Online Active Learning Beyond Realizability via Offline Regression

Shashaank Aiyer[1], Atul Ganju[1], Karthik Sridharan[1], and Ved Sriraman[1]

[1]Cornell University

Working Paper, January 28, 2025

Abstract

Active Learning is a supervised learning framework designed to model scenarios where acquiring labeled data is expensive. In the online case, the learner has access to a labeling oracle and engages in a learning protocol in which data is obtained sequentially. The learner's objective is to then output a classifier whose error is comparable to that of a passive learner, one that has access to labels for all data points, while minimizing the number of queries made to the labeling oracle.

Previous work demonstrates that, in the absence of noise, active learning algorithms cannot significantly reduce label complexity compared to passive learning. More recent approaches achieve improved label complexity by adopting some form of margin-based low-noise condition. These methods typically fall into one of two categories: (1) assuming access to a labeling oracle, which is widely considered intractable, or (2) assuming access to a regression oracle, which can be efficiently implemented for many benchmark function classes. However, the latter relies on the strong assumption that the benchmark regression class contains a function that perfectly models the data distribution.

In this work, we relax this strong realizability assumption. Specifically, we show that when the benchmark regression class is convex, it is possible to compete with it under much weaker assumptions about its relationship to the data distribution. Our algorithm achieves performance guarantees while making a polynomial improvement in label complexity, broadening the applicability of active learning in practical settings.

1 Introduction

Binary classification is a fundamental problem in machine learning. A classic result shows that the VC dimension governs learnability in this setting. Specifically, for hypothesis classes with finite VC dimension d, the excess-risk of the Empirical Risk Minimizer (ERM) is given by $\mathcal{O}(\sqrt{d/n})$. However, many relevant function classes have an infinite VC dimension, motivating the idea to develop techniques for learning that utilize additional information and learning assumptions.

An approach with great success has been reducing binary classification to regression — where the objective is to model the underlying conditional probability distribution over labels. Previous results show that by assuming the underlying conditional probability has low-noise and is realizable, meaning that it lies within a benchmark class accessible to the learner, regression-based algorithms can learn classifiers that achieve low excess-risk. Analyzing via regression induces additional structure onto the problem that facilitates the learning of model classes with infinite VC dimension, while having the added benefit of efficiency [due to the difficulty in optimizing the non-convex zero-one loss].

Reducing to regression has been explored in the active learning setting, which has gained recent interest due to its ability to capture environments where labels are often expensive to obtain. In this setting, the learner is presented with unlabeled data and is given access to a labeling oracle. The goal of the learner is to then output a classifier with low excess-risk on the underlying distribution while simultaneously making as few queries to the oracle as possible. Previous results have shown that under similar assumptions — low-noise and realizability — regression-based active learning algorithms provide label complexity bounds in terms of the *disagreement coefficient*, which, for many relevant hypothesis classes, is much easier to bound than

d. However, realizability from the standpoint of regression rarely holds in practice, as hidden variables not captured by the input space often shape real-world data, resulting in an inherently noisy data distribution. This raises a fundamental question: *under what minimal set of assumptions can regression-based algorithms effectively perform classification tasks in both passive and active learning?*

1.1 Related Work

To utilize the results from regression for classification settings, we need to establish that the regression solution can be carried over to the classification problem. This relies on the assumption that the classifier induced by the regression function best modeling the conditional probability is also optimal for the binary classification task. A central idea that establishes this is that the learner performs regression with respect to a surrogate loss that is *calibrated*. That is, the risk minimizing function over a calibrated surrogate loss induces the optimal binary classifier. This assumption alone facilitates slow rates on the order of ?? [**?**].

Subsequent works utilize realizability and low-noise assumptions to achieve fast rates on the order of ??. However, without this strong realizability assumption, our fast rates do not hold and we revert back to excess-risk bounds in terms of the VC dimension.

— Active learning — In the active learning setting, [it is generally impossible to get exponential speedup on the label complexity without additional assumptions], even if VC dimension is finite.

2 Learning Framework

Symbol	Meaning
X, Y, Z	attributes (variables)
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	sets of attributes
$Dom(X), Dom(\mathbf{X})$	their domains
$x \in Dom(X)$	an attribute value
$\mathbf{x} \in Dom(\mathbf{X})$	a tuple of attribute values
$\mathbf{k} \in Dom(\mathbf{K})$	a tuple of context attribute values
$\tilde{f}_m = \text{argmin}_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$	minimizer over the subdistribution
k_m	number of queries made in epoch m

Table 1: **Notation used in this section.**

2.1 Problem Setting

Let $\mathcal{X}$ denote the space of inputs and $\mathcal{Y}$ denote the space of labels. We focus on the problem of online binary classification, where $\mathcal{Y} = \{-1, +1\}$ and data points are generated i.i.d from a distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$. We consider two frameworks: in the first, learner is given access to a dataset $D = \{(x_1, y_1), \ldots, (x_T, y_T)\}$; and another where, in a T-round protocol, on each round t, nature samples $(x_t, y_t) \sim \mathcal{D}$, and the learner is given x_t and chooses to optionally observes y_t and incur a unit cost. In both settings, the goal of the learner is to output a classifier $\hat{h} : \mathcal{X} \to \mathcal{Y}$ that attains low 0-1 excess risk on the distribution $\mathcal{D}$ against a hypothesis class $\mathcal{H}$

$$\mathcal{E}(\hat{h}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}\{\hat{h}(x) \neq y\}] - \min_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}].$$

In the second framework however, the learner simultaneously wants to minimize the number of points it queries the label of. We denote the distribution over inputs as $\mathcal{D}_{\mathcal{X}}$ and the conditional probability function of each label given an input as $\eta(x) := \mathbb{P}_{X\sim\mathcal{D}_{\mathcal{X}}}[y = 1|X = x]$.

