Active Learning from Noisy and Abstention Feedback

Songbai Yan[1] Kamalika Chaudhuri[1] Tara Javidi[2]

***Abstract*— An active learner is given an instance space, a label space and a hypothesis class, where one of the hypotheses in the class assigns ground truth labels to instances. Additionally, the learner has access to a labeling oracle, which it can interactively query for the label of any example in the instance space. The goal of the learner is to find a good estimate of the hypothesis in the hypothesis class that generates the ground truth labels while making as few interactive queries to the oracle as possible.**

This work considers a more general setting where the labeling oracle can abstain from providing a label in addition to returning noisy labels. We provide a model for this setting where the abstention rate and the noise rate increase as we get closer to the decision boundary of the ground truth hypothesis. We provide an algorithm and an analysis of the number of queries it makes to the labeling oracle; finally we provide matching lower bounds to demonstrate that our algorithm has near-optimal estimation accuracy.

I. Introduction

An active learner is given an instance space, a label space and a hypothesis class, where one of the hypotheses in the class assigns ground truth labels to instances. Additionally, the learner has access to a labeling oracle, which it can interactively query for the label of any example in the instance space. The goal of the learner is to find a good estimate of the hypothesis in the hypothesis class that generates the ground truth labels while making as few interactive queries to the oracle as possible.

There is a great deal of literature on active learning [7], [9], [1], [13]. The problem is tractable when the oracle always outputs the ground truth labels [7], in which case, a generalized binary search-style approach results in an estimation error that decays exponentially with the number of labels. A more practical situation is when the oracle outputs a noisy label, and different noise models have been studied. [13], [12] consider active learning under random classification noise – when the probability that the oracle outputs an incorrect label for x is independent of x. [6] provides algorithms when the oracle outputs incorrect labels with higher probability closer to the decision boundary of the ground truth hypothesis. Finally, a long line of work [1], [9], [10], [2], [15] studies agnostic active learning, where a certain fraction of the labels can have arbitrary bias.

This work considers a more general setting where the labeling oracle can provide more general feedback than simply labels. Specifically, in addition to returning (a possibly incorrect) label, the oracle can also abstain from labeling. To model situations that arise in practice, we follow [6] and assume that the abstention rate of the labeling oracle as well as the noise rate of the label provided increases as we get closer to the decision boundary of the hypothesis that generates the ground truth labels. Finally, this work considers the case where the hypothesis class is the set of thresholds on $[0, 1]$; the ideas involved in our analysis are fairly general, and we leave the extension to d-dimensional spaces for future work.

We provide an algorithm for active learning in this setting, and provide an upper bound on the estimation accuracy achieved by this algorithm given a query budget. Our algorithm is a combination of the generlized binary search style algorithm that was proposed by [13] for learning from noisy labels, and repeatedly querying examples where the oracle abstains. We couple our upper bound with a lower bound on the achievable estimation accuracy given a fixed budget of queries in this setting. We show that our lower and upper bounds match up to constants and logarithmic factors, and therefore our algorithm achieves near-optimal estimation accuracy.

Our analysis yields two primary insights on the nature of active learning from noisy and abstention feedback. First, we show that given a fixed budget of queries, the estimation error due to abstention is much lower than that due to a comparable amount of label noise; in other words, abstention leads to significantly less estimation error than noise. Second, we show that an exponential decay in the estimation error can be obtained if and only if the maximum abstention rate is strictly less than 1 and the maximum noise rate is strictly less than $1/2$.

II. Related Work

Active learning is particularly challenging when the labels returned by the oracle are not necessarily the ground truth labels, and there is a lot of previous work on active learning in this setting. Three main types of noise models considered by the literature are random classification noise, agnostic active learning and variants of the Tsybakov noise conditions, where the noise rate is higher close to the decision boundary of the ground truth hypothesis. To the best of our knowledge, there has been no past theoretical work on active learning with abstention feedback.

In the random classification noise model, the probability that oracle returns an incorrect answer on a query x is independent of x, although it may depend on the label of x. [11] provides an algorithm for active learning from non-persistent random classification noise that proceeds by repeatedly querying the label of each examples. [13] provides

*This work was partially supported by NSF under IIS 1162581.

