Agnostic Active Learning via a Regression Oracle

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

[1]Cornell University

Working Paper, December 2, 2024

1 Introduction

High-quality labels are often hard to obtain. Active learning is an learning regime that is designed to reduce the need for expert advice by only querying for labels on selected data points. A standard assumption in the active learning literature is *realizability*—the notion that a "perfect model" is within our set of possible predictors, making it accessible to the algorithm. However, real-world complexities, such as noisy data and hidden variables, make it impractical to capture the true environment within a predefined hypothesis class. This crippling condition opens up a "computational gap" between theoretical algorithms and practical applications. We relax this assumption, instead only assuming that the samples are drawn $i.i.d.$ from a fixed distribution.

Active learning is a subset of supervised learning in which the learner is given a set of arbitrarily sampled, unlabeled contexts and chooses to obtain the label for a subset of these contexts. The idea here is to reduce the size of the training set by selecting contexts on which the learner is uncertain.

In this paper, we focus on an online variant of active learning settings, known as selective sampling. In the selective sampling setting, contexts are again sampled by some arbitrary stochastic process and presented sequentially in rounds. On each round, the learn predicts a label given the context but then chooses whether or not to query the label on that round. The goal of the learner is to model the conditional probability distribution of the underlying data, $\mathbb{P}[y_t = 1|x_t]$.

Our goal in this regime is to design algorithms that can leverage general function approximation and online regression oracles to achieve small regret on predicting the correct labels, and at the same time minimize the number of expert queries made (query complexity). We use online regression oracles because they are less computationally hard than classification oracles, which are often infeasible.

In addition, we make use of the margin of the noisy expert, which intuitively quantifies the confidence level of the expert. In particular, the margin is large for data points where the expert is very confident in terms of providing the correct labels, while on the other hand, the margin is small on the points where the expert is less confident and subsequently provides more noisy labels as feedback.

1.1 Related Work

Our paper builds off of the work of [Zhu and Nowak, 2022]

- Selective sampling and space of active learning strategies

- Many of the previously mentioned algorithms are analyzed in the agnostic learning model, where no assumption is made about the noise distribution (see also [Han07]). In this setting, the label complexity of active learning algorithms cannot generally improve over supervised learners by more than a constant factor [Kaa06, BDL09]. However, under a parameterization of the noise distribution related to Tsybakov's

low-noise condition [Tsy04], active learning algorithms have been shown to have improved label complexity bounds over what is achievable in the purely agnostic setting [CN06, BBZ07, CN07, Han09, Kol09]. We also consider this parameterization to obtain a tighter label complexity analysis.

2 Problem Setting

Let $\mathcal{X}$ denote the space of inputs and $\mathcal{Y}$ denote the space of labels. We focus on the problem of binary classification, where $\mathcal{Y} = \{-1, +1\}$, with data generated i.i.d from a distribution $\mathcal{D}$. Furthermore, we denote the distribution over inputs as $\mathcal{D}_{\mathcal{X}}$ and $\eta(x) = \mathbb{P}_{X \sim \mathcal{D}_{\mathcal{X}}}[y = 1 | X = x]$.

The selective sampling problem is then defined as the T-round learning protocol where, on each round $t = 1, \ldots, T$, the learner observes an input $x_t \in \mathcal{X}$ and determines whether or not to query the correct label y_t of x_t. After this protocol, the learner is expected to output a classifier $\hat{h} : \mathcal{X} \to \mathcal{Y}$. The learner's performance is then measured with respect to two metrics, the first being the expected 0-1 excess risk of the output classifier on the distribution $\mathcal{D}$ against a hypothesis class $\mathcal{H}$

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

and the second being the number of queries Q_T it makes of the true label of a data point during the T-round protocol. Note, sometimes we refer to the classifier that minimizer the expected 0-1 loss as $h^* \in \mathcal{H}$.

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 [Zhu and Nowak, 2022], we note $\mathcal{H} = \mathcal{H}_{\mathcal{F}} := \{h_f : f \in \mathcal{F}\}$ where $h_f(x) = \text{sign}(2f(x) - 1)$. Then, $h^* = h_{f^*}$ for some $f^* \in \mathcal{F}$, i.e. f^* is a function in $\mathcal{F}$ that induces the optimal classifier $h^* \in \mathcal{H}$.

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

Assumption 2.1 (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 2.2 (Massart's Noise Condition, [Massart and Nédélec, 2006])**.** *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 2.3 (Expressivity of $\mathcal{F}$)**.** *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 $\tilde{f} \in S_C$,

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}} \left[\mathbb{1} \left\{ \left| \frac{\tilde{f}(x) - \eta(x)}{\eta(x)} \right| \leq \frac{1}{2} \wedge x \in C \right\} \right] = 0.$$

This assumption implies that, for any algorithm with a deterministic query condition determined by the history, the optimal model in $\mathcal{H}$ on 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 with probability 1. We now show that with the assumptions given above, we can provide an algorithm that performs efficient selective sampling.

