Upper and Lower Error Bounds for Active Learning

Rui M. Castro and Robert D. Nowak

Abstract— **This paper analyzes the potential advantages and theoretical challenges of "active learning" algorithms. Active learning involves sequential, adaptive sampling procedures that use information gleaned from previous samples in order to focus the sampling and accelerate the learning process relative to "passive learning" algorithms, which are based on non-adaptive (usually random) samples. There are a number of empirical and theoretical results suggesting that in certain situations active learning can be significantly more effective than passive learning. However, the fact that active learning algorithms are feedback systems makes their theoretical analysis very challenging. It is known that active learning can provably improve on passive learning if the error or noise rate of the sampling process is bounded. However, if the noise rate is unbounded, perhaps the situation most common in practice, then no previously existing theory demonstrates whether or not active learning offers an advantage. To study this issue, we investigate the basic problem of learning a threshold function from noisy observations. We present an algorithm that provably improves on passive learning, even when the noise is unbounded. Moreover, we derive a minimax lower bound for this learning problem, demonstrating that our proposed active learning algorithm converges at the near-optimal rate.**

I. INTRODUCTION

In various learning tasks it is possible to use information gleaned from previous samples in order to focus the sampling process, in what is generally referred to as "active learning". These methods attempt to accelerate the learning task relative to "passive learning" algorithms, based on non-adaptive sampling. A prototypical example is document classification: suppose we are given a text document and want to decide if the contents pertain either finance, sports, or anything else. We are going learn how to do this task from examples, that is, we have access to a number of documents that have been labeled by an expert (usually a human), so we know the topics of those documents. In general, labeling examples is expensive and time consuming, so we would like to use as few examples as possible. In many applications we might have access to many un-labeled examples. This is the case for our prototypical scenario, where one has a virtually infinite supply of documents. Therefore, ways to automatically decide whether or not obtaining the label for an unlabeled example is worthwhile are crucial for the efficient design of good classifiers.

The interest in active learning in the machine learning community has increased greatly in the last few of years, in part due to the dramatic growth of data sets, and the high cost of labeling all the examples in such data sets. There are several empirical and theoretical results suggesting that in certain situations active learning can be significantly more effective than passive learning [2], [7], [8], [12], [17]. Many of these results pertain to the "noiseless" setting, in which the labels are a deterministic function of the features. In certain noiseless scenarios it has been shown that the number of labeled examples needed is to achieve a desired classification error rate is much smaller than what would be need using passive learning (in fact for those scenarios if passive learning requires n labeled examples then active learning requires only $O(\log n)$ labeled examples) [9]–[12]. Although this setting is interesting from a theoretical perspective, it is very restrictive, and seldom relevant for practical applications. Some of these results have been extended to the "bounded noise rate" setting. In this setting labels are no longer a deterministic function of the features, but for a given feature the probability of observing a particular label is significantly higher than the probability of observing any other label. In the case of binary classification this means that if (X, Y) is a feature-label pair, where $Y \in \{0, 1\}$, then $|\Pr(Y = 1|X = x) - 1/2| > \mu$ for every x in the feature space, with $\mu > 0$. Under this assumption it can be shown that results similar to the ones for the noiseless scenario can be achieved [1], [5], [7], [16]. These results are intimately related to adaptive sampling techniques in regression problems [3], [5], [13], [14], where similar performance gains have been reported.

In this paper we address scenarios where the noise rate is unbounded, perhaps the situation most common in practice. Previously existing theoretical studies have been unable to answer whether or not active learning is advantageous in this case. To study this issue, we investigate the basic problem of learning a threshold function from noisy observations. Since the noise rate is unbounded, the quality of the observations degrade when the samples are taken in the vicinity of the threshold location. In other words, the noise level tends to $1/2$, the flip of a fair coin, as the sampling locations tend to the threshold location. We present an algorithm that provably improves on passive learning, even when the noise is unbounded. Moreover, we derive a minimax lower bound for this learning problem, demonstrating that our proposed active learning algorithm converges at a near-optimal rate. To the best of our knowledge these are the first results of this kind.

The paper is organized as follows: in Section II we formalize the problem and setup. Section III presents the fundamental limits of active learning, in terms of minimax lower bounds. In Section IV we present various algorithms,

This work was partially supported by the National Science Foundation grants CCR-0310889, CNS-0519824, and ECS-0529381.

Rui Castro is with the Department of Electrical Engineering of University of Wisconsin - Madison, and Rice University. rcastro@cae.wisc.edu

Robert Nowak is with the Department of Electrical Engineering, University of Wisconsin - Madison. nowak@engr.wisc.edu

starting with an algorithm for the bounded noise rate scenario due to [3], which will be the building block for algorithms tackling unbounded noise rate problems. Section V presents some simulation results illustrating our methods and Section VI closes with some final remarks and open questions. We defer the proofs of the main results to the appendix.

II. Problem Formulation

Throughout this paper we focus on a relatively simple one-dimensional problem. Although this might seem a rather "toyish" scenario, it displays many of the features that make active learning appealing and theoretical analysis of algorithms challenging. Consider estimating a threshold function from noisy samples. This problem boils down to finding the threshold location. Adaptively sampling aims to find this location with a minimal number of strategically placed samples.

Let $(X, Y) \in [0,1] \times \{0,1\}$ be a random variable, with *unknown* distribution P_{XY}. Our goal is to construct a "good" classification rule, that is, given X we want to predict Y as accurately as possible, where our classification rule is a measurable function $f : [0,1] \to \{0,1\}$. The performance of the classifier is evaluated in terms of the expected loss, in particular a natural choice to consider is the 0/1-loss. With this choice the risk is simply the probability of classification error,

$$R(f) = \mathbb{E}[\mathbf{1}\{f(X) \neq Y\}] = \Pr(f(X) \neq Y) \ . \qquad (1)$$

Since we are considering only two classes there is a one-to-one correspondence between classifiers and sets: Any reasonable classifier is of the form $f(x) = \mathbf{1}_G(x)$, where G is a measurable subset of $[0,1]$. We use the term classifier interchangeably for both f and G. Define the optimal risk as

$$R^* = \inf_{G \text{measurable}} R(G) \ .$$

This minimum risk is attained by the *Bayes Classifier*

$$G^* = \{x \in [0,1] : \eta(x) \geq 1/2\} \ ,$$

where

$$\eta(x) = \mathbb{E}[Y|X = x] = \Pr(Y = 1|X = x) \ ,$$

is called the *conditional probability* (we use this term only if it is clear from the context). In general $R(G^*) > 0$, therefore even the optimal classifier misclassifies sometimes. The quantity of interest for the performance evaluation of a classifier G is therefore the *excess risk*

$$R(G) - R(G^*) = \int_{G \Delta G^*} |2\eta(x) - 1| \mathrm{d}P_X(x) \ ,$$

where Δ denotes the symmetric difference between two sets[1], and P_X is the marginal distribution of X.

If the distribution P_{XY} is known, then clearly we could just construct the Bayes classifier and we would be done. This is not the case for interesting problems, where we do not have direct access to P_{XY}. In most cases we have to construct a classifier based on a finite number of samples of P_{XY}, and this is where most differences between passive and active learning methods arise. For the purposes of this paper we consider particular classes of distributions P_{XY}. We assume that P_X is uniform of [0,1], but the results in this paper can easily be generalized to the case where the marginal density of X is bounded above and below. We assume furthermore that the Bayes classifier has the form $G^* = [\theta^*, 1]$, where $\theta^* \in [0,1]$. This means that there exists a threshold $\theta^* \in [0,1]$ such that $\eta(x)$ is less than $1/2$ for all $x < \theta^*$, and greater or equal to $1/2$ for all $x \geq \theta^*$.

We assume that we have a large (infinite) pool of examples we can select from. We can choose any feature point $X_i \in [0,1]$ and observe its label Y_i. The data collection operation has a temporal aspect to it, namely we collect the labeled examples one at the time, starting with (X_1, Y_1) and proceeding until (X_n, Y_n) is observed. Formally we have:

A1 - The observation Y_i, $i \in \{1, \ldots, n\}$ are distributed as

$$Y_i = \begin{cases} 1 & \text{, with probability } \eta(X_i) \\ 0 & \text{, with probability } 1 - \eta(X_i) \end{cases} \ .$$

The random variables $\{Y_i\}_{i=1}^n$ is conditionally independent given $\{X_i\}_{i=1}^n$.

