Active Learning for Classification With Abstention

Shubhanshu Shekhar, Mohammad Ghavamzadeh, and Tara Javidi

***Abstract*—We construct and analyze active learning algorithms for the problem of binary classification with abstention, in which the learner has an additional option to withhold its decision on certain points in the input space. We consider this problem in the *fixed-cost* setting, where the learner incurs a cost $\lambda \in (0, 1/2)$ every time the abstain option is invoked. Our proposed algorithm can work with the three most commonly used active learning query models, namely, *membership-query*, *pool-based*, and *stream-based* models. We obtain a high probability upper-bound on the excess risk of our algorithm, and establish its minimax near-optimality by deriving matching lower-bound (modulo poly-logarithmic factors). Since our algorithm relies on the knowledge of the smoothness parameters of the regression function, we also describe a new strategy to adapt to these unknown parameters in a data-driven manner under an additional *quality* assumption. We show that using this strategy our algorithm achieves the same performance in terms of excess risk as their counterparts with the knowledge of the smoothness parameters. We end the paper with a discussion about the extension of our results to the setting of *bounded rate* of abstention.**

***Index Terms*—Binary classification, minimax rates, abstention.**

I. Introduction

WE CONSIDER the problem of binary classification in which the learner has an additional provision of abstaining from declaring a label. This problem models several practical scenarios in which it is preferable to withhold a decision at the cost of some additional experimentation, instead of making an incorrect decision and incurring much higher costs. A canonical application of this problem is in automated medical diagnostic systems [21], where classifiers that defer to a human expert on uncertain inputs are more desirable than those that always make a decision. Other key applications include dialog systems and detecting harmful contents on the Web: it is costly for many companies to incorrectly label harmful (harmless) content as harmless (harmful) on their platform, when compared to the cost of gathering additional information.

Active learning is a learning paradigm in which the learner can sequentially request labels at certain input points selected based on the observed data. Existing results in the literature, such as [6], [11], have demonstrated the benefits of active (over passive) learning, in terms of improved sample complexity or equivalently, lower excess risk, in standard classification. However, in the case of classification with abstention, the design of active learning algorithms and their comparison with their passive counterparts have largely been unexplored. In this paper, we aim to fill this gap in the literature.

In this paper, we study the problem of classification with a *fixed-cost* of abstention, in which every usage of the abstain option results in a *known* cost $\lambda \in (0, 1/2)$. The fixed-cost setting is suitable for problems where a precise cost can be assigned to additional experimentation due to using the abstain option. The analysis of this problem was initiated by Chow [8], [9], who derived the Bayes optimal classifier for this setting in [8], and then studied the trade-off between the error rate and the rejection rate in [9]. More recently, Herbei and Wegkamp [14] obtained convergence rates for fixed-cost of abstention classifiers in a non-parametric framework, similar to our paper. The authors in [2] and [33] proposed calibrated convex surrogate loss functions for this problem, and obtained bounds on the excess risk of the classifiers constructed using these loss functions via empirical risk minimization. An ℓ_1-regularized version of this problem was studied in [31] and [32], while the authors in [10] introduced a new framework that involved learning a pair of functions, and proposed and analyzed convex surrogate loss functions.

An alternative to the fixed cost setting is the *bounded-rate* setting, in which the learner is allowed to abstain for up-to a given fraction $\delta \in (0, 1)$ of the input samples at no cost. This setting is more natural than fixed-cost in applications such as medical diagnostics, where the bottleneck is the processing speed of the human expert [20]. Binary classification with a *bounded-rate* of abstention has been studied less extensively than its fixed-cost counterpart. Pietraszek [20] proposed a method to construct abstaining classifiers using ROC analysis. The authors in [12] re-derived the Bayes optimal classifier for the bounded rate setting under the same assumptions as [8]. They further proposed a general plug-in strategy for constructing abstaining classifiers in a semi-supervised setting, and obtained an upper-bound on the excess risk.

However, all the prior work mentioned above study this problem in the *passive* setting, and thus a precise characterization of the potential benefits of active learning in this problem is not available. In this paper, we aim to address this issue.

Contributions: We now summarize the main contributions of the paper.

1) We begin by proposing an active learning algorithm for the *fixed-cost* setting with knowledge of the smoothness of the regression function, and obtain bounds on its excess risk. The proposed algorithm is general enough to work for the three most commonly used active learning query models: *membership query*, *pool-based*, and *stream-based* (Section III-A).

Manuscript received October 14, 2020; revised March 21, 2021 and May 13, 2021; accepted May 14, 2021. Date of publication May 19, 2021; date of current version June 21, 2021. *(Corresponding author: Shubhanshu Shekhar.)*

Shubhanshu Shekhar and Tara Javidi are with the Department of Electrical and Computer Engineering, University of California at San Diego, San Diego, CA 92093 USA (e-mail: shshekha@eng.ucsd.edu; tjavidi@eng.ucsd.edu).

Mohammad Ghavamzadeh is with Google Research, Mountain View, CA 94043 USA (e-mail: ghavamza@google.com).

Digital Object Identifier 10.1109/JSAIT.2021.3081433

2) Under an additional *quality* assumption [5], [24], we then propose an adaptive strategy that does not require the knowledge of the smoothness of the regression function, and achieves the same performance in terms of excess risk (Section III-C).

3) We then demonstrate the minimax near-optimality of our proposed algorithms by deriving matching (modulo logarithmic terms) lower-bound on the excess risk. The lower-bound proof relies on a new comparison inequality for classification with abstention, and a novel construction of a class of *hard* problems (Section III-D).

II. Preliminaries

Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{0, 1\}$ denote the set of labels to be assigned to points in $\mathcal{X}$. We assume that $\mathcal{X} = [0, 1]^D$ and d is the Euclidean metric on $\mathcal{X}$, i.e., for all $x, x' \in \mathcal{X}$, $d(x, x') := \sqrt{\sum_{i=1}^{D}(x_i - x_i')^2}$. A binary classification problem is completely specified by P_{XY}, i.e., the joint distribution of the input-label random variables. Equivalently, it can also be represented in terms of the marginal over the input space, P_X, and the *regression function* $\eta(x) := P_{Y|X}(Y = 1 \mid X = x)$. A (randomized) abstaining classifier is defined as a mapping $g : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Y}_1)$, where $\mathcal{Y}_1 = \mathcal{Y} \cup \{\Delta\}$, the symbol Δ represents the option of the classifier to abstain from declaring a label, and $\mathcal{P}(\mathcal{Y}_1)$ represents the set of probability distributions on $\mathcal{Y}_1$. Such a classifier g comprises of three functions $g_i : \mathcal{X} \to [0, 1]$, for $i \in \mathcal{Y}_1$, satisfying $\sum_{i\in\mathcal{Y}_1} g_i(x) = 1$, for each $x \in \mathcal{X}$. A classifier g is called *deterministic* if the functions g_i take values in $\{0, 1\}$. Every deterministic classifier g partitions $\mathcal{X}$ into three disjoint sets (G_0, G_1, G_Δ).

In this paper, we focus primarily on the **fixed-cost** model of classification with abstention, in which the abstain option can be employed with a fixed cost $\lambda \in (0, 1/2)$. In this setting, the classification risk is defined as $l_\lambda(g, x, y) := \mathbb{1}_{\{g(x)\neq\Delta\}}\mathbb{1}_{\{g(x)\neq y\}} + \lambda\mathbb{1}_{\{g(x)=\Delta\}}$, and the classification problem is stated as

$$\min_g R_\lambda(g) := \mathbb{E}\big[l_\lambda(g, X, Y)\big] = P_{XY}(g(X) \neq Y, g(X) \neq \Delta) + \lambda P_X(g(X) = \Delta). \tag{1}$$

The Bayes optimal classifier which achieves the minimum risk in (1) is defined as $g_\lambda^*(x) = 1$, 0, or Δ, depending on whether $1 - \eta(x)$, $\eta(x)$, or λ is the smallest.

Active Learning Models: For the problem of classification with a fixed cost of abstention, we propose active classification algorithms for three commonly used active learning models [22, Sec. 2]: (i) *membership query* (MQ), (ii) *pool-based* (PB), and (iii) *stream-based* (SB). MQ is the strongest query model, in which, given a total query budget n, the learner can sequentially request labels at arbitrary points of the input space. We use a slightly weaker version of MQ in this paper that only requires labels sampled from P_X restricted to certain partitions of $\mathcal{X}$, which we introduce in Definition 1. In the PB model, the learner is provided with a large pool of unlabelled samples and must request labels of a subset of size n of the pool. Finally, in the SB model, the learner receives a stream of samples and must decide whether to request a label or discard the sample until the total querying budget is exhausted.

A. Definitions

To construct our active classifier, we will require a hierarchical sequence of partitions of the input space, called the *tree of partitions* [4], [18]. Informally, the tree of partitions consists of a sequence of nested partitions of the input space, such that the diameter of any set in the h^{th} partition is of the order ρ^h for some $\rho < 1$ (i.e., geometrically decaying diameters).

Definition 1: A sequence of subsets $\{\mathcal{X}_h\}_{h\geq 0}$ of $\mathcal{X}$ is said to form a *tree of partitions* of $\mathcal{X}$, if they satisfy the following properties: **(i)** $|\mathcal{X}_h| = 2^h$ and we denote the elements of $\mathcal{X}_h$ by $x_{h,i}$, for $1 \leq i \leq 2^h$, **(ii)** for every $x_{h,i} \in \mathcal{X}_h$, we denote by $\mathcal{X}_{h,i}$, the *cell* associated with $x_{h,i}$, which is defined as $\mathcal{X}_{h,i} := \{x \in \mathcal{X} \mid d(x, x_{h,i}) \leq d(x, x_{h,j}),\ \forall j \neq i\}$, where ties are broken in an arbitrary but deterministic manner, and **(iii)** there exist constants $0 < \nu_2 \leq 1 \leq \nu_1$ and $\rho \in (0, 1)$, such that for all h and i, we have $B(x_{h,i}, \nu_2\rho^h) \subset \mathcal{X}_{h,i} \subset B(x_{h,i}, \nu_1\rho^h)$, where $B(x, a) := \{x' \in \mathcal{X} \mid d(x, x') < a\}$ is the open ball in $\mathcal{X}$ centered at x with radius a.

Remark 1: For the metric space $(\mathcal{X}, d)$ considered in our paper, i.e., $\mathcal{X} = [0, 1]^D$ and d being the Euclidean metric, we can construct a tree whose cells $\mathcal{X}_{h,i}$ are D-dimensional rectangles as follows.

- set $\mathcal{X}_0 = \{x_{0,1}\} = \{(0.5, \ldots, 0.5)\}$ and $\mathcal{X}_{0,1} = \mathcal{X}$.
- For $h \geq 1$, the set $\mathcal{X}_h$ consists of the center points of the cells obtained by slicing the cells of $\mathcal{X}_{h-1}$ into half along the longest side (breaking ties in a fixed manner). We will refer to this operation that takes a cell $\mathcal{X}_{h-1,i}$ and partitions it along it longest side to obtain two new cells, $\mathcal{X}_{h,2i-1}$ and $\mathcal{X}_{h,2i}$, as the cell-refinement operation.

For the tree of partitions so constructed, a suitable choice of parameter values for our algorithms are $\rho = 2^{-1/D}$, $\nu_1 = 2\sqrt{D}$, and $\nu_2 = 1/2$.

Next, we define the dimensionality of the region of the input space at which the regression function $\eta(\cdot)$ is close to some threshold value γ. This definition is motivated by similar notions used in bandit literature, such as the *near-optimality dimension* [4] and the *zooming dimension* [15].

Definition 2: For a function $\zeta : [0, \infty) \mapsto [0, \infty)$ and a threshold $\gamma \in (0, 1)$, we define the near-γ dimension associated with $(\mathcal{X}, d)$ and the regression function $\eta(\cdot)$ as

$$D_\gamma(\zeta) := \inf\big\{a \geq 0 \mid \exists C > 0 : M\big(\mathcal{X}_\gamma(\zeta(r)), r\big) \leq Cr^{-a},\ \forall r > 0\big\},$$

where $\mathcal{X}_\gamma(\zeta(r)) := \{x \in \mathcal{X} : |\eta(x) - \gamma| \leq \zeta(r)\}$ and $M(S, r)$ is the r packing number of $S \subseteq (\mathcal{X}, d)$.

To parse the above definition, consider an example with $\mathcal{X} = [0, 1]^D$, $x_0 = (0.5, \ldots, 0.5) \in \mathcal{X}$, $\eta(x) = \gamma + L\|x - x_0\|^c$ for some $L, c > 0$ and $\zeta(r) = r^b$ for some $b > 0$. Then for any $r > 0$, we have $\mathcal{X}_{\zeta(r)} = \{x : \|x - x_0\| \leq r^{b/c}L^{-1/c}\}$. As a result, the r packing number of the set $\mathcal{X}_{\zeta(r)}$ can be upper bounded by $Cr^{\max\{0, D(1-b/c)\}}$ for some $C < \infty$ by using standard volume arguments [30, Lemma 5.13]. As a result we observe that in this case $D_\gamma(\zeta) = \max\{0, D(1 - b/c)\}$. In general, for $\mathcal{X} = [0, 1]^D$ considered in this paper, the term $D_\gamma(\zeta)$ must be no

greater than D, i.e., $D_\gamma(\zeta) \leq D$ for any choice of $\zeta(\cdot)$. This is because $\mathcal{X}_\gamma(\zeta(r)) \subset \mathcal{X}$, for all $r > 0$, and there exists a constant $C_D < \infty$, such that $M(\mathcal{X}, r) \leq C_D r^{-D}$, for all $r > 0$.

Remark 2: We will use an instance of near-γ dimension for stating our results defined as $\tilde{D} = \max_{j=1,2}\{\tilde{D}_j\}$, where $\tilde{D}_j := D_{\gamma_j}(\zeta_1)$ with $\zeta_1(r) = 12(\frac{L_1 v_1}{v_2})^\beta r^\beta$ and $\gamma_j = \frac{1}{2} + (-1)^j(\frac{1}{2} - \lambda)$ in the fixed-cost setting.

This particular choice of $\tilde{D}$ is motivated by the fact the fact that Algorithm 1, described in Section III, refines cells of radius r only in those regions where $|\eta(x) - \gamma| \leq 12(\frac{L_1 v_1}{v_2})^\beta r^\beta$. Thus this instance of near$-\gamma$ dimension can be used to characterize the maximum number of cells of radius r refined (and hence the number of oracle queries) by the algorithm.

III. Fixed-Cost Setting

In this section, we design active learning strategies for the problem of classification with a fixed and known cost, $\lambda \in (0, 1/2)$, of abstention. We begin by describing an algorithm that requires the knowledge of the smoothness parameters of the regression function in Section III-A. Next, we describe an adaptive strategy that achieves similar performance without the knowledge of the smoothness parameters under an additional assumption in Section III-C. In Section III-D, we derive lower-bounds to demonstrate the minimax near-optimality of our algorithms. Finally, we conclude the section with some numerical experiments in Section III-E.