[1]Department of Computer Science, UC San Diego yansongbai,kamalika@eng.ucsd.edu

[2]Department of Electrical and Computer Engineering, UC San Diego tjavidi@ucsd.edu

a generalized binary search-style algorithm with improved label complexity, and [12] improves the label complexity further by an algorithm that queries the label of a sample that optimizes the step-wise EJS divergence.

In agnostic active learning, a fixed but arbitrary fraction of the labels output by the oracle may be arbitrarily biased, and the goal is to find the hypothesis in the hypothesis class that best fits the labels. As was shown by [8], active learning in this setting is particularly challenging, and a long line of work, from theoretical to practical, has specifically addressed this setting [3], [9], [10], [4], [15], [2].

Our work is most related to [5], which addresses active learning under noise that increases close to the decision boundary. They use a model which is very similar to ours, except that they only consider noise, and they provide upper and lower bounds on the label requirement of any active learner in this model. Our work can be viewed as an extension of theirs to include abstentions, and we show that abstentions affect the estimation error significantly less than a comparable amount of label noise.

III. Preliminaries

We consider active learning for binary classification. We are given an instance space $\mathcal{X} = [0,1]$, a hypothesis class $\mathcal{H}$ and label space $\mathcal{Y} = \{0, 1, \perp\}$, where $\perp$ means that the labeler abstains from providing a label. We are also given access to an oracle which we can query for labels. We assume that the ground truth labels are generated by a hypothesis θ^* in the hypothesis class, and our goal is to estimate this hypothesis accurately while making as few interactive queries to the labeling oracle as possible.

For the purpose of this paper, a hypothesis is a function $f : \mathcal{X} \to \{0, 1\}$ that belongs to a hypothesis class $\mathcal{H}$. Observe that we do not allow the hypothesis to output $\perp$. We also restrict the hypothesis class to the set of threshold functions over [0,1], i.e., $\mathcal{H} = \{f_\theta(x) = \mathbb{I}(x \geq \theta) : \theta \in [0, 1]\}$. Here $\mathbb{I}(A)$ is the indicator function: $\mathbb{I}(A) = 1$ if A is true, and $\mathbb{I}(A) = 0$ otherwise. Similar ideas apply to other geometric classifiers, such as surfaces in higher dimensional spaces. We leave this extension for future work. More specifically, we assume that there is an underlying $\theta^* \in [0, 1]$, and corresponding classifier $h^*(x) = \mathbb{I}(x \geq \theta^*)$ gives the ground-truth label.

In practice, ground truth labels are difficult to obtain and labels provided by oracles yield noisy information. We make the following assumptions about the oracle's response.

Assumption 1. *When queried on a point $x \in [0, 1]$, the oracle returns $y \in \mathcal{Y}$ such that:*

- *(abstention) $P(y = \perp |x) \leq C_1 - C_2 |x - \theta^*|^\alpha$.*
- *(flipping/noise) $P(y \neq \mathbb{I}(x \geq \theta^*)|x, y \neq \perp) \leq \frac{1}{2} - C_3 |x - \theta^*|^\beta$.*

Here $C_1, C_2, C_3, \alpha, \beta$ are non-negative constants and $C_2 \leq C_1$.

Thus, on every x, the oracle responds $\perp$ with some probability; moreover, even when the oracle returns a label, this label is incorrect with some probability. As expected in practice, the probabilities of abstention and returning an incorrect label increase as the query points get closer to the decision boundary.

Observe θ^* is unknown to the learner, and the goal is to estimate θ^* by interactively querying the oracle. Specifically, we aim to find a $\hat{\theta}$ such that $|\hat{\theta} - \theta^*|$ is as small as possible given a budget of n oracle queries.

a) Notation.: We next introduce some notations.

For any $\theta \in [0, 1]$, denote by $\eta_\theta(y|x)$ the probability that the oracle labels x as y when the ground-truth $\theta^* = \theta$, and by P_θ^n the distribution of n samples $\{(X_i, Y_i)\}_{i=1}^n$ where Y_i is drawn with probability $\eta_\theta(Y|X_i)$ and X_i is drawn by the active learning algorithm based solely on the knowledge of $\{(X_j, Y_j)\}_{j=1}^{i-1}$ (in other words, X_i is a (possibly randomized) function of $\{(X_j, Y_j)\}_{j=1}^{i-1}$).

