Active Learning from Imperfect Labelers

Songbai Yan
UC San Diego
yansongbai@eng.ucsd.edu

Kamalika Chaudhuri
UC San Diego
kamalika@eng.ucsd.edu

Tara Javidi
UC San Diego
tjavidi@ucsd.edu

Abstract

This paper studies the problem of active learning where the labeler can not only return an incorrect label but also abstain from labeling. Different noise and abstention models of the labeler are considered and the amount of queries to the labeler required to learn a good classifier is analyzed. An adaptive algorithm which automatically requests less queries with a more informative labeler is provided and is proved to have nearly optimal query complexity. The analysis also shows the gains of allowing a labeler to abstain from labeling by quantifying the reduction in the number of queries.

1 Introduction

In active learning, the learner is given an input space $\mathcal{X}$, a label space $\mathcal{L}$, and a hypothesis class $\mathcal{H}$ such that one of the hypotheses in the class generates ground truth labels. Additionally, the learner has at its disposal a labeler to which it can pose interactive queries about the labels of examples in the input space. Note that the labeler may output a noisy version of the ground truth label (flipped label). The goal of the learner is to learn a hypothesis in $\mathcal{H}$ which is close to the hypothesis that generates the ground truth labels.

There has been a significant amount of literature on active learning, both theoretical and practical. Previous theoretical work on active learning has mostly focused on the above basic setting [2, 4, 7, 10, 22] and has developed algorithms under a number of different models of label noise. A handful of exceptions include [3] which allows class conditional queries, [5] which allows requesting counterexamples to current version spaces, and [23] where the learner has access to a strong labeler and a weak labeler.

In this paper, we consider a more general setting where, in addition to a label, the labeler can sometimes abstain from prediction. This scenario arises naturally in difficult labeling tasks and has been considered in computer vision by [11, 15]. Our goal in this paper is to investigate this problem from a foundational perspective, and explore what kind of assumptions are needed, and how an abstaining labeler can affect properties such as consistency and convergence rates of active learning algorithms.

The setting of active learning with an abstaining noisy labeler was first considered by [21], who looked at learning binary threshold classifiers based on queries to an labeler whose abstention rate is higher closer to the decision boundary. They primarily looked at the case when the abstention rate at a distance Δ from the decision boundary is less than $1 - \Theta(\Delta^\alpha)$, and the rate of label flips at the same distance is less than $\frac{1}{2} - \Theta(\Delta^\beta)$; under these conditions, they provided an active learning algorithm that given parameters α and β, outputs a classifier with error ϵ using $O(\epsilon^{-\alpha-2\beta})$ queries to the labeler. However, there are several limitations to this work. The primary limitation is that parameters α and β need to be known to the algorithm, which is not usually the case in practice. A second major limitation is that even if the labeler has nice properties, such as, the abstention rates increase sharply close to the boundary, their algorithm is unable to exploit these properties to reduce the number of queries. A third and final limitation is that their analysis only applies to one dimensional thresholds, and not to more general decision boundaries.

In this work, we provide an algorithm which is completely adaptive in the sense that it does not need to know any parameters of the labeler's noise and abstention models. Additionally, we show that this algorithm is able to exploit nice properties of the labeler. Our algorithm is statistically consistent under very mild conditions – when the abstention rate is non-decreasing as we get closer to the decision boundary. Under the conditions of [21], we give the same rate of convergence. However, if the abstention rate of the labeler increases strictly monotonically close to the decision boundary – a condition that we call the c-growth property – then our algorithm adapts and does substantially better. It simply exploits the increasing abstention rate close to the decision boundary, and does not even have to rely on the noisy labels! Specifically, when applied to the case where the noise rate is at most $\frac{1}{2} - \Theta(\Delta^\beta)$ and the abstention rate is $1 - \Theta(\Delta^\alpha)$ at distance Δ from the decision boundary, our algorithm can output a classifier with error ϵ based on only $O(\epsilon^{-\alpha})$ queries. An important property of this algorithm is that all this is achieved in a *completely adaptive manner*; unlike previous work [21], our algorithm needs *no information whatsoever on the abstention rates or rates of label noise*. Thus our result also strengthens existing results on active learning from (non-abstaining) noisy labelers by providing an adaptive algorithm that achieves that same performance as [6] without knowledge of noise parameters.

We extend our algorithm so that it applies to any smooth d-dimensional decision boundary, not just one-dimensional thresholds, and we complement it with lower bounds on the number of queries that need to be made to any labeler. Our lower bounds generalize the lower bounds in [21], and shows that our upper bounds are nearly optimal. We also present an example that shows that at least a relaxed version of the monotonicity property is necessary to achieve this performance gain; if the abstention rate plateaus around the decision boundary, then our algorithm needs to query and rely on the noisy labels (resulting in higher query complexity) in order to find a hypothesis close to the one generating the ground truth labels.

1.1 Related work

There has been a considerable amount of work on active learning, most of which involves labelers that are not allowed to abstain. Theoretical work on this topic largely falls under two categories – the membership query model [6, 13, 17, 18], where the learner can request label of any example in the instance space, and the PAC model, where the learner is given a large set of unlabeled examples from an underlying unlabeled data distribution, and can request labels of a subset of these examples. Our work and also that of [21] builds on the membership query model.

There has also been a lot of work on active learning under different noise models. The problem is relatively easy when the labeler always provides the ground truth labels – see [8, 9, 12] for work in this setting in the PAC model, and [13] for the membership query model. Perhaps the simplest setting of label noise is random classification noise, where each label is flipped with a probability that is independent of the unlabeled instance. [14] shows how to address this kind of noise in the PAC model by repeatedly querying an example until the learner is confident of its label; [17, 18] provide more sophisticated algorithms with better query complexities in the membership query model. A second setting is when the noise rate increases closer to the decision boundary; this setting has been studied under the membership query model by [6] and in the PAC model by [10, 4, 22]. A final setting is agnostic PAC learning – when a fixed but arbitrary fraction of labels may disagree with the label assigned by the optimal hypothesis in the hypothesis class. Active learning is known to be particularly difficult in this setting; however, algorithms and associated label complexity bounds have been provided by [1, 2, 4, 10, 12, 22] among others.

Our work expands on the membership query model, and our abstention and noise models are related to a variant of the Tsybakov noise condition. A setting similar to ours was considered by [6, 21]. [6] considers a non-abstaining labeler, and provides a near-optimal binary search style active learning algorithm; however, their algorithm is non-adaptive. [21] gives a nearly matching lower and upper query complexity bounds for active learning with abstention feedback, but they only give a non-adaptive algorithm for learning one dimensional thresholds, and only study the situation where the abstention rate is upper-bounded by a polynomial function. Besides [21] , [11, 15] study active learning with abstention feedback in computer vision applications. However, these works are based on heuristics and do not provide any theoretical guarantees.

2 Settings

Notation $\mathbb{1}[A]$ is the indicator function: $\mathbb{1}[A] = 1$ if A is true, and 0 otherwise. For $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ $(d > 1)$, denote $(x_1, \ldots, x_{d-1})$ by $\tilde{\boldsymbol{x}}$. Define $\ln x = \log_e x$, $\log x = \log_{\frac{4}{3}} x$, $[\ln\ln]_+(x) = \ln\ln\max\{x, e^e\}$. We use $\tilde{O}$ and $\tilde{\Theta}$ to hide logarithmic factors in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, and d.

Definition 1. Suppose $\gamma \geq 1$. A function $g : [0,1]^{d-1} \to \mathbb{R}$ is (K, γ)-*Hölder smooth*, if it is continuously differentiable up to $\lfloor\gamma\rfloor$-th order, and for any $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^{d-1}$, $\left|g(\boldsymbol{y}) - \sum_{m=0}^{\lfloor\gamma\rfloor} \frac{g^{(m)}(x)}{m!}(\boldsymbol{y} - \boldsymbol{x})^m\right| \leq K \|\boldsymbol{y} - \boldsymbol{x}\|^\gamma$. We denote this class of functions by $\Sigma(K, \gamma)$.

Definition 2. A function $f : [0,1] \to [0,1]$ satisfies the *c-growth property* $(0 < c < 1)$ if f is nondecreasing and for any $0 < a \leq 1$ and $0 \leq b \leq \frac{2}{3}a$, $\frac{f(b)}{f(a)} \leq 1 - c$.

Note. The c-growth property prevents f from being too flat. For example, $f(x) = 1$ does not satisfy the c-growth property for any $0 < c < 1$, while $f(x) = x^\alpha$ $(\alpha > 0)$ satisfies the c-growth property with $c = 1 - \left(\frac{2}{3}\right)^\alpha$. Note that if f satisfies the c-growth property, then $f(0) = 0$ by letting $k \to \infty$ in $f\left(\left(\frac{2}{3}\right)^k\right) \leq (1-c)^k$.

We consider active learning for binary classification. We are given an instance space $\mathcal{X} = [0,1]^d$ and a label space $\mathcal{L} = \{0, 1\}$. Each instance $x \in \mathcal{X}$ is assigned to a label $l \in \{0, 1\}$ by an underlying function $h^* : \mathcal{X} \to \{0, 1\}$ in a hypothesis space of interest unknown to the learning algorithm. The learning algorithm has access to any $x \in \mathcal{X}$, but no access to their labels. Instead, it can only obtain label information through interactions with a labeler, whose relation to h^* is to be specified later. The objective of the algorithm is to output some classifier $\hat{h}$ that is close to h^* while making as few interactions with the labeler as possible.

We assume that the hypothesis space of interest is the *smooth boundary fragment* class $\mathcal{H} = \{h_g(\boldsymbol{x}) = \mathbb{1}[x_d > g(\tilde{\boldsymbol{x}})] \mid g : [0,1]^{d-1} \to [0,1]$ is (K, γ)-Hölder smooth$\}$. In other words, the decision boundaries of classifiers in this class are epigraph of smooth functions (see Figure 1 for example). We assume $h^*(\boldsymbol{x}) = \mathbb{1}[x_d > g^*(\tilde{\boldsymbol{x}})] \in \mathcal{H}$. Observe that when $d = 1$, our hypothesis space becomes the space of threshold functions $\mathcal{H} = \{h_\theta(x) = \mathbb{1}[x > \theta] : \theta \in [0,1]\}$.

The performance of a classifier $h(\boldsymbol{x}) = \mathbb{1}[x_d > g(\tilde{\boldsymbol{x}})]$ is evaluated by the volume of disagreement region which is the set of points that h and h^* assign different labels to. Mathematically, this is equal to the L^1 distance between the decision boundaries $\|g - g^*\| = \int_{[0,1]^{d-1}} |g(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})|\, d\tilde{\boldsymbol{x}}$.

The learning algorithm can only obtain label information by querying a labeler who is allowed to abstain from labeling or return an incorrect label (flipping between 0 and 1). For each query $\boldsymbol{x} \in [0,1]^d$, the labeler L will return $y \in \mathcal{Y} = \{0, 1, \perp\}$ ($\perp$ means that the labeler abstains from providing a 0/1 label) according to some distribution $P_L(Y = y \mid X = \boldsymbol{x})$. When it is clear from the context, we will drop the subscript of $P_L(Y \mid X)$. Note that while the labeler can declare its indecision by outputting $\perp$, we do not allow classifiers in our hypothesis space to output $\perp$.

In our active learning setting, our goal is to output a boundary g such that g is close to g^* while making as few interactive queries to the labeler as possible. In particular, we want to find an algorithm with low *query complexity* $\Lambda(\epsilon, \delta, \mathcal{A}, L, g^*)$, which is defined as the minimum number N such that if the ground truth is g^*, then with probability at least $1 - \delta$, the algorithm $\mathcal{A}$ can output a classifier $h(\boldsymbol{x}) = \mathbb{1}[x_d > g(\tilde{\boldsymbol{x}})]$ such that $\|g - g^*\| = \int_{[0,1]^{d-1}} |g(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})|\, d\tilde{\boldsymbol{x}} \leq \epsilon$, and the number of queries to the labeler L is at most N.

2.1 Assumptions

In this subsection, we introduce some assumptions on the response of the labeler.

Assumption 1. *The response distribution of the labeler $P(Y \mid X)$ satisfies:*

- *(abstention) For any $\tilde{\boldsymbol{x}} \in [0,1]^{d-1}$, $x_d, x'_d \in [0,1]$, if $|x_d - g^*(\tilde{\boldsymbol{x}})| \geq |x'_d - g^*(\tilde{\boldsymbol{x}})|$ then $P(\perp \mid (\tilde{\boldsymbol{x}}, x_d)) \leq P(\perp \mid (\tilde{\boldsymbol{x}}, x'_d))$;*
- *(noise) For any $\boldsymbol{x} \in [0,1]^d$, $P(Y \neq \mathbb{1}[x_d > g^*(\tilde{\boldsymbol{x}})] \mid \boldsymbol{x}, Y \neq \perp) \leq \frac{1}{2}$.*

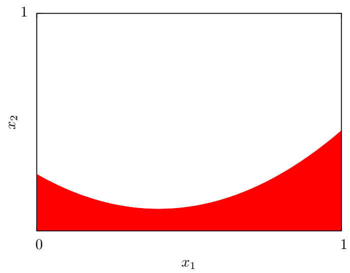

Figure 1: A classifier with boundary $g(\tilde{\boldsymbol{x}}) = (x - 0.4)^2 + 0.1$ for $d = 2$. Label 1 is assigned to the region above, 0 to the below (red region)

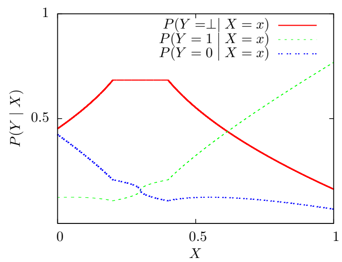

Figure 2: The distributions above satisfy Assumptions 1 and 2, but the abstention feedback is useless since $P(\perp \mid x)$ is flat between $x = 0.2$ and 0.4

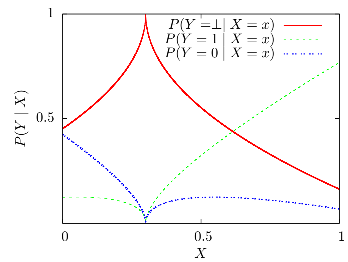

Figure 3: Distributions above satisfy Assumptions 1, 2, and 3.

In other words, we assume that the closer $\boldsymbol{x}$ is to the decision boundary $(\tilde{\boldsymbol{x}}, g^*(\tilde{\boldsymbol{x}}))$, the more likely the labeler will abstain from labeling; and the 0/1 labels can be flipped with probability as large as $\frac{1}{2}$ (note that when it is $\frac{1}{2}$, $P(0 \mid \boldsymbol{x}) = P(1 \mid \boldsymbol{x})$ so the 0/1 labels are completely uninformative).

Assumption 2. *When queried on a point $\boldsymbol{x} \in [0,1]^d$, the labeler returns $y \in \mathcal{Y}$ such that:*

- *(abstention)* $P(\perp \mid \boldsymbol{x}) \leq 1 - f\left(|x_d - g^*(\tilde{\boldsymbol{x}})|\right)$,
- *(noise)* $P(Y \neq \mathbb{1}\left[x_d > g^*(\tilde{\boldsymbol{x}})\right] \mid \boldsymbol{x}, Y \neq \perp) \leq \frac{1}{2}\left(1 - C_3 \left|x_d - g^*(\tilde{\boldsymbol{x}})\right|^\beta\right)$.

Here C_3, β are non-negative constants, and f is nondecreasing.

Assumption 2 gives upper bounds for the probabilities that the labeler returns abstentions or flipped labels, and these upper bounds decrease as $\boldsymbol{x}$ gets further away from the decision boundary. The assumption on the noise is a variant of the Tsybakov noise condition.

Assumption 3. *When queried on a point $\boldsymbol{x} \in [0,1]^d$, the labeler returns $y \in \mathcal{Y}$ such that $P(\perp \mid \boldsymbol{x}) = 1 - f\left(|x_d - g^*(\tilde{\boldsymbol{x}})|\right)$ where f satisfies the c-growth property for some constant $0 < c < 1$.*

Assumption 3 requires the abstention probability $P(\perp \mid (\tilde{\boldsymbol{x}}, x_d))$ not to be too flat with respect to x_d (Figure 3).