A. Algorithm With Known Smoothness Parameters

In this section, we propose an active learning algorithm, whose pseudo-code is shown in Algorithm 1, for the problem of binary classification with a fixed cost, λ, of abstention, and obtain theoretical bounds on its excess risk under the following two standard assumptions:

(MA): The joint distribution P_{XY} of the input-label pair satisfies the *margin assumption* with parameters $C_0 > 0$ and $\alpha_0 \geq 0$, for $\gamma \in \{\lambda, 1-\lambda\}$, which means that for any $0 < t \leq 1$, we have $P_X(|\eta(X) - \gamma| \leq t) \leq C_0 t^{\alpha_0}$.

(HÖ): The regression function η is Hölder continuous with parameters $L > 0$ and $0 < \beta \leq 1$, i.e., for all $x_1, x_2 \in (\mathcal{X}, d)$, we have $|\eta(x_1) - \eta(x_2)| \leq L \times d(x_1, x_2)^\beta$.

The *Hölder* continuity assumption (HÖ) ensures that points which are close to each other have similar distribution on the label set. It is a standard assumption employed in a large number of existing works in the nonparametric learning and estimation literature. Some examples of prior work using Hölder continuity assumption are [1], [6], [16], [17]. For simplicity, we restrict our attention to the case of $\beta \leq 1$ so that it suffices to consider piecewise constant estimators to achieve the minimax optimal rate. For Hölder functions with $\beta > 1$, our algorithms can be suitably modified by replacing the piece-wise constant estimators with local polynomial estimators as described in [19, Sec. 1.3].

The *margin assumption* (MA) controls the amount of P_X measure assigned to the regions of the input space with $\eta(\cdot)$ values in the vicinity of the threshold boundaries. The assumption (MA) as employed in this paper is a modification of Tsybakov's margin condition for binary classification [3, Definition 7] [27]. The original margin assumption for binary classification requires the condition $P_X(|\eta(X) - \gamma| \leq t) \leq C_o t^{\alpha_0}$ to hold only for $\gamma = 1/2$. In contrast, for the classification with abstention problem, the margin condition is required to hold at the threshold values λ and $1 - \lambda$ for the fixed-cost setting. As the abstention cost λ is changed, the threshold values at which the margin condition is required to hold also changes. Thus it is implicit in the definition that the parameters C_0 and α_0 are functions of λ in the fixed-cost setting. This modified margin condition is a natural generalization of the original margin assumption for the problem of classification with abstention, and it has been employed in several existing works in classification with abstention literature such as [2], [14], [33]. A similar modified margin condition was also employed in a related problem of Neyman-Pearson classification [25], [26].

Algorithm 1: An Active Learning Algorithm for Binary Classification With the Fixed-Cost $\lambda \in (0, 1/2)$ of Abstention, When the Smoothness Parameters, (L, β), Are Known

Input: n, λ, L, β, v_1, ρ, $h_{\max} = \log n$, Mode $\in \{MQ, PB, SB\}$
Initialize $t = 1$, $n_e = 0$, $\mathcal{X}_t = \{x_{0,1}\}$, $\mathcal{X}_t^{(u)} = \mathcal{X}_t$, $\mathcal{X}_t^{(c)} = \emptyset$, $\mathcal{X}_t^{(d)} = \emptyset$, $u_0(x_{0,1}) = 1$, $l_0(x_{0,1}) = 0$, $e_0(x_{0,1}) = +\infty$, $\mathcal{T} = \emptyset$
while $n_e \leq n$ **do**
  for $x_{h,i} \in \mathcal{X}_t^{(u)}$ **do**
    Define $J_t(x_{h,i}) \leftarrow [l_t(x_{h,i}), u_t(x_{h,i})]$
    if $J_t(x_{h,i}) \cap \{\lambda, 1-\lambda\} = \emptyset$ *or* $J_t(x_{h,i}) \subset [\lambda, 1-\lambda]$ **then**
      $\mathcal{X}_t^{(c)} \leftarrow \mathcal{X}_t^{(c)} \cup \{x_{h,i}\}$, $\mathcal{X}_t^{(u)} \leftarrow \mathcal{X}_t^{(u)} \setminus \{x_{h,i}\}$
    end
  end
  $x_{h_t,i_t} \in \arg\max_{x_{h,i} \in \mathcal{X}_t^{(u)}} I_t^{(1)}(x_{h,i}) = u_t(x_{h,i}) - l_t(x_{h,i})$
  if $\left(e_t(n_{h_t,i_t}(t)) < L(v_1\rho^{h_t})^\beta\right)$ *and* $(h_t < h_{\max})$ **then**
    $\mathcal{X}_t^{(u)} \leftarrow \mathcal{X}_t^{(u)} \setminus \{x_{h_t,i_t}\} \cup \{x_{h_t+1,2i_t-1}, x_{h_t+1,2i_t}\}$
    $u_t(x_{h_t+1,i'}) \leftarrow u_t(x_{h_t,i_t})$, $l_t(x_{h_t+1,i'}) \leftarrow l_t(x_{h_t,i_t})$, for $i' \in \{2i_t - 1, 2i_t\}$
  else
    $\mathcal{X}_t^{(u)}, \mathcal{X}_t^{(d)}, \tilde{x}_{h_t,i_t}, y_t$, Increment = REQUEST_LABEL$\left(\text{Mode}, x_{h_t,i_t}, \mathcal{X}_t^{(u)}, \mathcal{X}_t^{(d)}\right)$
  end
  if *Increment* **then**
    $n_e \leftarrow n_e + 1$
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{t\}$,
    $\mathcal{T}_{h_t,i_t} := \{s \in \mathcal{T} : \tilde{x}_{h_s,i_s} \in \mathcal{X}_{h_t,i_t}\}$, $n_{h_t,i_t}(t) := |\mathcal{T}_{h_t,i_t}|$
    $\hat{\eta}_t(x_{h_t,i_t}) = \frac{1}{n_{h_t,i_t}(t)} \sum_{s \in \mathcal{T}_{h_t,i_t}} \mathbb{1}_{\{\tilde{x}_{h_s,i_s} \in \mathcal{X}_{h_t,i_t}\}} y_s$
    $e_t(n_{h_t,i_t}(t)) \leftarrow \sqrt{\left(2\log(2\pi^2 t^3 n/3)\right)/n_{h_t,i_t}(t)}$, $V_{h_t} = L\left(v_1\rho^{h_t}\right)^\beta$
    $l_{t+1}(x_{h_t,i_t}) \leftarrow \max\left\{l_t(x_{h_t,i_t}),\ \hat{\eta}_t(x_{h_t,i_t}) - e_t(n_{h_t,i_t}(t)) - V_{h_t}\right\}$
    $u_{t+1}(x_{h_t,i_t}) \leftarrow \min\left\{u_t(x_{h_t,i_t}),\ \hat{\eta}_t(x_{h_t,i_t}) + e_t(n_{h_t,i_t}(t)) + V_{h_t}\right\}$
  end
  $t \leftarrow t + 1$
end
Output: $\hat{g}$ defined by Eq. (2)

Outline of Algorithm 1: At any time t, the algorithm maintains a set of active points $\mathcal{X}_t \subset \cup_{h \geq 0} \mathcal{X}_h$, such that the cells associated with the points in $\mathcal{X}_t$ partition the whole $\mathcal{X}$, i.e., $\cup_{x_{h,i} \in \mathcal{X}_t} \mathcal{X}_{h,i} = \mathcal{X}$. The set $\mathcal{X}_t$ is further divided into *classified* active points, $\mathcal{X}_t^{(c)}$, *unclassified* active points, $\mathcal{X}_t^{(u)}$, and *discarded* points, $\mathcal{X}_t^{(d)}$. The classified points are those at which the value of η has been estimated sufficiently well so

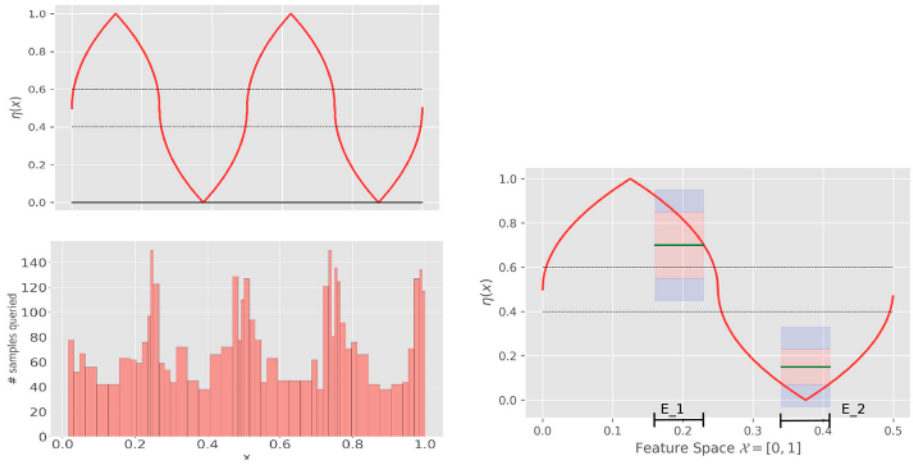

Fig. 1. (*Left*) The top left panel shows a synthetic regression function η with $\mathcal{X} = [0, 1]$ and the dotted horizontal lines represent the thresholds $\lambda = 0.4$ and $1 - \lambda = 0.6$. The figure on bottom left shows the histogram of the points queried by the Algorithm 1 (the bins of the histogram correspond to the final discretization of the input space constructed by the algorithm). As we can see, the active algorithm focuses sampling in the regions near the threshold boundaries that leads to a finer discretization, and hence more accurate estimates of η, in those regions. (*Right*) The figure on the right denotes the confidence intervals constructed for two cells, E_1 and E_2 in the unclassified set $\mathcal{X}_t^{(u)}$. The green horizontal line is the empirical estimate of η in these cells, while the red and blue shaded regions represent the terms e_t and V_h respectively.

that we do not need to evaluate them further. The unclassified points require further evaluation, and perhaps refinement before making a decision. The discarded points are those for which we do not have sufficiently many unlabelled samples in their cells (in the *stream-based* and *pool-based* settings). For every active point, the algorithm computes high probability upper and lower bounds on the maximum and minimum η values in the cell associated with the point. The difference of these upper and lower bounds can be considered as a surrogate for the uncertainty in the η value in a cell. In every round, the algorithm selects a candidate point from the unclassified set that has the largest value of this uncertainty. Having chosen the candidate point, the algorithm either refines the cell or asks for a label at that point.

At a high level, Algorithm 1 involves repeating the following two steps: 1) Maintaining a partition of the input space, and for each set in the partition, constructing upper and lower confidence bounds for the maximum and minimum (respectively) η values in the cell, and 2) Based on these confidence bounds, either refine the partition or request a label. Finally, when the sampling budget is exhausted, 3) Aggregate the information gathered by the sampling strategy to define an abstaining classifier. We now describe these three steps in more details.

a) *Confidence Interval Construction:* At $t \geq 1$, for any cell $\mathcal{X}_{h,i}$ associated with a point $x_{h,i} \in \mathcal{X}_t$, we compute an upper-bound on the maximum η value in the cell as $u_t(x_{h,i}) := \min\{u_{t-1}(x_{h,i}), \bar{u}_t(x_{h,i})\}$, where $\bar{u}_t(x_{h,i}) = \hat{\eta}_t(x_{h,i}) + e_t(n_{h,i}(t)) + V_h$. Here we have $\hat{\eta}_t(x_{h,i}) = \frac{1}{n_{h,i}(t)} \sum_{s=1}^{t-1} \mathbb{1}_{\{\tilde{x}_{h_s,i_s} \in \mathcal{X}_{h,i}\}} y_s$ with $n_{h,i}(t) := \sum_{s=1}^{t-1} \mathbb{1}_{\tilde{x}_{h_s,i_s} \in \mathcal{X}_{h,i}}$; $e_t(n_{h,i}(t))$ is the confidence interval length on the estimate of the average η value in the cell $\mathcal{X}_{h,i}$ (see Lemma 3), and $V_h = L(v_1 \rho^h)^\beta$ is an upper-bound on the maximum variation of the η value in a cell at level h of the tree of partitions $(\mathcal{X}_h)_{h \geq 0}$. We can define the lower-bound on the minimum η value in the cell in a similar manner, $l_t(x_{h,i}) := \max\{l_{t-1}(x_{h,i}), \bar{l}_t(x_{h,i})\}$, where $\bar{l}_t(x_{h,i}) := \hat{\eta}_t(x_{h,i}) - e_t(n_{h,i}(t)) - V_h$. We set $l_0(x_{h,i}) = 0$ and $u_0(x_{h,i}) = 1$ for all $x_{h,i}$.

b) *Refine or Request Label:* In order to select a candidate point, Algorithm 1 selects an *unclassified* point with maximum amount of uncertainty in its value. The uncertainty is measured by the index $I_t^{(1)}(x_{h,i}) = u_t(x_{h,i}) - l_t(x_{h,i})$ (Line 9). Having selected a candidate point x_{h_t,i_t} at time t, the algorithm either *refines* the cell (Lines 10-12) or requests a label depending on the relative magnitudes of $e_t(n_{h_t,i_t}(t))$ and V_{h_t} (Line 14). The label request depends on the query model and consists of the following steps: (i) In the *membership query model* (MQ), the point x_t for which we request the label is drawn from the distribution P_X restricted to the cell $\mathcal{X}_{h_t,i_t}$. (ii) In the *pool-based model* (PB), we request the label if there is an unlabelled sample remaining in the cell $\mathcal{X}_{h_t,i_t}$, otherwise, we remove x_{h_t,i_t} from $\mathcal{X}_t^{(u)}$ and add it to $\mathcal{X}_t^{(d)}$. (iii) In the *stream-based model* (SB), we discard the samples until a point in $\mathcal{X}_{h_t,i_t}$ arrives. If $N_n = n^2 \log(3n^2)$ samples have been discarded, we remove x_{h_t,i_t} from $\mathcal{X}_t^{(u)}$ and add it to $\mathcal{X}_t^{(d)}$. The pseudo-code of the above three steps is provided in the subroutine REQUEST_LABEL in Algorithm 2.

c) *Classifier Definition:* Let t_n denote the time at which the nth query is made and Algorithm 1 halts. We define the final estimate of the regression function as $\hat{\eta}(x) = \hat{\eta}_{t_n}(\pi_{t_n}(x))$, where $\pi_{t_n}(x) := \{x_{h,i} \in \mathcal{X}_{t_n} \mid x \in \mathcal{X}_{h,i}\}$, and the discarded region of the input space as $\tilde{\mathcal{X}}_n^{(d)} := \cup_{x_{h,i} \in \mathcal{X}_{t_n}^{(d)}} \mathcal{X}_{h,i}$. Finally, the classifier returned by the algorithm is defined as

$$\hat{g}(x) = \begin{cases} 1 & \text{if } u_{t_n}(\pi_{t_n}(x)) > 1 - \lambda \text{ or } x \in \tilde{\mathcal{X}}_n^{(d)}, \\ 0 & \text{if } l_{t_n}(\pi_{t_n}(x)) < \lambda \text{ and } x \notin \tilde{\mathcal{X}}_n^{(d)}, \\ \Delta & \text{otherwise.} \end{cases} \tag{2}$$