A2.1 - Passive Sampling: The sample locations X_i are independent of $\{Y_j\}_{j \neq i}$.

A2.2 - Active Sampling: The sampling location X_i depends only on $\{X_j, Y_j\}_{j<i}$. In other words

$$X_i|X_1 \ldots X_{i-1}, X_{i+1}, \ldots, X_n, Y_1 \ldots Y_{i-1}, Y_{i+1}, \ldots, Y_n \overset{\text{a.s.}}{=} X_i|X_1 \ldots X_{i-1}, Y_1 \ldots Y_{i-1} \ .$$

The conditional distribution on the right hand side (r.h.s) of the above expression is called the *sampling strategy* and is denoted by S_n. It completely defines our sampling schedule. After collecting the n examples, that is, after collecting $\{X_i, Y_i\}_{i=1}^n$, we construct a classifier $\widehat{G}_n$ that is desirably close to G^*. We use the subscript n to denote dependence on the data set, instead of writing it explicitly.

Under the passive sampling scenario (A2.1) the sample locations do not depend in any way on our observations, therefore the collection of sample points $\{X_i\}_{i=1}^n$ can be chosen before any observations are collected. On the other hand, the active sampling scenario (A2.2) allows for the i^{th} sample location to be chosen using all the information collected up to that point (the previous $i-1$ samples).

To be able to present results on rates of convergence of the excess risk we need to impose further conditions on $\eta(\cdot)$, namely on the behavior of the conditional probability around the 1/2 crossing point θ^*. Let $\kappa \geq 1$ and $\mu > 0$, then

$$|\eta(x) - 1/2| \geq \mu |x - \theta^*|^{\kappa - 1}, \quad \text{if } |x - \theta^*| \leq \epsilon_0 \ , \qquad (2)$$

$$|\eta(x) - 1/2| \geq \mu {\epsilon_0}^{\kappa - 1}, \quad \text{if } |x - \theta^*| > \epsilon_0 \ , \qquad (3)$$

for some $\epsilon_0 > 0$. Examples of such conditional probability functions are depicted in Figure 1. The conditions above are very similar to the so-called margin condition (or noise-condition) introduced by Tsybakov [18], although our condition is slightly more restrictive. Nevertheless the key aspect

[1] $A \Delta B \equiv (A \cap B^c) \cup (A^c \cap B)$, where A^c and B^c are the complement of A and B respectively.

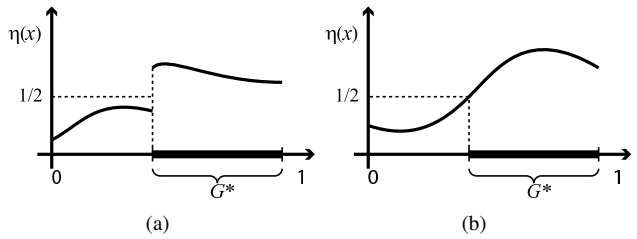

Fig. 1. *Examples of two conditional distributions $\eta(x) = \Pr(Y = 1|X = x)$. (a) In this case $\eta(\cdot)$ satisfies the margin condition with $\kappa = 1$; (b) Here the margin condition is satisfied for $\kappa = 2$.*

of these conditions is the behavior of $\eta(\cdot)$ near the Bayes decision boundary. Let $\mathcal{P}(\kappa, \mu)$ be the class of distributions P_{XY} such that the marginal P_X is uniform in $[0, 1]$ and $\eta(\cdot)$ satisfies (2) and (3). If $\kappa = 1$ then the $\eta(\cdot)$ function "jumps" across 1/2, that is $\eta(\cdot)$ is bounded away from the value $1/2$ (see Figure 1(a)). If $\kappa > 1$ then $\eta(\cdot)$ crosses the value $1/2$ at θ^*. Arguably the most interesting case corresponds to $\kappa = 2$ (Figure 1(b)). In this case the conditional probability behaves linearly around $1/2$. This means that the noise affecting observations that are made close to the decision boundary is roughly proportional to the distance to the boundary. This might be due to the fact that the feature X is not powerful enough to clearly distinguish the two classes near θ^*. Finally if $\kappa > 2$ then $\eta(\cdot)$ is very "flat" around θ^*.

III. Minimax Lower Bounds

In this section we present lower bounds on the performance of active and passive sampling methods, under the conditions described above. The proof techniques described here are relatively standard, and follow the approach in [19]. The key idea of the proof is to reduce the problem of estimating level sets of $\mathcal{P}(\kappa, \mu)$ to the problem of deciding among a finite number of such distributions. In other words we reduce the estimation problem to a hypotheses testing problem. In this case it suffices to consider only two different distributions P_{XY}.

Theorem 1 (Active Sampling Minimax Lower Bound): Let $\kappa > 1$. Under the assumptions (A1) and (A2.2) we have

$$\inf_{\widehat{G}_n, S_n} \sup_{P_{XY} \in \mathcal{P}(\kappa,\mu)} \mathbb{E}\left[R(\widehat{G}_n) - R(G^*)\right] \geq cn^{-\frac{\kappa}{2\kappa-2}} \ ,$$

for n large enough, where $c \equiv c(\kappa, \mu) > 0$ and the *infimum* is taken over the set of all possible classification rules $\widehat{G}_n$ and sampling strategies S_n.

Contrast this result with the one attained for passive sampling. Under the passive sampling scenario it is clear that the sample locations $\{X_i\}_{i=1}^n$ must be scattered around the interval $[0, 1]$ in a somewhat uniform manner. These can be deterministically placed, for example over a uniform grid, or simply taken uniformly distributed over $[0, 1]$. In [18] was shown that under assumptions (A1), (A2.1), and $\kappa \geq 1$, we have

$$\inf_{\widehat{G}_n} \sup_{P_{XY} \in \mathcal{P}(\kappa,\mu)} \mathbb{E}\left[R(\widehat{G}_n) - R(G^*)\right] \geq cn^{-\frac{\kappa}{2\kappa-1}} \ , \quad (4)$$

where the samples $\{X_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) uniformly over $[0, 1]$. Furthermore this bound is tight, in the sense that it is possible to devise classification strategies attaining the same asymptotic behavior.

We notice that under the passive sampling scenario the excess risk decays at a strictly slower rate than for the active sampling scenario. The difference is dramatic when $\kappa \to 1$ (Figure 1(a)). In that case it can actually be shown that an exponential rate of error decay is attainable when actively sampling [3]. When $\kappa \to \infty$ the excess risk decay rates are the same, regardless of the sampling paradigm. This indicates that if no assumptions are made on the conditional distribution $\Pr(Y = 1|X)$ then no advantage can be taken from the extra complexity of active sampling. As remarked before the most relevant case is $\kappa = 2$. In this case both passive and active sampling methods display polynomial rates of error decay, but the rate for active sampling is n^{-1}, which is significantly faster than $n^{-2/3}$, the best possible rate for passive sampling. In the rest of the paper we present algorithms showing that the rates of Theorem 1 are nearly achievable.

We finally point out that the result of Theorem 1 applies also the general setting presented in [18], and that using similar ideas minimax bounds can be derived for higher dimensional classifier classes, characterized in terms of metric entropy.

IV. Active Sampling Algorithms

In this section we present various active sampling algorithms that allow us to nearly achieve the lower bounds of Theorem 1. We start by presenting an algorithm proposed by Burnashev and Zigangirov [3], inspired by an idea of Horstein [15]. This algorithm is designed to work in the bounded noise rate case, that is, when $\kappa = 1$. This corresponds to a scenario where the conditional probability $\eta(x) = \Pr(Y = 1|X = x)$ is bounded away from $1/2$, $|\eta(x) - 1/2| \geq \mu$ for all $x \in [0, 1]$. The results and lessons learned from this algorithm yield some intuition for general margin parameters, namely a "back-of-the-envelope" analysis indicates that a similar algorithm can be used to nearly achieve the rates of Theorem 1.