We will show that there is a labeler (see Figure 2 for example) that satisfies Assumption 2 but the abstention feedback cannot be exploited to improve the query complexity, while under Assumption 3 the abstention feedback is always very useful.

Note that here c, f, C_3, β are unknown parameters that characterizes the complexity of the learning task. We want to design an algorithm that does not require knowledge of these parameters but still achieves nearly optimal query complexity.

3 Learning one-dimensional thresholds

In this section, we start with the one dimensional case ($d = 1$) to demonstrate the main idea. We will generalize these results to multidimensional instance space in the next section.

When $d = 1$, the decision boundary g^* becomes a point in $[0,1]$, and the corresponding classifier is a threshold function over [0,1]. In other words the hypothesis space becomes $\mathcal{H} = \{f_\theta(x) = \mathbb{1}\left[x > \theta\right] : \theta \in [0,1]\})$. We denote the ground truth decision boundary by $\theta^* \in [0,1]$. We want to find a $\hat{\theta} \in [0,1]$ such that $|\hat{\theta} - \theta^*|$ is small while making as few queries as possible.

3.1 Algorithm

The proposed algorithm is a binary search style algorithm shown as Algorithm 1. (For the sake of simplicity, we assume $\log \frac{1}{2\epsilon}$ is an integer.) Algorithm 1 takes a desired precision ϵ and confidence level δ as its input, and returns an estimation $\hat{\theta}$ of the decision boundary θ^*. The algorithm maintains an interval $[L_k, R_k]$ in which θ^* is believed to lie, and shrinks this interval iteratively. To find the subinterval that contains θ^*, Algorithm 1 relies on two auxiliary functions (marked in Procedure 2) to conduct adaptive sequential hypothesis tests regarding subintervals of interval $[L_k, R_k]$.

Algorithm 1 The active learning algorithm for learning thresholds

Input: δ, ϵ
$[L_0, R_0] \leftarrow [0, 1]$
for $k = 0, 1, 2, \dots, \log \frac{1}{2\epsilon} - 1$ **do**
  Define three quartiles: $U_k \leftarrow \frac{3L_k+R_k}{4}$, $M_k \leftarrow \frac{L_k+R_k}{2}$, $V_k \leftarrow \frac{L_k+3R_k}{4}$
  $A^{(u)}, A^{(m)}, A^{(v)}, B^{(u)}, B^{(v)} \leftarrow$ Empty Array
  for $n = 1, 2, \dots$ **do**
    Query at U_k, M_k, V_k, and receive labels $X_n^{(u)}, X_n^{(m)}, X_n^{(v)}$
    for $w \in \{u, m, v\}$ **do**
      ▷ We record whether $X^{(w)} = \perp$ in $A^{(w)}$, and the 0/1 label (as -1/1) in $B^{(w)}$ if $X^{(w)} \neq \perp$
      if $X^{(w)} \neq \perp$ **then**
        $A^{(w)} \leftarrow A^{(w)}$.append(1) , $B^{(w)} \leftarrow B^{(w)}$.append($2\mathbb{1}\left[X^{(w)} = 1\right] - 1$)
      else
        $A^{(w)} \leftarrow A^{(w)}$.append(0)
      end if
    end for
    ▷ Check if the differences of abstention feedbacks are statistically significant
    if CheckSignificant-Var($\left\{A_i^{(u)} - A_i^{(m)}\right\}_{i=1}^{n}$, $\frac{\delta}{4\log\frac{1}{2\epsilon}}$) **then**
      $[L_{k+1}, R_{k+1}] \leftarrow [U_k, R_k]$; break
    else if CheckSignificant-Var($\left\{A_i^{(v)} - A_i^{(m)}\right\}_{i=1}^{n}$, $\frac{\delta}{4\log\frac{1}{2\epsilon}}$) **then**
      $[L_{k+1}, R_{k+1}] \leftarrow [L_k, V_k]$; break
    end if
    ▷ Check if the differences between 0 and 1 labels are statistically significant
    if CheckSignificant($\left\{-B_i^{(u)}\right\}_{i=1}^{B^{(u)}.\text{length}}$, $\frac{\delta}{4\log\frac{1}{2\epsilon}}$) **then**
      $[L_{k+1}, R_{k+1}] \leftarrow [U_k, R_k]$; break
    else if CheckSignificant($\left\{B_i^{(v)}\right\}_{i=1}^{B^{(v)}.\text{length}}$, $\frac{\delta}{4\log\frac{1}{2\epsilon}}$) **then**
      $[L_{k+1}, R_{k+1}] \leftarrow [L_k, V_k]$; break
    end if
  end for
end for
Output: $\hat{\theta} = \left(L_{\log\frac{1}{2\epsilon}} + R_{\log\frac{1}{2\epsilon}}\right)/2$

Suppose $\theta^* \in [L_k, R_k]$. Algorithm 1 tries to shrink this interval to a $\frac{3}{4}$ of its length in each iteration by repetitively querying on quartiles $U_k = \frac{3L_k+R_k}{4}$, $M_k = \frac{L_k+R_k}{2}$, $V_k = \frac{L_k+3R_k}{4}$. To determine which specific subinterval to choose, the algorithm uses 0/1 labels and abstention feedbacks simultaneously. Since the ground truth labels are determined by $\mathbb{1}[x > \theta^*]$, one can infer that if the number of queries that return label 0 at U_k (V_k) is statistically significantly more (less) than label 1, then θ^* should be on the right (left) side of U_k (V_k). Similarly, from Assumption 1, if the number of non-abstention feedbacks at U_k (V_k) is statistically significantly more than non-abstention feedbacks at M_k, then θ^* should be closer to M_k than

Procedure 2 Adaptive sequential testing

▷ D_0, D_1 are absolute constants defined in Proposition 3 and Proposition 4
▷ $\{X_i\}$ are i.i.d. random variables bounded by 1. δ is the confidence level. Detect if $\mathbb{E}X > 0$
function CheckSignificant($\{X_i\}_{i=1}^n, \delta$)
  $p(n,\delta) \leftarrow D_0\left(1+\ln\frac{1}{\delta}+\sqrt{4n\left([\ln\ln]_+ 4n+\ln\frac{1}{\delta}\right)}\right)$
  Return $\sum_{i=1}^n X_i \geq p(n,\delta)$
end function
function CheckSignificant-Var($\{X_i\}_{i=1}^n, \delta$)
  Calculate the empirical variance $\mathrm{Var} = \frac{n}{n-1}\left(\sum_{i=1}^n {X_i}^2 - \frac{1}{n}\left(\sum_{i=1}^n X_i\right)^2\right)$
  $q(n,\mathrm{Var},\delta) \leftarrow D_1\left(1+\ln\frac{1}{\delta}+\sqrt{\left(\mathrm{Var}+\ln\frac{1}{\delta}+1\right)\left([\ln\ln]_+\left(\mathrm{Var}+\ln\frac{1}{\delta}+1\right)+\ln\frac{1}{\delta}\right)}\right)$
  Return $n \geq \ln\frac{1}{\delta}$ AND $\sum_{i=1}^n X_i \geq q(n,\mathrm{Var},\delta)$
end function

U_k (V_k).

Algorithm 1 relies on the ability to shrink the search interval via statistically comparing the numbers of obtained labels at locations U_k, M_k, V_k. As a result, a main building block of Algorithm 1 is to test whether i.i.d. bounded random variables Y_i are greater in expectation than i.i.d. bounded random variables Z_i with statistical significance. In Procedure 2, we have two test functions CheckSignificant and CheckSignificant-Var that take i.i.d. random variables $\{X_i = Y_i - Z_i\}$ ($|X_i| \leq 1$) and confidence level δ as their input, and output whether it is statistically significant to conclude $\mathbb{E}X_i > 0$.

CheckSignificant is based on the following uniform concentration result regarding the empirical mean:

Proposition 3. *Suppose $X_1, X_2, \ldots$ are a sequence of i.i.d. random variables with $X_1 \in [-2,2]$, $\mathbb{E}X_1 = 0$. Take any $0 < \delta < 1$. Then there is an absolute constant D_0 such that with probability at least $1-\delta$, for all n simultaneously,*

$$\left|\sum_{i=1}^n X_i\right| \leq D_0\left(1+\ln\frac{1}{\delta}+\sqrt{4n\left([\ln\ln]_+ 4n+\ln\frac{1}{\delta}\right)}\right)$$

In Algorithm 1, we use CheckSignificant to detect whether the expected number of queries that return label 0 at location U_k (V_k) is more/less than the expected number of label 1 with a statistical significance.

CheckSignificant-Var is based on the following uniform concentration result which further utilizes the empirical variance $V_n = \frac{n}{n-1}\left(\sum_{i=1}^n Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n Y_i\right)^2\right)$:

Proposition 4. *Take any $\delta > 0$. Then there is an absolute constant D_1 such that with probability at least $1-\delta$, for all $n \geq \ln\frac{1}{\delta}$ simultaneously,*

$$\left|\sum_{i=1}^n X_i\right| \leq D_1\left(1+\ln\frac{1}{\delta}+\sqrt{\left(1+\ln\frac{1}{\delta}+V_n\right)\left([\ln\ln]_+ (1+\ln\frac{1}{\delta}+V_n)+\ln\frac{1}{\delta}\right)}\right)$$

The use of variance results in a tighter bound when $\mathrm{Var}(X_i)$ is small.

In Algorithm 1, we use CheckSignificant-Var to detect the statistical significance of the relative order of the number of queries that return non-abstention feedbacks at U_k (V_k) compared to the number of non-abstention feedbacks at M_k. This results in a better query complexity than using CheckSignificant under Assumption 3, since the variance of difference in the abstention feedback approaches 0 when the interval $[L_k, R_k]$ zooms in on θ^*.[1]

[1] We do not apply CheckSignificant-Var to 0/1 labels because unlike the difference of numbers between non-abstention feedback at U_k (V_k) and M_k, the variance of the difference of numbers between 0 and 1 labels stays above a positive constant.

3.2 Analysis

For Algorithm 1 to be statistically consistent, we only need Assumption 1.

Theorem 5. *Let θ^* be the ground truth. If the labeler L satisfies Assumption 1 and Algorithm 1 stops to output $\hat{\theta}$, then $\left|\theta^* - \hat{\theta}\right| \leq \epsilon$ with probability at least $1 - \frac{\delta}{2}$.*

Under additional Assumptions 2 and 3, we can derive upper bounds of the query complexity for our algorithm.

Theorem 6. *Let θ^* be the ground truth, and $\hat{\theta}$ be the output of Algorithm 1. Under Assumptions 1 and 2, with probability at least $1 - \delta$, Algorithm 1 makes at most $\tilde{O}\left(\frac{1}{f(\frac{\epsilon}{2})}\epsilon^{-2\beta}\right)$ queries.*

Theorem 7. *Let θ^* be the ground truth, and $\hat{\theta}$ be the output of Algorithm 1. Under Assumptions 1 and 3, with probability at least $1 - \delta$, Algorithm 1 makes at most $\tilde{O}\left(\frac{1}{f(\frac{\epsilon}{2})}\right)$ queries.*

The query complexity given by Theorem 7 is independent of β that decides the flipping rate, and consequently smaller than the bound in Theorem 6. This improvement is due to the use of $\perp$ labels, which become much more informative under Assumption 3.

3.3 Lower Bounds

In this subsection, we give lower bounds of query complexity in the one-dimensional case and establish near optimality of Algorithm 1. We will give corresponding lower bounds for the high-dimensional case in the next section.

The lower bound in [21] can be easily generalized to Assumption 2:

Theorem 8. *([21]) There is a universal constant $\delta_0 \in (0, 1)$ and a labeler L satisfying Assumptions 1 and 2, for any active learning algorithm $\mathcal{A}$, there is a $\theta^* \in [0, 1]$, such that for small enough ϵ, $\Lambda(\epsilon, \delta_0, \mathcal{A}, L, \theta^*) \geq \Omega\left(\frac{1}{f(\epsilon)}\epsilon^{-2\beta}\right)$.*

Our query complexity (Theorem 7) for the algorithm is also almost tight under Assumptions 1 and 3 with a polynomial abstention rate.

Theorem 9. *There is a universal constant $\delta_0 \in (0, 1)$ and a labeler L satisfying Assumptions 1, 2, and Assumption 3 with $f(x) = C_2 x^\alpha$ ($C_2 > 0$ and $0 \leq \alpha \leq 2$ are constants), for any active learning algorithm $\mathcal{A}$, there is a $\theta^* \in [0, 1]$, such that for small enough ϵ, $\Lambda(\epsilon, \delta_0, \mathcal{A}, L, \theta^*) \geq \Omega\left(\epsilon^{-\alpha}\right)$.*

3.4 Remarks

Our results confirm the intuition that learning from abstention is easier than learning from noisy labels. This is true because a noisy label might mislead the learning algorithm, but an abstention response never does. Our analysis shows, in particular, that if the labeler never abstains, and outputs completely noisy labels with probability bounded by $1 - |x - \theta^*|^\gamma$ (i.e., $P(Y \neq \mathbb{I}[x > \theta^*] \mid x) \leq \frac{1}{2}(1 - |x - \theta^*|^\gamma)$), then the near optimal query complexity of $\tilde{O}\left(\epsilon^{-2\gamma}\right)$ is significantly larger than the near optimal $\tilde{O}\left(\epsilon^{-\gamma}\right)$ query complexity associated with a labeler who only abstains with probability $P(Y = \perp \mid x) \leq 1 - |x - \theta^*|^\gamma$ and never flips a label. More precisely, while in both cases the labeler outputs the same amount of corrupted labels, the query complexity of the abstention-only case is significantly smaller than the noise-only case.

Note that the query complexity of Algorithm 1 consists of two kinds of queries: queries which return 0/1 labels and are used by function CheckSignificant, and queries which return abstention and are used by function CheckSignificant-Var. Algorithm 1 will stop querying when the responses of one of the two kinds of query are statistically significant. Under Assumption 2, our proof actually shows that the optimal number of queries is dominated by the number of queries used by CheckSignificant function. In other words, a simplified variant of Algorithm 1 which excludes use of abstention feedback is near optimal. Similarly, under Assumption 3, the optimal query complexity is dominated by the number of queries used by CheckSignificant-Var function. Hence the variant of Algorithm 1 which disregards 0/1 labels would be near optimal.

4 The multidimensional case

We will follow [6] to generalize the results from one-dimensional thresholds to the d-dimensional ($d > 1$) smooth boundary fragment class $\Sigma(K, \gamma)$.

4.1 Lower bounds

Theorem 10. *There are universal constants $\delta_0 \in (0,1)$, $c_0 > 0$, and a labeler L satisfying Assumptions 1 and 2, for any active learning algorithm $\mathcal{A}$, there is a $g^* \in \Sigma(K,\gamma)$, such that for small enough ϵ, $\Lambda(\epsilon, \delta_0, \mathcal{A}, L, g^*) \geq \Omega\left(\frac{1}{f(c_0\epsilon)}\epsilon^{-2\beta-\frac{d-1}{\gamma}}\right)$.*

Theorem 11. *There is a universal constant $\delta_0 \in (0,1)$ and a labeler L satisfying Assumptions 1, 2, and Assumption 3 with $f(x) = C_2 x^\alpha$ ($C_2 > 0$ and $0 \leq \alpha \leq 2$ are constants), for any active learning algorithm $\mathcal{A}$, there is a $g^* \in \Sigma(K,\gamma)$, such that for small enough ϵ, $\Lambda(\epsilon, \delta_0, \mathcal{A}, L, g^*) \geq \Omega\left(\epsilon^{-\alpha-\frac{d-1}{\gamma}}\right)$.*

4.2 Algorithm and Analysis

Recall the decision boundary of the smooth boundary fragment class can be seen as the epigraph of a smooth function $[0,1]^{n-1} \to [0,1]$. For $d > 1$, we can reduce the problem to the one-dimensional problem by discretizing the first $d-1$ dimensions of the instance space and then perform a polynomial interpolation. The algorithm is shown as Algorithm 3. For the sake of simplicity, we assume γ, M/γ in Algorithm 3 are integers.