Analysis: Before stating an upper-bound on the excess risk of Algorithm 1, we show (Lemma 1) that it will suffice to prove this bound for the MQ model. Note that in MQ, the set $\tilde{\mathcal{X}}_n^{(d)}$ is empty. In Lemma 1 (proved in Appendix A-A), we show that under mild assumptions, the P_X measure of $\tilde{\mathcal{X}}_n^{(d)}$ in

Algorithm 2: REQUEST_LABEL Subroutine

Input: Mode, x_{h_t,i_t}, $\mathcal{X}_t^{(d)}$, $\mathcal{X}_t^{(u)}$

discard $\leftarrow$ False, Increment $\leftarrow$ False, $y_t = 1$ (arbitrary value)
if *Mode==MQ* then
  $\tilde{x}_{h_t,i_t} \sim P_X(\cdot \mid \mathcal{X}_{h_t,i_t})$, $y_t \sim$ Bernoulli$(\eta(\tilde{x}_{h_t,i_t}))$, Increment $\leftarrow$ True;
else if *Mode==PB* then
  if $Z_t \cap \mathcal{X}_{h_t,i_t} \neq \emptyset$ then
    choose $\tilde{x}_{h_t,i_t} \in Z_t \cap \mathcal{X}_{h_t,i_t}$ arbitrarily; $y_t \sim$ Bernoulli$(\eta(\tilde{x}_{h_t,i_t}))$, $Z_t \leftarrow Z_t \setminus \{\tilde{x}_{h_t,i_t}\}$, Increment $\leftarrow$ True;
  else
    discard $\leftarrow$ True;
  end
else
  counter $\leftarrow 1$, discard $\leftarrow$ True;
  while $(counter \leq N_n)$ do
    Observe next element of the stream $x \sim P_X$;
    if $x \in \mathcal{X}_{h_t,i_t}$ then
      $\tilde{x}_{h_t,i_t} \leftarrow x$, $y_t \sim$ Bernoulli$(\eta(x))$, discard $\leftarrow$ False, Increment $\leftarrow$ True;
      counter $\leftarrow N_n + 1$;
    end
    counter $\leftarrow$ counter $+1$;
  end
end
if *discard* then
  $\mathcal{X}_t^{(d)} \leftarrow \mathcal{X}_t^{(d)} \cup \{x_{h_t,i_t}\}$, $\mathcal{X}_t^{(u)} \leftarrow \mathcal{X}_t^{(u)} \setminus \{x_{h_t,i_t}\}$;
  $\tilde{x}_{h_t,i_t} = NULL$ /* A symbol indicating no label observed. */
end

Output: $\mathcal{X}_t^{(u)}$, $\mathcal{X}_t^{(d)}$, $\tilde{x}_{h_t,i_t}$, y_t, Increment

PB and SB models is no larger than $1/n$ with probability at least $(1 - 1/n)$. This implies that in these two models, with high probability, the misclassification risk of $\hat{g}$ can be upper-bounded by $1/n + P_{XY}(\hat{g}(X) \neq Y,\ \hat{g}(X) \neq \Delta,\ X \notin \tilde{\mathcal{X}}_n^{(d)})$, where the analysis of the second term is identical for all three active learning models.

Lemma 1: Assume that in the pool-based model, the pool size $M_n > \max\{2n^3, 16n^2 \log(n)\}$, and in the stream-based model, $N_n = n^2 \log(3n^2)$. Then, we have $\mathbb{P}(P_X(\tilde{\mathcal{X}}_n^{(d)}) > 1/n) \leq 1/n$.

As discussed above, given Lemma 1, we can carry out the rest of the analysis for the MQ model, with the knowledge that the same result holds for the other two models with an additional $1/n$ term. We now obtain an upper-bound on the excess risk of the classifier constructed by Algorithm 1 with a budget of n label queries in the MQ model.

Theorem 1: Suppose that the assumptions (MA) and (HÖ) hold, and let $\tilde{D}$ be the dimension term (first introduced in Remark 2) defined as $\tilde{D} = \max_{j=1,2}\{\tilde{D}_j\}$, where $\tilde{D}_j := D_{\gamma_j}(\zeta_1)$ with $\zeta_1(r) = 12(\frac{L_1 v_1}{v_2})^\beta r^\beta$ and $\gamma_j = \frac{1}{2} + (-1)^j(\frac{1}{2} - \lambda)$. For $a > \tilde{D}$ and the corresponding C_a, assume n is large enough to ensure $(\frac{n}{\log n}) \geq (\frac{64 C_a}{L^2 v_1^{2\beta} v_2^a})(\frac{8 L v_1^\beta}{\rho^\beta})^{(2\beta+a)/\beta}$. Then, for the classifier $\hat{g}$ defined by (2), with probability at least $1 - 2/n$, we have $R_\lambda(\hat{g}) - R_\lambda(g_\lambda^*) = \tilde{\mathcal{O}}(n^{-\beta(\alpha_0+1)/(2\beta+a)})$, where the hidden constant depends on the parameters L, β, v_1, v_2, ρ, C_0, and a.

The above result (proof in Appendix A-B) improves upon the convergence rate of the plug-in scheme of [14] in the passive setting, mirroring the benefits of active (over passive) learning in standard classification. We discuss this in more details next.

B. Performance of Algorithm 1

The convergence rates of the excess risk of our active learning algorithms improve upon those obtained for the passive case in the literature.

Achieved Rates: As shown in Theorem 1, Algorithm 1 achieves an excess risk convergence rate of $\tilde{\mathcal{O}}\big(n^{-\beta(1+\alpha_0)/(2\beta+a)}\big)$. Since $a \leq D$, this implies that $\alpha_0\beta > D/2$ is a sufficient condition for the convergence rate to be faster than $n^{-1/2}$ – the parametric rate. Furthermore, for a fixed $\beta > 0$, suppose the P_X measure is such that $P_X(\{x{:}|\eta(x) - \gamma| \leq \epsilon_0\}) = 0$ for some small $\epsilon_0 > 0$. Then we can see that α_0 is unbounded, and Theorem 1 implies that the excess risk of Algorithm 1 converges faster than any negative power of n.

Improvement Over Passive Algorithms: The minmax excess risk for classification with fixed cost of abstention in the passive case under the (MA) and (HÖ) assumptions is of the order $\Theta\big(n^{-\beta(1+\alpha_0)/(D+2\beta+\alpha_0\beta)}\big)$, where the upper bound of $\tilde{\mathcal{O}}\big(n^{-\beta(1+\alpha_0)/(D+2\beta+\alpha_0\beta)}\big)$ is achieved by the plug-in scheme of [14] using the estimators of [1]. The lower bound of $\Omega\big(n^{-\beta(1+\alpha_0)/(D+2\beta+\alpha_0\beta)}\big)$ can be proved by using Lemma 2 and the construction used in the proof of Theorem 3 and we omit the details to avoid repetition. Since the term a is never larger than D, and hence $D + \alpha_0\beta$, the achieved rates in the active setting are always faster than the corresponding passive convergence rates. If in addition, we assume that P_X satisfies the *strong density* assumption,[1] then the passive convergence rate improves to $\tilde{\mathcal{O}}\big(n^{-\beta(1+\alpha_0)/(2\beta+D)}\big)$. However, with this assumption, the active rate also improves further as we can show that $\tilde{D} \leq \max\{0, D - \alpha_0\beta\}$ in this case (see Appendix D for details).

Remark 3: We note that in this paper, we considered the active learning problem under the smoothness assumption (HÖ) on the regression function. An alternative approach is often taken in some other works on active learning, such as [6], in which complexity assumptions are placed on the decision boundary. Informally, under this assumption, the decision boundaries of the classification problem bisect the input space, and by placing regularity assumptions on the boundary, the classification problem can be reduced to a number of one-dimensional noisy binary search problems. However, as described by the authors in [1], the results under these two conditions are not directly comparable. For example, there exist smooth functions that induce very complex boundaries.

C. Adaptivity to Smoothness Parameters

The knowledge of the smoothness parameters, (L, β), is required by Algorithm 1 at three junctures: 1) to define the index $I_t^{(1)}$ for selecting a candidate point, 2) to decide the set of *classified* and *unclassified* active points, and 3) to decide

[1]That is, P_X admits a density p_X with respect to the Lebesgue measure, and furthermore, there exists a constant $\mu_0 > 0$ such that $p_X(x) \geq \mu_0$ for all $x \in \mathcal{X}$.

when to refine a cell. In this section, we describe a data-driven approach that can achieve similar convergence rates as Algorithm 1, without the knowledge of the smoothness parameters, but under an additional assumption.

Additional Notation: We need to introduce additional notation to describe the results of this section. For any cell $\mathcal{X}_{h,i}$, we define (i) the set $\mathcal{E}_j^{(h,i)} = \mathcal{X}_{h+j} \cap \mathcal{X}_{h,i}$ and the corresponding partition of $\mathcal{X}_{h,i}$, defined as $\mathcal{H}_j^{(h,i)} := \{\mathcal{X}_{h+j,i'} : x_{h+j,i'} \in \mathcal{E}_j^{(h,i)}\}$. In words, $\mathcal{E}_j^{(h,i)}$ is the set of points in the cell $\mathcal{X}_{h,i}$ that lie at level $h+j$ in the tree of partitions $(\mathcal{X}_h)_{h\geq 0}$, and (ii) $\tilde{\eta}(\mathcal{X}_{h,i}) = \tilde{\eta}(x_{h,i}) := \int_{\mathcal{X}_{h,i}} \eta d\nu$, where ν is the Lebesgue measure[2] on $[0,1]^D$. The empirical counterpart of $\tilde{\eta}(\mathcal{X}_{h,i})$ at time t is denoted by $\hat{\eta}_t(\mathcal{X}_{h,i}) = \hat{\eta}_t(x_{h,i})$. Next we introduce $\hat{\bar{\eta}}_j^{(h,i)}(t) := \max_{A\in\mathcal{H}_j^{(h,i)}} \hat{\eta}_t(A)$ and $\hat{\underline{\eta}}_j^{(h,i)}(t) := \min_{A\in\mathcal{H}_j^{(h,i)}} \hat{\eta}_t(A)$, which represent the maximum and minimum empirical average η values in cells in $\mathcal{H}_j^{(h,i)}$. We also define $w_j^{(h,i)} = \max_{A_1,A_2\in\mathcal{H}_j^{(h,i)}} (\tilde{\eta}(A_1) - \tilde{\eta}(A_2))$, and its empirical counterpart (at time t) as $\hat{w}_j^{(h,i)}(t) := \hat{\bar{\eta}}_j^{(h,i)}(t) - \hat{\underline{\eta}}_j^{(h,i)}(t)$. Finally, we define $V_{h,i} := \sup_{x_1,x_2\in\mathcal{X}_{h,i}} \eta(x_1) - \eta(x_2)$, which is the variation of the function $\eta(\cdot)$ in the cell $\mathcal{X}_{h,i}$. Note that under the assumption that the function is Hölder continuous with parameters (L,β) and that the cell $\mathcal{X}_{h,i}$ is contained in a ball of radius $\nu_1\rho^h$, we have $V_{h,i} \leq L(\nu_1\rho^h)^\beta$. This is equal to the term V_h that we previously used in Algorithm 1. At the end, we introduce $b_t(h,i,j) := \sqrt{\frac{8\log(1/\delta_t)}{n_{h,i}(t)(\nu_2/\nu_1)^D\rho^j}}$, for $1 \leq j \leq k_n := \lceil \frac{\log(\nu_1^D \log n)}{D\log(1/\rho)} \rceil$, where $\delta_t = \frac{12}{n^2 t^2 \pi^2 \log(n)}$. Note that by definition, we have $e_t(n_{h,i}(t)) \leq b_t(h,i,j)$, for all $1 \leq j \leq k_n$. Finally, for every $x_{h,i} \in \mathcal{X}_t^{(u)}$, we introduce the following two terms: $\hat{j}_t^{(h,i)} := \min\{1 \leq j_1 \leq k_n : |\hat{w}_{j_1}^{(h,i)}(t) - \hat{w}_{j_2}^{(h,i)}(t)| \leq 4b_t(h,i,j_2),\ \text{for all } j_1 \leq j_2 \leq k_n\}$ and $\hat{W}_t^{(h,i)} := 2(\hat{w}_{\hat{j}_t}^{(h,i)}(t) + 6b_t(h,i,k_n))$.

Next we recall the definition of *quality* from [24], suitably modified for our problem.

Definition 3: For a given $\mathcal{X} = [0,1]^D$, a regression function $\eta: \mathcal{X} \mapsto [0,1]$, and a tree of partitions $(\mathcal{X}_h)_{h\geq 0}$, we say the pair $(\eta, (\mathcal{X}_h)_{h\geq 0})$ have *quality* $q \in (0,1)$, if the following holds: for any cell $\mathcal{X}_{h,i}$, there exist two cells $\mathcal{X}_{h',i_1}$ and $\mathcal{X}_{h',i_2}$, both subsets of $\mathcal{X}_{h,i}$, such that 1) $\nu(\mathcal{X}_{h',i_j}) \geq q\nu(\mathcal{X}_{h,i})$, for $j=1,2$, and 2) $\tilde{\eta}(\mathcal{X}_{h',i_1}) - \tilde{\eta}(\mathcal{X}_{h',i_2}) \geq V_{h,i}/2$.

We now state the additional assumption required by our adaptive scheme.

(QU): The pair $(\eta, (\mathcal{X}_h)_{h\geq 0})$ has quality $q > 1/\log(n)$, where n is the label budget.

Adaptive Version of Algorithm 1: In the MQ model consists of the following steps.

- *Candidate Points Selection:* We select one candidate point for every h, such that $\mathcal{X}_h \cap \mathcal{X}_t \neq \emptyset$. Thus, Line 8 of Algorithm 1 changes to $x_{h,i_t} \in \arg\max_{x_{h,i}\in\mathcal{X}_t^{(u)}\cap\mathcal{X}_h} (\hat{\eta}_t(x_{h,i}) + e_t(n_{h,i}(t)))$, for all $h: \mathcal{X}_h \cap \mathcal{X}_t^{(u)} \neq \emptyset$.
- *Request Label:* For every candidate point, if the stopping rule (defined below) is not satisfied, we request the label at a point drawn uniformly from the cell. Thus, in each round, the algorithm may request up to $h_{\max} = \mathcal{O}(\log n)$ labels.
- *Stopping Rule:* We use the following rule for cell refinement: *Refine* a cell if $\hat{w}_{\hat{j}_t}^{(h,i)}(t) - 8b_t(h,i,k_n) \geq 0$. This modification is introduced in Line 9 of Algorithm 1.
- *Update $\mathcal{X}_t^{(u)}$ and $\mathcal{X}_t^{(c)}$:* We follow the same rule for updating the sets $\mathcal{X}_t^{(u)}$ and $\mathcal{X}_t^{(c)}$ as in Lines 10-11 of Algorithm 1, but with the data-driven construction of u_t and l_t, defined as $u_t(x_{h,i}) = \min\{\bar{u}_t(x_{h,i}),\ u_{t-1}(x_{h,i})\}$, where $\bar{u}_t(x_{h,i}) = \hat{\eta}_t(x_{h,i}) + e_t(n_{h,i}) + \hat{W}_t^{(h,i)}$, and $l_t(x_{h,i}) = \max\{\hat{l}_t(x_{h,i}),\ l_{t-1}(x_{h,i})\}$, where $\bar{l}_t(x_{h,i}) := \hat{\eta}_t(x_{h,i}) - e_t(n_{h,i}) - \hat{W}_t^{(h,i)}$.