2.2 Reducing Classification to Regression

We focus on the case where the hypothesis class $\mathcal{H}$ is induced by a class of regression functions $\mathcal{F} : \mathcal{X} \to [0, 1]$ which aim to model the conditional probability $\eta(x)$. Adopting the same notation as [9], we note $\mathcal{H} = \mathcal{H}_{\mathcal{F}} := \{h_f : f \in \mathcal{F}\}$ where $h_f(x) = \text{sign}(2f(x) - 1)$. The "size" of $\mathcal{F}$ is measured by the well-known complexity measure: the *Pseudo dimension* $\text{Pdim}(\mathcal{F})$ [6] [2] [3]. We assume $\text{Pdim}(\mathcal{F}) < \infty$ throughout the paper. In order to attain meaningful bounds on excess risk, it is necessary to make the following assumption.

Assumption 2.1 (Sign Assumption)**.** *Take* $f^* := \text{argmin}_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$*, then* $h_{f^*} = h_\eta$ *almost surely.*

Without this, the function minimizing squared loss would not induce an optimal classifier in $\mathcal{H}$, preventing us from interchanging between regression and classification in a useful manner. Past work relies on realizability to establish this relationship, but we instead assume it directly.

However, even with this assumption or realizability, if there is no bound on the noise in the data distribution, functions that are nearly identical to f^* can induce vastly different classifiers, leading to an arbitrarily high sample complexity for regression-based function approximation techniques. Hence, it becomes essential to adopt a low-noise assumption. Our result relies on the following one, although they can be extended to other margin-based noise conditions such as the Tsybakov noise condition [8].

Assumption 2.2 (Massart's Noise Condition, [5])**.** *For some* $\gamma > 0$*,* $\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[|\eta(x) - 1/2| < \gamma] = 0$.

We refer to γ as the hard margin of our distribution, as intuitively it establishes a separation between class probabilities.

However, note that what we actually need a $\gamma-$margin on f^*. In realizable settings, Assumption 2.7 trivially implies a $\gamma-$margin on f^*, since $f^* = \eta$, which is assumed to have $\gamma-$margin. In fact, this holds even in settings that slightly relax realizability, which we call ϵ-realizable.

Definition 2.3. *We call a setting $\epsilon-$realizable if, for some small ϵ, there exists $f \in \mathcal{F}$ such that*

$$\sup_{x \in \mathcal{D}_x} |f(x) - \eta(x)| < \epsilon$$

NOTE: Explain why we don't want to use ϵ-realizability

Without realizability or $\epsilon-$realizability, we no longer have an implied margin on f^*, which motivates the following assumption.

Assumption 2.4. *For some fixed $\alpha \in [0, 1]$,*

$$\left| f^*(x) - \frac{1}{2} \right| \geq \alpha \cdot \left| \eta(x) - \frac{1}{2} \right|$$

Intuitively, we require that f^* is some fraction as confident as η, which in turn implies a $\frac{\gamma}{\alpha}$-margin on f^*, as needed. The example below shows that Assumption 2.4 is in fact more general than assuming $\epsilon-$realizability.

Example 2.5 (Our condition holds but $\epsilon-$realizability does not)**.** *Our objective is to demonstrate that there exists a setting for which $\epsilon-$realizability does not hold but Assumption 2.4 does. Consider $\mathcal{X} = [-1, +1]$, $f_\eta(x) = x$, and $\mathcal{D}_\mathcal{X}$ to be the uniform distribution. The class of functions that we are trying to learn, $\mathcal{F}$, is the class of piecewise constant functions on the intervals $x > 0$ and $x < 0$, while discontinuous at $x = 0$. It is easy to verify that $f^*(x) = \text{argmin}_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}} \left[(f(x) - f_\eta(x))^2 \right]$ satisfies*

$$f^*(x) = \begin{cases} -\frac{(1+\gamma)}{2}, & x < 0, \\ \frac{(1+\gamma)}{2}, & x > 0. \end{cases}$$

First, we show that our condition is satisfied for some $\alpha \in [0, 1]$ and for all $x \in \mathcal{X}$. That is, we need to show that $|f^*(x)| \geq \alpha |\eta(x)|$, or $\frac{|f^*(x)|}{|\eta(x)|} \geq \alpha$, which implies $\frac{|\frac{1+\gamma}{2}|}{|x|} \geq \alpha$. Observe that for $|x| = 1, |\frac{1+\gamma}{2}| \geq \alpha$, and for $|x| < 1$ while x moves away from 1 toward 0, the fraction increases, and so the upper bound on α will become larger.

Hence, the minimal ratio $\frac{|f^*(x)|}{|\eta(x)|}$ over all $x \in [-1, +1] \setminus \{0\}$ is actually attained at $|x| = 1$, where it equals 0.5. Therefore, we can simply take $\alpha = 0.5$. Then for every $x \neq 0$,

$$|f^*(x)| = 0.5 \geq 0.5|x| = \alpha |f_\eta(x)|$$

Hence, our condition is satisfied for all $x \in \mathcal{X}$ for $\alpha = 0.5$.

However, to satisfy $\epsilon-$realizability, we would need

$$|f(x) - f_\eta(x)| = \begin{cases} |-\frac{(1+\gamma)}{2} - x|, & x < 0, \\ |\frac{(1+\gamma)}{2} - x|, & x > 0 \end{cases} \leq \epsilon \quad \text{for all} \quad x \in [-1, 1]$$

However, if we look at $x = 1$, then $|f(x) - f_\eta(x)| = |\frac{(1+\gamma)}{2} - 1| = |\frac{(\gamma-1)}{2}|$. So if $\epsilon-$realizability were to hold for all x, we would need $\epsilon \geq 0.5$.