IV. Results

We now provide bounds on estimation accuracy in terms of the label budget for active learning algorithms in this setting. First, in Section IV-A, we provide a lower bound on the accuracy of any active learning algorithm that makes n label queries to the oracle. Next in Section IV-B, we provide an algorithm and an analysis that demonstrates that this algorithm has optimal estimation accuracy (except for constant and logarithmic factors). We postpone all proofs to the appendix.

A. Lower Bounds

Theorem 2. *If Assumption 1 holds, then there is a universal constant $c \in (0, 1]$, such that for large enough n, for any active learning algorithm $\Psi : (\mathcal{X} \times \mathcal{Y})^n \to [0, 1]$ that makes at most n queries and outputs an estimation $\Psi((X, Y)^n)$ of θ^*, there is a $\theta \in [0, 1]$, such that $P_\theta^n \left(|\Psi((X, Y)^n) - \theta| > \left(\frac{1}{C_2 C_3^2 n} \right)^{\frac{1}{2\beta+\alpha}} \right) > c$ when $C_1 = 1$; $P_\theta^n \left(|\Psi((X, Y)^n) - \theta| > \left(\frac{1}{(1-C_1) C_3^2 n} \right)^{\frac{1}{2\beta}} \right) > c$ when $C_1 < 1$ and $\beta > 0$; and $P_\theta^n \left(|\Psi((X, Y)^n) - \theta| > 2^{-n} \right) > c$ when $C_1 < 1$ and $\beta = 0$.*

Theorem 2 gives lower bounds for the estimation accuracy of any active learning algorithms: roughly speaking, no algorithm can achieve a better accuracy than $O\left(\left(\frac{1}{n} \right)^{\frac{1}{2\beta+\alpha}} \right)$ for all ground truth θ^* when $P(y = \perp |x = \theta^*) = 1$, and $O\left(\left(\frac{1}{n} \right)^{\frac{1}{2\beta}} \right)$ when $P(y = \perp |x = \theta^*) < 1$. As a comparison, [5] studies learning thresholds with only noisy responses, and gives a lower bound of $O\left(\left(\frac{1}{n} \right)^{\frac{1}{2\beta}} \right)$, which can be seen as a special case of our result.

In particular, we note the significant difference between the $C_1 = 1$ case and the $C_1 < 1$ case. The $C_1 < 1$ case implies that the abstention rate at $\theta = \theta^*$ is < 1, and in this case, the loss of estimation accuracy due to abstention is a constant factor, which is to be expected. In contrast, the $C_1 = 1$ case occurs when the abstention rate at $\theta = \theta^*$ is 1. Theorem 2 shows that the estimation error due to abstention in this case is larger, and depends on the parameter α, which parameterizes the shape of the $P(y = \perp |x)$ curve around

θ^*. Moreover, observe that the dependence of the exponent of n on α is better than the dependence on β; thus, the estimation error due to abstention is considerably less than estimation error due to noise. Finally, the $C_1 < 1$ and $\beta = 0$ case implies that the noise rate and the abstention rate are constant; in this case, the lower bound is exponential.

The proof of Theorem 2 follows Le Cam's method, which reduces the problem to the error rate of binary hypothesis testing problems.

B. Algorithm and Analysis

In this section, we propose an algorithm (see Algorithm 1) that proceeds by repeatedly querying samples where the labeling oracle abstains, and we show that this algorithm achieves the optimal estimation error up to logarithmic factors and constants.

Algorithm 1 is motivated by the algorithm discussed in [6] which only deals with noisy oracles. It consists of two procedures: GBS and LearnThresholds. The GBS procedure is an iterative method. In each iteration, it first selects a sample to query the oracle, and then increases the weight of hypotheses that correctly label this sample and decreases the weight of those that make a mistake. The sampling strategy is generalized binary search: the algorithm selects the sample x that such that nearly half of total weights of the hypotheses assign x a label 0 and nearly half assign it label 1. If the oracle abstains from labeling, then the algorithm repeatedly queries the sample. Note that GBS will only query samples on a discrete grid Θ instead of [0,1] because it is easier to analyze the discrete case. In the LearnThresholds procedure, it runs GBS on three sets of grids to ensure that θ^* is far away from at least two sets of grids so that oracle's flipping probability on these two grids is low enough for GBS to work.

The following result is a direct corollary from [6].