A. Bounded noise rate: $\kappa = 1$

Under this scenario we assume that $|\eta(x) - 1/2| \geq \mu$ for all $x \in [0, 1]$. Notice that in this case the class $\mathcal{P}(1, \mu)$ is a quasi-parametric class: elements of the class are essentially characterized by the location of the point where $\eta(\cdot)$ "crosses" $1/2$, denoted by θ^*. If the observations are noiseless (that is $\eta(x) \in \{0, 1\}$ or equivalently $\mu = 1/2$) then it is clear that one can estimate θ^* very efficiently using binary bisection: start by taking a sample at $X_1 = 1/2$. Depending on the outcome we know if θ^* is to the left of X_1 (if $Y_1 = 1$) or to the right (if $Y_1 = 0$). Proceeding accordingly we can construct an estimate of θ^* denoted by $\widehat{\theta}_n$ and a corresponding classifier $\widehat{G}_n \equiv [\widehat{\theta}_n, 1]$ such that

$$\mathbb{E}[R(\widehat{G}_n) - R(G^*)] = \mathbb{E}[|\widehat{\theta}_n - \theta^*|] \leq 2^{-(n+1)}.$$

If there is noise ($\mu < 1/2$) things get a bit more complicated, in part because our decisions about the sampling depend on all the observations made in the past, which are noisy and therefore unreliable. Nevertheless there is a probabilistic bisection method, proposed in [15], that is suitable for this purpose. The key idea stems from Bayesian estimation. Suppose that we have a prior probability density function $P_0(x)$ on the unknown parameter θ^*, namely that θ^* is uniformly distributed over the interval $[0,1]$ (that is $P_0(x) = 1$ for all $x \in [0,1]$). To make the exposition clear assume a particular situation, namely that $\theta^* = 1/4$. Like before, we start by taking a measurement at $X_1 = 1/2$. With probability $\eta(X_1) \geq 1/2 + \mu$ we observe a one, and with probability $1-\eta(X_1) \leq 1/2-\mu$ we observe a zero. Therefore it is more likely to observe a one than a zero. Suppose a one was observed. Given these facts we can compute a "posterior" density simply by applying an approximate Bayes rule (we assume that a one is observed with probability $1/2 + \mu$). In this case we would get that

$$P_1(x|X_1, Y_1) = \begin{cases} 1+2\mu) & \text{, if } x \leq 1/2, \\ 1-2\mu) & \text{, if } x > 1/2, \end{cases} .$$

The next step is to choose the sample location X_2. We choose X_2 so that is *bisects* the posterior distribution, that is, we take X_2 such that $\Pr_{\theta\sim P_1(\cdot)}(\theta > X_2|X_1, Y_1) = \Pr_{\theta\sim P_1(\cdot)}(\theta < X_2|X_1, Y_1)$. In other words X_2 is just the median of the posterior distribution. If our model is correct, the probability of the event $\{\theta < X_2\}$ is identical to the probability of the event $\{\theta > X_2\}$ and therefore sampling Y_2 at $\boldsymbol{X}_2$ is most informative. We continue iterating this procedure until we have collected n samples. The estimate $\widehat{\theta}_n$ is defined as the median of the final posterior distribution. Figure 2 illustrates the procedure. Note that if $\mu = 1/2$ then probabilistic bisection is simply the binary bisection described above.

The above algorithm seems to work extremely well in practice, but it is hard to analyze and there are few theoretical guarantees for it, especially pertaining rates of error decay. In [3] a similar algorithm was proposed. Although its operation is slightly more complicated, it is easier to analyze. That algorithm uses essentially the same ideas, but enforces a parametric structure for the posterior by imposing that the sample locations X_i lie on a grid, in particular $X_i \in \{0, \Delta, 2\Delta, \ldots, 1\}$ where $m = \Delta^{-1} \in \mathbb{N}$. Furthermore in the application of the Bayes rule we use α instead of μ, where $0 < \alpha < \mu < 1/2$. A description of the algorithm can be found in [3] or in [4]. We call this method the BZ algorithm. We have the following remarkable result [3].

$$\Pr(|\widehat{\theta}_n - \theta^*| > \Delta) \leq \frac{1-\Delta}{\Delta}\left(\frac{1+2\mu}{2+4\alpha} + \frac{1-2\mu}{2-4\alpha}\right)^n . \quad (5)$$

From the estimate $\widehat{\theta}_n$ we construct a classifier $\widehat{G}_n \equiv [\widehat{\theta}_n, 1]$. To get a bound on the expected excess risk one proceeds by

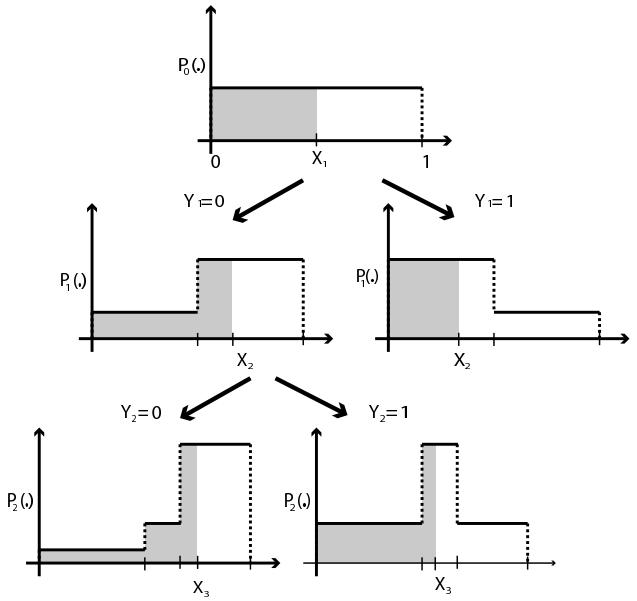

Fig. 2. *Illustration of the probabilistic bisection strategy. The shaded areas correspond to* $1/2$ *of the probability mass of the posterior densities.*

integration

$$\begin{aligned}
\mathbb{E}[R(\widehat{G}_n) - R(G^*)] &= \mathbb{E}\left[\int_{\widehat{G}_n \Delta G^*} |2\eta(x) - 1| \mathrm{d}x\right] \\
&\leq \mathbb{E}[|\widehat{G}_n \Delta G^*|] \\
&= \mathbb{E}[|\widehat{\theta}_n - \theta^*|] \\
&= \int_0^1 \Pr(|\widehat{\theta}_n - \theta^*| > t) \mathrm{d}t \\
&= \int_0^\Delta \Pr(|\widehat{\theta}_n - \theta^*| > t) \mathrm{d}t + \int_\Delta^1 \Pr(|\widehat{\theta}_n - \theta^*| > t) \mathrm{d}t \\
&\leq \Delta + (1-\Delta)\Pr(|\widehat{\theta}_n - \theta^*| > \Delta) \\
&\leq \Delta + \frac{(1-\Delta)^2}{\Delta}\left(\frac{1+2\mu}{2+4\alpha} + \frac{1-2\mu}{2-4\alpha}\right)^n .
\end{aligned}$$

Taking $\Delta = \left(\frac{1+2\mu}{2+4\alpha} + \frac{1-2\mu}{2-4\alpha}\right)^{n/2}$ and $\alpha = \frac{1-\sqrt{1-4\mu^2}}{4\mu}$ (to minimize the exponent) yields

$$\mathbb{E}[R(\widehat{G}_n) - R(G^*)] \leq 2\left(\frac{1}{2} + \frac{1}{2}\sqrt{1-4\mu^2}\right)^{n/2} .$$

Notice that the excess risk decays exponentially in the number of samples. This is much faster than what is attainable using only passive sampling, where the decay rate is $1/n$. Like in the noiseless scenario we obtain an exponential rate of error decay. The difference is that now the exponent depends on the noise margin μ, larger noise margins corresponding to faster error decay rates.

B. Unbounded rate noise: $\kappa > 1$

In this section we consider scenarios where the noise rate is not bounded, that is, as one makes observations closer to the transition point θ^* the observation noise becomes larger. This clearly hinders extremely fast excess risk decay rates,

since as one focus more and more on the Bayes decision boundary the quality of the observations degrades.