Now that we have established the power of regression in this setting, NOTE: Make the transition to the active learning setting

Diverging from the assumptions made in existing literature, we make the following structural assumption on the class $\mathcal{F}$.

Assumption 2.6 (Convexity of $\mathcal{F}$)**.** *The set of regression functions $\mathcal{F}$ is convex. That is, for any $\alpha \in [0, 1]$, and $f_1, f_2 \in \mathcal{F}$ the function $\alpha f_1 + (1 - \alpha) f_2 \in \mathcal{F}$.*

This assumption has previously been shown to reduce the problem of vanilla binary classification under indicator loss to squared loss regression when paired with the following assumption we will also make,

Assumption 2.7 (Massart's Noise Condition, [5]). *For some $\gamma > 0$, $\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[|\eta(x) - 1/2| < \gamma] = 0$.*

Essentially, we are saying that the probability under the input distribution $\mathcal{D}$ of sampling a point x for which the label is not γ-biased is 0.

Assumption 2.8 (Expressivity of $\mathcal{F}$ (more general version)). *For any set $C \subseteq \mathcal{X}$, take*

$$S_C := \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}\{x \in C\} \cdot (h_f(x) - y)^2\right]$$

Then, for any $\tilde{f} \in S_C$, for some fixed $\alpha \in [0, 1]$, and for all $x \in C$,

1. $h_{\tilde{f}}(x) = h_\eta(x)$
2. $\left|\tilde{f}(x) - \frac{1}{2}\right| \geq \alpha \cdot \left|\eta(x) - \frac{1}{2}\right|$

This assumption implies that, for any algorithm with a deterministic query condition determined by the history, the optimal model in $\mathcal{H}$ on the each point in the sub-distribution induced by the querying condition on a given round is biased in the same way as h_η on the data points observed. Additionally, this assumption combined with Assumption 2.7 implies that the optimal model ... We formalize these ideas in the following two lemmas.

2.3 Diverging from $\epsilon-$Realizability

Past work has aimed to address the impracticality of realizability by replacing it with a slightly weaker assumption known as ϵ-realizability (or misspecification), defined formally below.

Definition 2.9. *We say that $h \in \mathcal{H}$ is $\epsilon-$realizable if*

$$\sup_{x\in\mathcal{D}_x} |f(x) - f_\eta(x)| < \epsilon$$

We now claim that Assumption 2.8 is in fact more general than assuming $\epsilon-$realizability. This involves showing two things.

1. Assumption 2.8 implies ϵ-realizability
2. There exists a setting for which $\epsilon-$realizability does not hold but Assumption 2.8 does

We now show that with the assumptions given above, we can provide an algorithm that performs efficient selective sampling.

Example 2.10 (Our condition holds but $\epsilon-$realizability does not). *Our objective is to demonstrate that there exists a setting for which $\epsilon-$realizability does not hold but Assumption 2.8 does. Consider $\mathcal{X} = [-1, +1]$, $f_\eta(x) = x$, $\mathcal{D}_\mathcal{X}$ to be the uniform distribution, and $\mathcal{F}$ the class of piecewise constant functions on the intervals $x > 0$ and $x < 0$, while discontinuous at $x = 0$. It is easy to verify that $f^*(x) = \operatorname{argmin}_{f\in\mathcal{F}} \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}} \left[(f(x) - f_\eta(x))^2\right]$ satisfies*

$$f^*(x) = \begin{cases} -\frac{(1+\gamma)}{2}, & x < 0, \\ \frac{(1+\gamma)}{2}, & x > 0. \end{cases}$$

We must show that our condition is satisfied for some $\alpha \in [0, 1]$ and for all $x \in \mathcal{X}$. That is, $|f^*(x)| \geq \alpha|\eta(x)|$, or $\frac{|f^*(x)|}{|\eta(x)|} \geq \alpha$, which implies $\frac{|\frac{1+\gamma}{2}|}{|x|} \geq \alpha$. Observe that for $|x| = 1$, $|\frac{1+\gamma}{2}| \geq \alpha$, and for $|x| < 1$ while x moves away from 1 toward 0, the fraction increases, and so the upper bound on α will become larger.

Hence, the minimal ratio $\frac{|f^*(x)|}{|\eta(x)|}$ over all $x \in [-1, +1] \setminus \{0\}$ is actually attained at $|x| = 1$, where it equals 0.5. Therefore, we can simply take $\alpha = 0.5$. Then for every $x \neq 0$,

$$|f^*(x)| = 0.5 \geq 0.5|x| = \alpha|f_\eta(x)|$$

Hence, our condition is satisfied for all $x \in \mathcal{X}$ for $\alpha = 0.5$.

However, to satisfy $\epsilon-$realizability, we would need

$$|f(x) - f_\eta(x)| = \begin{cases} |-\frac{(1+\gamma)}{2} - x|, & x < 0, \\ |\frac{(1+\gamma)}{2} - x|, & x > 0 \end{cases} \leq \epsilon \quad \textit{for all} \quad x \in [-1, 1]$$

However, if we look at $x = 1$, then $|f(x) - f_\eta(x)| = |\frac{(1+\gamma)}{2} - 1| = |\frac{(\gamma-1)}{2}|$. So if $\epsilon-$realizability were to hold for all x, we would need $\epsilon \geq 0.5$.

