varepsilon < \gamma$ as before. □

Theorem 12. *Algorithm 3 can be efficiently implemented via the regression oracle and enjoys the same theoretical guarantees stated in Theorem 9. The number of oracle calls needed is $\widetilde{O}(\frac{\operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma^3})$ for a general set of regression functions $\mathcal{F}$, and $\widetilde{O}(\frac{\operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma})$ when $\mathcal{F}$ is convex and closed under pointwise convergence. The per-example inference time of the learned $\widehat{h}_M$ is $\widetilde{O}(\frac{1}{\gamma^2} \log^2(\frac{\operatorname{Pdim}(\mathcal{F})}{\varepsilon}))$ for general $\mathcal{F}$, and $\widetilde{O}(\log \frac{1}{\gamma})$ when $\mathcal{F}$ is convex and closed under pointwise convergence.*

Proof. Note that classifier $\widehat{h}_m$ and query function q_m in Algorithm 3 are constructed in the way as the ones in Algorithm 1, Thus, Algorithm 3 can be efficiently implemented in the same way as discussed in Theorem 5, and enjoys the same per-round computational complexities. The total computational complexity is then achieved by multiplying the per-round computational complexity by $T = \frac{\operatorname{Pdim}(\mathcal{F})}{\varepsilon \gamma}$. □

H.2 Discussion on $\kappa \le \varepsilon$

We provide guarantees (in Theorem 9) when $\kappa \le \varepsilon$ since the learned classifier suffers from an additive κ term in the excess error, as shown in the proof of Theorem 9. We next give preliminary discussions on this issue by relating active learning with to a (specific) regret minimization problem and connecting to existing lower bound in the literature. More specifically, we consider the perspective and notations discussed in Appendix G.1. Fix any epoch $m \ge 2$ and time step t within epoch m. We have

$$\mathbf{Regret}_t = \mathbb{E}[\ell_t(a_t) - \ell_t(a_t^\star) \mid \mathfrak{F}_{t-1}] = \operatorname{err}_\gamma(\widehat{h}_m) - \operatorname{err}(h^\star) = \mathsf{excess}_\gamma(\widehat{h}_m) = \widetilde{O}\left(\kappa + \frac{\bar{\theta}}{2^m \gamma}\right),$$

where the bound comes from similar analysis as in the proof of Theorem 9. Summing the instantaneous regret over T rounds, we have

$$\begin{aligned}
\mathbf{Regret}(T) &= \sum_{t=1}^{T} \mathbf{Regret}_t \\
&\le 2 + \sum_{m=2}^{M} (\tau_m - \tau_{m-1}) \cdot \mathsf{excess}_\gamma(\widehat{h}_m) \\
&\le \widetilde{O}\left(\kappa \cdot T + \frac{\bar{\theta}}{\gamma}\right).
\end{aligned}$$

The above bound indicates an additive regret term scales as $\kappa \cdot T$. On the other hand, it is known that an additive $\kappa \cdot T$ regret is in general unavoidable in linear bandits under model misspecification

(Lattimore et al., 2020). This connection partially explains/justifies why we only provide guarantee for Theorem 9 under $\kappa \leq \varepsilon$.

There are, however, many differences between the two learning problems. We list some distinctions below.

1. The regret minimization problem considered in Appendix G.1 only takes three actions $\mathcal{A} = \{+1, -1, \perp\}$, yet the lower bound in linear bandits is established with a large action set (Lattimore et al., 2020);
2. A standard contextual bandit problem will observe loss (with respect to the pulled action) at each step $t \in [T]$, however, the active learning problem will only observe (full) feedback at time steps when a query is issued, i.e., $\{t \in [T] : Q_t = 1\}$.

We leave a comprehensive study of the problem for feature work.

H.3 Supporting lemmas

We use the same notations defined in Appendix E, except $\widehat{h}_m$, g_m and β_m are defined differently. We adapt the proofs Theorem 4 (in Appendix E) to deal with model misspecification.

Note that although we do not have $f^\star \in \mathcal{F}$ anymore, one can still define random variables of the form $M_t(f)$, and guarantees in Lemma 4 still hold. We use $\mathcal{E}$ to denote the good event considered in Lemma 4, we analyze under this event through out the rest of this section. We also only analyze under the assumption of Theorem 9, i.e., $\kappa^2 \leq \varepsilon$.

Lemma 16. *Fix any epoch $m \in [M]$. We have*

$$\widehat{R}_m(\bar{f}) \leq \widehat{R}_m(f^\star) + \frac{3}{2} \cdot \kappa^2 \tau_{m-1} + C_\delta,$$

where $C_\delta := 8 \log\Big(\frac{|\mathcal{F}| \cdot T^2}{\delta}\Big)$.