Theorem 2: Suppose that the assumptions (MA), (HÖ), and (QU) hold, and let $\tilde{D}^{(a)} := \max\{D_1^{(a)}, D_2^{(a)}\}$, with $D_j^{(a)} = D_{1/2+(-1)^j(1/2-\lambda)}(\zeta_1^{(a)})$ and $\zeta_1^{(a)}(r) := 42(Lv_1/v_2)^\beta r^\beta$, for $r > 0$. Then, for large enough n, with probability at least $1-2/n$, for the classifier $\hat{g}$ defined by (2) and for any $a > \tilde{D}^{(a)}$, we have $R_\lambda(\hat{g}) - R_\lambda(g_\lambda^*) = \mathcal{O}(\frac{n}{\log^2(n)\log(n\log n)})^{-\beta(1+\alpha_0)/(a+2\beta)}$, where the hidden constant depends on the parameters L, β, v_1, v_2, ρ, C_0, and a, and is explicitly defined in (10) and (11) in Appendix B (where the proof of the theorem is given).

The result of Theorem 2 has two main differences with that of Theorem 1: 1) there is an additional polylogarithmic in n factor in the excess risk bound, and 2) the dimension term $\tilde{D}^{(a)}$ is larger than the corresponding dimension term $\tilde{D}$ in Theorem 1, as there is a factor of 42 in the definition of $\zeta_1^{(a)}$ compared to 12 in the definition of ζ_1. However, as we show in Section III-B, under an additional *strong density* assumption, both $\tilde{D}$ and $\tilde{D}^{(a)}$ can be upper-bounded by the same quantity, $\max\{0, D-\alpha_0\beta\}$, which can be much smaller than D.

Remark 4: We note that there are other adaptive schemes for active learning, such as [16], [17], that can also be suitably modified to apply to the problem studied in this paper. Our adaptive scheme allows us to obtain excess risk bounds that depend on the local dimensionality of the space near the λ and $1-\lambda$ level sets of η, and thus, are most directly comparable to the excess risk bounds of Algorithm 1. Moreover, we present the risk bound for the adaptive scheme under the (HÖ) assumption to facilitate comparison with Theorem 1. Our scheme can be easily modified to deal with *spatially inhomogeneous* η, as well as η with only *implicit* similarity information, as in [5], [24].

D. Lower Bound

We now derive minimax lower-bounds on the expected excess risk of the fixed-cost setting and the membership query model. Since this is the strongest active learning query model, the obtained lower-bounds are also true for the other two models. The proof follows the general outline for obtaining lower-bounds described in works, such as [1], [17], reducing the estimation problem to an appropriate multiple hypothesis testing problem, and then applying [28, Th. 2.5]. The novel elements of our proof are the construction of an

[2]To reduce notation in stating the adaptive scheme, we assume that P_X is the Lebesgue measure on $[0,1]^D$. The construction can be extended to general P_X that admit a density w.r.t. Lebesgue measure, by discarding regions where the density takes values below a threshold.

appropriate class of regression functions (see Appendix C) and the comparison inequality presented in Lemma 2 (proof is in Appendix C).

Lemma 2: In the fixed-cost of abstention setting with the cost $\lambda < 1/2$, let g represent any abstaining classifier and g^*_λ represent the Bayes optimal one. Then, we have $R_\lambda(g) - R_\lambda(g^*_\lambda) \geq cP_X((G^*_\lambda \setminus G_\lambda) \cup (G_\lambda \setminus G^*_\lambda))^{(1+\alpha_0)/\alpha_0}$, where $c > 0$ is a constant and α_0 is the parameter of the assumption (MA).

Lemma 2 aids our lower-bound proof in several ways: 1) it motivates our construction of *hard* problem instances in which it is difficult to distinguish between the 'abstain' and 'not-abstain' options, 2) it suggests a natural definition of pseudo-metric (see Theorem 4 in Appendix C-B), and 3) it allows us to convert the lower-bound on the hypothesis testing problem to that on the excess risk. We now state the main result of this section (see Appendix C for the proof).

Theorem 3: Let $\mathcal{A}$ be any active learning algorithm in the fixed-cost $\lambda < 1/2$ abstention setting and $\hat{g}_n$ be the abstaining classifier learned by $\mathcal{A}$ with n label queries. Let $\mathcal{P}(L, \beta, \alpha_0)$ represent the class of joint distributions P_{XY} satisfying the margin assumption (MA) with exponent $\alpha_0 > 0$, whose regression function is (L, β) Hölder continuous with $L \geq 3$ and $0 < \beta \leq 1$. Then, we have $\inf_{\mathcal{A}} \sup_{P_{XY} \in \mathcal{P}(L,\beta,\alpha_0)} (\mathbb{E}[R_\lambda(\hat{g}_n) - R_\lambda(g^*_\lambda)]) = \Omega(n^{-\beta(1+\alpha_0)/(2\beta+D)})$.

This result shows the minimax near-optimality of Algorithm 1, as its excess risk upper-bound matches the lower-bound up to poly-logarithmic factors in the worst case when $\tilde{D} = D$.

Remark 5: As mentioned above, the result in Theorem 3 is sufficient to demonstrate the minimax near-optimality of our proposed algorithm, since there exist problem instances for which $\tilde{D} = D$. An interesting question for future work is deriving more refined, instance dependent, lower bounds that actually depend on the near-γ dimension $\tilde{D}$ of the given problem instance. The techniques developed in multi-armed bandit literature, such as [13], may be helpful in this task.

E. Numerical Illustration

We now verify the advantages of active (over passive) learning shown by our theoretical results on a class of toy problems. In these problems, we fix $\mathcal{X} = [0, 1]$, P_X as the uniform distribution, and the cost of abstention at $\lambda = 0.4$, and set $\eta(x) = 0.5(1 + a(\sum_{k=0}^{3}(-1)^k(4x-k)^b))$, for $a, b \in [0, 1]$. The regression functions η are Hölder continuous with $L = 4^b$ and $\beta = b$. Moreover, with the above choice of η and P_X, the (MA) assumption holds with $\alpha_0 = 1/b$.

Algorithm 1 constructs a non-uniform partition of the input space, and then implements a piecewise constant estimator of η to decide the classification rule. To provide a benchmark for comparison, we consider a simple passive classifier which implements a classification rule based on a piecewise constant estimator of η using a uniform partition of $\mathcal{X}$ with a bandwidth $bw = 0.1$.

We ran the experiment for four pairs of (a, b) parameters: $(a, b) \in \{(1.0, 0.2), (1.0, 0.5), (0.8, 0.5), (0.5, 0.5)\}$. For every combination of parameters a, b and n, we ran 50 repetitions of the active and passive algorithms and computed the empirical risk with 10000 test samples. The variation of the empirical risk (on the test set) with the sample size n for the active and passive algorithms are plotted in Figure 2(b). We observe that in all the cases considered, the active algorithm outperforms the passive baseline.

IV. Extension to the Bounded-Rate Setting

In the **bounded-rate** setting, a classifier can abstain up to a given fraction $\delta \in (0, 1)$ of the input samples without incurring any cost. The misclassification risk of a classifier g in this setting can be defined as $R(g) := P_{XY}(g(X) \neq Y, g(X) \neq \Delta)$, and the corresponding classification problem is

$$\min_g \; R(g), \quad \text{subject to} \quad P_X(g(X) = \Delta) \leq \delta. \tag{3}$$

The Bayes optimal classifier for (3) is in general a randomized classifier. However, under some continuity assumptions on the joint distribution P_{XY}, it is again of a threshold type, $g^*_\delta(x) = 1$, 0, or Δ, depending on whether $1 - \eta(x)$, $\eta(x)$, or γ_δ is minimum, where $\gamma_\delta := \sup\{\gamma \geq 0 : P_X(|\eta(X) - 1/2| \leq \gamma) \leq \delta\}$ [8].

The main difference between (1) and (3) is that in the fixed-cost setting, the threshold levels are known beforehand, while in bounded-rate, the mapping $\delta \mapsto \gamma_\delta$ is unknown, and in general is quite complex. In order to construct a classifier that satisfies the constraint in (3), we need some information about the marginal P_X. Accordingly, this problem is studied in a *semi-supervised* framework in which the learner can request a limited number (polynomial in query budget n) of unlabelled samples to estimate the measure of any set of interest (details in [23, Appendix D]).

The optimal abstaining classifier in the *bounded-rate* setting with parameter δ corresponds to an optimal fixed-cost abstaining classifier with cost $\lambda = \lambda_\delta$ $(= 1/2 - \gamma_\delta)$. In this case, the idea underlying our fixed-cost algorithm (Algorithm 1) can be generalized with suitable modifications to construct an active classifier for the bounded rate setting. If the number of unlabelled samples available to the learner is sufficiently large, then the proposed classifier again achieves a $\tilde{\mathcal{O}}\big(n^{-\beta(1+\alpha_0)/(D+2\beta)}\big)$ upper bound on the excess risk under an additional *detectability* assumption. This assumption is a converse to the (MA) assumption stated earlier and has been employed in several works in nonparametric learning and estimation, such as [6], [7], [29].

Appendix A
Proofs for the Algorithm from Section III-A

A. Proof of Lemma 1

We begin with the proof of Lemma 1 which shows that with probability at least $1 - 1/n$, the P_X measure of the (random) set $\tilde{\mathcal{X}}^{(d)}_n$ is no larger than $1/n$.

Suppose the discarded region $\tilde{\mathcal{X}}^{(d)}_n := \cup_{x_{h,i} \in \mathcal{X}^{(d)}_{t_n}} \mathcal{X}_{h,i}$ consists of T components, i.e., $|\mathcal{X}^{(d)}_{t_n}| = T$. Since the algorithm only refines cells up to the depth $h_{\max} = \log(n)$, and the total number of cells in $\mathcal{X}_{h_{\max}}$ is $2^{h_{\max}} \leq e^{h_{\max}} = n$, we can trivially

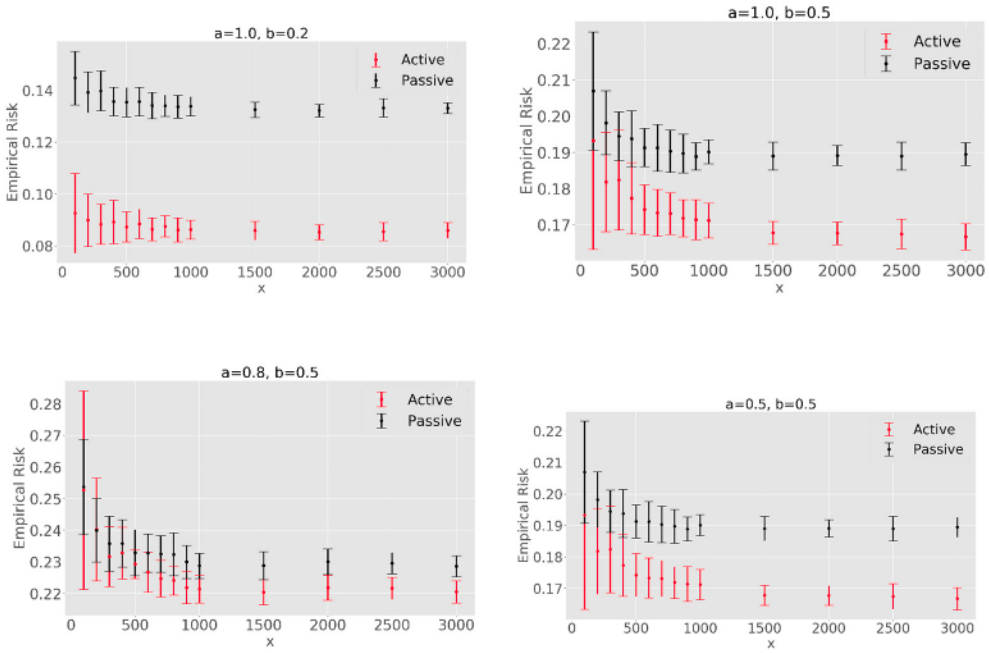

Fig. 2. The figure plots the average (over 50 runs) empirical risk of the active and passive algorithms for four different pairs of (a, b) values. The error bars show one standard deviation for the empirical risks.

upper bound the number of discarded cells/points with n, i.e., $T \leq n$.

a) *Stream-Based Setting:* In this case a cell $\mathcal{X}_{h,i}$ is discarded, if after N_n consecutive draws from P_X, none of the samples fall in $\mathcal{X}_{h,i}$. To write the steps of the proof, we introduce some new notation. We use $n_e^{(t)}$ to denote the number of oracle queries made by the algorithm prior to the iteration t (earlier in Algorithm 1, we simply used n_e to denote this quantity as the t dependence was not important). Also introduce the event $E_t = \{x_{h_t,i_t}$ is discarded at time t , and $P_X(\mathcal{X}_{h_t,i_t}) > 1/n^2\}$ and let $\mathcal{H}_{t-1}$ denote the sigma-field $\sigma\big(\{(x_{h_s,i_s}, y_s) : s \leq t-1\}\big)$. Note that x_{h_t,i_t} and $n_e^{(t)}$ are $\mathcal{H}_{t-1}$ measurable quantities.

We now proceed as follows:

$$\begin{aligned}
\mathbb{P}\left(P_X\left(\tilde{\mathcal{X}}_n^{(d)}\right) > \frac{1}{n}\right) &= \mathbb{P}\left(\sum_{x_{h,i}\in\mathcal{X}_{l_n}^{(d)}} P_X\left(\mathcal{X}_{h,i}\right) > 1/n\right) \\
&\overset{(a)}{\leq} \mathbb{P}\left(\exists x_{h,i} \in \mathcal{X}_{t_n}^{(d)} : P_X\left(\mathcal{X}_{h,i}\right) > 1/(nT)\right) \\
&\overset{(b)}{\leq} \mathbb{P}\left(\exists x_{h,i} \in \mathcal{X}_{t_n}^{(d)} : P_X\left(\mathcal{X}_{h,i}\right) > 1/n^2\right) \\
&\overset{(c)}{\leq} \mathbb{E}\left[\sum_{t=1}^{t_n} \mathbb{1}_{E_t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}_{\{n_e^{(t)}<n\}} \mathbb{1}_{E_t}\right] \\
&\overset{(d)}{=} \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{1}_{\{n_e^{(t)}<n\}} \mathbb{1}_{E_t} | \mathcal{H}_{t-1}\right]\right] \\
&\overset{(e)}{=} \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}_{\{n_e^{(t)}<n\}} \mathbb{E}\left[\mathbb{1}_{E_t} | \mathcal{H}_{t-1}\right]\right] \\
&\overset{(f)}{\leq} \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}_{\{n_e^{(t)}<n\}} \left(1 - \frac{1}{n^2}\right)^{N_n}\right] \\
&= \left(1 - \frac{1}{n^2}\right)^{N_n} \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbb{1}_{\{n_e^{(t)}<n\}}\right] \\
&= \left(1 - \frac{1}{n^2}\right)^{N_n} \mathbb{E}[t_n] \\
&\overset{(g)}{\leq} 3n\left(1 - \frac{1}{n^2}\right)^{N_n} \\
&\leq \exp\left(-\frac{N_n}{n^2} + \log(3n)\right) \overset{(h)}{=} \frac{1}{n}.
\end{aligned}$$

In the above display,

(a) follows from the pigeonhole principle.