Lemma 2.11. *Let $\mathcal{D} \in \Delta\mathcal{X} \times \mathcal{Y}$, $\mathcal{F}$ be a class of functions that satisfies Assumption 2.6, and $f^* := \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$. The following inequality holds for any $f \in \mathcal{F}$.*

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[(f(x) - f^*(x))^2] \leq \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2 - (f^*(x) - y)^2]$$

Proof. We can view f^* as the projection of $y \sim \mathcal{D}_\mathcal{Y}$ onto the convex set $\mathcal{F}$. It follows by the Law of Cosines that for any $f \in \mathcal{F}$

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2] = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[(f(x) - f^*(x))^2] + \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y^2)] - 2\mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - f^*(x))(f^*(x) - y)]$$

Rearranging, we get

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[(f(x) - f^*(x))^2] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2] - \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f^*(x) - y^2)] + 2\mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - f^*(x))(f^*(x) - y)]$$

Now, we use the fact that because $\mathcal{F}$ is a convex set, and f^* is interpreted as the closest point in the set to y, $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - f^*(x))(f^*(x) - y)] \leq 0$, which gives us our result. □

3 Agnostic Selective Sampling

In this section, we provide our main algorithm and prove, under the conditions outlined in **??**, that it achieves an excess risk of ϵ with a query complexity of TBD.

Offline Regression Oracle: Our algorithm makes use of the primitive of an *offline regression oracle* over $\mathcal{F}$. Specifically, for any set S of weighted examples $(w, x, y) \in \mathbb{R}^+ \times \mathcal{Y} \times \mathcal{Y}$, we have an oracle which outputs,

$$\hat{f} := \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{(w,x,y) \in S} w(f(x) - y)^2$$

This primitive has been studied extensively and is known to exist for function classes with low complexity [**?**]. As a result, we view a call to the oracle as an efficient operation and quantify the computational complexity of our algorithm in terms of the number of calls to this oracle.

3.1 Algorithm Overview

Algorithm 1 runs in epochs of geometrically increasing lengths and is a modification of the algorithm from [9] that performs active learning with abstention. At the beginning of each epoch $m \in [M]$, the offline regression oracle is used to obtain the function $\hat{f}_m \in \mathcal{F}$ with the smallest cumulative squared loss on the data points in the previous epoch whose labels were queried. Then, an implicit class of regression functions $\mathcal{F}_m \subseteq \mathcal{F}$ is constructed by including every function in $\mathcal{F}$ that attains a cumulative squared loss on the queried points in the previous epoch that is only β_m larger than the squared loss of $\hat{f}_m$. For every $x \in \mathcal{X}$, the algorithm uses the class of regression functions $\mathcal{F}_m$ to obtain both a new upper confidence bound $\text{ucb}(x, \mathcal{F}_m) = \sup_{f \in \mathcal{F}_m} f(x)$ and lower confidence bound $\text{lcb}(x, \mathcal{F}_m) = \inf_{f \in \mathcal{F}_m} f(x)$ on the probability $\eta(x)$. Intuitively, this confidence interval captures the disagreement among our remaining set of hypotheses on this particular x. From this, the algorithm amends its query condition $q_m : \mathcal{X} \to \{0, 1\}$. This query condition fires on any $x \in \mathcal{X}$ for which, for every $i \in [m]$, there exists a pair of functions $f, f' \in \mathcal{F}_i$ that induces classifiers $h_f, h_{f'}$ that classify x differently. Then, for each data point observed in epoch m, the classifier only queries its label if the query condition is satisfied. After all m epochs, the data of the last epoch is used one last time to create a final query condition q_{M+1}. Finally, the classifier $\hat{h}$ outputted by the algorithm is the one which, on any $x \in \mathcal{X}$, looks at the smallest i for which there did not exist a pair of functions $f, f' \in \mathcal{F}_i$ that induces classifiers $h_f, h_{f'}$ that classify x differently. If such an i exists, it outputs the classification of the consensus of the classifiers induced by the regression functions in $\mathcal{F}_i$; otherwise it outputs 1.

The algorithm follows a general design principle used when making selective sampling algorithms: specifically, on each round $t \in T$, if the algorithm has enough information to classify the point x_t with high probability, it will deterministically not query x_t. The query condition, $q_{m(t)}$, indicates whether or not we query for the expert label at round t, where $m(\cdot)$ is the function that maps round t to the epoch m it takes place in. As a result, for any epoch m, since the query condition remains constant, the observed data points can be thought of as coming i.i.d. from the same distribution over the input space. We will denote $\mathcal{D}_m$ to the distribution induced by the query condition q_m on the m-th epoch. This distribution would have a density function that is 0 on all points q_m tells the algorithm not to query and is proportional to the the original data distribution on points q_m tells the algorithm to query.

3.2 Overview of Algorithm Analysis

To see why the output classifier $\hat{h}$ can be shown to have low excess risk, consider the high probability event in which the minimizer of the expected squared loss on $\mathcal{D}_m$ is in $\mathcal{F}_m$ for all $m \in [M+1]$. Under this event, for any $x \in \mathcal{X}$, we are guaranteed that $\tilde{f}_m(x)$ is in the confidence interval $[\text{lcb}(x; \mathcal{F}_m), \text{ucb}(x; \mathcal{F}_m)]$ for all $m \in [M+1]]$. Then, error we will have on any $x \in \mathcal{X}$ will fall into one of two cases,

- **Case 1: (Label of x is not queried)** In this case, there must exist an $m \in [M+1]]$ for which $\frac{1}{2} \notin [\text{lcb}(x; \mathcal{F}_m), \text{ucb}(x; \mathcal{F}_m)]$. So, $\hat{h}(x) = h_{\hat{f}_m}(x) = h_{\tilde{f}_m}(x)$. Then, by Assumption 2.8, we know $h_{\tilde{f}_m}(x) = h_\eta(x)$ implying we make no error on x.