Proof. From Lemma 4 we know that

$$\begin{aligned}
\widehat{R}_m(\bar{f}) - \widehat{R}_m(f^\star) &\leq \sum_{t=1}^{\tau_{m-1}} \frac{3}{2} \cdot \mathbb{E}_t\Big[Q_t\big(\bar{f}(x_t) - f^\star(x_t)\big)^2\Big] + C_\delta \\
&\leq \frac{3}{2} \cdot \kappa^2 \tau_{m-1} + C_\delta,
\end{aligned}$$

where we use the fact that $\mathbb{E}_t[y_t \mid x_t] = f^\star(x_t)$ (and thus $\mathbb{E}_t[M_t(\bar{f})] = \mathbb{E}_t[Q_t(\bar{f}(x_t) - f^\star(x_t))^2]$) on the first line; and use the fact $\sup_x |\bar{f}(x) - f^\star(x)| \leq \kappa$ on the second line. □

Lemma 17. *The followings hold true:*

1. $\bar{f} \in \mathcal{F}_m$ *for any* $m \in [M]$.
2. $\sum_{t=1}^{\tau_{m-1}} \mathbb{E}_t[M_t(f)] \leq 4\beta_m$ *for any* $f \in \mathcal{F}_m$.
3. $\sum_{t=1}^{\tau_{m-1}} \mathbb{E}[Q_t(x_t)(f(x_t) - \bar{f}(x_t))^2] \leq 9\beta_m$ *for any* $f \in \mathcal{F}_m$.
4. $\mathcal{F}_{m+1} \subseteq \mathcal{F}_m$ *for any* $m \in [M-1]$.

Proof. 1. Fix any epoch $m \in [M]$. By Lemma 4, we have $\widehat{R}_m(f^\star) \leq \widehat{R}_m(f) + C_\delta/2$ for any $f \in \mathcal{F}$. Combining this with Lemma 16 leads to

$$\begin{aligned}
\widehat{R}_m(\bar{f}) &\leq \widehat{R}_m(f) + \frac{3}{2} \cdot \big(\kappa^2 \tau_{m-1} + C_\delta\big) \\
&\leq \widehat{R}_m(f) + \beta_m,
\end{aligned}$$

for any $f \in \mathcal{F}$, where the second line comes from the definition of β_m (recall that we have $\kappa \leq \varepsilon$ by assumption). We thus have $\bar{f} \in \mathcal{F}_m$ for any $m \in [M]$.

2. Fix any $f \in \mathcal{F}_m$. With Lemma 4, we have

$$\begin{aligned}
\sum_{t=1}^{\tau_{m-1}} \mathbb{E}_t[M_t(f)] &\le 2 \sum_{t=1}^{\tau_{m-1}} M_t(f) + C_\delta \\
&= 2\widehat{R}_m(f) - 2\widehat{R}_m(f^\star) + C_\delta \\
&\le 2\widehat{R}_m(f) - 2\widehat{R}_m(\bar{f}) + 3\kappa^2\tau_{m-1} + 3C_\delta \\
&\le 2\widehat{R}_m(f) - 2\widehat{R}_m(\widehat{f}_m) + 3\kappa^2\tau_{m-1} + 3C_\delta \\
&\le 2\beta_m + 3\kappa^2\tau_{m-1} + 3C_\delta \\
&\le 4\beta_m,
\end{aligned}$$

where the third line comes from Lemma 16; the fourth line comes from the fact that $\widehat{f}_m$ is the minimizer of $\widehat{R}_m(\cdot)$; and the fifth line comes from the fact that $f \in \mathcal{F}_m$.

3. Fix any $f \in \mathcal{F}_m$. With Lemma 4, we have

$$\begin{aligned}
\sum_{t=1}^{\tau_{m-1}} \mathbb{E}_t[Q_t(x_t)(f(x_t) - \bar{f}(x_t))^2] &= \sum_{t=1}^{\tau_{m-1}} \mathbb{E}_t[Q_t(x_t)((f(x_t) - f^\star(x_t)) + (f^\star(x_t) - \bar{f}(x_t)))^2] \\
&\le 2 \sum_{t=1}^{\tau_{m-1}} \mathbb{E}_t[Q_t(x_t)(f(x_t) - f^\star(x_t))^2] + 2\tau_{m-1}\kappa^2 \\
&= 2 \sum_{t=1}^{\tau_{m-1}} \mathbb{E}_t[M_t(f)] + 2\tau_{m-1}\kappa^2 \\
&\le 8\beta_m + 2\tau_{m-1}\kappa^2 \\
&\le 9\beta_m,
\end{aligned}$$

where we use $(a+b)^2 \le a^2 + b^2$ on the second line; and use statement 2 on the fourth line.