To gain some intuition we consider first the case where $|\eta(x) - 1/2| = \mu|x - \theta^*|^{\kappa-1}$, corresponding to a margin parameter κ. Proceed now as in the previous section, and collect samples over a grid, namely $X_i \in \{0, \Delta, 2\Delta, \dots, 1\}$ where $m = \Delta^{-1} \in \mathbb{N}$. If this grid does not line-up with the transition point θ^* then $|\eta(x) - 1/2| \geq \mu(\Delta/2)^{\kappa-1}$ for all $x \in \{0, \Delta, \dots, 1\}$. Of course this is in general an unrealistic assumption, but let us consider it for now. We can now proceed by using the method described in the previous section replacing μ by $\mu(\Delta/2)^{\kappa-1}$ and using (5). Begin by noticing that due to the special form of $\eta(\cdot)$ the behavior of the expected excess risk is related to the behavior of $|\widehat{\theta}_n - \theta^*|$ in an interesting way, namely

$$\begin{aligned}\mathbb{E}[R(\widehat{G}_n) - R(G^*)] &= \mathbb{E}\left[\int_{\widehat{G}_n \Delta G^*} |2\eta(x) - 1| \mathrm{d}x\right] \\ &= \mathbb{E}\left[\int_{\widehat{G}_n \Delta G^*} 2\mu|x - \theta^*|^{\kappa-1} \mathrm{d}x\right] \\ &\leq \mathbb{E}[|\widehat{\theta}_n - \theta^*|^{\kappa}] \ .\end{aligned}$$

We now proceed in a similar fashion as before

$$\begin{aligned}\mathbb{E}[R(\widehat{G}_n) - R(G^*)] &\leq \mathbb{E}[|\widehat{\theta}_n - \theta^*|^{\kappa}] \\ &= \int_0^1 \Pr(|\widehat{\theta}_n - \theta^*|^{\kappa} > t) \mathrm{d}t \\ &= \int_0^1 \Pr(|\widehat{\theta}_n - \theta^*| > t^{1/\kappa}) \mathrm{d}t \\ &= \int_0^{\Delta^{\kappa}} \Pr(|\widehat{\theta}_n - \theta^*| > t^{1/\kappa}) \mathrm{d}t \\ &\quad + \int_{\Delta^{\kappa}}^1 \Pr(|\widehat{\theta}_n - \theta^*| > t^{1/\kappa}) \mathrm{d}t \\ &\leq \Delta^{\kappa} + (1 - \Delta^{\kappa}) \Pr(|\widehat{\theta}_n - \theta^*| > \Delta) \\ &\leq \Delta^{\kappa} + \frac{1}{\Delta}\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\mu^2(\Delta/2)^{2\kappa-2}}\right)^n \\ &\leq \Delta^{\kappa} + \frac{1}{\Delta}\left(1 - \mu^2(\Delta/2)^{2\kappa-2}\right)^n \\ &\leq \Delta^{\kappa} + \frac{1}{\Delta}\exp\left(-n\mu^2(\Delta/2)^{2\kappa-2}\right) \ ,\end{aligned}$$

where the last two steps follow from the fact that $\sqrt{x} \leq (x+1)/2$ for all $x \geq 0$, and that $(1 + s(x))^x \leq \exp(xs(x))$ for all $x > 0$ and $s(x) > -1$. Finally, let

$$\Delta = \frac{1}{2}\left(\frac{\kappa + 1}{\mu^2(2\kappa - 2)} \ \frac{\log n}{n}\right)^{\frac{1}{2\kappa+2}} \ .$$

We conclude that

$$\mathbb{E}[R(\widehat{G}_n) - R(G^*)] \propto \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}} . \tag{6}$$

This is lower bound rate displayed in Theorem 1, apart from logarithmic factors. The result indicates that, in principle, a methodology similar to the Burnashev-Zigangirov algorithm might allow us to achieve the lower bound rates.

It is important to emphasize that the above result holds under the assumption that the sampling grid is not aligned with the unknown transition point θ^*. If that is not the case then we will have $|\eta(x) - 1/2| < \mu(\Delta/2)^{\kappa-1}$ for some sampling point, and the analysis above does not hold. One way to avoid this problem is to consider different sampling grids. A methodology that can be shown to work consists in dividing the available measurements into three equal sets, and use three different offset sampling grids. If we have a budget of n samples we allocate $n/3$ samples and run the BZ algorithm for one sampling grid. Then we use other $n/3$ samples and run the BZ algorithm for another sampling grid, essentially a slightly shifted version of the first sampling grid, and proceed in an analogous fashion with the remaining $n/3$ samples. The rational is that at most one of these sampling grids is going to be closely aligned with the unknown transition point θ^*, and therefore the other two are not going to display any problems. We then check for agreement amoung the three estimators and make a decision that provably attains the rate in (6). Although this methodology is satisfying from a theoretical point of view, it is somewhat wasteful (essentially only one third of the samples are effectively used), and does not generalize well to more complicated and realistic scenarios than the one considered in this paper. Therefore it is not a very relevant approach in practice. In the remaining of this document we present an alternative methodology that solves some of these problems and provides a practical algorithm capable of achieving the desired fast rates.

It is worth pointing out that, even when the assumption that the sampling grid does not line up with the transition θ^* does not hold, the BZ method still works extremely well in practice. The difficulties arise solely on the performance analysis of the algorithm.

C. Unbounded rate noise: A sample-efficient algorithm

The algorithm presented here is similar to the original BZ algorithm, although some modifications are made allowing the analysis of more general noise models than the bounded rate noise. Like in the original methodology we concentrate our sampling on a grid, in particular $X_i \in \{0, \Delta, 2\Delta, \dots, 1\}$ where $m = \Delta^{-1} \in \mathbb{N}$. The algorithm works by propagating a posterior-like density (we will denote this density as posterior hereafter). After j observation the posterior is described as $P_j : [0, 1] \to \mathbb{R}$,

$$P_j(x) = \Delta^{-1} \sum_{i=1}^{m} b_i(j) \mathbf{1}_{I_i}(x) \ ,$$

where $I_1 = [0, \Delta]$ and $I_i = (\Delta(i-1), \Delta i]$, for $i \in \{2, \dots, m\}$. We initialize this posterior by taking $b_i(0) = \Delta$ for all i. Note that the posterior is completely characterized by $\boldsymbol{b(j)} = \{b_1(j), \dots, b_m(j)\}$, and that $\sum_{i=1}^m b_i(j) = 1$. We select the sample location X_{j+1} using a randomized approach based on P_j: choose a number $k_{j+1} \in \{1, \dots, m\}$ according to the distribution $\boldsymbol{b(j)}$. In other words $\Pr(k_{j+1} = i) = b_i(j)$. Now let $W \in \{-2, -1, 0, 1\}$ be a independent random variable, taking one of the values -2,-1,0, or 1 with probability $1/4$. The sample location is given by $X_{j+1} = \min\left(\max\left(\Delta(k(j+1) - W), 0\right), 1\right)$. We collect a

label Y_{j+1} and update the posterior accordingly (implicitly assuming a noise margin $0 < \alpha < 1/2$), namely if $i \leq k_{j+1} - W$

$$b_i(j+1) = \begin{cases} \frac{(1/2-\alpha)b_i(j)}{1/2+\alpha-2\alpha\sum_{l=1}^{k_{j+1}-W} b_l(j)} & \text{, if } Y_{j+1} = 0 \\ \frac{(1/2+\alpha)b_i(j)}{1/2-\alpha+2\alpha\sum_{l=1}^{k_{j+1}-W} b_l(j)} & \text{, if } Y_{j+1} = 1 \end{cases},$$

and if $i > k_{j+1} - W$

$$b_i(j+1) = \begin{cases} \frac{((1/2+\alpha)b_i(j)}{1/2+\alpha-2\alpha\sum_{l=1}^{k_{j+1}-W} b_l(j)} & \text{, if } Y_{j+1} = 0 \\ \frac{(1/2-\alpha)b_i(j)}{1/2-\alpha+2\alpha\sum_{l=1}^{k_{j+1}-W} b_l(j)} & \text{, if } Y_{j+1} = 1 \end{cases}.$$

We call α the *update parameter*. Finally, at time j, the estimate of the true transition point θ^* is given by $\widehat{\theta}_j$ that is defined as the median of $P_j(\cdot)$.