- **Case 2: (Label of x is queried)** In this case, although we accumulate error, we show that given **??**, this event happens very infrequently.

Algorithm 1 Agnostic Selective Sampling in Epochs

Parameters: Learning rate $\gamma > 0$, error rate $\delta \in (0, 1)$
Define $\tau_m = 2^m - 1, \tau_{-1} = \tau_0 = 0$.
for $m = 1, \ldots, M + 1$ **do**
    Obtain the empirical risk minimizer on observed data in the previous epoch:

$$\hat{f}_m := \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{t=\tau_{m-2}+1}^{\tau_{m-1}} Q_t (f(x_t) - y_t)^2$$

    Implicitly construct the set of regression functions: $\mathcal{F}_m \subseteq \mathcal{F}$ as:

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

    Construct query function $q_m(x) : \mathcal{X} \to \{0, 1\}$ as:

$$q_m(x) := \prod_{i=1}^{m} \mathbb{1} \left\{ \frac{1}{2} \in [\mathrm{lcb}(x; \mathcal{F}_i), \mathrm{ucb}(x; \mathcal{F}_i)] \right\}$$

    if $m = M + 1$ **then**
        Define the function $\hat{f} : \mathcal{X} \to [0, 1]$ to be:

$$\hat{f}(x) = \begin{cases} 1 & \text{if } q_{M+1}(x) = 1 \\ \hat{f}_i(x) & \text{if } i := \min \left\{ m \in [M+1] : \frac{1}{2} \notin [\mathrm{lcb}(x; \mathcal{F}_m), \mathrm{ucb}(x; \mathcal{F}_m)] \right\} \end{cases}$$

        return $h_{\hat{f}}(x)$
    else
        for $t = \tau_{m-1} + 1, \ldots, \tau_m$ **do**
            Receive x_t for $(x_t, y_t) \sim \mathcal{D}$
            if $g_m(x_t) = 1$ **then**
                Query the label y_t of x_t

3.3 Analysis

Theorem 3.1. *MAIN THEOREM!*

Proof. Consider the classifier $h_{\hat{f}}$ outputted by Algorithm 1. By **??** we have,

$$\begin{aligned}\mathcal{E}(h_{\hat{f}}) &\leq \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{h_{\hat{f}}(x)\cdot h^*(x)=-1\}\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{q_{M+1}(x)=0, h_{\hat{f}}(x)\cdot h^*(x)=-1\}+\mathbb{1}\{q_{M+1}(x)=1, h_{\hat{f}}(x)\cdot h^*(x)=-1\}\right],\end{aligned}$$

where we split whether the classifier $h_{\hat{f}}$ would have queried. We will now separately bound the excess risk incurred when the classifier would have chosen not to query, i.e. when $q_{M+1}(x)=0$, and when it would have chosen to query, i.e. when $q_{M+1}(x)=1$, under the high probability event of Lemma 4.2.

For any $x\in\mathcal{X}$ such that $q_{M+1}(x)=0$, there exists an m such that every function in $\mathcal{F}_m$ agrees on the classification of x. Now take i to be the smallest such m.

We begin by bounding the excess risk incurred when the classifier would have chose not to query. For any $x\in\mathcal{X}$ such that $q_{M+1}(x)=0$, there exists an $m\in[M+1]$ such that $\frac{1}{2}\notin[\text{lcb}(x;\mathcal{F}_m),\text{ucb}(x;\mathcal{F}_m)]$ or in other words there exists an m such that every function in $\mathcal{F}_m$ agrees on the classification of x. Now take i to be the smallest such m. By Lemma 4.3, we know that $\tilde{f}_{i-1}\in\mathcal{F}_i$ and therefore that $h_{\tilde{f}_{i-1}}(x)=h_{\hat{f}_i}(x)=h_{\hat{f}}(x)$. Finally, by Assumption 2.1, we know that $h_{\tilde{f}_{i-1}}(x)=h_\eta(x)=h_{f^*}(x)$, implying that the classifier does not incur risk when it would not have queried.

The excess risk incurred when the classifier would have chosen to query can be bounded by the probability it would query a data point,

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{q_{M+1}(x)=1, h_{\hat{f}}(x)\cdot h^*(x)=-1\}\right]\leq\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathbb{1}\{q_{M+1}(x)=1\}\right]=\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_{M+1}(x)=1].$$

By Lemma 4.5, we know that

□

4 Intermediate Lemmas

4.1 Concentration Lemma

Lemma 4.1 ([1, 7]). *Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, $\mathcal{F} \subseteq \{f \mid f : \mathcal{X} \to [0,1]\}$, $n \geq 2$, and $\delta \in (0,1)$. Then with probability at least $1-\delta$ over samples $S = \{(x_i, y_i)\}_{i\in[n]}$ drawn i.i.d. from $\mathcal{D}$, the following inequalities hold for all $f, f' \in \mathcal{F}$:*

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[(f(x) - f'(x))^2] \leq 2 \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f'(x_i))^2\right) + \frac{C(\mathcal{F}, \delta, n)}{n},$$

and,

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \tilde{f}_m(x_i))^2 \leq 2 \cdot \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[(f(x) - f'(x))^2] + \frac{C(\mathcal{F}, \delta, n)}{n},$$

where $C(\mathcal{F}, \delta, n) = C'(C'' \log^3(n)\mathfrak{R}^2(\mathcal{F}) + \log(1/\delta) + \log\log n)$ for some absolute constant C', using the result of Lemma 8 part (i) from [7].