(b) uses the fact that $T \leq n$ almost surely.

(c) follows directly from the definition of the event E_t.

(d) uses the tower-rule for conditional expectations.

(e) uses the fact that $n_e^{(t)}$ is $\mathcal{H}_{t-1}-$measurable.

(f) follows from the rule used for discarding cells in the stream-based setting.

(g) follows from the fact that $t_n \leq 3n$ almost surely. This is because in every iteration the algorithm either queries the oracle, or discards a cell or expands a cell. Since each of these three operations can be performed no more than n times (as $h_{\max} = \log n$), we get that $t_n \leq 3n$ almost surely.

(h) follows from the choice of $N_n = n^2 \log(3n^2)$.

b) *Pool-Based Setting:* Let $\mathcal{Z} = \{X_1, X_2, \ldots, X_{M_n}\}$ denote the pool of unlabelled samples available to the learner, and for any $\mathcal{X}_{h,i}$ we introduce the notation $M_{h,i} := |\mathcal{Z} \cap \mathcal{X}_{h,i}|$ to represent the number of samples lying in the cell $\mathcal{X}_{h,i}$. Recall that a cell $\mathcal{X}_{h,i}$ is discarded if the number of unique unlabelled samples in the cell is smaller than the number of label requests in the cell, which can be trivially upper bounded by n, the

total budget. Thus, introducing the terms $\mathcal{C}_1 := \{x_{h,i} | M_{h,i} < n \text{ and } h \leq \log_2(n)\}$ and $\mathcal{C}_2 := \{x_{h,i} \in \mathcal{C}_1 | P_X(\mathcal{X}_{h,i}) \geq 1/(n^2)\}$, we get the following (for any realization of $\mathcal{Z}$):

$$P_X\left(\tilde{\mathcal{X}}_n^{(d)}\right) \leq P_X\left(\bigcup_{x_{h,i}\in\mathcal{C}_1} \mathcal{X}_{h,i}\right) \leq n\left(\frac{1}{n^2}\right) + P_X\left(\bigcup_{x_{h,i}\in\mathcal{C}_2} \mathcal{X}_{h,i}\right),$$

where in first term after the second inequality above, we use the fact that the total number of cells discarded up to the depth of $\log(n)$ cannot be larger than n.

Now, we claim that to complete the proof, it suffices to show that for any $\mathcal{X}_{h,i}$ such that $P_X(\mathcal{X}_{h,i}) > 1/n^2$, we have $\mathbb{P}(M_{h,i} < n) \leq 1/n^2$. To see this introduce the notation $\mathcal{C}_2' = \{x_{h,i} : P_X(\mathcal{X}_{h,i}) > 1/n^2, \text{ and } h \leq \log_2(n)\}$ and note that $\mathcal{C}_2 = \mathcal{C}_2' \cap \mathcal{C}_1$. Now, suppose it were true that for any $\mathcal{X}_{h,i} \in \mathcal{C}_2'$, we have $\mathbb{P}(M_{h,i} < n) \leq 1/n^2$. Then we have, by union bound, $\mathbb{P}(\{\exists \mathcal{X}_{h,i} \in \mathcal{C}_2' : M_{h,i} < n\}) \leq |\mathcal{C}_2'|/n^2$. Since $|\mathcal{C}_2'| \leq n$, as it contains only cells up to a depth of $\log_2 n$, it implies that $\mathbb{P}(\{\exists \mathcal{X}_{h,i} \in \mathcal{C}_2' : M_{h,i} < n\}) \leq 1/n$. In words, this means that with probability at least $1 - 1/n$, every cell $\mathcal{X}_{h,i}$ in $\mathcal{C}_2'$ satisfies that $M_{h,i} \geq n$. This in turn implies that with probability at least $1 - 1/n$, $\mathcal{C}_2 = \mathcal{C}_2' \cap \mathcal{C}_1$ must be empty.

It remains to show that $\mathbb{P}(M_{h,i} < n) \leq 1/n^2$ for any $\mathcal{X}_{h,i} \in \mathcal{C}_2'$. Consider any cell $\mathcal{X}_{h,i}$ such that $P_X(\mathcal{X}_{h,i}) = p \geq 1/n^2$. For points X_j in $\mathcal{Z}$ define the Bernoulli(p) random variable $U_j = \mathbb{1}_{\{X_j \in \mathcal{X}_{h,i}\}}$. Suppose $M_n = \max\{2n^3, 16n^2\log(n)\}$. Then we have the following:

$$\mathbb{P}(M_{h,i} < n) = \mathbb{P}\left(\sum_{j=1}^{M_n} U_j < n\right) \overset{(a)}{\leq} \mathbb{P}\left(\frac{1}{M_n}\sum_{j=1}^{M_n} U_j < \frac{1}{2n^2}\right) \overset{(b)}{\leq} \mathbb{P}\left(\frac{1}{M_n}\sum_{j=1}^{M_n} U_j \leq (1-1/2)p\right) \overset{(c)}{\leq} \exp(-M_n p/8) \overset{(d)}{\leq} \frac{1}{n^2}.$$

In the above display:

(a) follows from the fact that $M_n \geq 2n^3$,

(b) follows from the fact that $p > 1/n^2$,

(c) follows from the application of Chernoff inequality for the lower tail of Binomial,

(d) follows from the fact that $M_n \geq 16n^2\log(n)$ and $p \geq 1/n^2$.

B. Proof of Theorem 1

We first present a lemma that gives us high probability upper and lower bounds on the empirical estimates of the average η value in a cell $\mathcal{X}_{h,i}$ associated with a point $x_{h,i}$, denoted by $\tilde{\eta}(x_{h,i}) := \int_{\mathcal{X}_{h,i}} \eta(x) dP_X(x \mid \mathcal{X}_{h,i})$. The empirical estimate $\hat{\eta}_t(x_{h,i})$ is assumed to have been constructed from labels queried at samples drawn according to the distribution $P_X(\cdot \mid \mathcal{X}_{h,i})$ in an i.i.d. manner. In conjunction with Lemma 1, this next lemma provides a combined description of the confidence intervals of the empirical estimates of the average η value of cells in $\mathcal{X}_t^{(u)}$ or $\mathcal{X}_t^{(c)}$ constructed by any of the three active learning querying models.

Lemma 3: For $t \geq 1$, define the events $\Omega_{1,t}$ as follows:

$$\Omega_{1,t} := \left\{|\hat{\eta}_t(x_{h,i}) - \tilde{\eta}(x_{h,i})| \leq e_t(n_{h,i}),\ \forall x_{h,i} \in \mathcal{X}_t\right\},$$

with $e_t(n_{h,i}) := \sqrt{\frac{2\log(2\pi^2 t^3 n/3)}{n_{h,i}(t)}}$,

where $n_{h,i}(t)$ is the number of times that $x_{h,i}$ has been queried up until time t. Then the event $\Omega_1 := \cap_{t\geq 1}\Omega_{1,t}$ occurs with probability at least $1 - \frac{1}{n}$.

Proof: It suffices to show that $P(\Omega_{1,t}^c) \leq \frac{6}{n\pi^2 t^2}$. The result then follows from a union bound over all $t \geq 1$ and the fact that $\sum_{t\geq 1}\frac{1}{t^2} = \frac{\pi^2}{6}$. Now, for a given $x_{h,i} \in \mathcal{X}_t$ and for any $e_t(n_{h,i}(t)) > 0$, by Hoeffding-Azuma's inequality, we have

$$Pr\left(|\hat{\eta}_t(x_{h,i}) - \tilde{\eta}(x_{h,i})| > e_t(n_{h,i}(t))\right) \leq 2e^{-ne_t(n_{h,i}(t))^2/2}.$$

Finally, by selecting $e_t(n_{h,i}(t)) = \sqrt{\frac{2\log((2\pi^2 t^3 n)/3)}{n_{h,i}(t)}}$, we obtain

$$P(\Omega_{1,t}^c) \leq 2\sum_{(h,i):x_{h,i}\in\mathcal{X}_t} e^{-n_{h,i}(t)a_{h,i}^2/2} \leq \sum_{(h,i):x_{h,i}\in\mathcal{X}_t} \frac{3}{n\pi^2 t^3} \overset{(a)}{\leq} \frac{6}{n\pi^2 t^2}.$$

(a) follows from the fact that $|\mathcal{X}_t| \leq 2t$, for all $t \geq 1$. This is because of the following reasoning: $|\mathcal{X}_0| = 1$, and for any $1 \leq i \leq t$, we must have $|\mathcal{X}_i| \in \{|\mathcal{X}_{i-1}| + 1, |\mathcal{X}_{i-1}|\} \leq |\mathcal{X}_{i-1}| + 1$. Thus by induction, we get $|\mathcal{X}_t| \leq t + 1$, which is no larger than $2t$, for $t \geq 1$. ■

We now present a result on the monotonicity of the term $I_t^{(1)}(x_{h_t,i_t})$ which will be used in obtaining bounds on the estimation error of the regression function.

Lemma 4: $I_t^{(1)}(x_{h_t,i_t})$ is non-increasing in t.

Proof: The proof of this statement relies on the monotonic nature of $u_t(x_{h,i})$ and $l_t(x_{h,i})$. More specifically, for any $x_{h,i} \in \mathcal{X}_t^{(u)}$, we have $I_{t+1}^{(1)}(x_{h,i}) \leq I_t^{(1)}(x_{h,i})$ due to the definition of $u_t(x_{h,i})$ and $l_t(x_{h,i})$ given in Step 2 of Algorithm 1. Furthermore, if the algorithm refines the cell $\mathcal{X}_{h_t,i_t}$, then by definition, we also have $I_{t+1}^{(1)}(x_{h,i}) \leq I_t^{(1)}(x_{h_t,i_t})$, for $h = h_t + 1$ and $i \in \{2i_t - 1, 2i_t\}$, due to the cell refinement rule. These two statements together imply that the term $\sup_{x_{h,i}\in\mathcal{X}_t^{(u)}} I_t^{(1)}(x_{h,i})$ is also a non-increasing term. ■

We next derive a bound on the error in estimating the regression function at the cells close to the threshold values λ and $1 - \lambda$.

Lemma 5: Suppose t_n is the time at which Algorithm 1 stops (i.e., performs the n^{th} query) and $\mathcal{X}_{t_n}^{(u)}$ is the set of unclassified points at time t_n. Define the term $\tilde{D} = \max\{\tilde{D}_1, \tilde{D}_2\}$, where $\{\tilde{D}_j\}_{j=1}^2 := D_{1/2+(-1)^j(1/2-\lambda)}(\zeta_1)$ in which $\zeta_1(r) = 12L(v_1/(v_2\rho))^\beta r^\beta$ and $D_\lambda(\zeta)$ is from Definition 2. Then for large enough n and for any $a > \tilde{D}$, with probability at least $1 - \frac{1}{n}$, we have

$$\left|\eta(x_{h,i}) - \hat{\eta}(x_{h,i})\right| \leq b_n = \frac{4Lv_1^\beta}{\rho^\beta}\left(\frac{2C_a}{L^2 v_1^{2\beta} v_2^a}\right)^\beta \left(\frac{\log(2\pi n/3)}{n}\right)^{\frac{\beta}{(a+2\beta)}},$$

$$\text{for all } x_{h,i} \in \mathcal{X}_{t_n}^{(u)}.$$

Proof: First, note that $t_n \le n^2$, where t_n is the time step at which the algorithm halts. This follows from the fact that the maximum depth explored by the algorithm is $h_{\max} = \log n$, which implies that the maximum number of active points at any time is n. This implies that between any two label requests there can be at most n cell expansions/refinements. Together, these facts imply that $t_n \le n^2$.

Next, we recall that the algorithm refines the cell associated with a point $x_{h,i}$, if $e_t(n_{h,i}(t)) \le V_h = L(v_1\rho^h)^\beta$. The uncertainty of the estimate of $\eta(x_{h,i})$ can be further upper-bounded at any time t by setting $t = t_n$ in the expression of $e_t(n_{h,i}(t))$, i.e.,

$$e_t\big(n_{h,i}(t)\big) \le \sqrt{\frac{8\log\big(2\pi^2 n^7/3\big)}{n_{h,i}(t)}}.$$

Thus, to find an upper-bound on the number of times a point $x_{h,i}$ is queried by the algorithm, it suffices to find the number of queries sufficient to ensure that $\sqrt{(8\log(2\pi^2 n^7/3))/n_{h,i}(t)}$ is less than or equal to V_h. Equating this term with V_h, we obtain

$$n_{h,i}(t_n) \le \frac{8\log\big(2\pi^2 n^7/3\big)}{L^2 v_1^{2\beta}\rho^{2h\beta}}, \tag{4}$$

where t_n is the time at which the budget of n label queries is exhausted and the algorithm stops. Now, by definition, a point $x_{h,i}$ belongs to the set $\mathcal{X}_t^{(u)}$, only if $\{\lambda, 1-\lambda\} \cap [l_t(x_{h,i}), u_t(x_{h,i})] \ne \emptyset$. Suppose for a given $x_{h,i} \in \mathcal{X}_t$, the interval $[l_t(x_{h,i}), u_t(x_{h,i})]$ contains λ. This implies that for $h \ge 1$, we have

$$\begin{aligned}\sup_{x\in\mathcal{X}_{h,i}} |\eta(x)-\lambda| &\le \max\big\{u_t(x_{h,i}) + V_h - \lambda,\quad \lambda - l_t(x_{h,i}) - V_h\big\}\\ &\overset{(a)}{\le} u_t(x_{h,i}) - l_t(x_{h,i}) \overset{(b)}{\le} 4V_{h-1}\\ &= 4L\Big(v_1\rho^{h-1}\Big)^\beta.\end{aligned}$$

(a) follows from the condition that $l_t(x_{h,i}) \le \lambda \le u_t(x_{h,i})$.

(b) follows from the rule used for refining the parent cell of $x_{h,i}$, after which $x_{h,i}$ becomes active. More specifically, let $t_1 \le t$ be the time at which the parent cell of $x_{h,i}$ (denoted by $x_{h-1,i'}$) was refined to activate the point $x_{h,i}$. Then due to the monotonicity of u_t and l_t, we must have $u_t(x_{h,i}) \le u_{t_1}(x_{h-1,i'})$, and $l_t(x_{h,i}) \ge l_{t_1}(x_{h-1,i'})$. By definition we have $u_{t_1}(x_{h-1,i'}) - l_{t_1}(x_{h-1,i'}) \le 2(V_{h-1} + e_{t_1}(n_{h-1,i'}(t_1)))$. Finally, since the cell $\mathcal{X}_{h-1,i'}$ was refined at time t_1, we must have $e_{t_1}(n_{h-1,i'}(t_1)) \le V_{h-1}$, which implies the inequality **(b)** in the above display.