A couple of remarks are important at this point: (i) this method differs from the original BZ algorithm in that the latter chooses the next sample location to be one of the two grid points that are closest to the median of the posterior. Choosing the sample point according to this procedure, or "sampling" the posterior (as is the case with our algorithm) is essentially the same provided the posterior is somewhat concentrated around a point, which happens with high probability after a number of observations have been collected. (ii) Instead of choosing W to take values in $\{-2, -1, 0, 1\}$ one could take W to be Bernoulli with parameter $1/2$. This methodology works, provided we are working under the bounded noise rate assumption, or in general noise models if the transition point θ^* is not closely aligned with the sampling grid. If the last condition does not hold then the proof technique breaks down, and it is not possible to guarantee rates of error decay. This same problem arises in the original BZ method. (iii) When using four sample points around the chosen bin k_{j+1} (or in general points in the vicinity of the chosen bin) we can guarantee that most of the times we are taking a sample at a point that has a reasonable noise margin (that is, a point such that $|\eta(x)-1/2|$ is reasonably large).

The analysis of the algorithm above can be done in a similar fashion to the analysis of the original BZ algorithm, although further complications arise due to the inability to control the noise margin at one of the sampling points (see proof of Lemma 1 for details). We use the algorithm to construct the classifier $\widehat{G}_n = [\widehat{\theta}_n, 1]$ and have the following result.

Theorem 2: Let $P_{XY} \in \mathcal{P}(\kappa, \mu)$. Furthermore assume that P_{XY} does not satisfy the margin condition for any margin parameter κ_0 such that $\kappa_0 < \kappa$. Then there exists a bin size $\Delta \equiv \Delta(n, \kappa, \mu)$ and an update parameter $\alpha \equiv \alpha(n, \kappa, \mu)$ such that

$$\mathbb{E}[R(\widehat{G}_n) - R(G^*)] \leq C \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}}, \qquad (7)$$

where $C \equiv C(\kappa, \mu) > 0$.

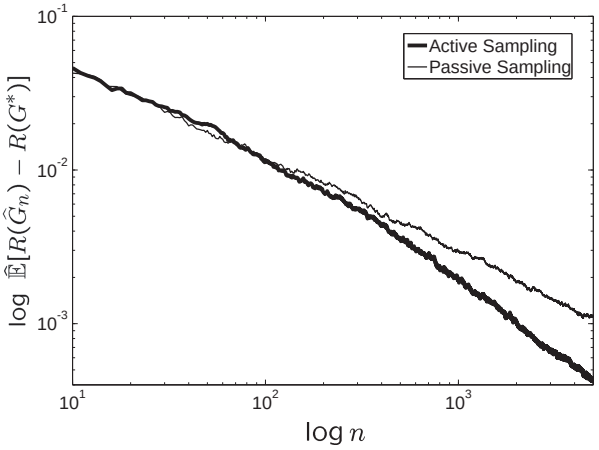

Fig. 3. *Simulation results for the case $\kappa = 2$. The empirical expected excess risk for the active sampling algorithm corresponds to the bold line, and the corresponding curve for passive sampling corresponds to the regular line.*

The proof of the theorem follows from an upper bound of $\Pr(|\widehat{\theta}_n - \theta^*| > \Delta)$, similar to (5), and the reasoning of Section IV-B. The key result enabling the proof is the following lemma.

Lemma 1: Let $P_{XY} \in \mathcal{P}(\kappa, \mu)$ and $\kappa \geq 2$. Then taking $\alpha = 0.09\mu(3\Delta/4)^{\kappa-1}$ yields

$$\Pr(|\widehat{\theta}_n - \theta^*| > \Delta) \leq \frac{1-2\Delta}{2\Delta}\left(1 - \frac{\mu^2(3\Delta/4)^{\kappa-1}}{50}\right)^n. \qquad (8)$$

Although this lemma is presented only for $\kappa \geq 2$ it is possible to obtain similar results for any $\kappa > 1$, identical to the above bounds apart from the constants. In the interest of clarity we consider only the case $\kappa \geq 2$.

Some remarks are important at this time: In the statement of Theorem 2 one notices that Δ and α are functions of κ and μ. This indicates that the proposed algorithm is not adaptive, that is, it cannot handle a scenario where κ and μ are unknown.

It turns out that the choice of Δ as a function of n, κ, and μ is not critical for the practical performance of the algorithm. Essentially finer sampling grids provide the same or better performance than the one predicted by the theoretical analysis. However, the choice of update parameter α is critical, since the wrong scaling of the parameter with n can lead into slower rates of error decay.

V. Simulation Results

In this section we present a simple simulation of the proposed algorithm. This does not constitute a thorough empirical study, but simply an illustration of the practical performance characteristics. We conducted various other simulation tests, all achieving results agreeing with the theoretical analysis above.

In this simulation we considered the case $\kappa = 2$, and the distribution P_{XY} is characterized by $\eta(x) = (x - \theta^*)/2 + 1/2$. We performed 1000 runs, in each one selecting θ^*

uniformly at random in the interval $[0,1]$. In Figure 3 we plot the empirical expected excess risk versus the number of samples n. The expected excess risk is estimated by averaging the excess risk over the 1000 trials. We display the results in a log-log plot. We observe that the active method outperforms the passive method for any number of samples. Furthermore the rates of error decay coincide with the rates predicted by the theory, that is, the excess risk decays at a rate n^{-1} for active sampling (in the log-log plot this corresponds to a slope of -1), and at a rate $n^{-2/3}$ for passive sampling (corresponding to a slope of $-2/3$ in the plot).

VI. Final Remarks and Open Questions

We present a formal analysis of active learning in an unbounded noise setting. We show that the extra flexibility of active sampling, when compared to passive sampling, allows for faster rates of error decay, but the errors depend on the behavior of the noise margin around the Bayes decision boundary. We present results characterizing the fundamental limits of active learning for the one-dimensional setting of this paper, and provide algorithms capable of nearly achieving those limits. The algorithms are practical and easily implemented. The algorithms presented are not adaptive to the noise conditions, in the sense that they require knowledge of margin parameters κ and μ in order to achieve the optimal rates, at least in theory. However, empirical evidence suggests that the proposed algorithms yield the optimal performance (in terms of error rates) solely with mild knowledge of κ, but the proof techniques used here are not powerful enough to demonstrate that. The ideas presented here can be extended to larger and more complicated classes of classifiers, characterized for example in terms of metric entropy. The development and analysis of such extensions is currently being investigated.

Appendix

A. Proof of Theorem 1

The proof strategy follows the basic idea behind standard minimax analysis methods, and consists in reducing the problem of classification in the class $\mathcal{P}(\kappa,\mu)$ to a hypothesis testing problem. In this case it suffices to consider two hypothesis.

$$\begin{aligned}\eta_0(x) &= \begin{cases} \frac{1}{2}-\mu(t-x)^{\kappa-1} & ,x\le t \\ 1 & ,x>t \end{cases} , \\ \eta_1(x) &= \begin{cases} \frac{1}{2}+\mu x^{\kappa-1} & ,x\le t \\ 1 & ,x>t \end{cases} .\end{aligned}$$

These are depicted in Figure 4. Note that $G_0^* = [t,1]$ and $G_1^* = [0,1]$ (provided t is small enough). Let $\widehat{G}_n$ be any classifier (i.e. a subset of $[0,1]$) and define $\psi_n = \arg\min_{i\in\{0,1\}} |\widehat{G}_n \Delta G_i^*|$, where $|\cdot|$ denotes the volume of a set. We have the following result.

Lemma 2: For $j\in\{0,1\}$

$$\psi_n \neq j \Rightarrow R_j(\widehat{G}_n) - R_j(G_0^*) \ge \frac{2\mu}{\kappa 2^\kappa} t^\kappa \triangleq s \propto \mu t^\kappa ,$$

where R_j denotes the risk (1) under hypothesis j.