Algorithm 3 The active learning algorithm for the smooth boundary fragment class

1: Input: δ, ϵ, γ

2: $M \leftarrow \Theta\left(\left(\frac{1}{\epsilon}\right)\right)^{1/\gamma}$. $\mathcal{L} \leftarrow \left\{\frac{0}{M}, \frac{1}{M}, \ldots, \frac{M-1}{M}\right\}^{d-1}$

3: For each $l \in \mathcal{L}$, apply Algorithm 1 with parameter (ϵ, δ/M^{d-1}) to learn a threshold g_l that approximates $g^*(l)$

4: Partition the instance space into cells $\{I_q\}$ indexed by $q \in \left\{0, 1, \ldots, \frac{M}{\gamma} - 1\right\}^{d-1}$, where

$$I_q = \left[\frac{q_1\gamma}{M}, \frac{(q_1+1)\gamma}{M}\right] \times \cdots \times \left[\frac{q_{d-1}\gamma}{M}, \frac{(q_{d-1}+1)\gamma}{M}\right]$$

5: For each cell I_q, perform a polynomial interpolation: $g_q(\tilde{\boldsymbol{x}}) = \sum_{l \in I_q \cap \mathcal{L}} g_l Q_{q,l}(\tilde{\boldsymbol{x}})$, where

$$Q_{q,l}(\tilde{\boldsymbol{x}}) = \prod_{i=1}^{d-1} \prod_{j=0, j \neq M l_i - \gamma q_i}^{\gamma} \frac{\tilde{\boldsymbol{x}}_i - (\gamma q_i + j)/M}{l_i - (\gamma q_i + j)/M}$$

6: Output: $g(\tilde{\boldsymbol{x}}) = \sum_{q \in \left\{0,1,\ldots,\frac{M}{\gamma}-1\right\}^{d-1}} g_q(\tilde{\boldsymbol{x}}) \mathbb{1}\left[\tilde{\boldsymbol{x}} \in q\right]$

We have similar consistency guarantee and upper bounds as in the one-dimensional case.

Theorem 12. *Let g^* be the ground truth, and g be the output of Algorithm 3. If the labeler L satisfies Assumption 1, then $\|g^* - g\| \leq \epsilon$ with probability at least $1 - \frac{\delta}{2}$.*

Theorem 13. *Let g^* be the ground truth, and g be the output of Algorithm 3. Under Assumptions 1 and 2, with probability at least $1-\delta$, Algorithm 3 makes at most $\tilde{O}\left(\frac{d}{f(\epsilon/2)}\epsilon^{-2\beta-\frac{d-1}{\gamma}}\right)$ queries.*

Theorem 14. *Let g^* be the ground truth, and g be the output of Algorithm 3. Under Assumptions 1 and 3, with probability at least $1-\delta$, Algorithm 3 makes at most $\tilde{O}\left(\frac{d}{f(\epsilon/2)}\epsilon^{-\frac{d-1}{\gamma}}\right)$ queries.*

References

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

[2] 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.

[3] Maria-Florina Balcan and Steve Hanneke. Robust interactive learning. In *Proceedings of The 25th Conference on Learning Theory*, 2012.

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

[5] Alina Beygelzimer, Daniel Hsu, John Langford, and Chicheng Zhang. Search improves label for active learning. *arXiv preprint arXiv:1602.07265*, 2016.

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

[7] Yuxin Chen, S Hamed Hassani, Amin Karbasi, and Andreas Krause. Sequential information maximization: When is greedy near-optimal? In *Proceedings of The 28th Conference on Learning Theory*, pages 338–363, 2015.

[8] D. A. Cohn, L. E. Atlas, and R. E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2), 1994.

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

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

[11] Meng Fang and Xingquan Zhu. I don't know the label: Active learning with blind knowledge. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2238–2241. IEEE, 2012.

[12] Steve Hanneke. Teaching dimension and the complexity of active learning. In *Learning Theory*, pages 66–81. Springer, 2007.

[13] Tibor Hegedűs. Generalized teaching dimensions and the query complexity of learning. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 108–117. ACM, 1995.

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

[15] Christoph Kading, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4343–4352. IEEE, 2015.

[16] Yuan-Chuan Li and Cheh-Chih Yeh. Some equivalent forms of bernoulli's inequality: A survey. *Applied Mathematics*, 4(07):1070, 2013.

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

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

[19] Maxim Raginsky and Alexander Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems*, pages 1026–1034, 2011.

[20] Aaditya Ramdas and Akshay Balsubramani. Sequential nonparametric testing with the law of the iterated logarithm. *arXiv preprint arXiv:1506.03486*, 2015.

[21] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning from noisy and abstention feedback. In *Communication, Control, and Computing (Allerton), 2015 53th Annual Allerton Conference on*. IEEE, 2015.

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

[23] Chicheng Zhang and Kamalika Chaudhuri. Active learning from weak and strong labelers. In *Advances in Neural Information Processing Systems*, pages 703–711, 2015.

A Proof of query complexities

A.1 Properties of adaptive sequential testing in Procedure 2

Lemma 15. *Suppose $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables such that $\mathbb{E}X_i \leq 0$, $|X_i| \leq 1$. Let $\delta > 0$. Then with probability at least $1-\delta$, for all $n \in \mathbb{N}$ simultaneously CheckSignificant($\{X_i\}_{i=1}^{n}, \delta$) in Procedure 2 returns false.*

Proof. This is immediate by applying Proposition 21 to $X_i - \mathbb{E}X_i$. □

Lemma 16. *Suppose $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables such that $\mathbb{E}X_i > \epsilon > 0$, $|X_i| \leq 1$. Let $\delta \in [0, \frac{1}{3}]$, $N \geq \frac{\xi}{\epsilon^2} \ln\frac{1}{\delta}[\ln\ln]_+\frac{1}{\epsilon}$ (ξ is an absolute constant specified in the proof). Then with probability at least $1-\delta$, CheckSignificant$\Big(\{X_i\}_{i=1}^{N}, \delta\Big)$ in Procedure 2 returns true.*

Proof. Let $S_N = \sum_{i=1}^{N} X_i$. CheckSignificant$\Big(\{X_i\}_{i=1}^{N}, \delta\Big)$ returns false if and only if $S_N \leq D_0\left(1 + \ln\frac{1}{\delta} + \sqrt{N\left([\ln\ln]_+ N + \ln\frac{1}{\delta}\right)}\right)$.

$$\begin{aligned}
&\Pr\left(S_N \leq D_0\left(1+\ln\frac{1}{\delta}+\sqrt{N\left([\ln\ln]_+N+\ln\frac{1}{\delta}\right)}\right)\right)\\
\leq&\Pr\left(S_N \leq D_0\left(1+\ln\frac{1}{\delta}+\sqrt{N[\ln\ln]_+N}+\sqrt{N\ln\frac{1}{\delta}}\right)\right)\\
\leq&\Pr\left(S_N - N\mathbb{E}X_i \leq D_0\left(1+\ln\frac{1}{\delta}+\sqrt{N[\ln\ln]_+N}+\sqrt{N\ln\frac{1}{\delta}}\right) - N\epsilon\right)
\end{aligned}$$

Suppose $N = \frac{c\xi}{\epsilon^2}\ln\frac{1}{\delta}[\ln\ln]_+\frac{1}{\epsilon}$ for constant $c \geq 1$ and ξ. ξ is set to be sufficiently large, such that (1) $\xi \geq 4D_0^2$; (2) $\frac{2D_0}{\sqrt{\xi}} + D_0\left(3+\sqrt{[\ln\ln]_+\xi}\right) + D_0 - \sqrt{\xi}/2 \leq -\sqrt{\frac{1}{2}}$; (3) $f(x) = D_0\sqrt{[\ln\ln]_+ x} - \sqrt{x}/2$ is decreasing when $x > \xi$. Here (2) is satisfiable since $\frac{D_0}{\sqrt{\xi}} + D_0\sqrt{[\ln\ln]_+\xi} - \sqrt{\xi}/2 \to -\infty$ as $\xi \to \infty$, (3) is satisfiable since $f'(x) \to -\infty$ as $x \to \infty$. (2) and (3) together implies $\frac{2D_0}{\sqrt{\xi}} + D_0\left(3+\sqrt{[\ln\ln]_+c\xi}\right) + D_0 - \sqrt{c\xi}/2 \leq -\sqrt{\frac{1}{2}}$.

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left(D_0\left(1+\ln\frac{1}{\delta}+\sqrt{N[\ln\ln]_+N}+\sqrt{N\ln\frac{1}{\delta}}\right)-N\epsilon\right)\\
=&\sqrt{\ln\frac{1}{\delta}}\left(\frac{D_0\epsilon(1+\ln\frac{1}{\delta})}{\sqrt{c\xi[\ln\ln]_+\frac{1}{\epsilon}}\ln\frac{1}{\delta}} + D_0\sqrt{\frac{[\ln\ln]_+\left(\frac{c\xi}{\epsilon^2}\ln\frac{1}{\delta}[\ln\ln]_+\frac{1}{\epsilon}\right)}{\ln\frac{1}{\delta}}} + D_0 - \sqrt{c\xi[\ln\ln]_+\frac{1}{\epsilon}}\right)
\end{aligned}$$

Since $[\ln\ln]_+\frac{1}{\epsilon}, c, \ln\frac{1}{\delta} \geq 1$ and $\epsilon < 1$, we have $\frac{D_0\epsilon(1+\ln\frac{1}{\delta})}{\sqrt{c\xi[\ln\ln]_+\frac{1}{\epsilon}}\ln\frac{1}{\delta}} \leq \frac{2D_0}{\sqrt{\xi}}$.

Since $[\ln\ln]_+ x \geq 1$ if $x \geq 1$, we have $[\ln\ln]_+ \frac{1}{\epsilon} \leq \frac{1}{\epsilon}$, and thus

$$\begin{aligned}
\sqrt{[\ln\ln]_+ \left(\frac{c\xi}{\epsilon^2}\ln\frac{1}{\delta}[\ln\ln]_+\frac{1}{\epsilon}\right)} &= \sqrt{\ln\left[\max\left\{e, 2\ln\frac{1}{\epsilon} + \ln c\xi + \ln\ln\frac{1}{\delta} + \ln[\ln\ln]_+\frac{1}{\epsilon}\right\}\right]} \\
&\leq \sqrt{\ln\left[\max\left\{e, 3\ln\frac{1}{\epsilon} + \ln c\xi + [\ln\ln]_+\frac{1}{\delta}\right\}\right]} \\
&\overset{(a)}{\leq} \sqrt{\ln\left[\max\left\{e, 9\ln\frac{1}{\epsilon}\ln c\xi[\ln\ln]_+\frac{1}{\delta}\right\}\right]} \\
&\leq \sqrt{3 + [\ln\ln]_+\frac{1}{\epsilon} + [\ln\ln]_+ c\xi + \ln[\ln\ln]_+\frac{1}{\delta}} \\
&\overset{(b)}{\leq} \sqrt{3} + \sqrt{[\ln\ln]_+ c\xi} + \sqrt{[\ln\ln]_+\frac{1}{\epsilon}} + \sqrt{\ln[\ln\ln]_+\frac{1}{\delta}}
\end{aligned}$$

where (a) follows by $a+b+c \leq 3abc$ if $a,b,c \geq 1$, and (b) follows by $\sqrt{\sum_i x_i} \leq \sum_i \sqrt{x_i}$ if $x_i \geq 0$. Thus, we have

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left(D_0\left(1 + \ln\frac{1}{\delta} + \sqrt{N[\ln\ln]_+ N} + \sqrt{N\ln\frac{1}{\delta}}\right) - N\epsilon\right) \\
\leq& \sqrt{\ln\frac{1}{\delta}}\left(\frac{2D_0}{\sqrt{\xi}} + D_0\frac{\sqrt{3} + \sqrt{[\ln\ln]_+ c\xi} + \sqrt{[\ln\ln]_+\frac{1}{\epsilon}} + \sqrt{\ln[\ln\ln]_+\frac{1}{\delta}}}{\sqrt{\ln\frac{1}{\delta}}} + D_0 - \sqrt{c\xi[\ln\ln]_+\frac{1}{\epsilon}}\right) \\
\overset{(c)}{\leq}& \sqrt{\ln\frac{1}{\delta}}\left(\frac{2D_0}{\sqrt{\xi}} + D_0\left(3 + \sqrt{[\ln\ln]_+ c\xi}\right) + D_0 - \sqrt{c\xi}/2\right) \\
\overset{(d)}{\leq}& -\sqrt{\ln\frac{1}{\delta}}/2
\end{aligned}$$

(c) follows by $\sqrt{\ln\frac{1}{\delta}} \geq \max\left\{1, \sqrt{\ln[\ln\ln]_+\frac{1}{\delta}}\right\}$, $D_0 \geq 1$, and $\sqrt{[\ln\ln]_+\frac{1}{\epsilon}}\left(\frac{D_0}{\sqrt{\ln\frac{1}{\delta}}} - \sqrt{c\xi}\right) \leq D_0 - \sqrt{c\xi} \leq -\sqrt{c\xi}/2$ if $c\xi \geq 4D_0^2$. (d) follows by our choose of ξ.

Therefore,

$$\begin{aligned}
&\Pr\left(S_N - N\mathbb{E}X_i \leq D_0\left(1 + \ln\frac{1}{\delta} + \sqrt{N[\ln\ln]_+ N} + \sqrt{N\ln\frac{1}{\delta}}\right) - N\epsilon\right) \\
\leq& \Pr\left(S_N - N\mathbb{E}X_i \leq -\sqrt{N\ln\frac{1}{\delta}}/2\right)
\end{aligned}$$

which is at most δ by Hoeffding Bound. □

Lemma 17. *Suppose $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables such that $\mathbb{E}X_i \leq 0$, $|X_i| \leq 1$. Let $\delta > 0$. Then with probability at least $1-\delta$, for all n simultaneously CheckSignificant-Var($\{X_i\}_{i=1}^{n}, \delta$) in Procedure 2 returns false.*

Proof. Define $Y_i = X_i - \mathbb{E}X_i$. It is easy to check $\frac{n}{n-1}\left(\sum_{i=1}^n Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n Y_i\right)^2\right) = \frac{n}{n-1}\left(\sum_{i=1}^n X_i^2 - \frac{1}{n}\left(\sum_{i=1}^n X_i\right)^2\right)$. The result is immediate from Proposition 4. □

Lemma 18. *Suppose $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables such that $\mathbb{E}X_i > \tau\epsilon$, $|X_i| \le 1$, Var$(X_i) \le 2\epsilon$ where $0 < \epsilon \le 1$, $\tau > 0$. Let $\delta < 1$, $N = \frac{\xi}{\tau\epsilon}\ln\frac{2}{\delta}$ (ξ is a constant specified in the proof). Then with probability at least $1-\delta$, CheckSignificant-Var$\left(\{X_i\}_{i=1}^{N}, \delta\right)$ in Procedure 2 returns true.*

Proof. Let $Y_i = X_i - \mathbb{E}X_i$, η be the constant η in Lemma 29. Set $\xi = \max(\eta, \frac{16}{\tau} + \frac{4}{3})$.

CheckSignificant-Var$\left(\{X_i\}_{i=1}^{N}, \delta\right)$ returns false if and only if $\sum_{i=1}^{N} X_i \le q(N, \mathrm{Var}, \delta/2)$.

By applying Lemma 29 to X_i, $\frac{q(N,\mathrm{Var},\delta)}{N} - \mathbb{E}X_i \le -\tau\epsilon/2$ with probability at least $1-\delta/2$.