Lemma 4.2. *Fix any $\delta \in (0,1)$. Then, for all $m \in [M]$, if we denote k_m as the number of queries made by Algorithm 1 in epoch m, with probability $1-\delta$, we have:*

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}\left[(f(x) - \tilde{f}_m(x))^2\right] \leq 2 \cdot \left(\frac{1}{k_m}\sum_{t=\tau_{m-1}+1}^{\tau_m}\mathbb{1}\{x_t \in \mathcal{X}_m\} \cdot (f(x_t) - \tilde{f}_m(x_t))^2\right) + \frac{C(\mathcal{F}, \delta, k_m)}{k_m}.$$

and,

$$\frac{1}{k_m} \cdot \sum_{t=\tau_{m-1}+1}^{\tau_m}\mathbb{1}\{x_t \in \mathcal{X}_m\} \cdot (f(x_t) - \tilde{f}_m(x_t))^2 \leq 2 \cdot \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[(f(x) - f'(x))^2] + \frac{C(\mathcal{F}, \delta, k_m)}{k_m},$$

where $C(\mathcal{F}, \delta, k_m) = C'(C'' \log^3(k_m)\mathfrak{R}^2(\mathcal{F}) + \log\log T + \log(1/\delta) + \log\log k_m)/k_m$ for some absolute constant C'.

Proof. For any $f \in \mathcal{F}_m$, consider the average squared distance between f and $\tilde{f}_m$ over labeled data observed during epoch m. Since, each data point (x_t, y_t) of this epoch is sampled from $\mathcal{D}$, and its label y_t is observed exactly when $q_m(x_t) = 1$, we can imagine each data point whose label was observed as being sampled from $\mathcal{D}_m$ – the original data distribution $\mathcal{D}$ normalized after being restricted to the set $\mathcal{X}_m$. Therefore, if we denote the rounds for which Algorithm 1 did query in epoch m as $t_1^m, \ldots, t_{k_m}^m$ we have:

$$\frac{1}{k_m} \cdot \sum_{t=\tau_{m-1}+1}^{\tau_m}\mathbb{1}\{x_t \in \mathcal{X}_m\} \cdot (f(x_t) - \tilde{f}_m(x_t))^2 = \frac{1}{k_m} \cdot \sum_{i=1}^{k_m}(f(x_{t_i^m}) - \tilde{f}_m(x_{t_i^m}))^2.$$

Applying the upper and lower bounds on this quantity by Lemma 4.1, we get our desired result. □

4.2 Other Lemmas

Lemma 4.3. *Under the high probability event of Lemma 4.2, it is true that for any $m \in [M]$, $\tilde{f}_m \in \mathcal{F}_{m+1}$.*

Proof. From Lemma 4.2, we have,

$$\begin{aligned}\sum_{t=\tau_{m-1}+1}^{\tau_m} Q_t(\hat{f}_{m+1}(x_t) - \tilde{f}_m(x_t))^2 &= \sum_{i=1}^{k_m}(\hat{f}_{m+1}(x_{t_i^m}) - \tilde{f}_m(x_{t_i^m}))^2 + C_2(\delta, \mathcal{F}) \\ &\leq 2k_m \cdot \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}[(\tilde{f}_m(x) - \hat{f}_{m+1}(x))^2] + C(\mathcal{F}, \delta, k_m) \\ &\leq 2k_m \cdot \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}_m}}[(\hat{f}_{m+1}(x) - y)^2 - (\tilde{f}_m(x) - y)^2] + C(\mathcal{F}, \delta, k_m) \\ &= B_{m+1},\end{aligned}$$

where the second inequality comes from convexity of our function class and the final inequality comes from a bound on the excess risk of the ERM. □

Lemma 4.4. *For all $m \in M$, with probability $1-\delta$, $k_m \geq (1-\epsilon)\cdot\mathbb{E}[k_m] = (1-\epsilon)\cdot(\tau_m - \tau_{m-1})\cdot\mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_m(x) = 1]$, where $\epsilon = \sqrt{\frac{-2\log(\delta/M)}{\mathbb{E}[k_m]}} \in (0,1)$.*

Proof. Let k_m denote the random variable of interest, which is the number of times the learner queries for the label in epoch m, and note that its expected value satisfies

$$\frac{\mathbb{E}[k_m]}{(\tau_m - \tau_{m-1})} = \mathbb{P}_{x\sim\mathcal{D}_{\mathcal{X}}}[q_m(x) = 1].$$

By the Chernoff tail bound for a sum of independent Bernoulli random variables, we have $\Pr\left(k_m < (1-\epsilon)\cdot\mathbb{E}[k_m]\right) \leq \exp\left(-\frac{\epsilon^2\cdot\mathbb{E}[k_m]}{2}\right)$, where $\epsilon \in (0,1)$. We can define the bad event for each m as $E_m = \{k_m < (1-\epsilon)\cdot\mathbb{E}[k_m]\}$, so the bound controls the probability of each individual bad event. To ensure the inequality holds for all $m \in M$, we use the union bound over the bad events, substituting the Chernoff bound for each $\Pr(E_m)$:

$$\Pr\left(\bigcup_{m\in M} E_m\right) \leq \sum_{m\in M} \exp\left(-\frac{\epsilon^2\cdot\mathbb{E}[k_m]}{2}\right).$$