Lemma 3. *Let $\hat{\theta}$ be the output of LearnThreasholds($\alpha, \beta, C_1, C_3, n, \delta$) in Algorithm 1, and c be some constant. If $C_1 = C_2 = 0$, then with probability at least $1-\delta$, $|\hat{\theta} - \theta^*| \le c\left(\frac{\log n(1+\log\frac{1}{\delta})}{C_3^2 n}\right)^{\frac{1}{2\beta}}$. If further $C_1 = C_2 = \beta = 0$, then with probability at least $1-\delta$, $|\hat{\theta} - \theta^*| \le \frac{c}{\delta} 2^{-C_3^2 n}$. .*

In a general setting where the oracle can abstain, we have the following upper bound on the estimation error that matches the lower bound in Theorem 2 up to logarithmic factors and constants.

Theorem 4. *Let $\hat{\theta}$ be the output of LearnThreasholds($\alpha, \beta, C_1, C_3, n, \delta$) in Algorithm 1.*

With probability at least $1-\delta$, $|\hat{\theta} - \theta^*| \le \tilde{O}\left(\left(\frac{1}{C_2 C_3^2 n}\right)^{\frac{1}{\alpha+2\beta}}\right)$ if $C_1 = 1$, $|\hat{\theta} - \theta^*| \le \tilde{O}\left(\left(\frac{1}{(1-C_1)C_3^2 n}\right)^{\frac{1}{2\beta}}\right)$ if $C_1 < 1$, $\beta \neq 0$, and $|\hat{\theta} - \theta^*| \le O\left(\frac{1}{\delta} 2^{-(1-C_1)C_3^2 n}\right)$ if $C_1 < 1$, $\beta = 0$. (The $\tilde{O}(\cdot)$ hides logarithmic factors of n, δ).

Algorithm 1 A repetitive querying learning algorithm with a generlzied binary search subroutine.

procedure GBS($\gamma, \Theta = \{\theta_1, \dots, \theta_m\}, T$)
  $p_i \leftarrow 1/m$ for $i = 0 \dots m-1$
  $t \leftarrow 0$
  $N \leftarrow 0$
  while $t < T$ **do**
    $x_N \leftarrow \arg\min_i \left|\sum_{j=0}^{i} p_j - 1/2\right|$
    repeat
      Query x_N and receive y_t
      $t \leftarrow t+1$
    until $y_t \neq \perp$ or $t > T$
    for $i = 1, 2, \dots, m$ **do**
      $p_i \leftarrow \begin{cases} p_i * (1+2\gamma) & \text{if } \mathbb{I}\{x_N \ge \theta_i\} = y_t \\ p_i * (1-2\gamma) & \text{if } \mathbb{I}\{x_N \ge \theta_i\} \neq y_t \end{cases}$
    end for
    Normalize $\boldsymbol{p}$
    $N \leftarrow N+1$
  end while
  Output: θ_{opt} where opt $= \arg\max_i p_i$
end procedure
procedure LEARNTHRESHOLDS($\alpha, \beta, C_1, C_3, n, \delta$)
  if $C_1 = 1$ **then**
    $\epsilon \leftarrow \left(\frac{6^{\alpha+1} * 4^{2\beta} * \log n(1+\log\frac{1}{\delta})}{C_2 C_3^2 n}\right)^{\frac{1}{\alpha+2\beta}}$
  else if $C_1 < 1$ and $\beta > 0$ **then**
    $\epsilon \leftarrow \left(\frac{6 * 4^{2\beta} * \log n \log\log n(1+\log\frac{1}{\delta})}{(1-C_1)C_3^2 n}\right)^{\frac{1}{2\beta}}$
  else if $C_1 < 1$ and $\beta = 0$ **then**
    $\epsilon \leftarrow \frac{2}{\delta} 2^{-\frac{(1-C_1)C_3^2 n}{6}}$
  end if
  $\gamma \leftarrow C_3 (6\epsilon)^{\beta}$
  for $i = 0, 1, 2$ **do**
    $\Theta_i \leftarrow \{0 + \frac{i}{3}\epsilon, \epsilon + \frac{i}{3}\epsilon, 2\epsilon + \frac{i}{3}\epsilon, 3\epsilon + \frac{i}{3}\epsilon, \dots, \lfloor\frac{1}{\epsilon}\rfloor \epsilon + \frac{i}{3}\epsilon\}$
    $\theta_i \leftarrow$ GBS($\gamma, \Theta_i, n/3$)
  end for
  for $i, j = 0, 1, 2$ **do**
    if $i \neq j$ and $|\theta_i - \theta_j| < \epsilon/3$ **then**
      Output: $(\theta_i + \theta_j)/2$
    end if
  end for
end procedure