Applying Bernstein's inequality to Y_i, we have

$$\begin{aligned}\Pr\left(\frac{1}{N}\sum_{i=1}^{N} Y_i \le -\tau\epsilon/2\right) &\le \exp\left(-\frac{N\left(-\tau\epsilon\right)^2/4}{4\epsilon + \tau\epsilon/3}\right)\\ &= \exp\left(-\frac{\xi\ln\frac{2}{\delta}}{16/\tau + 4/3}\right)\\ &\le \delta/2\end{aligned}$$

Thus, by a union bound,

$$\begin{aligned}&\Pr\left(\sum_{i=1}^{N} X_i \le q(N,\mathrm{Var},\delta)\right)\\ \le&\Pr\left(\frac{q(N,\mathrm{Var},\delta)}{N} - \mathbb{E}X_i \ge -\tau\epsilon/2\right)\\ &+\Pr\left(\frac{q(N,\mathrm{Var},\delta)}{N} - \mathbb{E}X_i \le -\tau\epsilon/2 \text{ and } \frac{1}{N}\sum_{i=1}^{N} X_i \le \frac{q(N,\mathrm{Var},\delta)}{N}\right)\\ \le&\delta/2 + \Pr\left(\frac{q(N,\mathrm{Var},\delta)}{N} - \mathbb{E}X_i \le -\tau\epsilon/2 \text{ and } \frac{1}{N}\sum_{i=1}^{N} Y_i \le \frac{q(n,\mathrm{Var},\delta)}{N} - \mathbb{E}X_i\right)\\ \le&\delta/2 + \Pr\left(\frac{1}{N}\sum_{i=1}^{N} Y_i \le -\tau\epsilon/2\right)\\ \le&\delta\end{aligned}$$

□

A.2 The one-dimensional case

Proof of Theorem 5. Since $\hat{\theta} = \left(L_{\log\frac{1}{2\epsilon}} + R_{\log\frac{1}{2\epsilon}}\right)/2$ and $R_{\log\frac{1}{2\epsilon}} - L_{\log\frac{1}{2\epsilon}} = 2\epsilon$, $\left|\hat{\theta} - \theta^*\right| > \epsilon$ is equivalent to $\theta^* \notin [L_{\log\frac{1}{2\epsilon}}, R_{\log\frac{1}{2\epsilon}}]$. We have

$$\begin{aligned}\Pr\left(\left|\hat{\theta} - \theta^*\right| > \epsilon\right) &= \Pr\left(\theta^* \notin [L_{\log\frac{1}{2\epsilon}}, R_{\log\frac{1}{2\epsilon}}]\right)\\ &= \Pr\left(\exists k : \theta^* \in [L_k, R_k] \text{ and } \theta^* \notin [L_{k+1}, R_{k+1}]\right)\\ &\le \sum_{k=0}^{\log\frac{1}{2\epsilon}-1} \Pr\left(\theta^* \in [L_k, R_k] \text{ and } \theta^* \notin [L_{k+1}, R_{k+1}]\right)\end{aligned}$$

For any $k = 0, \ldots, \log\frac{1}{2\epsilon} - 1$, define $\mathbb{Q}_k = \left\{(p,q) : p, q \in \mathbb{Q}\cap[0,1] \text{ and } q - p = \left(\frac{3}{4}\right)^k\right\}$ where $\mathbb{Q}$ is the set of rational numbers. Note that $L_k, R_k \in \mathbb{Q}_k$, and $\mathbb{Q}$ is countable. So we have

$$
\begin{aligned}
&\Pr\left(\theta^* \in [L_k, R_k] \text{ and } \theta^* \notin [L_{k+1}, R_{k+1}]\right) \\
=& \sum_{(p,q)\in\mathbb{Q}_k : p \le \theta^* \le q} \Pr\left(L_k = p, R_k = q \text{ and } \theta^* \notin [L_{k+1}, R_{k+1}]\right) \\
=& \sum_{(p,q)\in\mathbb{Q}_k : p \le \theta^* \le q} \Pr\left(\theta^* \notin [L_{k+1}, R_{k+1}] | L_k = p, R_k = q\right) \Pr\left(L_k = p, R_k = q\right)
\end{aligned}
$$

Define event $E_{k,p,q}$ to be the event $L_k = p, R_k = q$. To show $\Pr\left(\left|\hat{\theta} - \theta^*\right| > \epsilon\right) \le \frac{\delta}{2}$, it suffices to show $\Pr\left(\theta^* \notin [L_{k+1}, R_{k+1}] | E_{k,p,q}\right) \le \frac{\delta}{2\log\frac{1}{2\epsilon}}$ for any $k = 0, \ldots, \log\frac{1}{2\epsilon} - 1$, $(p,q) \in \mathbb{Q}_k$ and $p \le \theta^* \le q$.

Conditioning on event $E_{k,p,q}$, event $\theta^* \notin [L_{k+1}, R_{k+1}]$ happens only if some calls of CheckSignificant and CheckSignificant-Var between Line 16 and 27 of Algorithm 1 return true incorrectly. In other words, at least one of following events happens for some n:

- $O^{(1)}_{k,p,q}$: $\theta^* \in [L_k, U_k]$ and CheckSignificant-Var$(\left\{A^{(u)}_i - A^{(m)}_i\right\}_{i=1}^n, \frac{\delta}{4\log\frac{1}{2\epsilon}})$ returns true;
- $O^{(2)}_{k,p,q}$: $\theta^* \in [V_k, R_k]$ and CheckSignificant-Var$(\left\{A^{(v)}_i - A^{(m)}_i\right\}_{i=1}^n, \frac{\delta}{4\log\frac{1}{2\epsilon}})$ returns true;
- $O^{(3)}_{k,p,q}$: $\theta^* \in [L_k, U_k]$ and CheckSignificant$(\left\{-B^{(u)}_i\right\}_{i=1}^n, \frac{\delta}{4\log\frac{1}{2\epsilon}})$ returns true;
- $O^{(4)}_{k,p,q}$: $\theta^* \in [V_k, R_k]$ and CheckSignificant$(\left\{B^{(v)}_i\right\}_{i=1}^n, \frac{\delta}{4\log\frac{1}{2\epsilon}})$ returns true;

Note that since $[U_k, V_k] \subset [L_{k+1}, R_{k+1}]$ for any k by our construction, if $\theta^* \in [U_k, V_k]$ then $\theta^* \in [L_{k+1}, R_{k+1}]$. Besides, event $\theta^* \in [L_k, U_k]$ and event $\theta^* \in [V_k, R_k]$ are mutually exclusive.

Conditioning on event $E_{k,p,q}$, suppose for now $\theta^* \in [L_k, U_k]$.

$$
\begin{aligned}
&\Pr\left(O^{(1)}_{k,p,q} \mid E_{k,p,q}\right) \\
=&\Pr\left(\exists n : \text{CheckSignificant-Var}(\left\{D^{(u,m)}_i\right\}_{i=1}^n, \frac{\delta}{4\log\frac{1}{2\epsilon}}) \text{ returns true} \mid \theta^* \in [L_k, U_k], E_{k,p,q}\right)
\end{aligned}
$$

On event $\theta^* \in [L_k, U_k]$ and $E_{k,p,q}$, the sequences $\left\{A^{(u)}_i\right\}$ and $\left\{A^{(m)}_i\right\}$ are i.i.d., and $\mathbb{E}\left[A^{(u)}_i - A^{(m)}_i \mid \theta^* \in [L_k, U_k], E_{k,p,q}\right] \le 0$. By Lemma 17, the probability above is at most $\frac{\delta}{4\log\frac{1}{2\epsilon}}$.

Likewise,

$$
\begin{aligned}
&\Pr\left(O^{(3)}_{k,p,q} \mid E_{k,p,q}\right) \\
=&\Pr\left(\exists n : \text{CheckSignificant}(\left\{-B^{(u)}_i\right\}_{i=1}^n, \frac{\delta}{4\log\frac{1}{2\epsilon}}) \text{ returns true} \mid \theta^* \in [L_k, U_k], E_{k,p,q}\right)
\end{aligned}
$$

On event $\theta^* \in [L_k, U_k]$ and $E_{k,p,q}$, the sequence $\left\{B^{(u)}_i\right\}$ is i.i.d., and $\mathbb{E}\left[-B^{(u)}_i \mid \theta^* \in [L_k, U_k], E_{k,p,q}\right] \le 0$. By Lemma 15, the probability above is at most $\frac{\delta}{4\log\frac{1}{2\epsilon}}$.

Thus, $\Pr\left(\theta^* \notin [L_{k+1}, R_{k+1}] \mid E_{k,p,q}\right) \le \frac{\delta}{2\log\frac{1}{2\epsilon}}$ when $\theta^* \in [L_k, U_k]$. Similarly, when $\theta^* \in [V_k, R_k]$, we can show $\Pr\left(\theta^* \notin [L_{k+1}, R_{k+1}] \mid E_{k,p,q}\right) \le \Pr\left(O^{(2)}_{k,p,q} \mid E_{k,p,q}\right) + \Pr\left(O^{(4)}_{k,p,q} \mid E_{k,p,q}\right) \le \frac{\delta}{2\log\frac{1}{2\epsilon}}$.

Therefore, $\Pr\left(\theta^* \notin [L_{k+1}, R_{k+1}] \mid E_{k,p,q}\right) \le \frac{\delta}{2\log\frac{1}{2\epsilon}}$, and thus $\Pr\left(\left|\hat{\theta} - \theta^*\right| > \epsilon\right) \le \delta/2$. □

Proof of Theorem 6. Define T_k to be the number of iterations of the loop at Line 6, $T = \sum_{k=0}^{\log \frac{1}{2\epsilon}-1} T_k$. For any numbers $m_1, m_2, \ldots, m_{\log \frac{1}{2\epsilon}-1}$, we have:

$$\begin{aligned}
\Pr(T \geq m) &\leq \Pr\left(\left|\hat{\theta} - \theta^*\right| > \epsilon\right) + \Pr\left(\left|\hat{\theta} - \theta^*\right| < \epsilon \text{ and } T \geq \sum_{k=0}^{\log \frac{1}{2\epsilon}-1} m_k\right) \\
&\leq \frac{\delta}{2} + \Pr\left(T \geq \sum_{k=0}^{\log \frac{1}{2\epsilon}-1} m_k \text{ and } \left|\hat{\theta} - \theta^*\right| < \epsilon\right) \\
&\leq \frac{\delta}{2} + \sum_{k=0}^{\log \frac{1}{2\epsilon}-1} \Pr\left(T_k \geq m_k \text{ and } \left|\hat{\theta} - \theta^*\right| < \epsilon\right) \\
&\leq \frac{\delta}{2} + \sum_{k=0}^{\log \frac{1}{2\epsilon}-1} \Pr\left(T_k \geq m_k \text{ and } \theta^* \in [L_k, R_k]\right)
\end{aligned} \tag{1}$$

The first and the third inequality follows by union bounds. The second follows by Theorem 5. The last follows since $\left|\hat{\theta} - \theta^*\right| < \epsilon$ is equivalent to $\theta^* \in [L_{\log \frac{1}{2\epsilon}}, R_{\log \frac{1}{2\epsilon}}]$, which implies $\theta^* \in [L_k, R_k]$ for all $k = 0, \ldots, \log \frac{1}{2\epsilon} - 1$.

We define $\mathbb{Q}_k$ as in the previous proof. For all $k = 0, \ldots, \log \frac{1}{2\epsilon} - 1$,

$$\begin{aligned}
&\Pr(T_k \geq m_k \text{ and } \theta^* \in [L_k, R_k]) \\
= &\sum_{(p,q)\in\mathbb{Q}_k : p \leq \theta^* \leq q} \Pr(T_k \geq m_k, L_k = p, R_k = q) \\
= &\sum_{(p,q)\in\mathbb{Q}_k : p \leq \theta^* \leq q} \Pr(T_k \geq m_k | L_k = p, R_k = q) \Pr(L_k = p, R_k = q)
\end{aligned}$$

Thus, in order to prove the query complexity of Algorithm 1 is $O\left(\sum_{k=0}^{\log \frac{1}{2\epsilon}-1} m_k\right)$, it suffices to show that $\Pr(T_k \geq m_k \mid L_k = p, R_k = q) \leq \frac{\delta}{2 \log \frac{1}{2\epsilon}}$ for any $k = 0, \ldots, \log \frac{1}{2\epsilon} - 1$, $(p, q) \in \mathbb{Q}_k$ and $p \leq \theta^* \leq q$.

For each k, p, q, define event $E_{k,p,q}$ to be the event $L_k = p, R_k = q$. Define $l_k = q - p = \left(\frac{3}{4}\right)^k$, N_k to be $\tilde{\Theta}\left(\frac{1}{f(l_k/4)} l_k^{-2\beta}\right)$. The logarithm factor of N_k is to be specified later. Define $S_n^{(u)}$ and $S_n^{(v)}$ to be the size of array $B^{(u)}$ and $B^{(v)}$ before Line 16 respectively.

To show $\Pr(T_k \geq N_k \mid E_{k,p,q}) \leq \frac{\delta}{2 \log \frac{1}{2\epsilon}}$, it suffices to show that on event $E_{k,p,q}$, with probability at least $1 - \frac{\delta}{2 \log \frac{1}{2\epsilon}}$, if $n = N_k$ then at least one of the two calls to CheckSignificant between Line 22 and Line 27 will return true.

On event $E_{k,p,q}$, if $\theta^* \in [L_k, M_k]$ (note that on event $E_{k,p,q}$, L_k and M_k are deterministic), then $|V_k - \theta^*| \geq \frac{l_k}{4}$. We will show

$$p_1 := \Pr\left(\text{CheckSignificant}\left(\left\{B_i^{(v)}\right\}_{i=1}^{S_{N_k}^{(v)}}, \frac{\delta}{4 \log \frac{1}{2\epsilon}}\right) \text{ returns false} \mid E_{k,p,q}\right) \leq \frac{\delta}{2 \log \frac{1}{2\epsilon}}$$

To prove this, we will first show that $S_{N_k}^{(v)}$, the length of the array $B^{(v)}$, is large with high probability, and then apply Lemma 16 to show that CheckSignificant will return true if $S_{N_k}^{(v)}$ is large.

By definition, $S_{N_k}^{(v)} = \sum_{i=1}^{N_k} A_i^{(v)}$. By Assumption 2, $\mathbb{E}\left[A_i^{(u)} \mid E_{k,p,q}\right] = \Pr(Y \neq \perp \mid X = U_k, E_{k,p,q}) \geq f\left(\frac{l_k}{4}\right)$.

On event $E_{k,p,q}$, $\left\{A_i^{(v)}\right\}$ is a sequence of i.i.d. random variables. By the multiplicative Chernoff bound, $\Pr\left(S_{N_k}^{(v)} \le \frac{1}{2}N_k f\left(\frac{l_k}{4}\right) \mid E_{k,p,q}\right) \le \exp\left(-N_k f\left(\frac{l_k}{4}\right)/8\right)$.

Now,

$$\begin{aligned} p_1 \le & \Pr\left(\text{CheckSignificant}\left(\left\{B_i^{(v)}\right\}_{i=1}^{S_{N_k}^{(v)}}, \frac{\delta}{4\log\frac{1}{2\epsilon}}\right) \text{ returns false}, S_{N_k}^{(v)} \ge \frac{1}{2}N_k f\left(\frac{l_k}{4}\right) \mid E_{k,p,q}\right) \\ & + \Pr\left(S_{N_k}^{(v)} < \frac{1}{2}N_k f\left(\frac{l_k}{4}\right) \mid E_{k,p,q}\right) \end{aligned}$$

By Assumption 2 and $|V_k - \theta^*| \ge \frac{l_k}{4}$, $\mathbb{E}\left[B_i^{(v)} \mid E_{k,p,q}\right] \ge C_2\left(\frac{l_k}{4}\right)^\beta$. On event $E_{k,p,q}$, $\left\{B_i^{(v)}\right\}$ is a sequence of i.i.d. random variables. Thus, On event $E_{k,p,q}$, by Lemma 16, with probability at least $1 - \frac{\delta}{4\log\frac{1}{2\epsilon}}$, CheckSignificant will return true if $\frac{1}{2}N_k f\left(\frac{l_k}{4}\right) = \Theta\left(\frac{1}{l_k^{2\beta}}\ln\frac{\ln 1/\epsilon}{\delta}[\ln\ln]_+ \frac{1}{l_k^{2\beta}}\right)$. We have already proved $\Pr\left(S_{N_k}^{(v)} \le \frac{1}{2}N_k f\left(\frac{l_k}{4}\right) \mid E_{k,p,q}\right) \le \exp\left(-N_k f\left(\frac{l_k}{4}\right)/8\right)$. By setting $N_k = \Theta\left(\frac{1}{f(l_k/4)} l_k^{-2\beta}\ln\frac{\ln 1/\epsilon}{\delta}[\ln\ln]_+ \frac{1}{l_k^{2\beta}}\right)$, we can ensure p_1 is at most $\delta/2\log\frac{1}{2\epsilon}$.