Setting $\epsilon = \sqrt{\frac{-2\log(\delta/M)}{\mathbb{E}[k_m]}}$ and substituting this ϵ into the bound for $\Pr(E_m)$, we get:

$$\Pr(E_m) \leq \exp\left(-\frac{\left(\sqrt{\frac{-2\log(\delta/M)}{\mathbb{E}[k_m]}}\right)^2\mathbb{E}[k_m]}{2}\right) = \frac{\delta}{M}.$$

Thus, using the union bound:

$$\Pr\left(\bigcup_{m\in M} E_m\right) \leq \sum_{m\in M}\frac{\delta}{M} = \delta.$$

The complement of the union of bad events is the event where all k_m satisfy $k_m \geq (1-\epsilon)\cdot\mathbb{E}[k_m]$. Therefore:

$$\Pr\left(\bigcap_{m\in M}\{k_m \geq (1-\epsilon)\cdot\mathbb{E}[k_m]\}\right) = 1 - \Pr\left(\bigcup_{m\in M} E_m\right) \geq 1-\delta.$$

□

Lemma 4.5. *With probability $1-\delta$, for all $m \geq 2$, we have:*

$$\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1] \leq \frac{C_m^*}{\alpha\beta\gamma\cdot 2^{m-2}}\cdot\theta^{1/2}\left(\mathcal{F}_m,\alpha\gamma,\frac{C_m^*}{k_{m-1}}\right)$$

Proof. By the construction of Algorithm 1, we can decompose the query condition on epoch m as,

$$\begin{aligned}
\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1] &= \mathbb{P}_{x\sim D_\mathcal{X}}\left[\frac{1}{2}\in[\mathrm{lcb}(x;\mathcal{F}_m),\mathrm{ucb}(x;\mathcal{F}_m)] \;\middle|\; q_{m-1}(x)=1\right]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]\\
&= \mathbb{P}_{x\sim D_\mathcal{X}}\left[\frac{1}{2}\in[\mathrm{lcb}(x;\mathcal{F}_m),\mathrm{ucb}(x;\mathcal{F}_m)] \;\middle|\; x\in\mathcal{X}_{m-1}\right]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]\\
&= \mathbb{P}_{x\sim D_{\mathcal{X}_{m-1}}}\left[\frac{1}{2}\in[\mathrm{lcb}(x;\mathcal{F}_m),\mathrm{ucb}(x;\mathcal{F}_m)]\right]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1].
\end{aligned}$$

We now bound the first term in this product. Recall that from Lemma 4.3, we have $\tilde{f}_{m-1}\in\mathcal{F}_m$ for all $m\in[M+1]$. Then, for any $x\in\mathcal{X}_{m-1}$ for which $\frac{1}{2}\in[\mathrm{lcb}(x;\mathcal{F}_m)],\mathrm{ucb}(x;\mathcal{F}_m)]$, there must exist a function $f\in\mathcal{F}_m$ for which $|\tilde{f}_{m-1}(x)-f(x)|\geq|\tilde{f}_{m-1}(x)-\frac{1}{2}|$. Furthermore, we know by Assumption 2.8 that $|\tilde{f}_{m-1}(x)-\frac{1}{2}|\geq\alpha\cdot|\eta(x)-\frac{1}{2}|>\alpha\gamma$. However, since $f\in\mathcal{F}_m$, by Lemma 4.6 we also know an upper bound on $\|\tilde{f}_m-f\|_{\mathcal{D}_{\mathcal{X}_{m-1}}}$. Therefore, we can bound by the probability these two events happen simultaneously. Specifically, take $C_m^*:=8B_m+C(\mathcal{F},\delta,k_{m-1})$, then

$$\begin{aligned}
&\mathbb{P}_{x\sim D_{\mathcal{X}_{m-1}}}\left[\frac{1}{2}\in[\mathrm{lcb}(x;\mathcal{F}_m),\mathrm{ucb}(x;\mathcal{F}_m)]\right]\\
&\leq \mathbb{P}_{x\sim D_{\mathcal{X}_{m-1}}}\left[\exists f\in\mathcal{F}_m:|f(x)-\tilde{f}_{m-1}(x)|>\alpha\gamma,\|f-\tilde{f}_{m-1}\|_{\mathcal{D}_{\mathcal{X}_{m-1}}}\leq\frac{C_m^*}{k_{m-1}}\right]\\
&\leq\left(\frac{C_m^*}{\alpha\gamma k_{m-1}}\right)^2\cdot\theta\left(\mathcal{F}_m,\alpha\gamma,\frac{C_m^*}{k_{m-1}}\right).
\end{aligned}$$

Now, from Lemma 4.4, we know that, with probability at least $1-\delta$, if $\epsilon=\sqrt{-\frac{2\log(\delta/M)}{\mathbb{E}[k_m]}}$, then the following inequality holds:

$$k_{m-1}\geq(1-\epsilon)\cdot\mathbb{E}[k_m]=(1-\epsilon)\cdot(\tau_{m-1}-\tau_{m-2})\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1].$$

If $\epsilon\leq\frac{1}{2}$, then:

$$\left(\frac{C_m^*}{\alpha\gamma k_{m-1}}\right)^2\leq\left(\frac{2C_m^*}{\alpha\gamma\cdot(\tau_{m-1}-\tau_{m-2})\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]}\right)^2.$$

