Agnostic Active Learning via Regression

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

[1]Cornell University

Working Paper, January 15, 2025

1 Introduction

Active learning is a protocol for supervised learning in which the learner is presented with unlabeled data and is given access to an oracle for labeling data. The goal of the learner is to then output a classifier with low risk on the underlying data distribution while making as few queries to the oracle as possible. It is a model for learning that has gained recent interest due to its ability to capture environments where labels are often expensive to obtain, and thus there is value in algorithms that can learn well from less data. The active learning paradigm stands in contrast to passive supervised learning in which models are trained on a pre-labeled dataset.

Early works in this field focus on the noise-free setting, in which the label of a data point is a deterministic function of its features. In certain noiseless cases it has been shown that the number of labeled examples needed to achieve a desired classification error rate by an active learning algorithm is logarithmic in the number of examples that would be needed by a passive learning algorithm. This exponential improvement in query complexity shows that active learning is both a promising and motivated direction for future research.

Although the noiseless setting is interesting from a theoretical perspective, relevant hidden variables not captured by the input space often shape real-world data, resulting in an inherently noisy data distribution. As a result, more recent work tackles the problem of active learning with noisy labels. In this setting, relevant results typically follow a *disagreement-based approach*, where the learning algorithm maintains both a version space, which consists of all plausible models for the underlying data distribution, and a region of uncertainty, which is the subset of the input space where these models in the version space disagree. These algorithms then make predictions by evaluating a function in the version space on new data points and are therefore only incorrect in the region of uncertainty. The ways in which these works track the region of uncertainty all fall into one of three categories: (1) explicitly enumerating the version space (2) reducing to classification (3) reducing to regression

1. **Explicitly Enumerating the Version Space:** The earliest works employing a disagreement-based approach update the region of uncertainty through enumeration, where each time the version space is updated, the new region of uncertainty is determined by evaluating every function in the version space on each point in the old region of uncertainty. This algorithmic approach, developed by [Balcan et al., 2006], was shown by them to have exponential improvements on query complexity over passive learning algorithms NOTE: Determine if this is indeed the model class for model classes such as linear threshold function. Following this result, the work of [Hanneke, 2007] derives a general upper bound on the query complexity of this algorithm. However, it was observed that, without noise assumptions on the data distribution, active learning algorithms generally cannot achieve a significant improvement over supervised learners in terms of query complexity [Hsu, 2010], [Hanneke, 2014]. Moreover, since enumerating the version space is computationally expensive for large input spaces and model classes, this algorithm is impractical for most real-world applications.

2. **Reduction to Classification:** Results that update the region of uncertainty using a classification oracle do so by NOTE: Determine how they do this.

   Results that follow this approach assume access to a classification/labeling oracle [several works] or explicitly enumerate the version space [Balcan, '06] [Chaudhuri, '14]. Early works using this approach consider active learning with arbitrary noise and provide algorithms with guarantees even in the agnostic case [Balcan et al., 2006], [Hanneke, 2007]. However, it was discerned that with no noise assumptions on the data distribution, active learning algorithms cannot generally improve over supervised learners by more than a constant factor in query complexity [Hsu, 2010], [Hanneke, 2014]. To address this, the work of [Castro and Nowak, 2006] and [Castro and Nowak, 2007] initiated the analysis of active learning under low noise conditions. These results obtain significant improvements in query complexity [Hanneke, '09] [Koltchinskii, '10] [Hanneke, '11] [Hanneke, '12] and serve as a natural bridge between the no-noise and arbitrary noise settings. The main limitation of this approach is that classification is widely believed to be a computationally hard problem in its own right as it is an NP-hard problem.

3. **Reduction to Regression:** Results that update the region of uncertainty using a classification oracle do so by NOTE: Determine how they do this. This avenue has been explored more recently to address the intractability of the previous two methods. Offline and online regression algorithms are well studied and that follow this approach assume access to a regression oracle. These works consider the case where the benchmark hypothesis class is induced by a class of regression functions that attempt to model the underlying conditional distribution over labels and performance is measured via some surrogate loss [Hanneke and Yang, '12] [Dekel, '12] [Krishnamurthy, '19] [Zhu, '22] [Sekhari, '23]. Results that take this approach are therefore constrained to data distributions with low noise as otherwise regression function which are arbitrarily close to the optimal one can classify inputs very differently. However, the key advantage of this approach is that regression is often a tractable problem, making active learning tractable in these cases as well.

Our Contribution:

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 the conditional probability distribution of each label given an input as $\eta(x) = \mathbb{P}_{X \sim \mathcal{D}_{\mathcal{X}}}[y = 1 | X = x]$.

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 $\tilde{f} \in S_C$ *and for all* $x \in C$,

$$\left|\frac{\tilde{f}(x) - \eta(x)}{\eta(x)}\right| \leq \frac{1}{2}.$$

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. 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 **??**, 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 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 2.3, 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:

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

    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 $x_t \sim \mathcal{D}_{\mathcal{X}}$
            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:

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

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 $q_{M+1}(x) = 0$. 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:

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

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

[Balcan et al., 2006] Balcan, M.-F., Beygelzimer, A., and Langford, J. (2006). Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 65–72.

[Castro and Nowak, 2006] Castro, R. M. and Nowak, R. D. (2006). Upper and lower error bounds for active learning. In *Proceedings of the 44th Annual Allerton Conference on Communication, Control, and Computing*.

[Castro and Nowak, 2007] Castro, R. M. and Nowak, R. D. (2007). Minimax bounds for active learning. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT 2007)*, pages 19–34.

[Hanneke, 2007] Hanneke, S. (2007). A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pages 353–360.

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

[Hsu, 2010] Hsu, D. J. (2010). *Algorithms for Active Learning*. PhD thesis, University of California, San Diego.

[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.