Now, we define the function $\zeta_1(r) = 12L(v_1/(v_2\rho))^\beta r^\beta$ and use it[3] to define the term $\tilde{D}_1 = D_\lambda(\zeta_1)$ (see Definition 2). Similarly, we define $\tilde{D}_2 = D_{1-\lambda}(\zeta_1)$ at the other threshold value and introduce the notation $\tilde{D} = \max\{\tilde{D}_1, \tilde{D}_2\}$. Thus, the total number of points that are activated by the algorithm at level h of the tree, denoted by N_h, can be upper-bounded by the packing number of the set $\mathcal{X}_\lambda(\zeta_1(v_2\rho^h)) \cup \mathcal{X}_{1-\lambda}(\zeta_1(v_2\rho^h))$ with balls of radius $v_2\rho^h$. Now, by the definition of $\tilde{D}$, for any $a > \tilde{D}$, there exists a $C_a < \infty$ such that we can upper-bound N_h with the term $2C_a(v_2\rho^h)^a$. Using the bound on N_h and $n_{h,i}(t_n)$, we observe that the number of queries made by the algorithm at level h of the tree is no more than $N_h n_{h,i}(t_n)$. Hence, for any $H \ge 1$, we have

$$\begin{aligned}\sum_{h=0}^{H} N_h n_{h,i}(t_n) &\le \frac{8\log\big(2\pi^2 n^7/3\big)C_a v_2^{-a}}{L^2 v_1^{2\beta}} \sum_{h=0}^{H}\left(\frac{1}{\rho}\right)^{h(a+2\beta)}\\ &\le \frac{8\log\big(2\pi^2 n^7/3\big)C_a v_2^{-a}}{L^2 v_1^{2\beta}}\left(\frac{1}{\rho}\right)^{H(a+2\beta)}.\end{aligned} \tag{5}$$

Next, we need to find a lower-bound on the depth in the tree that has been explored by the algorithm. This can be done by finding the largest H for which (5) is smaller than or equal to n. By equating (5) with n, we obtain the following relation for the largest such value of H, denoted by H_0,

$$\left(\frac{1}{\rho}\right)^{H_0} = \left(\frac{L^2 v_1^{2\beta} v_2^a}{8C_a}\right)^{1/(a+2\beta)}\left(\frac{n}{\log\big(2\pi^2 n^7/3\big)}\right)^{1/(a+2\beta)}. \tag{6}$$

Now, for any $x \in \cup_{x_{h,i}\in\mathcal{X}_{t_n}^{(u)}} \mathcal{X}_{h,i}$, we must have

$$\begin{aligned}\big|\hat{\eta}(x) - \eta(x)\big| &= \big|\hat{\eta}_{t_n}(\pi_{t_n}(x)) - \eta(x)\big| \le u_{t_n}(x) - l_{t_n}(x)\\ &\overset{(a)}{\le} I_{t_n}^{(1)}\big(x_{h_{t_n},i_{t_n}}\big).\end{aligned}$$

(a) follows from the point selection rule of the algorithm.

Lemma 6 implies that if the algorithm is evaluated at a point at level H_0 at some time $t \le t_n$, then we have

$$\sup_{x_{h,i}\in\mathcal{X}_{t_n}^{(u)}} I_{t_n}^{(1)}(x_{h,i}) \le 4V_{H_0-1} = 4L(v_1\rho^{H_0-1})^\beta := b_n,$$

where

$$\begin{aligned}b_n &= \frac{4Lv_1^\beta}{\rho^\beta}\left(\frac{8C_a}{L^2 v_1^{2\beta} v_2^a}\right)^{\beta/(a+2\beta)}\left(\frac{\log(2\pi^2 n^7/3)}{n}\right)^{\beta/(a+2\beta)}\\ &= \mathcal{O}\left(\left(\frac{n}{\log n}\right)^{-\beta/(a+2\beta)}\right).\end{aligned}$$

■

We note that the our classifier is well defined only when $b_n \le 1 - 2\lambda$, a sufficient condition for which is that n is large enough to ensure that

$$\left(\frac{n}{\log n}\right) \ge \left(\frac{64C_a}{L^2 v_1^{2\beta} v_2^a}\right)\left(\frac{4Lv_1^\beta}{(1-2\lambda)\rho^\beta}\right)^{(2\beta+a)/\beta}. \tag{7}$$

Finally, we combine Lemma 5 with the margin assumptions to obtain the required result.

Lemma 6: The excess risk of the classifier $\hat{g}$ in (2), learned by Algorithm 1, w.r.t. the optimal classifier in the fixed cost of abstention setting, with the fixed abstention cost $\lambda \in (0, 1/2)$, satisfies $R_\lambda(\hat{g}) - R_\lambda(g_\lambda^*) \le \tilde{\mathcal{O}}(n^{-\beta(\alpha_0+1)/(2\beta+a)})$.

Proof: By definition of the classifier $\hat{g} = (\hat{G}_0, \hat{G}_1, \hat{G}_\Delta)$, under the event Ω_1, the set $\hat{G}_\Delta \subset G_\Delta^*$.

Now, by Lemma 5, we know that $\sup_{x_{h,i}\in\mathcal{X}_{t_n}^{(u)}} I_t^{(1)}(x_{h,i}) \le b_n$, which for n large enough ensures that $b_n \le \lambda$ leading to

[3]Actually, a factor of 4 instead of 12 suffices, but we use 12 so that the same $\tilde{D}$ can be used for stating the result of the bounded-rate setting as well.

$\hat{G}_0 \subset \{x \in \mathcal{X} | \eta(x) \geq 1/2\}$. This implies that $\hat{G}_0 \cap G_1^* = \emptyset$. Similarly, we can obtain $\hat{G}_1 \cap G_0^* = \emptyset$. Thus, the excess risk of the estimated classifier can be written as

$$\begin{aligned} R_\lambda(\hat{g}) - R_\lambda(g_\lambda^*) &= \int_{\hat{G}_0} \eta(x) dP_X + \int_{\hat{G}_1} (1-\eta(x)) dP_X \\ &\quad + \lambda P_X\left(\hat{G}_\Delta\right) \\ &\quad - \int_{G_0^*} \eta(x) dP_X - \int_{G_1^*} (1-\eta(x)) dP_X \\ &\quad - \lambda P_X(G_\Delta^*) \\ &= \int_{\hat{G}_0 \cap G_\Delta^*} (\eta(x) - \lambda) dP_X \\ &\quad + \int_{\hat{G}_1 \cap G_\Delta^*} (1 - \lambda - \eta(x)) dP_X \\ &\quad + \int_{\hat{G}_\Delta \cap G_0^*} (\lambda - \eta(x)) dP_X \\ &\quad + \int_{\hat{G}_\Delta \cap G_\Delta^*} (\eta(x) - 1 + \lambda) dP_X \\ &\leq b_n P_X(|\eta(X) - \lambda| \leq b_n) \\ &\quad + b_n P_X(|\eta(X) - 1 + \lambda| \leq b_n) \\ &\leq 2 C_0 b_n^{1+\alpha_0}. \end{aligned}$$

■

APPENDIX B
PROOF OF THEOREM 2 (THE ADAPTIVE SCHEME)

In this section, we elaborate on the adaptive scheme introduced in Section III-C of the main text. Before describing the adaptive routine, we first state the following concentration result.

Proposition 1: For a cell $\mathcal{X}_{h,i}$ and $1 \leq j \leq k_n$, and time $t \geq 1$, we define the event $\Theta(t,h,i,j)$ as follows:

$$\Theta(t,h,i,j) := \left\{ |\hat{\eta}_t(A) - \tilde{\eta}(A)| \leq b_t(h,i,j)\ \forall A \in \mathcal{H}_j^{(h,i)} \right\}$$

$$\text{where } b(h,i,j) := \sqrt{\frac{8\log(\delta_t)}{n_{h,i}(t)(v_2/v_1)^D \rho^j}}$$

$$\text{and } \delta_t = \frac{12}{n^2 \log(n) t^2 \pi^2}.$$

Then the event $\Theta := \{\cap \Theta(t,h,i,j) | t \geq 1, (h,i) : x_{h,i} \in \mathcal{X}_t, 1 \leq j \leq k_n\}$ occurs with probability at least $1 - 1/n$.

The proof of this result follows in an analogous manner to the proof of Lemma 3, and we omit the details.

We next state the lemma, which tells us that the adaptive scheme ensures that the number of samples allocated to a cell $\mathcal{X}_{h,i}$ before refining, denoted by $n_{h,i}(t)$, satisfies a condition analogous to that derived in (4) for the known smoothness case. As a consequence of this, we can also get an upper bound on the total number of samples allocated up to a level H of the tree of partition, similar to (5) in the known smoothness case.

Lemma 7: If the adaptive scheme refines a cell $\mathcal{X}_{h,i}$ at time t, and if $n_{h,i}$ denotes the number of labels that were requested in the cells $\mathcal{X}_{h,i}$ before refining, then we have the following:

$$\frac{32\log(1/\delta_t)\log(n)}{V_{h,i}^2} \leq n_{h,i}(t) \leq \frac{6273\log(1/\delta_t)\log n}{V_{h,i}^2}.$$

Proof: We will drop the superscript and denote the terms such as $w_j^{(h,i)}$ with $w_j^{(h,i)}$ for this proof. Since the cell was refined at time $t \geq 2$, the following is true

$$\begin{aligned} \left|\hat{w}_{j_t} - \hat{w}_{k_n}\right| &\leq 4b_t(h,i,k_n) \Rightarrow \hat{w}_{k_n} \geq \hat{w}_{j_t} - 4b_t(h,i,k_n) \\ &\Rightarrow V_{h,i} \geq w_{k_n} \geq \hat{w}_{k_n} - 2b_t(h,i,k_n) \\ &\geq \hat{w}_{j_t} - 6b_t(h,i,k_n) \geq 2b_t(h,i,k_n) \end{aligned} \tag{8}$$

Since $b_t(h,i,k_n) \leq \sqrt{\frac{8\log(1/\delta_t)\log(n)}{n_{h,i}(t)}}$, this implies that

$$n_{h,i}(t) \geq \frac{32\log(1/\delta_t)\log n}{V_{h,i}^2}.$$

Next, let t_1 denote the time at which a label was requested in the cell $\mathcal{X}_{h,i}$. Since it was not refined at time t_1, the following sequence is true.

$$\begin{aligned} \left|\hat{w}_{j_{t_1}} - \hat{w}_{k_n}\right| &\leq 4b_{t_1}(h,i,k_n) \Rightarrow \hat{w}_{k_n} \leq \hat{w}_{j_{t_1}} + 4b_t(h,i,k_n) \\ &\Rightarrow \hat{w}_{k_n} + 2b_{t_1}(h,i,k_n) \leq 14 b_{t_1}(h,i,k_n) \\ &\Rightarrow \frac{V_{h,i}}{2} \leq w_{k_n} \leq \hat{w}_{k_n} + 2b_{t_1}(h,i,k_n) \\ &\leq 14 b_{t_1}(h,i,k_n). \end{aligned}$$

This implies the following for $N_1 = \log(n)$:

$$\begin{aligned} \frac{V_{h,i}}{2} &\leq 14\sqrt{\frac{8\log\left(1/\delta_{t_1}\right)\log n}{\left(n_{h,i}(t)-1\right)}} \\ &\Rightarrow n_{h,i}(t) \leq 1 + \frac{6272\log(1/\delta_t)\log(n)}{V_{h,i}^2} \\ &\leq \frac{6273\log(1/\delta_t)\log(n)}{V_{h,i}^2}. \end{aligned}$$

■

Next we present a lemma which obtains a bound on the maximum deviation of $\eta(x)$ from λ or $1-\lambda$ for x lying in the subset of the input space covered by the cells of the unclassified active points.

Lemma 8: If a cell $x_{h,i} \in \mathcal{X}_t^{(u)}$ for some $h \geq 1$, then we must have for $i' := \lfloor (i+1)/2 \rfloor$,

$$\min\left\{\left|\eta(x_{h,i}) - 1 + \lambda\right|,\ \left|\eta(x_{h,i}) - \lambda\right|\right\} \leq 42 V_{h-1,i'}. \tag{9}$$

Proof: Let $t_1 \leq t$ be the time at which the parent cell of $\mathcal{X}_{h,i}$ was expanded to include $x_{h,i}$ in the active unclassified set, and let $t_2 \leq t_1$ be the previous time instant at which the cell $\mathcal{X}_{h-1,i'}$ was queried. Since $x_{h,i} \in \mathcal{X}_t^{(u)}$, the interval $[l_t(x_{h,i}), u_t(x_{h,i})]$ must contain either $1-\lambda$ or λ. Without loss of generality assume that $[l_t(x_{h,i}), u_t(x_{h,i})]$ contains λ (The other case can be handled in exactly the same way.). Then we have the following:

$$\begin{aligned} |\eta(x) - \lambda| &\leq u_t(x_{h,i}) - l_t(x_{h,i}) \leq u_{t_1}(x_{h-1,i'}) - l_{t_1}(x_{h-1,i'}) \\ &\overset{(a)}{\leq} u_{t_2}(x_{h-1,i'}) - l_{t_2}(x_{h-1,i'}) \leq \bar{u}_{t_2}(x_{h-1,i'}) \\ &\quad - \bar{l}_{t_2}(x_{h-1,i'}) = 2\left(e_{t_2}(n_{h-1,i'}) + \hat{W}_{t_2}^{(h-1,i')}\right) \end{aligned}$$

$$\begin{aligned}
&\overset{(b)}{\leq} 2e_{t_2}\big(n_{h-1,i'}\big) \\
&\quad + 4\big(8b_{t_2}\big(h-1,i',k_n\big) + 6b_{t_2}\big(h-1,i',k_n\big)\big) \\
&\overset{(c)}{\leq} 2b_{t_2}\big(h-1,i',k_n\big) \\
&\quad + 4\big(8b_{t_2}\big(h-1,i',k_n\big) + 6b_{t_2}\big(h-1,i',k_n\big)\big) \\
&\overset{(d)}{\leq} \sqrt{2}\big(2b_{t_1}\big(h-1,i',k_n\big) \\
&\qquad + 4\big(8b_{t_1}\big(h-1,i',k_n\big) + 6b_{t_1}\big(h-1,i',k_n\big)\big)\big) \\
&\leq 84b_{t_1}\big(h-1,i',k_n\big) \overset{(e)}{\leq} 42V_{h-1,i'}.
\end{aligned}$$

In the above display,

(a) follows from the definition of the terms $\bar{l}_t$ and $\bar{u}_t$, and the fact that $t_1 \leq t$,

(b) follows from the fact that $t_2 \leq t_1$ and the monotonicity of u_t and l_t,

(c) follows from the fact that $e_{t_2}\big(n_{h-1,i'}\big)\big) \leq b_{t_2}\big(h-1,i',k_n\big)$,

(d) uses the fact that $n_{h-1,i'}(t_2) \geq n_{h-1,i'}(t_1)/2$,

(e) uses the fact that $V_{h-1,i'} \geq 2b_{t_1}\big(h-1,i',k_n\big)$ as shown in (8). ∎

The rest of the proof follows along the lines of the proof of Theorem 1. We first present a lemma, which is analogous to Lemma 5 and introduces an appropriate notion of dimensionality $\tilde{D}^{(a)}$ for the adaptive scheme.