Plugging this upper bound back in for the first term gives us,

$$\begin{aligned}
&\mathbb{P}_{x\sim D_{\mathcal{X}_{m-1}}}\left[\frac{1}{2}\in[\mathrm{lcb}(x;\mathcal{F}_m),\mathrm{ucb}(x;\mathcal{F}_m)]\right]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]\\
&\leq\left(\frac{2C_m^*}{\alpha\gamma\cdot(\tau_{m-1}-\tau_{m-2})\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]}\right)^2\cdot\theta\left(\mathcal{F}_m,\alpha\gamma,\frac{C_m^*}{k_{m-1}}\right)\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]\\
&=\left(\frac{2C_m^*}{\alpha\gamma\cdot 2^{m-2}\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]}\right)^2\cdot\theta\left(\mathcal{F}_m,\alpha\gamma,\frac{C_m^*}{k_{m-1}}\right)\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1].
\end{aligned}$$

Multiplying both sides by $\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]$ yields,

$$\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1]\cdot\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]\leq\left(\frac{C_m^*}{\alpha\gamma\cdot 2^{m-3}}\right)^2\cdot\theta\left(\mathcal{F}_m,\alpha\gamma,\frac{C_m^*}{k_{m-1}}\right),$$

where since $\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1] \leq \mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_{m-1}(x)=1]$, we can replace the latter with the former. Then, by taking the square root of both sides, we have,

$$\mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1] \leq \frac{C_m^*}{\alpha\gamma\cdot 2^{m-3}}\cdot\theta^{1/2}\left(\mathcal{F}_m, \alpha\gamma, \frac{C_m^*}{k_{m-1}}\right),$$

On the other hand, if $\epsilon > \frac{1}{2}$, then:

$$\frac{1}{2} < \sqrt{-\frac{2\log(\delta/M)}{\mathbb{E}[k_m]}} \implies \mathbb{P}_{x\sim\mathcal{D}_\mathcal{X}}[q_m(x)=1] = \mathbb{E}[k_m] \leq 8\log M/\delta.$$

Where □

Lemma 4.6. *Under the high probability event of Lemma 4.2, for any $m \in [M]$ and $f \in \mathcal{F}_{m+1}$, we have,*

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_m}} \left[(\tilde{f}_m(x) - f(x))^2 \right] \leq \frac{8B_{m+1} + C(\mathcal{F}, \delta, k_m)}{k_m}.$$

Proof. Suppose that the active learning algorithm queried the labeling oracle for k_m samples in epoch m. Then by **??**, we have:

$$\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}_m}} \left[(f(x) - \tilde{f}_m(x))^2 \right] \leq 2 \cdot \left(\frac{1}{k_m} \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{x \in \mathcal{X}_m\} \cdot (f(x_t) - \tilde{f}_m(x_t))^2 \right) + \frac{C(\mathcal{F}, \delta, k_m)}{k_m}.$$

To bound the summation term, we apply a basic triangle inequality, $(a-b)^2 \leq 2(a-c)^2 + 2(b-c)^2$,

$$\begin{aligned}
&\sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{x \in \mathcal{X}_m\} \cdot (f(x_t) - \tilde{f}_m(x_t))^2 \\
&\leq \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{x \in \mathcal{X}_m\} \cdot \left(2(f(x_t) - \hat{f}_{m+1}(x))^2 + 2(\hat{f}_{m+1}(x) - \tilde{f}_m(x_t))^2 \right) \\
&\leq 4 \cdot \sup_{f' \in \mathcal{F}_m} \sum_{t=\tau_{m-1}+1}^{\tau_m} \mathbb{1}\{x \in \mathcal{X}_m\} \cdot (f'(x_t) - \hat{f}_{m+1}(x))^2 \\
&\leq 4B_{m+1}.
\end{aligned}$$

Finally, by plugging this bound back in, we achieve our desired result. □

Lemma 4.7 (Rademacher Complexity Integration)**.** *Consider a convex function class $\mathcal{F}$, and the ... With probability $1-\delta$,*

$$\mathbb{P}(\mathcal{E}(\hat{f}) > 4t) \leq 8\delta$$

for any

$$t > u \cdot \inf_{\alpha \in [0,\gamma]} \left(4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{\gamma} \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)}\, d\delta \right) + \frac{2}{C} \frac{\log \mathcal{N}_2(\mathcal{G}, \gamma) + u}{n},$$

where $u \leq \left(\log\left(\frac{2}{\delta}\right) + \sqrt{\frac{-1}{c} \log\left(\frac{\delta(1-e^{-2})}{4}\right)} \right)$.

Proof. Setting both terms on the right hand side of Lemma 7 in [4] to $\frac{\delta}{2}$ and solving for u yields $u \leq \log(\frac{2}{\delta}) + \sqrt{-\frac{1}{c} \log\left(\frac{\delta(1-e^{-2})}{4}\right)}$. This gives us a bound on the offset Rademacher complexity term, which we can use in the event to get a high probability bound on the excess risk via Theorem 4 of [4], giving us:

$$\mathbb{P}\left(\mathcal{E}(\hat{f}) > 4u\right) \leq 4\delta + 4\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \xi_i h(X_i) - \tilde{c} \cdot h(X_i)^2 > t \right) = 4\delta + 4\delta = 8\delta,$$

where $t > u \cdot \inf_{\alpha \in [0,\gamma]} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{\gamma} \sqrt{\log \mathcal{N}_2(\mathcal{G}, \delta)}\, d\delta \right\} + \frac{2}{C} \frac{\log \mathcal{N}_2(\mathcal{G}, \gamma) + u}{n}$.

□

References

[1] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Biologische Kybernetik, 2002.

[2] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[3] D. 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.

[4] T. Liang, A. Rakhlin, and K. Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory*, pages 1260–1285, 2015.

[5] P. Massart and Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5), Oct. 2006.

[6] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[7] A. Rakhlin, K. Sridharan, and A. B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2), May 2017.

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

[9] Y. Zhu and R. Nowak. Efficient active learning with abstention, 2022.