Now we have proved on event $E_{k,p,q}$, if $\theta^* \in [L_k, M_k]$, then

$$\Pr\left(\text{CheckSignificant}\left(\left\{B_i^{(v)}\right\}_{i=1}^{S_{N_k}^{(v)}}, \frac{\delta}{4\log\frac{1}{2\epsilon}}\right) \text{ returns true} \mid E_{k,p,q}\right) \ge 1 - \frac{\delta}{2\log\frac{1}{2\epsilon}}$$

Likewise, on event $E_{k,p,q}$, if $\theta^* \in [M_k, R_k]$, then

$$\Pr\left(\text{CheckSignificant}\left(\left\{-B_i^{(u)}\right\}_{i=1}^{S_{N_k}^{(u)}}, \frac{\delta}{4\log\frac{1}{2\epsilon}}\right) \text{ returns true} \mid E_{k,p,q}\right) \ge 1 - \frac{\delta}{2\log\frac{1}{2\epsilon}}$$

Therefore, we have shown $\Pr\left(T_k \ge N_k \mid E_{k,p,q}\right) \le \frac{\delta}{2\log\frac{1}{2\epsilon}}$ for any k, p, q. By (1), with probability at least $1-\delta$, the number of samples queried is at most

$$\begin{aligned} & \sum_{k=0}^{\log\frac{1}{2\epsilon}-1} O\left(\frac{1}{f(\left(\frac{3}{4}\right)^k/4)}\left(\frac{3}{4}\right)^{-2\beta k}\ln\frac{\ln 1/\epsilon}{\delta}[\ln\ln]_+\left(\frac{3}{4}\right)^{-2k\beta}\right) \\ =& O\left(\frac{\epsilon^{-2\beta}}{f(\epsilon/2)}\ln\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln\ln\frac{1}{\epsilon}\right)[\ln\ln]_+\frac{1}{\epsilon}\right) \end{aligned}$$

□

Proof of Theorem 7. For each k in Algorithm 1 at Line 3, Let $l_k = R_k - L_k$. Let $N_k = \eta\frac{1}{f(l_k/4)}\ln\frac{4\log\frac{1}{2\epsilon}}{\delta}$, where η is a constant to be specified later. As with the previous proof, it suffices to show $\Pr\left(T_k \ge N_k \mid E_{k,p,q}\right) \le \frac{\delta}{2\log\frac{1}{2\epsilon}}$ where event $E_{k,p,q}$ is defined to be $L_k = p, R_k = q$, T_k is the number of iterations at the loop at Line 6.

On event $E_{k,p,q}$, we will show that the loop at Line 6 will terminate after $n = N_k$ with probability at least $1 - \frac{\delta}{2\log\frac{1}{2\epsilon}}$.

Suppose for now $\theta^* \in [M_k, R_k]$. Let $Z_i = A_i^{(u)} - A_i^{(m)}$, $\zeta = \theta^* - M_k$. Clearly, $|Z_i| \le 1$. On event $E_{k,p,q}$, sequence $\{Z_i\}$ is i.i.d.. By the c-growth property, $\mathbb{E}\left[Z_i \mid E_{k,p,q}\right] = f(\zeta + \frac{l_k}{4}) - f(\zeta) \ge cf(\zeta + \frac{l_k}{4})$ since $\zeta \le \frac{2}{3}(\zeta + \frac{l_k}{4})$. $\mathrm{Var}\left[Z_i | E_{k,p,q}\right] = \mathrm{Var}\left[A_i^{(u)} \mid E_{k,p,q}\right] + \mathrm{Var}\left[A_i^{(m)} \mid E_{k,p,q}\right] \overset{(a)}{\le} \mathbb{E}\left[A_i^{(u)} \mid E_{k,p,q}\right] + \mathbb{E}\left[A_i^{(m)} \mid E_{k,p,q}\right] = f(\zeta + \frac{l_k}{4}) + f(\zeta) \overset{(b)}{\le} 2f(\zeta + \frac{l_k}{4})$ where (a) follows by $A_i \in \{0,1\}$ and (b) follows by the monotonicity of

f . Thus, on event $E_{k,p,q}$, by Lemma 18, if we set η sufficiently large (independent of l_k, ϵ, δ), then with probability at least $1 - \frac{\delta}{4\log\frac{1}{2\epsilon}}$ CheckSignificant-Var$\left(\{Z_i\}_{i=1}^{N_k}, \frac{\delta}{4\log\frac{1}{2\epsilon}}\right)$ in Procedure 2 returns true.

Similarly, we can show that on event $E_{k,p,q}$, if $\theta^* \in [L_k, M_k]$, by Lemma 18, with probability at least $1 - \frac{\delta}{4\log\frac{1}{2\epsilon}}$, CheckSignificant-Var$\left(\left\{A_i^{(v)} - A_i^{(m)}\right\}_{i=1}^{N_k}, \frac{\delta}{4\log\frac{1}{2\epsilon}}\right)$ returns true.

Therefore, the loop at Line 6 will terminate after $n = N_k$ with probability at least $1 - \frac{\delta}{4\log\frac{1}{2\epsilon}}$ on event $E_{k,p,q}$. Therefore, with probability at least $1-\delta$, the number of samples queried is at most $\sum_{k=0}^{\log\frac{1}{2\epsilon}-1} \frac{1}{f((\frac{3}{4})^k/4)} \ln\frac{\ln 1/\epsilon}{\delta} = O\left(\frac{1}{f(\epsilon/2)}\ln\frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln\ln\frac{1}{\epsilon}\right)\right)$. □

A.3 The d-dimensional case

To prove the d-dimensional case, we only need to use a union bound to show that with high probability all calls of Algorithm 1 succeed, and consequently the output boundary g produced by polynomial interpolation is close to the true underlying boundary due to the smoothness assumption of g^*.

Proof of Theorem 12. For $q \in \left\{0, 1, \ldots, \frac{M}{\gamma} - 1\right\}^{d-1}$, define the "polynomial interpolation" version of g^* as

$$g_q^*(\tilde{\boldsymbol{x}}) = \sum_{l \in I_q \cap \mathcal{L}} g^*(l) Q_{q,l}(\tilde{\boldsymbol{x}})$$

Recall that we choose $M = O\left(\epsilon^{-1/\gamma}\right)$.

By Theorem 5, each run of Algorithm 1 at the line 3 of Algorithm 3 will return a g_l such that $\left|g_l - g_q^*(l)\right| \leq \epsilon$ with probability at least $1 - \delta/2M^{d-1}$.

$$\begin{aligned}
&\|g - g^*\| \\
=& \sum_{q \in \{0,\ldots,M/\gamma-1\}^{d-1}} \|(g_q - g^*)\,\mathbb{1}\{\tilde{\boldsymbol{x}} \in I_q\}\| \\
\leq& \sum_{q \in \{0,\ldots,M/\gamma-1\}^{d-1}} \left\|\left(g_q - g_q^*\right)\mathbb{1}\{\tilde{\boldsymbol{x}} \in I_q\}\right\| + \left\|\left(g_q^* - g^*\right)\mathbb{1}\{\tilde{\boldsymbol{x}} \in I_q\}\right\|
\end{aligned}$$

$$\begin{aligned}
\left\|\left(g_q^* - g^*\right)\mathbb{1}\{\tilde{\boldsymbol{x}} \in I_q\}\right\| &= \int_{I_q} \left|g_q^*(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})\right| d\tilde{\boldsymbol{x}} \\
&= O\left(\int_{I_q} M^{-\gamma} d\tilde{\boldsymbol{x}}\right) \\
&= O\left(M^{-\gamma-d+1}\right)
\end{aligned}$$

The second equality follows from Lemma 3 of [6] that $|g_q(\tilde{\boldsymbol{x}}) - g^*(\tilde{\boldsymbol{x}})| = O\left(M^{-\gamma}\right)$ since g^* is γ-Hölder smooth.

$$\begin{aligned}
&\left\|\left(g_q - g_q^*\right)\mathbb{1}\{\tilde{\boldsymbol{x}} \in I_q\}\right\| \\
=& \sum_{l \in I_q \cap \mathcal{L}} \left|g_l - g_q^*(l)\right| \|Q_{q,l}\| \\
\leq& \sum_{l \in I_q \cap \mathcal{L}} \epsilon \|Q_q\| \\
=& O(\epsilon M^{-d+1})
\end{aligned}$$

Therefore, overall we have $\|g - g^*\| \le O\left(M^{-\gamma-d+1} + \epsilon M^{-d+1}\right)\left(\frac{M}{\gamma}\right)^{d-1} = O(\epsilon)$. ☐

Proof of Theorem 13. By Theorem 6, each run of Algorithm 1 at the line 3 of Algorithm 3 will make $\tilde{O}\left(\frac{d}{f(\epsilon/2)}\epsilon^{-2\beta}\right)$ queries with probability at least $1 - \delta/M^{d-1}$, thus by a union bound, the total number of queries made is $\tilde{O}\left(\frac{d}{f(\epsilon/2)}\epsilon^{-2\beta-\frac{d-1}{\gamma}}\right)$ with probability at least $1-\delta$. ☐

Proof of Theorem 14. The proof is similar to the previous proof. ☐

B Proof of lower bounds

First, we introduce some notations for this section. Given a labeler L and an active learning algorithm $\mathcal{A}$, denote by $P^n_{L,\mathcal{A}}$ the distribution of n samples $\{(X_i, Y_i)\}_{i=1}^n$ where Y_i is drawn from distribution $P_L(Y|X_i)$ and X_i is drawn by the active learning algorithm based solely on the knowledge of $\{(X_j, Y_j)\}_{j=1}^{i-1}$. We will drop the subscripts of $P^n_{L,\mathcal{A}}$ and $P_{\mathcal{L}}(Y|X)$ when it is clear from the context. For a sequence $\{X_i\}_{i=1}^\infty$ denote by X^n the subsequence $\{X_1, \ldots, X_n\}$.

Definition 19. For any distributions P, Q on a countable support, define KL-divergence as $d_{\mathrm{KL}}(P, Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)}$. For two random variables X, Y, define the mutual information as $I(X;Y) = d_{\mathrm{KL}}(P(X,Y) \parallel P(X)P(Y))$.

We will use Fano's method shown as below to prove the lower bounds.

Lemma 20. *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. Let $\mathcal{V} = \{\theta_1, \ldots, \theta_M\} \subseteq \Theta$ such that $\forall i \ne j$, $d(\theta_i, \theta_j) \ge 2s > 0$. Let $\bar{P} = \frac{1}{M}\sum_{\theta\in\mathcal{V}} P_\theta$. If $d_{KL}\left(P_\theta \parallel \bar{P}\right) \le \delta$ for any $\theta \in \mathcal{V}$, then for any algorithm $\hat{\theta}$ that given a sample X drawn from P_θ outputs $\hat{\theta}(X)$, the following inequality holds:*

$$\sup_\theta P_\theta\left(d(\theta, \hat{\theta}(X)) \ge s\right) \ge 1 - \frac{\delta + \ln 2}{\ln M}$$

Proof. For any algorithm $\hat{\theta}$, define a test function $\hat{\Psi} : \mathcal{X} \to \{1, \ldots, M\}$ such that $\hat{\Psi}(X) = \arg\min_{i\in\{1,\ldots,M\}} d(\hat{\theta}(X), \theta_i)$. We have

$$\sup_\theta P_\theta\left(d(\theta, \hat{\theta}(X)) \ge s\right) \ge \max_{\theta\in\mathcal{V}} P_\theta\left(d(\theta, \hat{\theta}(X)) \ge s\right) \ge \max_{i\in\{1,\ldots,M\}} P_{\theta_i}\left(\hat{\Psi}(X) \ne i\right)$$

Let V be a random variable uniformly taking values from $\mathcal{V}$, and X be drawn from P_V. By Fano's Inequality, for any test function $\Psi : \mathcal{X} \to \{1, \ldots, M\}$

$$\max_{i\in\{1,\ldots,M\}} P_{\theta_i}(\Psi(X) \ne i) \ge 1 - \frac{I(V;X) + \ln 2}{\ln M}$$

The desired result follows by the fact that $I(V;X) = \frac{1}{M}\sum_{\theta\in\mathcal{V}} d_{\mathrm{KL}}\left(P_\theta \parallel \bar{P}\right)$. ☐

B.1 The one dimensional case

Proof of Theorem 9. [2] Without lose of generality, let $C_3 = 1$ in Assumption 2. Let $\epsilon \le \frac{1}{4}\min\left\{\left(\frac{1}{2}\right)^{1/\beta}, \left(\frac{4}{5}\right)^{1/\alpha}, \frac{1}{4}\right\}$. We will prove the desired result using Lemma 20.

First, we construct $\mathcal{V}$ and P_θ. For any $k \in \{0, 1, 2, 3\}$, let $P_{L_k}(Y \mid X)$ be the distribution of the labeler L_k's response with the ground truth $\theta_k = k\epsilon$:

[2] Actually we can use Le Cam's method to prove this one dimensional case (which only needs to construct 2 distributions instead of 4 here), but this proof can be generalized to the multidimensional case more easily.

$$
\begin{aligned}
P_{L_k}(Y=\perp|x) &= 1-\left|x-\frac{1}{2}-k\epsilon\right|^{\alpha} \\
P_{L_k}(Y=0|x) &= \begin{cases} \left(x-\frac{1}{2}-k\epsilon\right)^{\alpha}\left(1-\left(x-\frac{1}{2}-k\epsilon\right)^{\beta}\right)/2 & x>\frac{1}{2}+k\epsilon \\ \left(\frac{1}{2}+k\epsilon-x\right)^{\alpha}\left(1+\left(\frac{1}{2}+k\epsilon-x\right)^{\beta}\right)/2 & x\le\frac{1}{2}+k\epsilon \end{cases} \\
P_{L_k}(Y=1|x) &= \begin{cases} \left(x-\frac{1}{2}-k\epsilon\right)^{\alpha}\left(1+\left(x-\frac{1}{2}-k\epsilon\right)^{\beta}\right)/2 & x>\frac{1}{2}+k\epsilon \\ \left(\frac{1}{2}+k\epsilon-x\right)^{\alpha}\left(1-\left(\frac{1}{2}+k\epsilon-x\right)^{\beta}\right)/2 & x\le\frac{1}{2}+k\epsilon \end{cases}
\end{aligned}
$$

Clearly, P_{L_k} complies with Assumptions 1, 2 and 3.

Define P_k^n to be the distribution of n samples $\{(X_i,Y_i)\}_{i=1}^n$ where Y_i is drawn from distribution $P_{L_k}(Y|X_i)$ and X_i is drawn by the active learning algorithm based solely on the knowledge of $\{(X_j,Y_j)\}_{j=1}^{i-1}$.

Define $\bar{P}_L=\frac{1}{4}\sum_j P_{L_j}$ and $\bar{P}^n=\frac{1}{4}\sum_j P_k^n$. We take Θ to be $[0,1]$, and $d(\theta_1,\theta_2)=|\theta_1-\theta_2|$ in Lemma 20. To use Lemma 20, we need to bound $d_{\mathrm{KL}}\left(P_k^n \parallel \bar{P}^n\right)$ for $k\in\{0,1,2,3\}$.