This theorem shows that Algorithm 1 is nearly optimal (up to logarithmic and constant factors) for a variety of settings. In the easiest case where there is no noise or abstention ($C_1 = C_2 = \beta = 0$, $C_3 = \frac{1}{2}$), the estimation error is $O\left(2^{-n}\right)$. In the hardest case where there can be both abstentions and non-trivial noise ($C_1, \beta > 0$), the estimation error is $\tilde{O}\left(\left(\frac{1}{n}\right)^{\frac{1}{\alpha+2\beta}}\right)$ when $C_1 = 1$ and $\tilde{O}\left(\left(\frac{1}{n}\right)^{\frac{1}{2\beta}}\right)$ when $C_1 < 1$. When the oracle never abstains ($C_1 = C_2 = 0$), the estimation error is $\tilde{O}\left(\left(\frac{1}{n}\right)^{\frac{1}{2\beta}}\right)$. When there is no noise ($\beta = 0$, $C_3 = \frac{1}{2}$) or the noise is bounded ($\beta = 0$,

$0 < C_3 < \frac{1}{2}$), the estimation error is $\tilde{O}\left(\left(\frac{1}{n}\right)^{\frac{1}{\alpha}}\right)$.

It is interesting to note that our algorithm will just repeat querying when receiving an abstention feedback from the labeler. It seems that one can get information of θ^* from abstention since it is more likely to get an abstention feedback when querying a sample near the decision boundary (in other words, if we receive an abstention feedback on x we could "infer" that θ^* is close to x). But our near matching upper bounds and lower bounds show that there is little information in an abstention feedback.

It is also interesting to note that in nontrival cases (i.e., $C_1 = 1$ or $\beta > 0$), our upper bounds match lower bounds (up to constants and logarithmic factors) with respect to noise and abstention parameters $C_1, C_2, C_3, \alpha, \beta$.

Moreover, our results reveal that abstention leads to significantly less estimation error than noise. More specifically, if the oracle never abstains, and the flipping rate $P(y \neq \mathbb{I}(x \geq \theta^*)|x) \leq \frac{1}{2} - C\left|x - \theta^*\right|^{\gamma}$, the best (and also achievable) estimation error is $\tilde{O}\left(\left(\frac{1}{n}\right)^{\frac{1}{2\gamma}}\right)$. Whereas, if the oracle never flips the label, and the abstention rate $P(y = \perp |x) \leq 1 - C\left|x - \theta^*\right|^{\gamma}$, the best (and also achievable) estimation error is $\tilde{O}\left(\left(\frac{1}{n}\right)^{\frac{1}{\gamma}}\right)$. This result agrees with our intuition that learning from abstention is easier than learning from noisy labels.

V. Conclusion

In conclusion, we introduce a new model for active learning where the labeling oracle can abstain from providing a label in addition to providing noisy labels. Our model considers a setting where the rate of abstentions as well as the noise rate increase as we get closer to the decision boundary of the ground truth hypothesis. We provide an algorithm for this setting and analyze the estimation error given a query budget. Finally we provide matching lower bounds that demonstrate that our algorithm has near-optimal estimation error.

APPENDIX

A. Proofs for the Lower Bound

The proof is similar to the one in [6]. We use the following formulation of Le Cam's method ([14]):

Lemma 5. *Let Θ be a class of parameters, and $\{P_\theta : \theta \in \Theta\}$ be a class of probability distributions indexed by Θ over some sample space $\mathcal{X}$. Let $d : \Theta \times \Theta \to \mathbb{R}$ be a semi-metric. If there exist $\theta_0, \theta_1 \in \Theta$, such that $KL(P_{\theta_0}||P_{\theta_1}) \leq \alpha$ and $d(\theta_0, \theta_1) \geq 2s > 0$, then for any algorithm $\hat{\theta}$ that given a sample X outputs $\hat{\theta}(X)$ as an estimation of θ, the following inequality holds:*

$$\sup_{\theta \in \Theta} \underset{X \sim P_\theta}{P_\theta} \left(d(\theta, \hat{\theta}(X)) \geq s \right) \geq \max \left\{ \frac{e^{-\alpha}}{4}, \frac{1 - \sqrt{\alpha/2}}{2} \right\}$$

We need the following lemma in the proof of lower bounds.