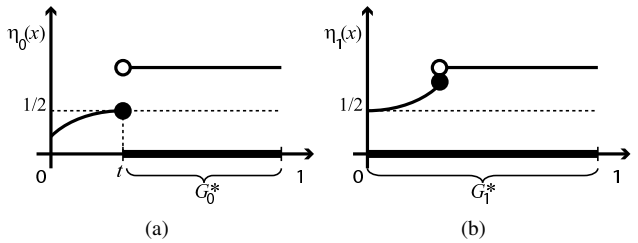

Fig. 4. *The two conditional distributions used for the proof of Theorem 1.*

Proof: Assume first that $j=0$, the case $j=1$ is analogous. Define $\tilde{G}_n = \widehat{G}_n \cap [0,t]$, and $\tilde{G}_n^c = \widehat{G}_n^c \cap [0,t]$. Since $\psi_n = 1$ we have

$$\begin{aligned} |\widehat{G}_n \Delta G_1^*| \le |\widehat{G}_n \Delta G_0^*| &\Leftrightarrow |\tilde{G}_n \Delta G_1^*| \le |\tilde{G}_n \Delta G_0^*| \\ &\Leftrightarrow |\tilde{G}_n^c| \le |\tilde{G}_n| . \end{aligned}$$

Therefore

$$\begin{aligned} R_0(\widehat{G}_n) - R_0(G_0^*) &\ge \int_{\tilde{G}_n} |2\eta_0(x)-1| \mathrm{d}x \\ &\ge \int_{t/2}^{t} |2\eta_0(x)-1| \mathrm{d}x \\ &= \frac{2\mu}{\kappa 2^\kappa} t^\kappa . \end{aligned}$$

□

A consequence of Lemma 2 is that

$$\begin{aligned} &\inf_{\widehat{G}_n} \sup_{P_{XY}\in\mathcal{P}(\kappa,a)} \Pr(R(\widehat{G}_n) - R(G^*) \ge s) \\ &\quad \ge \max_j \mathcal{P}_j(\psi_n \neq j) \\ &\quad \ge \inf_{\phi_n} \sup_j \mathcal{P}_j(\phi_n \neq j) \triangleq p_e , \end{aligned} \tag{9}$$

where the *infimum* on the r.h.s. is taken with respect to all functions of the data onto $\{0,1\}$. All that remains to be done now is to construct a lower bound for p_e. To this we use a result from [19]. Let $P_{0,n} \equiv P^{(0)}_{X_1,\ldots,X_n,Y_1,\ldots,Y_n}$ be the probability measure of the random variables $\{X_i,Y_i\}_{i=1}^n$ under hypothesis 0 and define analogously $P_{1,n} \equiv P^{(1)}_{X_1,\ldots,X_n,Y_1,\ldots,Y_n}$. Theorem 2.2 of [19] says that if $\mathrm{K}(P_{1,n}\|P_{0,n}) \le \alpha < \infty$ then

$$p_e \ge \max\left(\frac{1}{4}\exp(-\alpha), \frac{1-\sqrt{\alpha/2}}{2}\right) ,$$

where $\mathrm{K}(\cdot\|\cdot)$ denotes the Kullback divergence. Now

$$\begin{aligned} \mathrm{K}(P_{1,n}\|P_{0,n}) &= \mathbb{E}_1\left[\log\frac{P_{1,n}}{P_{0,n}}\right] \\ &= \mathbb{E}_1\left[\mathbb{E}_1\left[\log\frac{P_{1,n}}{P_{0,n}}\middle| X_1,\ldots,X_n\right]\right] \\ &= n\mathbb{E}_1\left[\mathbb{E}_1\left[\log\frac{P^{(1)}_{Y_i|X_i}}{P^{(0)}_{Y_i|X_i}}\middle| X_1,\ldots,X_n\right]\right] \\ &\le n\,\mathbb{E}_1\left[\log\frac{P^{(1)}_{Y_i|X_i}}{P^{(0)}_{Y_i|X_i}}\middle| X_i=0,\ \forall i\right] \\ &\le n\left((2at^{\kappa-1})^2 + o(2at^{\kappa-1})^2\right) , \end{aligned}$$

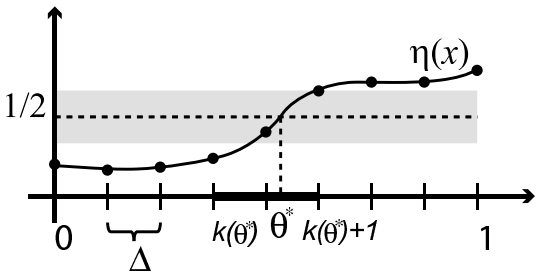

Fig. 5. *Illustration of case 2 in the proof of Theorem 2. Only one sample point is lying inside the gray area (corresponding to a noise margin less than* $\mu(\Delta/4)^{\kappa-1}$.

as $t \to 0$. In the above $\mathbb{E}_1$ denotes the expectation taken with respect to measure $P_{1,n}$. Therefore $\mathrm{K}(P_{1,n}\|P_{0,n}) \leq 4\mu^2 n t^{2\kappa-2}$. Taking $t \propto (\mu^2 n)^{-\frac{1}{2\kappa-2}}$ and using (9) we conclude that

$$\inf_{\widehat{G}_n} \sup_{P_{XY}\in\mathcal{P}(\kappa,a)} \Pr\left(R(\widehat{G}_n) - R(G^*) \geq \mu^{-\frac{1}{\kappa}} n^{-\frac{\kappa}{2\kappa-2}}\right) \geq c > 0 \ ,$$

where $c > 0$ is a constant. The statement of the theorem now follows from the application of Markov's inequality to the above expression,

$$\mathbb{E}\left[R(\widehat{G}_n) - R(G^*)\right] \geq \mu^{-\frac{1}{\kappa}} n^{-\frac{\kappa}{2\kappa-2}} \Pr\left(R(\widehat{G}_n) - R(G^*) \geq \mu^{-\frac{1}{\kappa}} n^{-\frac{\kappa}{2\kappa-2}}\right) .$$

Remark: Notice that, when bounding the Kullback divergence, we considered all the feature examples to be taken at $X_i = 0$, the most beneficial place to take a sample. If instead we assume X_i i.i.d. uniformly over $[0, 1]$ the Kullback divergence is approximately $nt^{2\kappa-2}t = nt^{2\kappa-1}$, since roughly only a fraction t of the samples are informative (any sample taken in $(t, 1]$ is non-informative). Proceeding as before we obtain the passive sampling minimax bound (4), proved in [18] for a more general setting. □

B. Proof of Lemma 1

The proof of Lemma 1 follows closely the approach taken in [3], but some complications arise from the fact that the noise margin for one of the grid sample points (namely the point closest to θ^*) cannot be lower bounded. For a fix distribution $P_{XY} \in \mathcal{P}(\kappa, \mu)$ and a given parameter Δ two different scenarios can happen. For all the results below we assume that Δ is small enough so that condition (3) does not play a critical role. The proof also assumes that $\kappa \geq 2$. It is possible generalize the argument for any $\kappa > 1$, (by changing the definition of case 1 and case 2 below), but we do not consider this for the sake of clarity.

Case 1: *The sampling grid points are not aligned with the transition point* θ^*. We consider that this happens if the point θ^* is in the middle half of one of the bins. Formally this means that for some $i \in \mathbb{N}$ $\theta^* - i\Delta > \Delta/4$ and $(i+1)\Delta - \theta^* > \Delta/4$. This implies that for all sampling points the noise margin can be lower bounded, in particular $|\eta(x) - 1/2| \geq \mu(\Delta/4)^{\kappa-1}$ for all $x = \Delta i$, $i \in \{0, \ldots, m\}$. Therefore this case is essentially like the bounded noise margin scenario, and the analysis can be carried out using techniques in [3].

Define $k(\theta^*)$ to be the index of the bin I_i containing θ^*, that is $\theta^* \in I_{k(\theta^*)}$. Define

$$M(j) = \frac{1 - b_{k(\theta^*)}(j)}{b_{k(\theta^*)}(j)},$$

and

$$N(j+1) = \frac{M(j+1)}{M(j)} = \frac{b_{k(\theta^*)}(j)(1 - b_{k(\theta^*)}(j+1))}{b_{k(\theta^*)}(j+1)(1 - b_{k(\theta^*)}(j))}.$$

The reasoning behind these definitions is made clear later. For now, notice that $M(j)$ is a decreasing function of $b_{k(\theta^*)}(j)$.