For any $k\in\{0,1,2,3\}$,

$$
\begin{aligned}
& d_{\mathrm{KL}}\left(P_k^n \parallel \bar{P}_0^n\right) \\
=& \mathbb{E}_{P_k^n}\left(\ln\frac{P_k^n\left(\{(X_i,Y_i)\}_{i=1}^n\right)}{\bar{P}^n\left(\{(X_i,Y_i)\}_{i=1}^n\right)}\right) \\
=& \mathbb{E}_{P_k^n}\left(\ln\frac{P_k^n(X_1)\,P_k^n(Y_1\mid X_1)\,P_k^n(X_2\mid X_1,Y_1)\cdots P_k^n(Y_n\mid X_1,Y_1,\ldots,X_n)}{\bar{P}^n(X_1)\,\bar{P}^n(Y_1\mid X_1)\,\bar{P}^n(X_2\mid X_1,Y_1)\cdots \bar{P}^n(Y_n\mid X_1,Y_1,\ldots,X_n)}\right) \\
\overset{(a)}{=}& \mathbb{E}_{P_k^n}\left(\ln\frac{\Pi_{i=1}^n P_{L_k}(Y_i|X_i)}{\Pi_{i=1}^n \bar{P}_L(Y_i|X_i)}\right) \\
=& \sum_{i=1}^n \mathbb{E}_{P_k^n}\left(\mathbb{E}_{P_k^n}\left(\ln\frac{P_{L_k}(Y_i|X_i)}{\bar{P}_L(Y_i|X_i)}\mid X^n\right)\right) \\
\le& n\max_{x\in[0,1]} d_{\mathrm{KL}}\left(P_{L_k}(Y\mid x)\parallel \bar{P}_L(Y\mid x)\right)
\end{aligned}
\tag{2}
$$

(a) follows by the fact that $P_k^n(X_{i+1}\mid X_1,Y_1,\ldots X_i,Y_i)=\bar{P}^n(X_{i+1}\mid X_1,Y_1,\ldots,X_i,Y_i)$ since X_{i+1} is drawn by the same active learning algorithm based solely on the knowledge of $\{(X_j,Y_j)\}_{j=1}^i$ regardless of the labeler's response distribution, and the fact that $P_k^n(Y_i\mid X_1,Y_1,\ldots,X_i)=P_{\mathcal{L}_k}(Y_i|X_i)$ and $\bar{P}^n(Y_i\mid X_1,Y_1,\ldots,X_i)=\bar{P}_L(Y_i|X_i)$ by definition.

For any $k\in\{1,2,3\}, x\in[0,1]$,

$$
\bar{P}_L(\cdot\mid x)\ge\frac{P_{L_0}(\cdot\mid x)+P_{L_k}(\cdot\mid x)}{4} \tag{3}
$$

For any $k\in\{0,1,2,3\}, x\in[0,1]$, $y\in\{1,-1,\perp\}$

$$
\begin{aligned}
&\left(\bar{P}_L(Y=y\mid x)-P_{L_k}(Y=y\mid x)\right)^2\\
=&\left(\sum_j \frac{1}{4}\left(P_{L_j}(Y=y\mid x)-P_{L_0}(Y=y\mid x)\right)+\left(P_{L_0}(Y=y\mid \boldsymbol{x})-P_{L_k}(Y=y\mid x)\right)\right)^2\\
\leq&\left(\frac{5}{16}\sum_{j>0}\left(P_{L_j}(Y=y\mid x)-P_{L_0}(Y=y\mid x)\right)^2+5\left(P_{L_0}(Y=y\mid x)-P_{L_k}(Y=y\mid x)\right)^2\right)\\
\leq&6\sum_{j>0}\left(P_{L_j}(Y=y\mid x)-P_{L_0}(Y=y\mid x)\right)^2 \qquad (4)
\end{aligned}
$$

where the first inequality follows by $\left(\sum_{i=0}^4 a_i\right)^2 \leq 5\sum_{i=0}^4 a_i^2$ by letting $a_j=\frac{1}{4}\left(P_{L_j}(Y=y\mid x)-P_{L_0}(Y=y\mid x)\right)$ for $j=0,\ldots,3$ and $a_4=P_{L_0}(Y=y\mid \boldsymbol{x})-P_{L_k}(Y=y\mid x)$, and noting that $a_0=0$ under this setting.

Thus,

$$
\begin{aligned}
&d_{\mathrm{KL}}\left(P_{L_k}(Y\mid x)\parallel \bar{P}_L(Y\mid x)\right)\\
\leq&\sum_y \frac{1}{\bar{P}_L(Y=y\mid \boldsymbol{x})}\left(P_{L_k}(Y=y\mid x)-\bar{P}_L(Y=y\mid x)\right)^2\\
\leq&24\sum_{j>0}\sum_y \frac{1}{P_{L_j}(y\mid x)+P_{L_0}(y\mid x)}\left(P_{L_j}(Y=y\mid x)-P_{L_0}(Y=y\mid x)\right)^2\\
\leq&O(\epsilon^\alpha)
\end{aligned}
$$

The first inequality follows from Lemma 25. The second inequality follows by (3) and (4). The last inequality follows by applying Lemma 26 to $P_{L_0}(\cdot\mid x)$ and $P_{L_j}(\cdot\mid x)$ and the assumption $\alpha\leq 2$.

Therefore, we have $d_{\mathrm{KL}}\left(P_k^n\parallel \bar{P}_0^n\right)=nO(\epsilon^\alpha)$. By setting $n=\epsilon^{-\alpha}$, we get $d_{\mathrm{KL}}\left(P_k^n\parallel \bar{P}_0^n\right)\leq O(1)$, and thus by Lemma 20,

$$
\sup_\theta P_\theta\left(d(\theta,\hat{\theta}(X))\geq \Omega(\epsilon)\right)\geq 1-\frac{O(1)+\ln 2}{\ln 4}=O(1)
$$

□

B.2 The d-dimensional case

Again, we will use Lemma 20 to prove the lower bounds for d-dimensional cases. We first construct $\{P_\theta:\theta\in\Theta\}$ using a similar idea with [6], and then use Lemma 27 to select a subset $\tilde{\Theta}\subset\Theta$ to apply Lemma 20.

Proof of Theorem 10. Recall that for $\boldsymbol{x}=(x_1,\ldots,x_d)\in\mathbb{R}^d$, we have defined $\tilde{\boldsymbol{x}}$ to be $(x_1,\ldots,x_{d-1})$. Define $m=\left(\frac{1}{\epsilon}\right)^{1/\gamma}$. $\mathcal{L}=\left\{0,\frac{1}{m},\ldots,\frac{m-1}{m}\right\}^{d-1}$, $h(\tilde{\boldsymbol{x}})=\Pi_{i=1}^{d-1}\exp\left(-\frac{1}{1-4x_i^2}\right)\mathbb{1}\left\{|x_i|<\frac{1}{2}\right\}$, $\phi_l(\tilde{\boldsymbol{x}})=Km^{-\gamma}h(m(\tilde{\boldsymbol{x}}-l)-\frac{1}{2})$ where $l\in\mathcal{L}$. It is easy to check $\phi_l(\tilde{\boldsymbol{x}})$ is (K,γ)-Hölder smooth and has bounded support $[l_1,l_1+\frac{1}{m}]\times\cdots\times[l_{d-1},l_{d-1}+\frac{1}{m}]$, which implies that for different $l_1,l_2\in\mathcal{L}$, the support of ϕ_{l_1} and ϕ_{l_2} do not intersect.

Let $\Omega=\{0,1\}^{m^{d-1}}$. For any $\boldsymbol{\omega}\in\Omega$, define $g_\omega(\tilde{\boldsymbol{x}})=\sum_{l\in\mathcal{L}}\omega_l\phi_l(\tilde{\boldsymbol{x}})$. For each $\boldsymbol{\omega}\in\Omega$, define the conditional distribution of labeler $L_{\boldsymbol{\omega}}$'s response as follows:

For $x_d\leq A$, $P_{L_{\boldsymbol{\omega}}}(y=\perp|\boldsymbol{x})=1-f(A)$, $P_{L_{\boldsymbol{\omega}}}(y\neq\mathbb{I}(x_d>g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}}))|\boldsymbol{x},y\neq\perp)=\frac{1}{2}\left(1-C_3\left|x_d-g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}})\right|^\beta\right)$;

For $x_d\geq A$, $P_{L_{\boldsymbol{\omega}}}(y=\perp|\boldsymbol{x})=1-f(x_d)$, $P_{L_{\boldsymbol{\omega}}}(y\neq\mathbb{I}(x_d>g_\omega(\tilde{\boldsymbol{x}}))|\boldsymbol{x},y\neq\perp)=\frac{1}{2}\left(1-C_3x_d^\beta\right)$.

Here, $A=c\max\phi(\tilde{\boldsymbol{x}})=c'\epsilon$ for some constants c,c'.

It can be easily verified that $P_{L_{\boldsymbol{\omega}}}$ satisfies Assumptions 1 and 2. Note that $g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}})$ can be seen as the underlying decision boundary for labeler $P_{L_{\boldsymbol{\omega}}}$.

Define $P_{\boldsymbol{\omega}}^n$ to be the distribution of n samples $\{(X_i, Y_i)\}_{i=1}^n$ where Y_i is drawn from distribution $P_{L_{\boldsymbol{\omega}}}(Y|X_i)$ and X_i is drawn by the active learning algorithm based solely on the knowledge of $\{(X_j, Y_j)\}_{j=1}^{i-1}$.

By Lemma 27, when ϵ is small enough so that m^{d-1} is large enough, there is a subset $\left\{\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(M)}\right\} \subset \Omega$ such that $\left\|\boldsymbol{\omega}^{(i)} - \boldsymbol{\omega}^{(j)}\right\|_0 \geq m^{d-1}/12$ for any $0 \leq i < j \leq M$ and $M \geq 2^{m^{d-1}/48}$. Define $P_i^n = P_{\boldsymbol{\omega}^{(i)}}^n, \bar{P}^n = \frac{1}{M}\sum_{i=1}^M P_i^n$.

Next, we will apply Lemma 20 to $\left\{\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(M)}\right\}$ with $d(\boldsymbol{\omega}^{(i)}, \boldsymbol{\omega}^{(j)}) = \|g_{\boldsymbol{\omega}^{(i)}} - g_{\boldsymbol{\omega}^{(j)}}\|$. We will lower-bound $d(\boldsymbol{\omega}^{(i)}, \boldsymbol{\omega}^{(j)})$ and upper-bound $d_{\mathrm{KL}}\left(P_i^n \parallel \bar{P}^n\right)$.

For any $1 \leq i < j \leq M$,

$$\begin{aligned}
&\|g_{\boldsymbol{\omega}^{(i)}} - g_{\boldsymbol{\omega}^{(j)}}\| \\
=&\sum_{l \in \{1,\ldots,m\}^{d-1}} \left|\omega_l^{(i)} - \omega_l^{(j)}\right| K m^{-\gamma-(d-1)} \|h\| \\
\geq& m^{d-1}/12 * K m^{-\gamma-(d-1)} \|h\| \\
=& K m^{-\gamma} \|h\| / 12 \\
=& \Theta(\epsilon)
\end{aligned}$$

By the convexity of KL-divergence, $d_{\mathrm{KL}}\left(P_i^n \parallel \bar{P}^n\right) \leq \frac{1}{M}\sum_{j=1}^M d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right)$, so it suffices to upper-bound $d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right)$ for any i, j.

For any $1 < i, j \leq M$,

$$\begin{aligned}
&d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right) \\
\leq& n \max_{\boldsymbol{x} \in [0,1]^d} d_{\mathrm{KL}}\left(P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \mid \boldsymbol{x}) \parallel P_{L_{\boldsymbol{\omega}^{(j)}}}^n(Y \mid \boldsymbol{x})\right) \\
=& n \max_{\boldsymbol{x} \in [0,1]^d} P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \neq \perp \mid \boldsymbol{x}) d_{\mathrm{KL}}\left(P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp) \parallel P_{L_{\boldsymbol{\omega}^{(j)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp)\right)
\end{aligned}$$

The inequality follows as (2) in the proof of Theorem 9. The equality follows since $P_{\boldsymbol{\omega}}(y = \perp | \boldsymbol{x})$ is the same for all $\boldsymbol{\omega} \in \Omega$.

If $x_d \geq A$, then $P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp) = P_{L_{\boldsymbol{\omega}^{(j)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp)$, so $d_{\mathrm{KL}}\left(P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp) \parallel P_{L_{\boldsymbol{\omega}^{(j)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp)\right) = 0$. If $x_d < A$, then $P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \neq \perp \mid \boldsymbol{x}) = f(A)$. Therefore,

$$d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right) \leq n f(A) \max_{\boldsymbol{x} \in [0,1]^d} d_{\mathrm{KL}}\left(P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp) \parallel P_{L_{\boldsymbol{\omega}^{(j)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp)\right)$$

.

Apply Lemma 25 to $P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp)$ and $P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp)$, and noting they are bounded above by a constant, we have $\max_{\boldsymbol{x} \in [0,1]^d} d_{\mathrm{KL}}\left(P_{L_{\boldsymbol{\omega}^{(i)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp) \parallel P_{L_{\boldsymbol{\omega}^{(j)}}}^n(Y \mid \boldsymbol{x}, Y \neq \perp)\right) = O\left(A^{2\beta}\right)$. Thus,

$$d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right) \leq n f(A) O\left(A^{2\beta}\right) = n f(c'\epsilon) O(\epsilon^{2\beta})$$

By setting $n = \frac{1}{f(c'\epsilon)}\epsilon^{-2\beta-\frac{d-1}{\gamma}}$, we get $d_{\mathrm{KL}}\left(P_i^n \parallel P_j^n\right) \leq O\left(\epsilon^{-\frac{d-1}{\gamma}}\right)$. The desired results follows by Lemma 20. $\square$

The proof of Theorem 11 follows the same structure.

Proof of Theorem 11. As in the proof of Theorem 10, define $m = \left(\frac{1}{\epsilon}\right)^{1/\gamma}$. $\mathcal{L} = \left\{0, \frac{1}{m}, \ldots, \frac{m-1}{m}\right\}^{d-1}$, $h(\tilde{\boldsymbol{x}}) = \Pi_{i=1}^{d-1} \exp\left(-\frac{1}{1-4x_i^2}\right) \mathbb{1}\left\{|x_i| < \frac{1}{2}\right\}$, $\phi_l(\tilde{\boldsymbol{x}}) = K m^{-\gamma} h(m(\tilde{\boldsymbol{x}} - l) - \frac{1}{2})$ where $l \in \mathcal{L}$. Let $\Omega = \{0,1\}^{m^{d-1}}$. For any $\boldsymbol{\omega} \in \Omega$, define $g_\omega(\tilde{\boldsymbol{x}}) = \frac{1}{2} + \sum_{l \in \mathcal{L}} \omega_l \phi_l(\tilde{\boldsymbol{x}})$, which can be seen as a decision boundary. $A = \max \phi(\tilde{\boldsymbol{x}}) = c'\epsilon$ for some constants c'.

Let $g_+(\tilde{\boldsymbol{x}}) = g_{(1,1,\dots,1)}(\tilde{\boldsymbol{x}}) = \sum_{l\in\mathcal{L}} \phi_l(\tilde{\boldsymbol{x}})$, $g_-(\tilde{\boldsymbol{x}}) = g_{(0,0,\dots,0)}(\tilde{\boldsymbol{x}}) = 0$. In other words, g_+ is the "highest" boundary, and g_- is the "lowest" boundary.

For each $\boldsymbol{\omega} \in \Omega$, define the conditional distribution of labeler $L_{\boldsymbol{\omega}}$'s response as follows:

$$P_{L_{\boldsymbol{\omega}}}(y = \perp | \boldsymbol{x}) = 1 - C_2 \left| x_d - g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}}) \right|^{\alpha}$$

$$P_{L_{\boldsymbol{\omega}}}(y \neq \mathbb{I}(x_d > g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}}))|\boldsymbol{x}, y \neq \perp) = \frac{1}{2}\left(1 - C_3 \left| x_d - g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}}) \right|^{\beta}\right)$$

It can be easily verified that $P_{L_{\boldsymbol{\omega}}}$ satisfies Assumptions 1, 2, and 3. Without lose of generality, let $C_2 = C_3 = 1$.