4. Fix any $f \in \mathcal{F}_{m+1}$. We have

$$\begin{aligned}
\widehat{R}_m(f) - \widehat{R}_m(\widehat{f}_m) &\le \widehat{R}_m(f) - \widehat{R}_m(f^\star) + \frac{C_\delta}{2} \\
&= \widehat{R}_{m+1}(f) - \widehat{R}_{m+1}(f^\star) - \sum_{t=\tau_{m-1}+1}^{\tau_m} M_t(f) + \frac{C_\delta}{2} \\
&\le \widehat{R}_{m+1}(f) - \widehat{R}_{m+1}(\bar{f}) + \frac{3}{2}\kappa^2\tau_m + C_\delta - \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{E}_t[M_t(f)]/2 + C_\delta \\
&\le \widehat{R}_{m+1}(f) - \widehat{R}_{m+1}(\widehat{f}_{m+1}) + \frac{3}{2}\kappa^2\tau_m + 2C_\delta \\
&\le \beta_{m+1} + \frac{3}{2}\kappa^2\tau_m + 2C_\delta \\
&\le \beta_m,
\end{aligned}$$

where the first line comes from Lemma 4; the third line comes from Lemma 16 and Lemma 4; the fourth line comes from the fact that $\widehat{f}_{m+1}$ is the minimizer with respect to $\widehat{R}_{m+1}$ and Lemma 4; the last line comes from the definition of β_m.

□

Since the classifier $\widehat{h}_m$ and query function g_m are defined in the same way as in Algorithm 1, Lemma 6 holds true for Algorithm 3 as well. As a result of that, Lemma 7 and Lemma 8 hold true with minor modifications. We present the modified versions below, whose proofs follow similar steps as in Lemma 7 and Lemma 8 but replace $f^\star$ with $\widehat{f}$ (and thus using concentration results derived in Lemma 17).

Lemma 18. *Fix any epoch $m \geq 2$. We have*

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x)=1)] \leq \frac{36\beta_m}{\tau_{m-1}\,\gamma^2}\cdot\theta^{\mathrm{val}}_{\bar{f}}\Big(\mathcal{F},\gamma/2,\sqrt{\beta_m/\tau_{m-1}}\Big).$$

Lemma 19. *Fix any epoch $m \geq 2$. We have*

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathbb{1}(g_m(x)=1)\cdot w(x;\mathcal{F}_m)] \leq \frac{36\beta_m}{\tau_{m-1}\gamma}\cdot\theta^{\mathrm{val}}_{\bar{f}}\Big(\mathcal{F},\gamma/2,\sqrt{\beta_m/\tau_{m-1}}\Big).$$

Lemma 20. *Fix any $m \in [M]$. We have $\mathsf{excess}_\gamma(\widehat{h}_m;x) \leq 2\kappa$ if $g_m(x)=0$.*

Proof. Recall that

$$\begin{aligned}\mathsf{excess}_\gamma(\widehat{h};x) = {}& \mathbb{1}\big(\widehat{h}(x)\neq\perp\big)\cdot\big(\mathbb{P}_{y|x}\big(y\neq\mathrm{sign}(\widehat{h}(x))\big)-\mathbb{P}_{y|x}\big(y\neq\mathrm{sign}(h^\star(x))\big)\big)\\ &+\mathbb{1}\big(\widehat{h}(x)=\perp\big)\cdot\big(\big(1/2-\gamma\big)-\mathbb{P}_{y|x}\big(y\neq\mathrm{sign}(h^\star(x))\big)\big).\end{aligned}$$

We now analyze the event $\{g_m(x)=0\}$ in two cases.

Case 1: $\widehat{h}_m(x)=\perp$.

Since $\bar{f}(x)\in[\mathsf{lcb}(x;\mathcal{F}_m),\mathsf{ucb}(x;\mathcal{F}_m)]$ by Lemma 17, we know that $\eta(x)=f^\star(x)\in[\frac{1}{2}-\gamma-\kappa,\frac{1}{2}+\gamma+\kappa]$ and thus $\mathbb{P}_y\big(y\neq\mathrm{sign}(h^\star(x))\big)\geq\frac{1}{2}-\gamma-\kappa$. As a result, we have $\mathsf{excess}_\gamma(\widehat{h}_m;x)\leq\kappa$.

Case 2: $\widehat{h}_m(x)\neq\perp$ but $\frac{1}{2}\notin(\mathsf{lcb}(x;\mathcal{F}_m),\mathsf{ucb}(x;\mathcal{F}_m))$.

We clearly have $\mathsf{excess}_\gamma(\widehat{h}_m;x)\leq 0$ if $\mathrm{sign}(\widehat{h}_m(x))=\mathrm{sign}(h^\star(x))$. Now consider the case when $\mathrm{sign}(\widehat{h}_m(x))\neq\mathrm{sign}(h^\star(x))$. Since $\bar{f}(x)\in[\mathsf{lcb}(x;\mathcal{F}_m),\mathsf{ucb}(x;\mathcal{F}_m)]$ and $|\bar{f}(x)-f^\star(x)|\leq\kappa$, we must have $|f^\star(x)-1/2|\leq\kappa$ in that case, which leads to $\mathsf{excess}_\gamma(\widehat{h}_m;x)\leq 2|f^\star(x)-1/2|\leq 2\kappa$. □