After n observations our estimate of θ^* is the median of the posterior density P_n. Taking that into account we conclude that

$$\begin{aligned} \Pr(|\widehat{\theta}_n - \theta^*| > \Delta) &\leq \Pr(b_{k(\theta^*)}(n) < 1/2) \\ &= \Pr(M(n) > 1) \\ &\leq \mathbb{E}[M(n)], \end{aligned}$$

where the last step follows from Markov's inequality. The definition of $M(j)$ above is meant to get more leverage out of Markov's inequality, in a similar spirit of Chernoff bounding techniques. Using the definition of $N(j)$ and some conditioning we get

$$\begin{aligned} \mathbb{E}[M(n)] &= \mathbb{E}[M(n-1)N(n)] \\ &= \mathbb{E}\left[\mathbb{E}[M(n-1)N(n)|\boldsymbol{b}(n-1)]\right] \\ &= \mathbb{E}\left[M(n-1)\mathbb{E}[N(n)|\boldsymbol{b}(n-1)]\right] \\ &\vdots \\ &= M(0)E\left[\mathbb{E}[N(1)|\boldsymbol{b}(0)]\cdots\mathbb{E}[N(n)|\boldsymbol{b}(n-1)]\right] \\ &\leq M(0) \times \\ &\quad \left\{\max_{j\in\{0,\ldots,n-1\}} \sup_{\boldsymbol{b}(j)} \mathbb{E}[N(j+1)|\boldsymbol{b}(j)]\right\}^n . \end{aligned} \tag{10}$$

The rest of the proof consists of upper bounding $\mathbb{E}[N(j+1)|\boldsymbol{b_j}] < 1$, showing that it is always less than 1. Before proceeding we make some remarks about the above technique. Note that $M(j)$ measures how much mass is on the bin containing θ^* (if $M(j) = 0$ all the mass in our posterior is in the bin containing θ^*, the least error scenario). The ratio $N(j)$ is a measure of the improvement (in terms of concentrating the posterior around the bin containing θ^*) by sampling at X_j and observing Y_j. This is strictly less than one when an improvement is made. The bound (10) above is therefore only useful if, no matter what happened in the past, a measurement made with the proposed algorithm always leads to a performance improvement on average.

It is possible to show that

$$\mathbb{E}[N(j+1)|\boldsymbol{b_j}] \leq \frac{1}{2}\left(\frac{1 + 2\mu(\Delta/4)^{\kappa-1}}{1+2\alpha} + \frac{1 - 2\mu(\Delta/4)^{\kappa-1}}{1-2\alpha}\right)$$

Since this derivation is cumbersome, and very similar to the one of case 2, we present only the derivation for that case. One can check that using the update parameter α as prescribed in the lemma statement and (10) satisfies the desired bound (8).

Case 2: *The grid sampling points are aligned with the transition point* θ^*. Formally, there is a grid point $i\Delta$, $i \in \mathbb{N}$, such that $|\theta^* - i\Delta| \leq \Delta/4$ (see Figure 5). Therefore, for that particular sampling point we have $|\eta(i\Delta) - 1/2| \leq \mu(\Delta/4)^{\kappa-1}$. For all the other grid-points $x \neq i\Delta$ we have $|\eta(x) - 1/2| > \mu(3\Delta/4)^{\kappa-1}$. Unlike in case 1 it is not sufficient to keep track of the bin containing θ^*. This is in part due to the proof strategy, but also due to the fact that if θ^* coincides exactly with a sampling point the two bins of the posterior that share that point might both get significant probability mass after several measurements. There is the need to keep track of both bins, which complicates matters a bit. To simplify the notation below let $p = \mu(3\Delta/4)^{\kappa-1}$ and $q = \eta(i\Delta) - 1/2$, so that $|q| \leq \mu(\Delta/4)^{\kappa-1}$.

Define $k(\theta^*) = \arg\min_{i\in\mathbb{N}} |\theta^* - i\Delta|$ so that θ^* is either in bin $I_{k(\theta^*)}$ or bin $I_{k(\theta^*)+1}$. In this presentation we ignore "edge-effects", that is we assume that θ^* is not close to 0 or 1. These cases can be handled in a similar fashion. Define

$$M(j) = \frac{1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)}{b_{k(\theta^*)}(j) + b_{k(\theta^*)+1}(j)},$$

and

$$\begin{aligned} N(j+1) &= \frac{M(j+1)}{M(j)} \\ &= \frac{b_{k(\theta^*)}(j) + b_{k(\theta^*)+1}(j)}{1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)} \times \\ &\quad \frac{1 - b_{k(\theta^*)}(j+1) - b_{k(\theta^*)+1}(j+1)}{b_{k(\theta^*)}(j+1) + b_{k(\theta^*)+1}(j+1)}. \end{aligned}$$

These are very similar to the definitions in the proof of case 1, except that now we are keeping track of two bins. We proceed exactly as in case 1 and obtain

$$\begin{aligned} &\Pr(|\widehat{\theta}_n - \theta^*| > \Delta) \leq \\ &\quad \Pr(b_{k(\theta^*)}(n) + b_{k(\theta^*)+1}(n) < 1/2) \\ &\leq M(0) \left\{ \max_{j\in\{0,\ldots,n-1\}} \sup_{\boldsymbol{b}(j)} \mathbb{E}[N(j+1)|\boldsymbol{b}(j)] \right\}^n \end{aligned}$$

To bound $\mathbb{E}[N(j+1)|\boldsymbol{b}(j)]$ we are going to condition on W (recall the algorithm described in Section IV-C). For the purposes of illustration we present here only the case $W = 0$. The procedure for the remaining scenarios is analogous. If $W = 0$ then the sample location at time $j+1$ is $X(j+1) = \Delta k_(j+1)$. Next we evaluate $N(j+1)$ for three different cases: (i) $k_{j+1} < k(\theta^*)$; (ii) $k_{j+1} > k(\theta^*)$; $k_{j+1} = k(\theta^*)$. For case (i) we have

$$\begin{aligned} N(j+1) &\leq 1 - 2\alpha \left(\frac{1+2p}{1+2\alpha} - \frac{1-2p}{1-2\alpha} \right) \times \\ &\quad \frac{\sum_{i=1}^{k_{j+1}} b_i(j)}{1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)}. \end{aligned}$$

For case (ii) we have

$$\begin{aligned} N(j+1) &\leq 1 - 2\alpha \left(\frac{1+2p}{1+2\alpha} - \frac{1-2p}{1-2\alpha} \right) \times \\ &\quad \frac{\sum_{i=k_{j+1}+1}^{m} b_i(j)}{1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)}. \end{aligned}$$

Finally for case (iii) we have (11), displayed in the next page.

To compute $\mathbb{E}[N(j+1)|\boldsymbol{b}(j)|W = 0]$ we first need to evaluate $\mathbb{E}[\sum_{i=1}^{k_{j+1}} b_i(j)|k_{j+1} < k(\theta^*)]$. It is not hard to show that

$$\begin{aligned} &\mathbb{E}\left[\sum_{i=1}^{k_{j+1}} b_i(j) | k_{j+1} < k(\theta^*) \right] = \\ &\quad \frac{\sum_{i=1}^{k(\theta^*)} b_i(j)}{2} - \frac{\sum_{i=1}^{k(\theta^*)} b_i^2(j)}{2\sum_{i=1}^{k(\theta^*)} b_i(j)}. \end{aligned}$$

Similarly we have

$$\begin{aligned} &\mathbb{E}\left[\sum_{i=k_{j+1}+1}^{m} b_i(j) | k_{j+1} > k(\theta^*) \right] = \\ &\quad \frac{\sum_{i=k(\theta^*)+1}^{m} b_i(j)}{2} - \frac{\sum_{i=k(\theta^*)+1}^{m} b_i^2(j)}{2\sum_{i=k(\theta^*)+1}^{m} b_i(j)}. \end{aligned}$$

These results can be used to easily evaluate $\mathbb{E}[N(j+1)|\boldsymbol{b}(j), W = 0]$. We do not display this expression since it is cumbersome.