Let $P_+(\cdot \mid \boldsymbol{x}) = P_{L_{(1,1,\dots,1)}}(\cdot \mid \boldsymbol{x})$, $P_-(\cdot \mid \boldsymbol{x}) = P_{L_{(0,0,\dots,0)}}(\cdot \mid \boldsymbol{x})$. By the construction of g, for any $\boldsymbol{x} \in [0,1]^d$, any $\boldsymbol{\omega} \in \Omega$, $P_{L_{\boldsymbol{\omega}}}(\cdot \mid \boldsymbol{x})$ equals either $P_+(\cdot \mid \boldsymbol{x})$ or $P_-(\cdot \mid \boldsymbol{x})$.

Define $P^n_{\boldsymbol{\omega}}$ to be the distribution of n samples $\{(X_i, Y_i)\}_{i=1}^n$ where Y_i is drawn from distribution $P_{L_{\boldsymbol{\omega}}}(Y|X_i)$ and X_i is drawn by the active learning algorithm based solely on the knowledge of $\{(X_j, Y_j)\}_{j=1}^{i-1}$.

By Lemma 27, when ϵ is small enough so that m^{d-1} is large enough,, there is a subset $\Omega' = \left\{\boldsymbol{\omega}^{(1)}, \dots, \boldsymbol{\omega}^{(M)}\right\} \subset \Omega$ such that (i) (well-separated) $\left\|\boldsymbol{\omega}^{(i)} - \boldsymbol{\omega}^{(j)}\right\|_0 \geq m^{d-1}/12$ for any $0 \leq i < j \leq M$, $M \geq 2^{m^{d-1}/48}$; and (ii) (well-balanced) for any $j = 1, \dots, m^{d-1}$, $\frac{1}{24} \leq \frac{1}{M}\sum_{i=1}^M \boldsymbol{\omega}_j^{(i)} \leq \frac{3}{24}$.

Define $P_i^n = P^n_{\boldsymbol{\omega}^{(i)}}, \bar{P}^n = \frac{1}{M}\sum_{i=1}^M P_i^n$. Define $P_{L_i} = P_{L_{\boldsymbol{\omega}^{(i)}}}$, $\bar{P}_L = \frac{1}{M}\sum_{i=1}^M P_{L_i}$. By the well-balanced property, for any $\boldsymbol{x} \in [0,1]^d$, $\bar{P}_L(\cdot \mid \boldsymbol{x})$ is between $\frac{1}{24}P_+(\cdot \mid \boldsymbol{x}) + \frac{23}{24}P_-(\cdot \mid \boldsymbol{x})$ and $\frac{3}{24}P_+(\cdot \mid \boldsymbol{x}) + \frac{21}{24}P_-(\cdot \mid \boldsymbol{x})$. Therefore

$$\bar{P}_L(\cdot \mid \boldsymbol{x}) \geq \frac{1}{24}\left(P_+(\cdot \mid \boldsymbol{x}) + P_-(\cdot \mid \boldsymbol{x})\right) \tag{5}$$

Moreover, since $P_{L_i}(\cdot \mid \boldsymbol{x})$ can only take $P_+(\cdot \mid \boldsymbol{x})$ or $P_-(\cdot \mid \boldsymbol{x})$ for any $\boldsymbol{x}$,

$$\left|P_{L_i}(\cdot \mid \boldsymbol{x}) - \bar{P}_{\mathcal{L}}(\cdot \mid \boldsymbol{x})\right| \leq \left|P_+(\cdot \mid \boldsymbol{x}) - P_-(\cdot \mid \boldsymbol{x})\right| \tag{6}$$

Next, we will apply Lemma 20 to $\left\{\boldsymbol{\omega}^{(1)}, \dots, \boldsymbol{\omega}^{(M)}\right\}$ with $d(\boldsymbol{\omega}^{(i)}, \boldsymbol{\omega}^{(j)}) = \|g_{\boldsymbol{\omega}^{(i)}} - g_{\boldsymbol{\omega}^{(j)}}\|$. We already know from the proof of Theorem 10 $\|g_{\boldsymbol{\omega}^{(i)}} - g_{\boldsymbol{\omega}^{(j)}}\| = \Omega(\epsilon)$.

For any $0 < i \leq M$, $d_{\mathrm{KL}}\left(P_i^n \parallel \bar{P}_0^n\right) \leq n \max_{\boldsymbol{x}\in[0,1]^d} d_{\mathrm{KL}}\left(P_{L_i}(Y \mid \boldsymbol{x}) \parallel \bar{P}_L(Y \mid \boldsymbol{x})\right)$. For any $\boldsymbol{x} \in [0,1]^d$,

$$\begin{aligned}
&d_{\mathrm{KL}}\left(P_{L_i}(Y \mid \boldsymbol{x}) \parallel \bar{P}_L(Y \mid \boldsymbol{x})\right) \\
\leq &\sum_y \frac{1}{\bar{P}_L(Y = y \mid \boldsymbol{x})}\left(P_{L_i}(Y = y \mid \boldsymbol{x}) - \bar{P}_L(Y = y \mid \boldsymbol{x})\right)^2 \\
\leq &\sum_y \frac{24}{P_+(y \mid \boldsymbol{x}) + P_-(y \mid \boldsymbol{x})}\left(P_+(Y = y \mid \boldsymbol{x}) - P_-(Y = y \mid \boldsymbol{x})\right)^2 \\
\leq &O(A^{\alpha})
\end{aligned}$$

The first inequality follows from Lemma 25. The second inequality follows by (5) and (6). The last inequality follows by applying Lemma 26 to $P_+(\cdot \mid \boldsymbol{x})$ and $P_-(\cdot \mid \boldsymbol{x})$, setting the ϵ in Lemma 26 to be $g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}})$, and using $g_{\boldsymbol{\omega}}(\tilde{\boldsymbol{x}}) \leq A$ and the assumption $\alpha \leq 2$.

Therefore, we have

$$d_{\mathrm{KL}}\left(P_i^n \parallel P_0^n\right) \leq nO\left(A^{\alpha}\right) = nO(\epsilon^{\alpha})$$

By setting $n = \epsilon^{-\alpha - \frac{d-1}{\gamma}}$, we get $d_{\mathrm{KL}}\left(P_i^n \parallel P_0^n\right) \leq O\left(\epsilon^{-\frac{d-1}{\gamma}}\right)$. Thus by Lemma 20,

$$\sup_{\theta} P_{\theta}\left(d(\theta, \hat{\theta}(X)) \geq \Omega\left(\epsilon\right)\right) \geq 1 - \frac{O\left(\epsilon^{-\frac{d-1}{\gamma}}\right) + \ln 2}{\epsilon^{-\frac{d-1}{\gamma}}/48} = O\left(1\right)$$

, from which the desired result follows. ∎

C Technical lemmas

C.1 Concentration bounds

In this subsection, we define $Y_1, Y_2, \dots$ to be a sequence of i.i.d. random variables. Assume $Y_1 \in [-2, 2]$, $\mathbb{E}Y_1 = 0$, $\mathrm{Var}(Y_1) = \sigma^2 \leq 4$. Define $V_n = \frac{n}{n-1}\left(\sum_{i=1}^n Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n Y_i\right)^2\right)$. It is easy to check $\mathbb{E}V_n = n\sigma^2$.

We need following two results from [20]

Lemma 21. *([20], Theorem 2) Take any $0 < \delta < 1$. Then there is an absolute constant D_0 such that with probability at least $1 - \delta$, for all n simultaneously,*

$$\left|\sum_{i=1}^n Y_i\right| \leq D_0 \left(1 + \ln \frac{1}{\delta} + \sqrt{n\sigma^2 \left[\ln \ln\right]_+ (n\sigma^2) + n\sigma^2 \ln \frac{1}{\delta}}\right)$$

Lemma 22. *([20], Lemma 3) Take any $0 < \delta < 1$. Then there is an absolute constant K_0 such that with probability at least $1 - \delta$, for all n simultaneously,*

$$n\sigma^2 \leq K_0 \left(1 + \ln \frac{1}{\delta} + \sum_{i=1}^n Y_i^2\right)$$

We note that Proposition 3 is immediate from Lemma 21 since $\mathrm{Var}(Y_i) \leq 4$.

Lemma 23. *Take any $\delta > 0$. Then there is an absolute constant K_3 such that with probability at least $1-\delta$, for all $n \geq \ln \frac{1}{\delta}$ simultaneously,*

$$n\sigma^2 \leq K_3 \left(1 + \ln \frac{1}{\delta} + V_n\right)$$

Proof. By Lemma 22, with probability at least $1 - \delta/2$,

$$n\sigma^2 \leq K_0 \left(\sum_{i=1}^n Y_i^2 + \ln \frac{2}{\delta} + 1\right) = K_0 \left(\frac{n-1}{n} V_n + \frac{1}{n}\left(\sum_{i=1}^n Y_i\right)^2 + \ln \frac{2}{\delta} + 1\right)$$

By Theorem 21, with probability at least $1 - \delta/2$,

$$\begin{aligned}
\frac{1}{n}\left(\sum_{i=1}^n Y_i\right)^2 &< \frac{1}{n}\left(D_0\left(1 + \ln\frac{2}{\delta} + \sqrt{n\sigma^2 \left[\ln\ln\right]_+ (n\sigma^2) + n\sigma^2 \ln \frac{2}{\delta}}\right)\right)^2 \\
&= \frac{D_0^2}{n}\left(1 + \ln \frac{2}{\delta}\right)^2 + D_0^2\sigma^2 \left[\ln\ln\right]_+ (n\sigma^2) + D_0^2 \sigma^2 \ln \frac{2}{\delta} \\
&\quad + 2D_0^2 \left(1 + \ln \frac{2}{\delta}\right)\sqrt{\frac{\sigma^2 \left[\ln\ln\right]_+ (n\sigma^2) + \sigma^2 \ln \frac{2}{\delta}}{n}} \\
&\leq K_1 \left(1 + \ln \frac{1}{\delta} + \left[\ln\ln\right]_+ (n\sigma^2)\right)
\end{aligned}$$

for some absolute constant K_1. The last inequality follows by $n > \ln \frac{1}{\delta}$.

Thus, by a union bound, with probability at least $1-\delta$, $n\sigma^2 \leq K_0 V_n + K_0(K_1+2)\ln\frac{1}{\delta} + K_0 K_1 \left[\ln\ln\right]_+ (n\sigma^2) + K_0(K_1+3)$.

Let $K_2 > 0$ be an absolute constant such that $\forall x \geq K_2$, $K_0 K_1 \left[\ln\ln\right]_+ x \leq \frac{x}{2}$.

Now if $n\sigma^2 \geq K_2$, then $n\sigma^2 \leq K_0 V_n + K_0(K_1+2)\ln\frac{1}{\delta} + \frac{n\sigma^2}{2} + K_0(K_1+3)$, and thus

$$n\sigma^2 \leq 2K_0 V_n + 2K_0(K_1+2)\ln\frac{1}{\delta} + 2K_0(K_1+3) + K_2 \tag{7}$$

If $n\sigma^2 \leq K_2$, clearly (7) holds. This concludes the proof. □

We note that Proposition 4 is immediate by applying above lemma to Lemma 21.

Lemma 24. *Take any $\delta > 0$. Then with probability at least $1-\delta$,*

$$V_n \leq 4n\sigma^2 + 8\ln\frac{1}{\delta}$$

Proof. Applying Bernstein's Inequality to Y_i^2, and noting that $\mathrm{Var}(Y_i^2) \leq 4\sigma^2$ since $|Y_i| \leq 2$, we have with probability at least $1-\delta$,

$$\begin{aligned}\sum_{i=1}^{n} Y_i^2 &\leq \frac{4}{3}\ln\frac{1}{\delta} + n\sigma^2 + \sqrt{8n\sigma^2\ln\frac{1}{\delta}} \\ &\leq 4\ln\frac{1}{\delta} + 2n\sigma^2\end{aligned}$$

The last inequality follows by the fact that $\sqrt{4ab} \leq a+b$.

The desired result follows by noting that $V_n = \frac{n}{n-1}\left(\sum_{i=1}^n Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n Y_i\right)^2\right) \leq 2\sum_{i=1}^n Y_i^2$. □

C.2 Bounds of distances among probability distributions

Lemma 25. *If P, Q are two probability distributions on a countable support $\mathcal{X}$, then*

$$d_{KL}\left(P \parallel Q\right) \leq \sum_x \frac{\left(P(x)-Q(x)\right)^2}{Q(x)}$$

Proof.

$$\begin{aligned}d_{\mathrm{KL}}\left(P \parallel Q\right) &= \sum_x P(x)\ln\frac{P(x)}{Q(x)} \\ &\leq \sum_x P(x)\left(\frac{P(x)}{Q(x)} - 1\right) \\ &= \sum_x \frac{\left(P(x)-Q(x)\right)^2}{Q(x)}\end{aligned}$$

The first inequality follows by $\ln x \leq x - 1$. The second equality follows by $\sum_x P(x)\left(\frac{P(x)}{Q(x)} - 1\right) = \sum_x\left(\frac{P^2(x)-P(x)Q(x)}{Q(x)} - P(x) + Q(x)\right) = \sum_x \frac{(P(x)-Q(x))^2}{Q(x)}$. □

Define

$$\begin{aligned}P_0\left(Y = \perp | x\right) &= 1 - \left|x - \frac{1}{2}\right|^\alpha \\ P_0\left(Y = 0 | x\right) &= \begin{cases}\left(x-\frac{1}{2}\right)^\alpha\left(1-\left(x-\frac{1}{2}\right)^\beta\right)/2 & x > \frac{1}{2} \\ \left(\frac{1}{2}-x\right)^\alpha\left(1+\left(\frac{1}{2}-x\right)^\beta\right)/2 & x \leq \frac{1}{2}\end{cases} \\ P_0\left(Y = 1 | x\right) &= \begin{cases}\left(x-\frac{1}{2}\right)^\alpha\left(1+\left(x-\frac{1}{2}\right)^\beta\right)/2 & x > \frac{1}{2} \\ \left(\frac{1}{2}-x\right)^\alpha\left(1-\left(\frac{1}{2}-x\right)^\beta\right)/2 & x \leq \frac{1}{2}\end{cases}\end{aligned}$$

and

$$
\begin{aligned}
P_1\left(Y=\perp|x\right) &= 1-\left|x-\epsilon-\frac{1}{2}\right|^{\alpha} \\
P_1\left(Y=0|x\right) &= \begin{cases} \left(x-\epsilon-\frac{1}{2}\right)^{\alpha}\left(1-\left(x-\epsilon-\frac{1}{2}\right)^{\beta}\right)/2 & x>\epsilon+\frac{1}{2} \\ \left(\epsilon+\frac{1}{2}-x\right)^{\alpha}\left(1+\left(\epsilon+\frac{1}{2}-x\right)^{\beta}\right)/2 & x\le\epsilon+\frac{1}{2} \end{cases} \\
P_1\left(Y=1|x\right) &= \begin{cases} \left(x-\epsilon-\frac{1}{2}\right)^{\alpha}\left(1+\left(x-\epsilon-\frac{1}{2}\right)^{\beta}\right)/2 & x>\epsilon+\frac{1}{2} \\ \left(\epsilon+\frac{1}{2}-x\right)^{\alpha}\left(1-\left(\epsilon+\frac{1}{2}-x\right)^{\beta}\right)/2 & x\le\epsilon+\frac{1}{2} \end{cases}
\end{aligned}
$$

Lemma 26. *Let P_0, P_1 be the distributions defined above. If $x\in[0,1]$, $\epsilon\le\min\left\{\left(\frac{1}{2}\right)^{1/\beta},\left(\frac{4}{5}\right)^{1/\alpha},\frac{1}{4}\right\}$, then*

$$
\sum_y \frac{\left(P_0(Y=y|x)-P_1(Y=y|x)\right)^2}{P_0(Y=y|x)+P_1(Y=y|x)} = O\left(\epsilon^{\alpha}+\epsilon^2\right) \tag{8}
$$

Proof. By symmetry, it suffices to show for $0\le x\le\frac{1+\epsilon}{2}$. Let $t=\frac{1}{2}+\epsilon-x$.