Lemma 6. *If P, Q are two Bernoulli random variables with parameter p, q respectively and $\frac{1}{4} < p, q < \frac{1}{2}$, then $KL(P||Q) \leq 8(p-q)^2$.*

Proof.

$$\begin{aligned} \mathrm{KL}(P||Q) &= \int_q^p \left(\frac{p}{x} - \frac{1-p}{1-x} \right) dx \\ &= \int_q^p \frac{p-x}{x(1-x)}\, dx \\ &\leq 16 \int_q^p p - x\, dx \\ &= 8(p-q)^2 \end{aligned}$$

The inequality in line 3 follows from the fact that $x(1-x) > \frac{1}{16}$ when $\frac{1}{4} < x < \frac{1}{2}$. □

Proof of the Theorem 2. We first consider the $C_1 < 1$ and $\beta = 0$ case. It is well known that in the noise-free and abstention-free setting, there is no algorithm such that the following statement holds: there exists a positive constant δ (which does not decay as n increases) such that for any $\theta^* \in [0, 1]$ the algorithm could output an estimation $\hat{\theta}$ satisfying $|\hat{\theta} - \theta^*| \leq o\left(2^{-n}\right)$ with probability at least δ. If there is an algorithm that can achieve $|\hat{\theta} - \theta^*| = o(2^{-n})$ for all θ^* with at least some positive constant probability in the $C_1 < 1$ and $\beta = 0$ case, then this algorithm could also achieve $|\hat{\theta} - \theta^*| = o(2^{-n})$ with at least some positive constant probability for the noise-free and abstention-free case, since the noise-free and abstention-free case (i.e., $C_1 = C_2 = \beta = 0$ and $C_3 = \frac{1}{2}$) is a special case of the Assumption 1 when $C_1 < 1$ and $\beta = 0$. This contradicts with the fact in the beginning.

Next we consider the more general case: $C_1 = 1$ or $\beta > 0$.

We take Θ be $[0, 1]$, and $d(\theta_1, \theta_2) = |\theta_1 - \theta_2|$ in Lemma 5. We consider two thresholds $\theta_0 = 0$ and $\theta_1 = t$ where $t \in [0, 1]$ is to be chosen later. Next, we will define two distributions P_0 and P_1 corresponding to P_{θ_0} and P_{θ_1} in Lemma 5 respectively.

For $\theta_0 = 0$, we define the distribution of oracle's response as follows:

$$P_0(Y = \perp |x) = \begin{cases} C_1 - C_2 x^\alpha & x > t \\ C_1 - C_2 t^\alpha - C_2 (t-x)^\alpha & x \leq t \end{cases}$$

$$P_0(Y = 0 | x, Y \neq \perp) = \frac{1}{2} - C_3 x^\beta$$

This distribution complies with Assumption 1 in that $P_0(y = \perp |x) \leq C_1 - C_2 |x - 0|^\alpha$ and $P_0(y \neq \mathbb{I}(x \geq 0)|x, y \neq \perp) = \frac{1}{2} - C_3 |x - 0|^\beta$.

For $\theta_1 = t$, we define the distribution of oracle's response as follows:

$$\begin{aligned} P_1(Y = \perp |x) &= P_0(Y = \perp |x) \\ &= \begin{cases} C_1 - C_2 x^\alpha & x > t \\ C_1 - C_2 t^\alpha - C_2 (t-x)^\alpha & x \leq t \end{cases} \end{aligned}$$

$$P_1(Y=0|x, Y\neq\perp) = \begin{cases} \frac{1}{2} - C_3 x^\beta & x > t \\ \frac{1}{2} + C_3 (t-x)^\beta & x \leq t \end{cases}$$

This distribution complies with Assumption 1 in that $P_1(y = \perp | x) \leq C_1 - C_2 |x-t|^\alpha$ and $P_1(y \neq \mathbb{I}(x \geq t)|x, y \neq \perp) \leq \frac{1}{2} - C_3 |x-t|^\beta$.