Lemma 9: Suppose t_n is the time at which the adaptive algorithm stops (i.e., performs the n^{th} query) and $\mathcal{X}_{t_n}^{(u)}$ is the set of unclassified points at time t_n. Define the term $\tilde{D}^{(a)} = \max\{\tilde{D}_1^{(a)}, \tilde{D}_2^{(a)}\}$, where $\{\tilde{D}_j^{(a)}\}_{j=1}^2 := D_{1/2+(-1)^j(1/2-\lambda)}(\zeta_1^{(a)})$ in which $\zeta_1^{(a)}(r) = 36L(v_1/(v_2\rho))^\beta r^\beta$ and $D_{1/2+(-1)^j(1/2-\lambda)}(\zeta)$ is from Definition 2. Then for large enough n and for any $a > \tilde{D}^{(a)}$, with probability at least $1 - \frac{1}{n}$, we have

$$\big|\eta(x_{h,i}) - \hat{\eta}(x_{h,i})\big| \leq b_n^{(a)} = \mathcal{O}\left(\frac{n}{\log^2(n)}\right)^{-\beta(a+2\beta)}, \quad \text{for all } x_{h,i} \in \mathcal{X}_{t_n}^{(u)}.$$

Proof: We know from Lemma 7 that we have $N_{h,i} \leq \frac{6273\log(n)\log(1/\delta_{t_n})}{V_{h,i}^2}$, where we used the fact that δ_t is decreasing in t. Since the maximum depth is $h_{\max} = \log(n)$, we must have $t_n \leq n^3$. Thus we can obtain the following bound:

$$N_{h,i} \leq \frac{6273\log(n)\log(1/\delta_t)}{V_{h,i}^2} \leq \frac{6273\log(n)\log\big(n^5\log(n)\big)}{V_{h,i}^2} := \frac{C_n}{V_{h,i}^2}. \tag{10}$$

Also from Lemma 9, we know that any point in $\mathcal{X}_t^{(u)}$ at level h satisfies $\min\{|\eta(x_{h,i}) - 1 + \lambda|,\ |\eta(x_{h,i}) - \lambda|\} \leq 42V_{h-1,i}$.

Due to the Holder continuity assumption on η, we again have $V_{h,i} \leq L\big(v_1\rho^h\big)^\beta$ for all h, i. The rest of the proof follows the steps of the proof of Lemma 5, and we get that

$$\begin{aligned}
\big|\hat{\eta}(x_{h,i}) - \eta(x_{h,i})\big| &\leq L\left(\frac{v_1}{\rho}\right)^\beta \left(\frac{nv_2^aL^2v_1^{2\beta}}{2C_nC_a}\right)^{-\beta/(a+2\beta)} \\
&:= b_n^{(a)} = \mathcal{O}\left(\frac{n}{\log^2(n)}\right)^{-\beta/(a+2\beta)}
\end{aligned} \tag{11}$$

A sufficient condition for this bound to be non-trivial (i.e., for the RHS to be less than 1) is if the following holds:

$$\frac{n}{\log n\log(n\log n)} \geq L^{(a+2\beta)/\beta}\left(\frac{62730C_a}{v_2^aL^2v_1^{2\beta}}\right)\left(\frac{v_1}{\rho}\right)^{a+2\beta}. \tag{12}$$

∎

Having obtained the result of Lemma 9, the result in the statement of Theorem 2 follows by an application of Lemma 6.

Appendix C
Proof of Lower Bound

A. Proof of Lemma 2

[In this section, we use the notation $\int_A f d\mu$ as a shorthand for $\int_A f(x)d\mu(x)$ for the integral of function f with respect to some measure μ over some set A.]

We first observe the following:

$$\begin{aligned}
R_\lambda(g) - R_\lambda(G_\lambda^*) &= \int_{G_\lambda} \lambda dP_X + \int_{G_0} \eta dP_X + \int_{G_1} (1-\eta)dP_X \\
&\quad - \int_{G_\lambda^*} \lambda dP_X - \int_{G_0^*} \eta dP_X \\
&\quad - \int_{G_1^*} (1-\eta)dP_X \\
&= \int_{G_\lambda \cap G_0^*} (\lambda - \eta)dP_X \\
&\quad + \int_{G_\lambda \cap G_1^*} (\lambda - 1 + \eta)dP_X \\
&\quad + \int_{G_\lambda^* \cap G_0} (\eta + \lambda)dP_X \\
&\quad + \int_{G_\lambda^* \cap G_1} (1 - \eta - \lambda)dP_X \\
&\quad + \int_{G_0 \cap G_1^*} (2\eta - 1)dP_X \\
&\quad + \int_{G_0^* \cap G_1} (1 - 2\eta)dP_X \\
&:= T_1 + T_2 + T_3 + T_4 + T_5 + T_6.
\end{aligned}$$

We now consider the six terms separately.

- By definition of G_1^*, we know that $\eta \geq 1 - \lambda$ in this set. This implies that the integrand in T_5 is at least $1 - 2\lambda \geq 0$. Thus we can lower bound T_5 with 0. The term T_6 can similarly be shown to be non-negative.
- To lower bound the term T_1, we partition G_0^* into two regions: $G_{0,a}^*$ which is close to the boundary, and $G_{0,b}^*$ which is the region away from the boundary.

$$G_{0,a}^* := \big\{x \in G_0^* | \eta(x) \geq \lambda - t\big\}, \text{ and}$$
$$G_{0,b}^* := G_0^* \setminus G_{0,a}^*,$$

where $t > 0$ will be decided later. In the set $G_\lambda \cap G_{0,b}^*$, we have $\lambda - \eta \geq t$, which implies that

$$\begin{aligned}
T_1 = \int_{G_\lambda \cap G_0^*} (\lambda - \eta)dP_X &\geq \int_{G_\lambda \cap G_{0,b}^*} (\lambda - \eta)dP_X \\
&\geq tP_X\big(G_\lambda \cap G_{0,b}^*\big)
\end{aligned}$$

$$\geq t\big(P_X\big(G_\lambda \cap G_0^*\big) - P_X\big(G_{0,a}^*\big)\big)$$
$$\overset{(i)}{\geq} tP_X\big(G_\lambda \cap G_0^*\big) - C_0 t^{1+\alpha_0},$$

where the inequality (i) follows from the margin condition.

- To lower bound the term T_2, we introduce the sets G_1^* into $G_{1,a}^* \cup G_{1,b}^*$ where $G_{1,a}^* := \{x \in G_1^* \mid \eta(x) \leq 1 - \lambda + t\}$ and $G_{1,b}^* := G_1^* \setminus G_{1,a}^*$. We then have:

$$T_2 = \int_{G_\lambda \cap G_1^*} (\lambda - 1 + \eta) dP_X \geq \int_{G_\lambda \cap G_{1,b}^*} (\lambda - 1 + \eta) dP_X$$
$$\geq tP_X\big(G_\lambda \cap G_{1,b}^*\big)$$
$$\geq t\big(P_X\big(G_\lambda \cap G_1^*\big) - P_X\big(G_{1,a}^*\big)\big)$$
$$\geq tP_X\big(G_\lambda \cap G_1^*\big) - C_0 t^{1+\alpha_0}.$$

- To lower bound T_3 we introduce $G_{\lambda,a}^* := \{x \in G_\lambda^* \mid \eta(x) \leq \lambda + t\}$, and $G_{\lambda,b}^* := G_\lambda^* \setminus G_{\lambda,a}^*$. Then we have the following:

$$T_3 := \int_{G_0 \cap G_\lambda^*} (\eta - \lambda) dP_X \geq \int_{G_0 \cap G_{\lambda,b}^*} (\eta - \lambda) dP_X$$
$$\geq tP_X\big(G_0 \cap G_{\lambda,b}^*\big)$$
$$\geq t\big(P_X\big(G_0 \cap G_\lambda^*\big) - P_X\big(G_{\lambda,a}^*\big)\big)$$
$$\geq tP_X\big(G_0 \cap G_\lambda^*\big) - C_0 t^{\alpha_0+1}.$$

- Finally, to lower bound the term T_4, we introuce $G_{\lambda,c}^* := \{x \in G_\lambda^* \mid \eta(x) \geq 1 - \lambda - t\}$, and $G_{\lambda,d}^* = G_\lambda^* \setminus G_{\lambda,c}^*$. Then we have

$$T_4 := \int_{G_1 \cap G_\lambda^*} (1 - \eta - \lambda) dP_X \geq \int_{G_1 \cap G_{\lambda,d}^*} (1 - \eta - \lambda) dP_X$$
$$\geq tP_X\big(G_1 \cap G_{\lambda,d}^*\big)$$
$$\geq t\big(P_X\big(G_1 \cap G_\lambda^*\big) - P_X\big(G_{\lambda,c}^*\big)\big)$$
$$\geq tP_X\big(G_1 \cap G_\lambda^*\big) - C_0 t^{\alpha_0+1}.$$

Combining the above we have the following:

$$R_\lambda(g) - R_\lambda\big(g_\lambda^*\big) \geq t\big(P_X\big(G_\lambda \cap (G_\lambda^*)^c\big) + P_X\big(G_\lambda^* \cap G_\lambda^c\big)\big) - 4C_0 t^{1+\alpha_0}$$
$$= tP_X\big(G_\lambda \triangle G_\lambda^*\big) - 4C_0 t^{1+\alpha_0}. \tag{13}$$

The result then follows by setting t such that $tP_X\big(G_\lambda \triangle G_\lambda^*\big) = 5C_0 t^{1+\alpha_0}$, which leads to the following:

$$R_\lambda(g) - R_\lambda\big(g_\lambda^*\big) \geq C_0 \left(\frac{P_X\big(G_\lambda \triangle G_\lambda^*\big)}{5C_0}\right)^{(1+\alpha_0)/\alpha_0}$$
$$= \left(\frac{1}{5}\right)^{(1+\alpha_0)/\alpha_0} \left(\frac{1}{C_0}\right)^{1/\alpha_0} \times P_X\big(G_\lambda \triangle G_\lambda^*\big)^{(1+\alpha_0)/\alpha_0}$$
$$:= cP_X\big(G_\lambda \triangle G_\lambda^*\big)^{(1+\alpha_0)/\alpha_0}.$$

B. Proof of Theorem 3

We follow the general scheme for obtaining lower bounds in nonparametric learning problems used in prior work such as [1], [17]. This method involves constructing a set of *hard* problem instances which are (1) sufficiently well separated in terms of some *pseudo-metric*, and (2) sufficiently close together in terms of some statistical distance (such as KL divergence or χ^2 distance). Once we have such a construction, we can employ [28, Th. 2.5] (recalled below as Theorem 4) to get a lower bound on the distance in terms of the pseudo-metric for any estimator. Finally, we can use the comparison lemma (Lemma 2) to convert this to a lower bound on the excess risk.

Theorem 4 [28, Th. 2.5]: Assume that for $\tilde{M} \geq 2$, $\Theta = \{\theta_1, \ldots, \theta_{\tilde{M}}\}$, $\tilde{d}$ is a pseudo-metric on Θ, and $\{P_{\theta_j} \mid \theta_j \in \Theta\}$ is a collection of probability measures such that:

- $\tilde{d}\big(\theta_i, \theta_j\big) \geq 2s > 0$ for all $1 \leq i, j \leq \tilde{M}$.
- $P_{\theta_i} << P_{\theta_0}$ for all $1 \leq i \leq \tilde{M}$.
- $\frac{1}{\tilde{M}} \sum_{j=1}^{\tilde{M}} D_{KL}\big(P_{\theta_j}, P_{\theta_0}\big) \leq a \log\big(\tilde{M}\big)$ for $0 < a < 1/8$.

Then we have for $\tilde{M} \geq 10$,

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta\Big(\tilde{d}\Big(\hat{\theta}, \theta\Big) \geq s\Big) \geq \frac{1}{4}$$

where the infimum is over all estimators $\hat{\theta}$ constructed using samples from P_θ.

We now describe the construction of the regression functions. First, given $\mathcal{X} = [0, 1]^D$, for some $\epsilon > 0$ to be decided later, we partition $\mathcal{X}$ into hypercubes of side ϵ, and denote by $M = (1/\epsilon)^D$ the number of such hypercubes. Let V be the set of centers of the hypercubes, i.e., $V = \{z_1, z_2, \ldots, z_M\}$, and let $\pi : \mathcal{X} \mapsto V$ denote the projection operator onto V.

a) *Choose Appropriate Subsets of the Input Space:* Assuming $D \geq 2$, let e_1, e_2, e_3 and e_4 denote any four corner points of $\mathcal{X} = [0, 1]^D$. We define the following subsets of the space $\mathcal{X}$

$$Q_j := \big\{x \in \mathcal{X} \mid \|x - e_j\| \leq 1/3\big\} \quad \text{for } j = 1, 2, 3 \text{ and } 4.$$

For ϵ small enough, we note that there exists a constant $c_1 > 0$ such that the number of hypercubes contained inside each Q_j, denoted by M_j, can be lower bounded by $c_1 M$. (Note that by symmetry $M_1 = M_2 = M_3 = M_4$, so we will use $\tilde{M}$ to denote any of M_j). We will use $V_j = \{z_{j,1}, z_{j,2}, \ldots, z_{j,\tilde{M}}\}$ to denote the centers of the hypercubes contained in Q_j, and $Y_j := \bigcup_{z \in V_j} B_\infty(z, \epsilon/2)$ to denote the union of all the hypercubes strictly contained in Q_j. Here $B_\infty(z, \epsilon/2)$ denotes the hypercube with center z and side ϵ.

b) *Define the Regression Function:* Let $u : [0, \infty) \mapsto [0, 1]$ be a function defined as $u(z) = \min\{(1 - z)^\beta, 0\}$. Note that u satisfies the following properties: (1) $u(0) = 1 - u(1) = 1$, (2), $u(z) = 0$ for $z \geq 1$, and (3) u is $(1, \beta)$ Hölder continuous for $0 < \beta \leq 1$.

For any $z \in S$, we define the function $\varphi_z(x) = L(\epsilon/2)^\beta u((2/\epsilon)\|x - z\|)$. By construction, the function φ_z is (L, β) Hölder continuous. Furthermore, we assume that ϵ is small enough to ensure that $L(\epsilon/2)^\beta < 1/2 - \lambda$.

For any $\vec{\sigma}^{(j)} \in \{-1, 1\}^{\tilde{M}}$, for $j = 1, 2$ we introduce the notation $\vec{\sigma} = (\vec{\sigma}^{(1)}, \vec{\sigma}^{(2)}) \in \{-1, 1\}^{2\tilde{M}}$. Next we define $\eta_{\vec{\sigma}}(x) = \lambda + \sum_{i=1}^{\tilde{M}} \sigma_i^{(1)} \varphi_{z_{1,i}}(x)$ for $x \in Y_1$ and $1 - \lambda + \sum_{i=1}^{\tilde{M}} \sigma_i^{(2)} \varphi_{z_{2,i}}(x)$ for $x \in Y_2$. For x lying in $Q_1 \setminus Y_1$ and $Q_2 \setminus Y_2$, we assign $\eta_{\vec{\sigma}}(x)$ the values λ and $1 - \lambda$ respectively.