We first show (8) holds for $\frac{\epsilon}{2}\le t\le\epsilon$ (i.e. $\frac{1}{2}\le x\le\frac{1+\epsilon}{2}$).

We claim $\min_y\left(P_0(Y=y|X=t)+P_1(Y=y|X=t)\right)\ge\frac{1}{2}\left(\frac{\epsilon}{2}\right)^{\alpha}$. This is because:

- $P_0(Y=\perp|X=t)+P_1(Y=\perp|X=t)=1-(\epsilon-t)^{\alpha}+1-t^{\alpha}\ge 2-2\epsilon^{\alpha}\ge\frac{1}{2}\left(\frac{\epsilon}{2}\right)^{\alpha}$ where the last inequality follows by $\epsilon\le\left(\frac{4}{5}\right)^{1/\alpha}$;
- $2\left(P_0(Y=0|X=t)+P_1(Y=0|X=t)\right)=(\epsilon-t)^{\alpha}\left(1-(\epsilon-t)^{\beta}\right)+t^{\alpha}\left(1+t^{\beta}\right)\ge t^{\alpha}\left(1+t^{\beta}\right)\ge\left(\frac{\epsilon}{2}\right)^{\alpha}$. Therefore, $P_0(Y=0|X=t)+P_1(Y=0|X=t)\ge\frac{1}{2}\left(\frac{\epsilon}{2}\right)^{\alpha}$.
- Similarly, $P_0(Y=1|X=t)+P_1(Y=1|X=t)\ge\frac{1}{2}\left(\frac{\epsilon}{2}\right)^{\alpha}$.

Besides,

$$
\begin{aligned}
&\sum_y\left(P_0(Y=y|X=t)-P_1(Y=y|X=t)\right)^2 \\
=&\left(t^{\alpha}-(\epsilon-t)^{\alpha}\right)^2+\frac{1}{4}\left(t^{\alpha}\left(1-t^{\beta}\right)-(\epsilon-t)^{\alpha}\left(1+(\epsilon-t)^{\beta}\right)\right)^2 \\
&+\frac{1}{4}\left(t^{\alpha}\left(1+t^{\beta}\right)-(\epsilon-t)^{\alpha}\left(1-(\epsilon-t)^{\beta}\right)\right)^2 \\
=&\left(t^{\alpha}-(\epsilon-t)^{\alpha}\right)^2+\frac{1}{4}\left(t^{\alpha}-(\epsilon-t)^{\alpha}-t^{\alpha+\beta}-(\epsilon-t)^{\alpha+\beta}\right)^2 \\
&+\frac{1}{4}\left(t^{\alpha}-(\epsilon-t)^{\alpha}+t^{\alpha+\beta}+(\epsilon-t)^{\alpha+\beta}\right)^2 \\
\overset{(a)}{\le}&\left(t^{\alpha}-(\epsilon-t)^{\alpha}\right)^2+\frac{1}{2}\left(t^{\alpha}-(\epsilon-t)^{\alpha}\right)^2+\frac{1}{2}\left(t^{\alpha+\beta}+(\epsilon-t)^{\alpha+\beta}\right)^2 \\
&+\frac{1}{2}\left(t^{\alpha}-(\epsilon-t)^{\alpha}\right)^2+\frac{1}{2}\left(t^{\alpha+\beta}+(\epsilon-t)^{\alpha+\beta}\right)^2 \\
=&2\left(t^{\alpha}-(\epsilon-t)^{\alpha}\right)^2+\left(t^{\alpha+\beta}+(\epsilon-t)^{\alpha+\beta}\right)^2 \\
\le&2\epsilon^{2\alpha}+4\epsilon^{2\alpha+2\beta} \\
\le&6\epsilon^{2\alpha}
\end{aligned}
$$

where (a) follows by the inequality $(a+b)^2\le 2a^2+2b^2$ for any a,b.

Therefore, we get $\sum_y \frac{(P_0(Y=y|x)-P_1(Y=y|x))^2}{P_0(Y=y|x)+P_1(Y=y|x)} \le \frac{\sum_y (P_0(Y=y|x)-P_1(Y=y|x))^2}{\min_y (P_0(Y=y|x)+P_1(Y=y|x))} \le 12 * 2^\alpha \epsilon^\alpha$ when $\frac{1}{2} \le x \le \frac{1+\epsilon}{2}$.

Next, We show (8) holds for $\epsilon \le t \le \frac{1}{2}+\epsilon$ (i.e. $0 \le x \le \frac{1}{2}$). We will show $\frac{(P_0(Y=y|x)-P_1(Y=y|x))^2}{P_0(Y=y|x)+P_1(Y=y|x)} = O\left(\epsilon^\alpha+\epsilon^2\right)$ for $Y=\perp, 1, 0$.

For $Y=\perp$, for the denominator,

$$P_0(Y=\perp|X=t)+P_1(Y=\perp|X=t) = 2 - t^\alpha - (t-\epsilon)^\alpha \ge 2 - \left(\frac{3}{4}\right)^\alpha - \left(\frac{1}{2}\right)^\alpha$$

For the numerator,

$$(P_0(Y=\perp|X=t) - P_1(Y=\perp|X=t))^2 = (t^\alpha - (t-\epsilon)^\alpha)^2 = t^{2\alpha}\left(1-\left(1-\frac{\epsilon}{t}\right)^\alpha\right)^2$$

By Lemma 28, if $\alpha \ge 1$, $t^{2\alpha}\left(1-\left(1-\frac{\epsilon}{t}\right)^\alpha\right)^2 \le t^{2\alpha}\left(\alpha\frac{\epsilon}{t}\right)^2 = t^{2\alpha-2}(\alpha\epsilon)^2 = O\left(\epsilon^2\right)$. If $0 \le \alpha \le 1$, $t^{2\alpha}\left(1-\left(1-\frac{\epsilon}{t}\right)^\alpha\right)^2 \le t^{2\alpha}\left(\frac{\epsilon}{t}\right)^2 = t^{2\alpha-2}\epsilon^2 \le \epsilon^{2\alpha}$.

Thus, we have $\frac{(P_0(Y=\perp|x)-P_1(Y=\perp|x))^2}{P_0(Y=\perp|x)+P_1(Y=\perp|x)} = O\left(\epsilon^{2\alpha}+\epsilon^2\right)$.

For $Y=1$, for the denominator,

$$\begin{aligned} 2\left(P_0(Y=1|X=t)+P_1(Y=1|X=t)\right) &= t^\alpha\left(1-t^\beta\right) + (t-\epsilon)^\alpha\left(1-(t-\epsilon)^\beta\right) \\ &\ge t^\alpha\left(1-t^\beta\right) \\ &\ge t^\alpha\left(1-\left(\frac{3}{4}\right)^\beta\right) \end{aligned}$$

For the numerator,

$$\begin{aligned} &(P_0(Y=1|X=t) - P_1(Y=1|X=t))^2 \\ =&\frac{1}{4}\left(t^\alpha\left(1-t^\beta\right) - (t-\epsilon)^\alpha\left(1-(t-\epsilon)^\beta\right)\right)^2 \\ \le&\frac{1}{2}\left(t^\alpha-(t-\epsilon)^\alpha\right)^2 + \frac{1}{2}\left(t^{\alpha+\beta}-(t-\epsilon)^{\alpha+\beta}\right)^2 \\ =&\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\epsilon}{t})^\alpha\right)^2 + \frac{1}{2}t^{2\alpha+2\beta}\left(1-(1-\frac{\epsilon}{t})^{\alpha+\beta}\right)^2 \\ \le&\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\epsilon}{t})^\alpha\right)^2 + \frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\epsilon}{t})^{\alpha+\beta}\right)^2 \end{aligned}$$

If $\alpha \ge 1$, by Lemma 28, $\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\epsilon}{t})^\alpha\right)^2 + \frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\epsilon}{t})^{\alpha+\beta}\right)^2 \le \frac{1}{2}t^{2\alpha}\left(\alpha\frac{\epsilon}{t}\right)^2 + \frac{1}{2}t^{2\alpha}\left((\alpha+\beta)\frac{\epsilon}{t}\right)^2 = \left(\frac{1}{2}\alpha^2+\frac{1}{2}(\alpha+\beta)^2\right)t^{2\alpha-2}\epsilon^2$. Thus, $\frac{(P_0(Y=1|x)-P_1(Y=1|x))^2}{P_0(Y=1|x)+P_1(Y=1|x)} \le \left(\frac{1}{2}\alpha^2+\frac{1}{2}(\alpha+\beta)^2\right)t^{\alpha-2}\epsilon^2/\left(1-\left(\frac{3}{4}\right)^\beta\right)$ which is $O(\epsilon^2)$ if $\alpha \ge 2$ and $O\left(\epsilon^\alpha\right)$ if $\alpha \le 2$.

If $\alpha \le 1$ and $\alpha+\beta \ge 1$, by Lemma 28, $\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\epsilon}{t})^\alpha\right)^2 + \frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\epsilon}{t})^{\alpha+\beta}\right)^2 \le \frac{1}{2}t^{2\alpha}\left(\frac{\epsilon}{t}\right)^2 + \frac{1}{2}t^{2\alpha}\left((\alpha+\beta)\frac{\epsilon}{t}\right)^2 = \left(\frac{1}{2}+\frac{1}{2}(\alpha+\beta)^2\right)t^{2\alpha-2}\epsilon^2 \le \left(\frac{1}{2}+\frac{1}{2}(\alpha+\beta)^2\right)t^{2\alpha-2}\epsilon^2$. Thus, $\frac{(P_0(Y=1|x)-P_1(Y=1|x))^2}{P_0(Y=1|x)+P_1(Y=1|x)} \le \left(\frac{1}{2}+\frac{1}{2}(\alpha+\beta)^2\right)t^{\alpha-2}\epsilon^2/\left(1-\left(\frac{3}{4}\right)^\beta\right) = O\left(\epsilon^\alpha\right)$.

If $\alpha \le 1$, $\alpha+\beta \le 1$, by Lemma 28, $\frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\epsilon}{t})^\alpha\right)^2 + \frac{1}{2}t^{2\alpha}\left(1-(1-\frac{\epsilon}{t})^{\alpha+\beta}\right)^2 \le \frac{1}{2}t^{2\alpha}\left(\frac{\epsilon}{t}\right)^2 + \frac{1}{2}t^{2\alpha}\left(\frac{\epsilon}{t}\right)^2 = t^{2\alpha-2}\epsilon^2$. Thus, $\frac{(P_0(Y=1|x)-P_1(Y=1|x))^2}{P_0(Y=1|x)+P_1(Y=1|x)} \le t^{\alpha-2}\epsilon^2/\left(1-\left(\frac{3}{4}\right)^\beta\right) = O\left(\epsilon^\alpha\right)$.

Therefore, we have $\frac{(P_0(Y=1|x)-P_1(Y=1|x))^2}{P_0(Y=1|x)+P_1(Y=1|x)} = O\left(\epsilon^\alpha+\epsilon^2\right)$.

Likewise, we can get $\frac{(P_0(Y=0|x)-P_1(Y=0|x))^2}{P_0(Y=0|x)+P_1(Y=0|x)} = O\left(\epsilon^\alpha+\epsilon^2\right)$. So we prove $\sum_y \frac{(P_0(Y=y|x)-P_1(Y=y|x))^2}{P_0(Y=y|x)+P_1(Y=y|x)} = O\left(\epsilon^\alpha+\epsilon^2\right)$ when $x \le \frac{1}{2}$. This concludes the proof. □

C.3 Other lemmas

Lemma 27. *([19], Lemma 4) For sufficiently large* $d > 0$*, there is a subset* $M \subset \{0,1\}^d$ *with following properties: (i)* $|M| \geq 2^{d/48}$*; (ii)* $\|v - v'\|_0 > \frac{d}{12}$ *for any two distinct* $v, v' \in M$*; (iii) for any* $i = 1, \dots, d$*,* $\frac{1}{24} \leq \frac{1}{M} \sum_{v \in M} v_i \leq \frac{3}{24}$*.*

Lemma 28. *If* $x \leq 1, r \geq 1$*, then* $(1-x)^r \geq 1 - rx$ *and* $1 - (1-x)^r \leq rx$*.*
If $0 \leq x \leq 1, 0 \leq r \leq 1$*, then* $(1-x)^r \geq \frac{1-x}{1-x+rx}$ *and* $1 - (1-x)^r \leq \frac{rx}{1-(1-r)x} \leq x$*.*

Inequalities above are know as Bernoulli's inequalities. One proof can be found in [16].

Lemma 29. *Suppose* ϵ, τ *are positive numbers and* $\delta \leq \frac{1}{2}$*. Suppose* $\{Z_i\}_{i=1}^{\infty}$ *is a sequence of i.i.d random variables bounded by 1,* $\mathbb{E}Z_i \geq \tau\epsilon$*, and* $Var(Z_i) = \sigma^2 \leq 2\epsilon$*. Define* $V_n = \frac{n}{n-1}\left(\sum_{i=1}^n Z_i - \frac{1}{n}\left(\sum_{i=1}^n Z_i\right)^2\right)$*,* $q_n = q(n, V_n, \delta)$ *as Procedure 2. If* $n \geq \frac{\eta}{\tau\epsilon} \ln\frac{1}{\delta}$ *for some sufficiently large number* η *(to be specified in the proof), then with probability at least* $1 - \delta$ *,* $\frac{q_n}{n} - \mathbb{E}Z_i \leq -\tau\epsilon/2$*.*

Proof. By Lemma 24, with probability at least $1-\delta$, $V_n \leq 4n\sigma^2 + 8\ln\frac{1}{\delta}$, which implies

$$q_n \leq D_1\left(1 + \ln\frac{1}{\delta} + \sqrt{\left(4n\sigma^2 + 9\ln\frac{1}{\delta} + 1\right)\left([\ln\ln]_+ (4n\sigma^2 + 9\ln\frac{1}{\delta} + 1) + \ln\frac{1}{\delta}\right)}\right)$$

We denote the RHS by q.

On this event, we have

$$\begin{aligned}
\frac{q_n}{n} - \mathbb{E}Z_i &\leq \frac{q}{n} - \tau\epsilon \\
&= \tau\epsilon\left(\frac{q}{n\tau\epsilon} - 1\right) \\
&\overset{(a)}{\leq} \tau\epsilon\left(\frac{2D_1}{\eta} + \frac{D_1}{\eta\ln\frac{1}{\delta}}\sqrt{\frac{9\eta}{\tau}\ln\frac{1}{\delta}\left([\ln\ln]_+ (\frac{9\eta}{\tau}\ln\frac{1}{\delta}) + \ln\frac{1}{\delta}\right)} - 1\right) \\
&= \tau\epsilon\left(\frac{2D_1}{\eta} + D_1\sqrt{\frac{9}{\eta\tau\ln\frac{1}{\delta}}[\ln\ln]_+ (\frac{9\eta}{\tau}\ln\frac{1}{\delta}) + \frac{9}{\eta\tau}} - 1\right)
\end{aligned}$$

where (a) follows from $\frac{q}{n}$ being monotonically decreasing with respect to n. By choosing η sufficiently large, we have $\frac{2D_1}{\eta} + D_1\sqrt{\frac{9}{\eta\tau\ln\frac{1}{\delta}}[\ln\ln]_+ (\frac{9\eta}{\tau}\ln\frac{1}{\delta}) + \frac{9}{\eta\tau}} - 1 \leq -\frac{1}{2}$, and thus $\frac{q_n}{n} - \mathbb{E}Z_i \leq -\tau\epsilon/2$. □