Proceeding in a similar fashion for all the remaining possibilities for W yields the bound for $\mathbb{E}[N(j+1)|\boldsymbol{b}(j)]$ given by (12), displayed in the next page. Although this expression looks terribly complicated, it is possible to explicitly find the posterior $\boldsymbol{b}(j)$ maximizing it. This posterior only has mass in the vicinity of the true transition point θ^*. The intuitive reason for that is due to the fact that collecting samples at the $k(\theta^*)\Delta$ (the sampling point with a "bad" noise margin) yields the worst possible scenario. This can be formally shown using Lagrange multipliers. We do not include that derivation here, but it is available in a technical report [6]. For illustration purposes suppose that $q > 0$. Then $\mathbb{E}[N(j+1)|\boldsymbol{b}(j)]$ is largest when the posterior is of the form $b_{k(\theta^*)-1} = \delta - \xi/2$, $b_{k(\theta^*)+2} = 1 - \delta - \xi/2$, and $b_{k(\theta^*)+1} = \xi$, and all the other entries of $(b)(j)$ are zero. The largest value is attained taking $\xi \to 0$, and choosing δ maximizing the bound.

$$\begin{aligned} &\mathbb{E}[N(j+1)|\boldsymbol{b}(j)] \leq \max_{\delta} \Bigg\{ 1 - \frac{1}{4} - \\ &\quad \frac{1-2\alpha}{4(1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j))} \left(\frac{1+2p}{1+2\alpha} - \frac{1-2p}{1-2\alpha} \right) \times \\ &\quad (\delta^2 + (1-\delta)^2) + \frac{1}{4} \Bigg((1/2+q) \frac{1 - \alpha - (1-2\alpha)\delta}{\alpha} + \\ &\quad (1/2-q) \frac{\alpha + (1-2\alpha)\delta}{1-\alpha} \Bigg) \Bigg\} \end{aligned}$$

This bound can be computed explicitly, but the expressions are cumbersome and do not provide much insight. Instead we provide a simpler characterization of the bounds, namely

$$N(j+1) = \frac{b_{k(\theta^*)}(j) + b_{k(\theta^*)}(j)}{1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)} \times \left[(1/2+q) \frac{(1/2+\alpha)(1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)) - 2\alpha \sum_{i=1}^{k(\theta^*)-1} b_i(j)}{(1/2+\alpha) b_{k(\theta^*)}(j) + (1/2-\alpha) b_{k(\theta^*)}(j)} + \right.$$
$$\left. (1/2-q) \frac{(1/2-\alpha)(1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)) + 2\alpha \sum_{i=1}^{k(\theta^*)-1} b_i(j)}{(1/2-\alpha) b_{k(\theta^*)}(j) + (1/2+\alpha) b_{k(\theta^*)}(j)} \right] \quad (11)$$

$$\begin{aligned}
\mathbb{E}[N(j+1)|\boldsymbol{b}(j)] \quad \leq \quad & 1 - \frac{1}{4}(b_{k(\theta^*)-1}(j) + b_{k(\theta^*)}(j) + b_{k(\theta^*)+1}(j) + b_{k(\theta^*)+2}(j)) \\
& - \frac{\alpha}{2(1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j))} \left(\frac{1+2p}{1+2\alpha} - \frac{1-2p}{1-2\alpha} \right) \times \\
& \left[\left(\sum_{i=1}^{k(\theta^*)-2} b_i(j) \right) \left(\sum_{i=1}^{k(\theta^*)+1} b_i(j) \right) - b_{k(\theta^*)-2}(j) \cdot b_{k(\theta^*)-1}(j) + b_{k(\theta^*)-1}(j) \cdot b_{k(\theta^*)+1}(j) + \right. \\
& \left(\sum_{k(\theta^*)+3}^{m} b_i(j) \right) \left(\sum_{k(\theta^*)}^{m} b_i(j) \right) - b_{k(\theta^*)+2}(j) \cdot b_{k(\theta^*)+3}(j) + b_{k(\theta^*)}(j) \cdot b_{k(\theta^*)+2}(j) + \\
& \left. \left(\sum_{i=1}^{k(\theta^*)-1} b_i(j) \right) \left(\sum_{i=1}^{k(\theta^*)} b_i(j) \right) + \left(\sum_{k(\theta^*)+2}^{m} b_i(j) \right) \left(\sum_{k(\theta^*)+1}^{m} b_i(j) \right) \right] + \\
& \frac{1}{4}(b_{k(\theta^*)-1}(j) + b_{k(\theta^*)}(j) + b_{k(\theta^*)+1}(j) + b_{k(\theta^*)+2}(j)) \times \\
& b_{k(\theta^*)}(j) \frac{b_{k(\theta^*)}(j) + b_{k(\theta^*)}(j)}{1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)} \times \\
& \left[(1/2+q) \frac{(1/2+\alpha)(1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)) - 2\alpha \sum_{i=1}^{k(\theta^*)-1} b_i(j)}{(1/2+\alpha) b_{k(\theta^*)}(j) + (1/2-\alpha) b_{k(\theta^*)}(j)} + \right. \\
& \left. (1/2-q) \frac{(1/2-\alpha)(1 - b_{k(\theta^*)}(j) - b_{k(\theta^*)+1}(j)) + 2\alpha \sum_{i=1}^{k(\theta^*)-1} b_i(j)}{(1/2-\alpha) b_{k(\theta^*)}(j) + (1/2+\alpha) b_{k(\theta^*)}(j)} \right] . \quad (12)
\end{aligned}$$

that, for $\kappa \geq 2$ and $\alpha = 0.09 \times p$ we obtain $\mathbb{E}[N(j+1)|\boldsymbol{b}(j)] \leq 1 - (1/2-p)^2/50$. □

REFERENCES

[1] N. Balcan, A. Beygelzimer, and J. Langford. Agostic active learning. available at http://hunch.net/∼jl/projects/agnostic_active/active.pdf.

[2] G. Blanchard and D. Geman. Hierarchical testing designs for pattern recognition. to appear in Annals of Statistics, 2005.

[3] M. V. Burnashev and K. Sh. Zigangirov. An interval estimation problem for controlled observations. *Problems in Information Transmission*, 10:223–231, 1974. (Translated from *Problemy Peredachi Informatsii,* 10(3):51–61, July-September, 1974. Original article submitted June 25, 1973).

[4] R. Castro and R. Nowak. *Foundations and Applications of Sensor Management*, chapter Active Learning and Sampling. Springer-Verlag, 2006. Available at http://homepages.cae.wisc.edu/∼rcastro/active_sensing_chapter.pdf.

[5] R. Castro, R. Willett, and R. Nowak. Faster rates in regression via active learning. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2005. longer Technical Report available at http://homepages.cae.wisc.edu/∼rcastro/ECE-05-3.pdf.

[6] R. M. Castro and R. D. Nowak. Error bounds for active learning. Technical report, University of Wisconsin - Madison, ECE Dept., 2006. (available at http://homepages.cae.wisc.edu/∼rcastro/bounds_active.pdf).

[7] N. Cesa-Bianchi, A. Conconi, and C. Gentile. Learning probabilistic linear-threshold classifiers via selective sampling. In *The Sixteenth Annual Conference on Learning Theory. LNAI 2777, Springer*, 2003.

[8] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, pages 129–145, 1996.

[9] S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing (NIPS)*, 2004.

[10] S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing (NIPS)*, 2005.

[11] S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Eighteen Annual Conference on Learning Theory (COLT)*, 2005.

[12] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, August 1997.

[13] G. Golubev and B. Levit. Sequential recovery of analytic periodic edges in the binary image models. *Mathematical Methods of Statistics*, 12:95–115, 2003.

[14] P. Hall and I. Molchanov. Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. *The Annals of Statistics*, 31(3):921–941, 2003.

[15] M. Horstein. Sequential decoding using noiseless feedback. *IEEE Trans. Info. Theory*, 9(3):136–143, 1963.

[16] Matti Kääriäinen. Active learning in the non-realizable case. Personal communication, 2006.

[17] D. J. C. Mackay. Information-based objective functions for active data selection. *Neural Computation*, 4:698–714, 1991.

[18] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[19] Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Mathématiques et Applications, 41. Springer, 2004.