3 Agnostic Selective Sampling

In this section, we provide our main algorithm and prove, under the conditions outlined in Section 1, 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 [Zhu and Nowak, 2022] 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 label was 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 $g_m : \mathcal{X} \to \{0, 1\}$. 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 g_{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, $g_{m(t)}$, 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 g_m on the m-th epoch. This distribution would have a density function that is 0 on all points g_m tells the algorithm not to query and is proportional to the the original data distribution on points g_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 2.3, 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 Assumption 2.2, 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:

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

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

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

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

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

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

3.3 Analysis

The excess risk of a classifier $h_f \in \mathcal{H}_{\mathcal{F}}$ can be decomposed in the following way,

$$\begin{aligned}
\mathcal{E}(h_f) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{h_f(x) \neq y\} - \mathbb{1}\{h^*(x) \neq y\}\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{h_f(x) \cdot h^*(x) = -1\}(1 - 2 \cdot \Pr(h^*(x) \neq y))\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{h_f(x) \cdot h^*(x) = -1\}(1 - 2 \cdot \Pr(h_\eta(x) \neq y))\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{h_f(x) \cdot h^*(x) = -1\} \cdot |2\eta(x) - 1|\right] \\
&\leq \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{h_f(x) \cdot h^*(x) = -1\}\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{(2f(x) - 1)(2f^*(x) - 1) \leq 0\}\right],
\end{aligned}$$

Where the third equality comes from an application of Assumption 2.3 with $C = \mathcal{X}$ and the inequality comes from bounding $|2\eta(x) - 1|$ by 1 since $\eta(\cdot)$ represents a probability. Now, considering the classifer $h_{\hat{f}}$ outputted by Algorithm 1, we can decompose this into the following:

$$= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{g_{M+1}(x) = 0, (2f(x) - 1)(2f^*(x) - 1) \leq 0\} + \mathbb{1}\{g_{M+1}(x) = 1, (2f(x) - 1)(2f^*(x) - 1) \leq 0\}\right].$$

Consider the high probability event of Lemma 4.1. Then, by Lemma 4.2, we know that $\tilde{f}_{m-1} \in \mathcal{F}_m$ for all $m \in [M+1]$. We will first bound the value of the first term.

Suppose $g_{M+1}(x) = 0$. Then, 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. Taking i to be the smallest such m, since $\text{sign}(2\tilde{f}_{i-1}(x) - 1) = \text{sign}(2f^*(x) - 1)$, we have that $\text{sign}(2f^*(x) - 1) = \text{sign}(2\hat{f}(x) - 1)$, or in other words, their product is greater than 0 and we do not make an error.

To bound the second term, we rewrite it in the following way:

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{g_{M+1}(x) = 1, (2f(x) - 1)(2f^*(x) - 1) \leq 0\}\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{(2f(x) - 1)(2f^*(x) - 1) \leq 0\} \cdot \prod_{m=1}^{M+1} \mathbb{1}\left\{\frac{1}{2} \in [\text{lcb}(x; \mathcal{F}_m), \text{ucb}(x; \mathcal{F}_m)]\right\}\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\prod_{m=1}^{M+1} \mathbb{1}\left\{\frac{1}{2} \in [\text{lcb}(x; \mathcal{F}_m), \text{ucb}(x; \mathcal{F}_m)], (2f(x) - 1)(2\tilde{f}_m(x) - 1) \leq 0\right\}\right],
\end{aligned}$$

where it might be useful to bound in terms of the margins of the subdistributions.

4 Supporting Lemmas

For any time step $t \in [T]$ and function $f \in \mathcal{F}$, define $M_t(f) := Q_t \left((f(x_t) - y_t)^2 - (\tilde{f}_{m(t)-1}(x_t) - y_t)^2 \right)$, where $m(t)$ denotes the epoch to which t belongs, and $Q_t = \mathbb{1}\{g_{m(t)}(x) = 1\}$. Furthermore, define the filtration $\mathfrak{F}_t := \sigma\left((x_1, y_1), \ldots, (x_t, y_t)\right)$ and denote $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot | \mathfrak{F}_{t-1}]$. Then, from [**?**], we have that,

Lemma 4.1. *[**?**] Suppose* $\mathrm{Pdim}(\mathcal{F}) < \infty$. *For any fixed* $\delta \in (0,1)$, *for any* $\tau, \tau' \in [T]$ *such that* $\tau < \tau'$, *with probability at least* $1 - \delta$, *we have:*

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

and

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

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

Lemma 4.2. *Under the high probability event of Lemma 4.1, it is true that for any* $m \in [M+1]$, $\tilde{f}_{m-1} \in \mathcal{F}_m$

Proof. For any $f \in \mathcal{F}$, under the high probability event of Lemma 4.1, we have,

$$\sum_{t=\tau_{m-2}+1}^{\tau_{m-1}} \mathbb{E}_t \left[Q_t \left((f(x_t) - y_t)^2 - (\tilde{f}_{m-1}(x_t) - y_t)^2 \right) \right]$$
$$\leq 2 \sum_{t=\tau_{m-2}+1}^{\tau_{m-1}} Q_t \left((f(x_t) - y_t)^2 - (\tilde{f}_{m-1}(x_t) - y_t)^2 \right) + C_\delta(\mathcal{F}).$$

Now, by the convexity of $\mathcal{F}$ and the definition of $\tilde{f}_{m-1}$, we can lower bound the left hand side,

$$\sum_{t=\tau_{m-2}+1}^{\tau_{m-1}} \mathbb{E}_t \left[Q_t \left((f(x_t) - y_t)^2 - (\tilde{f}_{m-1}(x_t) - y_t)^2 \right) \right] \geq \sum_{t=\tau_{m-2}+1}^{\tau_{m-1}} \mathbb{E}_t \left[Q_t \left(\left(f(x_t) - \tilde{f}_{m-1}(x_t) \right)^2 \right) \right] \geq 0.$$

So we have,

$$0 \leq 2 \sum_{t=\tau_{m-2}+1}^{\tau_{m-1}} Q_t \left((f(x_t) - y_t)^2 - (\tilde{f}_{m-1}(x_t) - y_t)^2 \right) + C_\delta(\mathcal{F}),$$

where rearranging gives us our desired result. □

References

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

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