Furthermore, we assign $\eta_{\vec{\sigma}}(x) = 1$ for $x \in Q_3$ and $\eta_{\vec{\sigma}}(x) = 0$ for $x \in Q_4$.

It remains to specify the values of $\eta_{\vec{\sigma}}(\cdot)$ in the region $\mathcal{X} \setminus (\bigcup_{j=1}^{4} Q_j)$. For any $A \subset \mathcal{X}$ and $x \in \mathcal{X}$, we use $d_A(x) := \inf\{\|y - x\| | y \in A\}$ to represent the distance of the point x from the set A. We also introduce the terms $z_1 = (\frac{\lambda}{L})^{1/\beta}$ and $z_2 = (\frac{1}{2L})^{1/\beta}$, and assume that $L \geq 3$ which ensures that $z_1 \leq z_2 \leq 1/6$. Now for all $x \in \mathcal{X} \setminus \bigcup_{j=1}^{4} Q_j$, we define

$$\eta_{\vec{\sigma}}(x) = \begin{cases} \lambda + Lu\big(1 - d_{Q_1}(x)\big) & \text{if } x : d_{Q_1}(x) \leq z_1 \\ 1 - \lambda - Lu\big(1 - d_{Q_2}(x)\big) & \text{if } x : d_{Q_2}(x) \leq z_1 \\ 1 - Lu\big(1 - d_{Q_3}(x)\big) & \text{if } x : d_{Q_3}(x) \leq z_2 \\ Lu\big(1 - d_{Q_4}(x)\big) & \text{if } x : d_{Q_4}(x) \leq z_2 \\ 1/2 & \text{otherwise} \end{cases}$$

This completes the definition of the regression function at all points in $\mathcal{X}$. By construction, we have that for any $\vec{\sigma} \in \{-1, 1\}^{2\tilde{M}}$, the regression function $\eta_{\vec{\sigma}}$ is (L, β) Hölder continuous for $0 < \beta \leq 1$ and $L \geq 3$.

c) *Define the Marginal P_X:* Next, we need to define a marginal such that the margin condition is satisfied with exponent $\alpha_0 > 0$. For this we can proceed as in [1, Sec. 6.2] and for some $w < (1/(2\tilde{M}))$, define the density of the marginal w.r.t. the Lebesgue measure as follows:

$$p_X(x) = \begin{cases} \frac{w \mathbb{1}_{B(\pi(x), \epsilon/4)}(x)}{\mathrm{Vol}(B(\pi(x), \epsilon/4))} & \text{for } x \in Y_1 \cup Y_2 \\ \frac{1 - 2\tilde{M}w}{2\mathrm{Vol}(Q_j)} & \text{for } x \in Q_j, \text{ for } j = 3, 4 \\ 0 & \text{otherwise.} \end{cases}$$

We can now check that the joint distribution thus defined satisfied the Margin condition for a given exponent $\alpha_0 > 0$ with constant $C_0 = (8/3)^{\beta\alpha_0}$, if we have $\tilde{M}w = \mathcal{O}\big(\epsilon^{\alpha_0\beta}\big)$.

d) *Apply Theorem 4:* In order to apply Theorem 4, we proceed as follows:

- Let Σ denote the set $\{-1, 1\}^{2\tilde{M}}$. Then by *Gilbert-Varshamov bound* [28, Lemma 2.9], we know that there exists a subset of Σ, denoted by $\tilde{\Sigma}$, such that $|\tilde{\Sigma}| \geq 2^{\tilde{M}/4}$, $\vec{\sigma}_0 = (1, 1, \ldots, 1) \in \tilde{\Sigma}$, and for any $\vec{\sigma}_1, \vec{\sigma}_2 \in \tilde{\Sigma}$, we have $d_H(\vec{\sigma}_1, \vec{\sigma}_2) \geq \tilde{M}/4$. Here $d_H(\cdot, \cdot)$ denotes the Hamming distance.
- Let $\mathcal{P}'$ denote the class of joint distributions $P_{\vec{\sigma}}$ with marginal P_X, and conditional distribution $\eta_{\vec{\sigma}}$ for $\vec{\sigma} \in \tilde{\Sigma}$. For any two $P_{\vec{\sigma}_1}$ and $P_{\vec{\sigma}_2}$ in $\mathcal{P}'$, we introduce the pseudo-metric $\tilde{d}$ defined as $\tilde{d}\big(P_{\vec{\sigma}_1}, P_{\vec{\sigma}_2}\big) := P_X\big(\mathrm{sign}\big(\eta_{\vec{\sigma}_1} - \lambda\big) \neq \mathrm{sign}\big(\eta_{\vec{\sigma}_2} - \lambda\big)\big) + P_X\big(\mathrm{sign}\big(\eta_{\vec{\sigma}_1} - 1 + \lambda\big) \neq \mathrm{sign}\big(\eta_{\vec{\sigma}_2} - 1 + \lambda\big)\big)$.
  Thus, by the properties of $\tilde{\Sigma}$, we get that for any $\vec{\sigma}_1, \vec{\sigma}_2 \in \tilde{\Sigma}$, we have
  $$\tilde{d}\big(P_{\vec{\sigma}_1}, P_{\vec{\sigma}_2}\big) \geq \frac{\tilde{M}w}{4}.$$
- Next, by using [17, eq. (10)], we can upper bound the average KL divergence between the distributions in $\mathcal{P}'$ after n label requests by any active learning algorithm:
  $$D_{KL}\big(P_{\vec{\sigma}_1}, P_{\vec{\sigma}_2}\big) \leq 32nL^2\Big(\frac{\epsilon}{2}\Big)^{2\beta}.$$
  If we select, $\epsilon = c_2 n^{-1/(D+2\beta)}$, with c_2 small enough (a suitable value is $c_2 = \big((4^\beta c_1)/(32^2 L^2)\big)^{1/(D+2\beta)}$), we have
  $$D_{KL}\big(P_{\vec{\sigma}_1}, P_{\vec{\sigma}_2}\big) \leq \frac{\tilde{M}}{4} \leq \frac{1}{8}\log\Big(\big|\tilde{\Sigma}\big|\Big),$$
  as required by Theorem 4.

Since all the conditions of Theorem 4 are satisfied by our construction, we can conclude that for any active learning algorithm $\hat{\eta}$, we have

$$\inf_{\hat{\eta}} \sup_{(\eta, P_X) \in \mathcal{P}'} \mathbb{P}\Big(P_X\big(\mathrm{sign}\big(\hat{\eta} - \kappa\big) \neq \mathrm{sign}(\eta - \kappa) \text{ for } \kappa \in \{\lambda, 1 - \lambda\}\big) \geq c_3 n^{-(\alpha_0\beta)/(D+2\beta)}\Big) \geq \frac{1}{4}.$$

e) *Apply the Comparison Inequality (Lemma 2):* Finally, by employing the comparison inequality (Lemma 2), we obtain the following:

$$\inf_{\hat{g}} \sup_{(\eta, P_X) \in \mathcal{P}'} \mathbb{P}\Big(R_\lambda\big(\hat{g}\big) - R_\lambda\big(^*\big) \geq c_4 n^{-\beta(1+\alpha_0)/(D+2\beta)}\Big) \geq \frac{1}{4},$$

which gives us the required bound:

$$\inf_{\hat{g}} \sup_{(\eta, P_X) \in \mathcal{P}'} \mathbb{E}\big[R_\lambda\big(\hat{g}\big) - R_\lambda\big(g^*\big)\big] \geq \tfrac{c_4}{4} n^{-\beta(1+\alpha_0)/(D+2\beta)}.$$

Appendix D
Details From Section III-B

Suppose that the marginal P_X has a density p_X w.r.t. the Lebesgue measure, and that the density is bounded below by a constant $c_0 > 0$ almost surely. This implies that for any set $A \subset \mathcal{X}$, we have $\mathbb{P}(X \in A) = P_X(A) \geq c_0 \mathrm{Vol}(A)$.

Here we show that under this assumption, we have $\tilde{D}^{(a)} \leq \max\{0, D - \alpha_0\beta\}$ which also implies that $\tilde{D} \leq \max\{0, D - \alpha_0\beta\}$ as we know that $\tilde{D} \leq \tilde{D}^{(a)}$ by definition. Recall that $\tilde{D}^{(a)}$ was introduced in Theorem 2 and $\tilde{D}$ was introduced in Remark 2.

Define $\lambda_j = 1/2 + (-1)^j(1/2 - \lambda)$ for $j = 1, 2$, and the set $\mathcal{X}_{\lambda_j}(\zeta_3(r)) := \{x \in \mathcal{X} \mid |\eta(x) - \lambda_j| \leq 42L(v_1/(v_2\rho))^\beta r^\beta\}$. Then by the assumption (MA), we have the following

$$P_X\big(\mathcal{X}_{\lambda_j}(\zeta_1(r))\big) \leq C_0 L^{\alpha_0}\left(\frac{v_1 r}{v_2 \rho}\right)^{\beta\alpha_0} \leq \tilde{C}_1 r^{\beta\alpha_0}$$

for some constant $\tilde{C}_1 > 0$ depending on $L, v_1, v_2, \rho, C_0, \alpha_0, \beta$. Furthermore, by the additional assumption on P_X, for any $x \in \mathcal{X}$ and $r > 0$, we have

$$P_X(B(x, r)) \geq c_0 \mathrm{Vol}(B(x, r)) = \tilde{C}_2 r^D$$

for some constant $\tilde{C}_2 > 0$ depending on c_0 and D. Thus for $r > 0$, the r-packing number of the set $\mathcal{Z}_r := \mathcal{X}_{\lambda_1}(\zeta_3(r)) \cup \mathcal{X}_{\lambda_2}(\zeta_3(r))$ can be upper bounded as follows:

$$\begin{aligned} \tilde{C}_1 r^{\beta\alpha_0} &\geq P_X(\mathcal{Z}_r) \geq M(\mathcal{Z}_r, r)\tilde{C}_2 r^D \\ &\Rightarrow M(\mathcal{Z}_r, r) \leq \frac{\tilde{C}_1}{\tilde{C}_2} r^{-(D - \beta\alpha_0)}. \end{aligned}$$

Finally, by the definition of *near-λ dimension* we observe that $\tilde{D}^{(a)} \leq \max\{0, D - \beta\alpha_0\}$.

References

[1] J.-Y. Audibert and A. B. Tsybakov, "Fast learning rates for plug-in classifiers," *Ann. Stat.*, vol. 35, no. 2, pp. 608–633, 2007.

[2] P. L. Bartlett and M. H. Wegkamp, "Classification with a reject option using a hinge loss," *J. Mach. Learn. Res.*, vol. 9, pp. 1823–1840, Aug. 2008.

[3] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Summer School on Machine Learning*. Heidelberg, Germany: Springer, 2003, pp. 169–207.

[4] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "X-armed bandits," *J. Mach. Learn. Res.*, vol. 12, pp. 1655–1695, May 2011.
[5] A. D. Bull, "Adaptive-treed bandits," *Bernoulli*, vol. 21, no. 4, pp. 2289–2307, 2015.
[6] R. M. Castro and R. D. Nowak, "Minimax bounds for active learning," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2339–2353, May 2008.
[7] L. Cavalier, "Nonparametric estimation of regression level sets," *Stat. A, J. Theor. Appl. Stat.*, vol. 29, no. 2, pp. 131–160, 1997.
[8] C.-K. Chow, "An optimum character recognition system using decision functions," *IRE Trans. Electron. Comput.*, vol. EC-6, no. 4, pp. 247–254, Dec. 1957.
[9] C.-K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 1, pp. 41–46, Jan. 1970.
[10] C. Cortes, G. DeSalvo, and M. Mohri, "Learning with rejection," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2016, pp. 67–82.
[11] S. Dasgupta, "Coarse sample complexity bounds for active learning," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006, pp. 235–242.
[12] C. Denis and M. Hebir, "Consistency of plug-in confidence sets for classification in semi-supervised learning," 2015, [Online]. Available: arXiv:1507.07235.
[13] A. Garivier, P. Ménard, and G. Stoltz, "Explore first, exploit next: The true shape of regret in bandit problems," *Math. Oper. Res.*, vol. 44, no. 2, pp. 377–399, 2019.
[14] R. Herbei and M. Wegkamp, "Classification with reject option," *Can. J. Stat.*, vol. 34, no. 4, pp. 709–721, 2006.
[15] R. Kleinberg, A. Slivkins, and E. Upfal, "Bandits and experts in metric spaces," *J. ACM*, vol. 66, no. 4, p. 30, May 2019.
[16] C. A. L. Andrea and K. Samory, "Adaptivity to noise parameters in non-parametric active learning," in *Proc. 30th Conf. Learn. Theory (COLT)*, vol. 65, 2017, pp. 1383–1416.
[17] S. Minsker, "Plug-in approach to active learning," *J. Mach. Learn. Res.*, vol. 13, pp. 67–90, Jan. 2012.
[18] R. Munos, "From bandits to Monte-Carlo tree search: The optimistic principle applied to optimization and planning," *Found. Trends Mach. Learn.*, vol. 7, no. 1, pp. 1–129, 2014.
[19] A. Nemirovski, "Topics in non-parametric statistics," in *École d'Eté de Probabilités de Saint-Flour*, vol. 28. New York, NY, USA: Springer, 2000, p. 85.
[20] T. Pietraszek, "Optimizing abstaining classifiers using ROC analysis," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 665–672.
[21] P. Rubegni *et al.*, "Automated diagnosis of pigmented skin lesions," *Int. J. Cancer*, vol. 101, no. 6, pp. 576–580, 2002.
[22] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Rep. 1648, 2009.
[23] S. Shekhar, M. Ghavamzadeh, and T. Javidi, "Active learning for binary classification with abstention," 2019, [Online]. Available: arXiv:1906.00303.
[24] A. Slivkins, "Multi-armed bandits on implicit metric spaces," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Cuuran, 2011, pp. 1602–1610.
[25] X. Tong, "A plug-in approach to Neyman-Pearson classification," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3011–3040, 2013.
[26] X. Tong, Y. Feng, and A. Zhao, "A survey on Neyman-Pearson classification and suggestions for future research," *Wiley Interdiscipl. Rev. Comput. Stat.*, vol. 8, no. 2, pp. 64–81, 2016.
[27] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *Ann. Stat.*, vol. 32, no. 1, pp. 135–166, 2004.
[28] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, 1st ed. New York, NY, USA: Springer, 2009.
[29] A. B. Tsybakov, "On nonparametric estimation of density level sets," *Ann. Stat.*, vol. 25, no. 3, pp. 948–969, 1997.
[30] R. van Handel, "Probability in high dimension," ORF Dept., Princeton Univ., Princeton, NJ, USA, Rep. ADA623999, 2014.
[31] M. Wegkamp, "Lasso type classifiers with a reject option," *Electron. J. Stat.*, vol. 1, pp. 155–168, May 2007.
[32] M. Wegkamp and M. Yuan, "Support vector machines with a reject option," *Bernoulli*, vol. 17, no. 4, pp. 1368–1385, 2011.
[33] M. Yuan and M. Wegkamp, "Classification methods with reject option based on convex risk minimization," *J. Mach. Learn. Res.*, vol. 11, no. 5, pp. 111–130, 2010.