Next, we consider P_0^n and P_1^n, the distributions of n samples $\{(X_i, Y_i)\}_{i=1}^n$ where Y_i is drawn with conditional probability P_0 and P_1 respectively, and X_i is drawn by the active learning algorithm.

$$\begin{aligned} \mathrm{KL}\left(P_1^n || P_0^n\right) &= \mathop{\mathbb{E}}_{P_1}\left(\log \frac{P_1^n\left(\{(X_i,Y_i)\}_{i=1}^n\right)}{P_0^n\left(\{(X_i,Y_i)\}_{i=1}^n\right)}\right) \\ &= \mathop{\mathbb{E}}_{P_1}\left(\log \frac{\prod_{i=1}^n P_1\left(Y_i|X_i\right)}{\prod_{i=1}^n P_0\left(Y_i|X_i\right)}\right) \\ &= \mathop{\mathbb{E}}_{P_1}\left(\mathop{\mathbb{E}}_{P_1}\left(\log \frac{\prod_{i=1}^n P_1\left(Y_i|X_i\right)}{\prod_{i=1}^n P_0\left(Y_i|X_i\right)}\middle| X_1, ..., X_n\right)\right) \\ &\leq n \max_{x\in[0,1]} \mathop{\mathbb{E}}_{P_1}\left(\log \frac{P_1\left(Y|x\right)}{P_0\left(Y|x\right)}\middle| x\right) \end{aligned}$$

where the second equality follows from the fact that Y_i is conditional independent with X_j ($j \neq i$) given X_i and that the active learner will draw X_i based solely on the knowledge of $\{(X_j, Y_j)\}_{j=1}^{i-1}$, and hence $P_0\left(X_i|X_1, Y_1, X_2, Y_2, \ldots, X_{i-1}, Y_{i-1}\right) = P_1\left(X_i|X_1, Y_1, X_2, Y_2, \ldots, X_{i-1}, Y_{i-1}\right)$.

$$\begin{aligned} &\mathop{\mathbb{E}}_{P_1}\left(\log \tfrac{P_1(Y|x)}{P_0(Y|x)}\middle| x\right) \\ =\ & P_1(Y=\perp|x)\log\tfrac{P_1(Y=\perp|x)}{P_0(Y=\perp|x)} + P_1(Y=1|x)\log\tfrac{P_1(Y=1|x)}{P_0(Y=1|x)} \\ & + P_1(Y=0|x)\log\tfrac{P_1(Y=0|x)}{P_0(Y=0|x)} \\ =\ & 0 + P_0(Y\neq\perp|x)\mathrm{KL}\left(P_1(Y|x,Y\neq\perp)||P_0(Y|x,Y\neq\perp)\right) \\ \leq\ & \left(1 - C_1 + 2C_2 t^\alpha\right)\mathrm{KL}\left(P_1(Y|x,Y\neq\perp)||P_0(Y|x,Y\neq\perp)\right) \end{aligned}$$

When $x \geq t$, $\mathrm{KL}\left(P_1(Y|x, Y\neq\perp)||P_0(Y|x, Y\neq\perp)\right)$=0. When $x < t$, we can apply Lemma 6 and have

$$\begin{aligned} & \mathrm{KL}\left(P_1(Y|x,Y\neq\perp)||P_0(Y|x,Y\neq\perp)\right) \\ \leq\ & 8\left(\left(\tfrac{1}{2} + C_3(t-x)^\beta\right) - \left(\tfrac{1}{2} - C_3 x^\beta\right)\right)^2 \\ \leq\ & 8C_3^2 t^{2\beta} \end{aligned}$$

Therefore, in either case, we have $\mathrm{KL}\left(P_1^n||P_0^n\right) \leq 8n\left(1 - C_1 + 2C_2 t^\alpha\right) C_3^2 t^{2\beta}$.

Recall that we are looking at $C_1 = 1$ or $\beta = 0$ case. When $C_1 = 1$, we set $t = \left(\frac{1}{C_2 C_3^2 n}\right)^{\frac{1}{2\beta+\alpha}}$. Then we will have $\mathrm{KL}(P_1^n||P_0^n) \leq 16$, and $d(\theta_0, \theta_1) = \left(\frac{1}{C_2 C_3^2 n}\right)^{\frac{1}{2\beta+\alpha}}$. By Lemma 5, for any active learning algorithm $\Psi : (\mathcal{X}\times\mathcal{Y})^n \to [0,1]$, there is a $\theta \in [0,1]$, such that $P_\theta^n\left(|\Psi(X^n) - \theta| > \left(\frac{1}{C_2 C_3^2 n}\right)^{\frac{1}{2\beta+\alpha}}\right) > e^{-16}/4$.

Likewise, when $C_1 < 1$, we can set $t = \left(\frac{1}{(1-C_1)C_3^2 n}\right)^{\frac{1}{2\beta}}$. Then we will have $\mathrm{KL}(P_1^n||P_0^n) \leq 8$, and $d(\theta_0, \theta_1) = \left(\frac{1}{(1-C_1)C_3^2 n}\right)^{\frac{1}{2\beta}}$. By Lemma 5, for any active learning algorithm $\Psi : (\mathcal{X}\times\mathcal{Y})^n \to [0,1]$, there is a $\theta \in [0,1]$, such that $P_\theta^n\left(|\Psi(X^n) - \theta| > \left(\frac{1}{(1-C_1)C_3^2 n}\right)^{\frac{1}{2\beta}}\right) > e^{-8}/4$. . This concludes the proof. □

B. the Proof for the Upper Bound

Proof of the Theorem 4. It is easy to see there are at least 2 sets of grids (WLOG say the 2 sets of grids are Θ_1 and Θ_2) that $\theta - \theta^* > \frac{\epsilon}{6}$ for any $\theta \in \Theta_1 \cup \Theta_2$. On these two sets of grids, each query in line 8 will return a non-abstaining label with probability at least $1 - C_1 + C_2\left(\frac{\epsilon}{6}\right)^\alpha$. By the union bound, we will have with probability at least $1-\delta$, $N \geq T\left(1 - C_1 + C_2\left(\frac{\epsilon}{6}\right)^\alpha\right)/\log\frac{\delta}{2T}$ in the GBS procedure for Θ_1 and Θ_2 .

Therefore, if we set the label budget

$$n = \frac{6 * 4^{2\beta} * \frac{1}{C_3^2}\left(\frac{1}{\epsilon}\right)^{2\beta}}{\left(1 - C_1 + C_2\left(\frac{\epsilon}{6}\right)^\alpha\right)} \log\frac{1}{\epsilon\delta} \log\left(\frac{1}{C_3^2}\left(\frac{1}{\epsilon}\right)^{2\beta}\log\frac{1}{\epsilon\delta}\right) \tag{1}$$

, for Θ_1 and Θ_2, the number of non-abstaining responses $N \geq 6 * 4^{2\beta} * \frac{1}{C_3^2}\left(\frac{1}{\epsilon}\right)^{2\beta}\log\frac{1}{\epsilon\delta}$ with probability at least $1 - \delta/2$. Consequently by Lemma 3 we will have $|\theta_1 - \theta^*| \leq \epsilon$ and $|\theta_2 - \theta^*| \leq \epsilon$ with probability at least $1-\delta$. Thus, LearnThreasholds in Algorithm 1 will output a $\hat{\theta}$ such that $|\hat{\theta} - \theta^*| \leq \epsilon$ with probability at least $1-\delta$.

What remains is to solve ϵ from (1). When $C_1 = 1$, $\epsilon = \tilde{O}\left(\left(\frac{1}{C_2 C_3^2 n}\right)^{\frac{1}{\alpha+2\beta}}\right)$. When $C_1 < 1$ and $\beta \neq 0$, $\epsilon = \tilde{O}\left(\left(\frac{1}{(1-C_1)C_3^2 n}\right)^{\frac{1}{2\beta}}\right)$. When $C_1 < 1$ and $\beta = 0$, $\epsilon = O\left(\frac{1}{\delta}2^{-(1-C_1)C_3^2 n}\right)$. This concludes the proof. □

REFERENCES

[1] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *J. Comput. Syst. Sci.*, 75(1):78–89, 2009.

[2] M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *COLT*, 2013.

[3] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72. ACM, 2006.

[4] A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.

[5] Rui Castro and Robert D. Nowak. Minimax bounds for active learning. In *COLT*, pages 5–19, 2007.

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

[7] S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

[8] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *ICML*, 2008.

[9] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *NIPS*, 2007.

[10] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

[11] M. Kääriäinen. Active learning in the non-realizable case. In *ALT*, 2006.

[12] Mohammad Naghshvar, Tara Javidi, and Kamalika Chaudhuri. Bayesian active learning with non-persistent noise. *IEEE Transactions on Information Theory*, 61(7):4080–4098, 2015.

[13] R. D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906, 2011.

[14] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[15] Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 442–450, 2014.