

Noisy Bayesian Active Learning

Mohammad Naghshvar, Tara Javidi, and Kamalika Chaudhuri

Abstract

We consider the problem of noisy Bayesian active learning, where we are given a finite set of functions \mathcal{H} , a sample space \mathcal{X} , and a label set \mathcal{L} . One of the functions in \mathcal{H} assigns labels to samples in \mathcal{X} . The goal is to identify the function that generates the labels even though the result of a label query on a sample is corrupted by independent noise. More precisely, the objective is to declare one of the functions in \mathcal{H} as the true label generating function with high confidence using as few label queries as possible, by selecting the queries adaptively and in a strategic manner.

Previous work in Bayesian active learning considers Generalized Binary Search, and its variants for the noisy case, and analyzes the number of queries required by these sampling strategies. In this paper, we show that these schemes are, in general, suboptimal. Instead we propose and analyze an alternative strategy for sample collection. Our sampling strategy is motivated by a connection between Bayesian active learning and active hypothesis testing, and is based on querying the label of a sample which maximizes the Extrinsic Jensen–Shannon divergence at each step. We provide upper and lower bounds on the performance of this sampling strategy, and show that these bounds are better than previous bounds.

Index Terms

Bayesian active learning, hypothesis testing, generalized binary search, Extrinsic Jensen–Shannon divergence.

This paper was presented in part at Allerton 2012 and 2013.

This work was done while M. Naghshvar was with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093 USA. He is now with Qualcomm Technologies Inc., San Diego, CA 92121 USA (e-mail: mnaghshvar@qti.qualcomm.com). T. Javidi is with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093 USA (e-mail: tjavidi@ucsd.edu). K. Chaudhuri is with the Department of Computer Science, University of California San Diego, La Jolla, CA 92093 USA. (e-mail: kchaudhuri@ucsd.edu).

The work of M. Naghshvar and T. Javidi was partially supported by the industrial sponsors of UCSD Center for Wireless Communication (CWC), Information Theory and Applications Center (ITA), and Center for Networked Systems (CNS), as well as NSF Grants CCF-0729060 and CCF-1018722. The work of K. Chaudhuri was partially supported by NSF under IIS-1162581.

I. INTRODUCTION

We consider the problem of noisy Bayesian active learning, where we are given a finite set of functions \mathcal{H} , a sample space \mathcal{X} , and a label set \mathcal{L} . One of the functions in \mathcal{H} assigns labels to samples in \mathcal{X} , and our goal is to identify this function when the result of a label query on a sample is corrupted by independent noise. The objective is to declare one of the functions in \mathcal{H} as the true label generating function with high confidence using as few label queries as possible, by selecting the queries adaptively and in a strategic manner.

A special case of the problem, first considered by [1], arises when the label set is binary and the natural sampling strategy for Bayesian active learning becomes closely related to Generalized Binary Search (GBS). In the binary label setting, GBS queries the label of a sample x for which the size of the subsets of functions that label x as $+1$ and -1 respectively, are as balanced as possible. A variant of GBS is Modified Soft-Decision Generalized Binary Search (MSGBS), which was introduced by [1] to address the case when the observed labels may be noisy. [1] analyzes the performance of MSGBS, under a symmetric and non-persistent noise model which flips the labels randomly, and shows that the number of samples required to identify the correct function with probability of error satisfying $\text{Pe} \leq \epsilon$ is $O\left(\frac{\log M + \log \frac{1}{\epsilon}}{\lambda}\right)$, where M is the number of functions in the class \mathcal{H} , and λ is a parameter which depends on the structure of the function class, the sample space, and the noise rate. The first contribution of this paper is to generalize the above problem to the case of general (non-binary) label set with general (and potentially non-symmetric) non-persistent observation noise.

By allowing for the number of samples collected to be determined in a sequential manner (according to a random stopping time as a function of past observations), we draw a parallel between active sequential hypothesis testing and Bayesian active learning. In active sequential hypothesis testing, we are given a set of M hypotheses, and a set of actions; each action, conditioned on the true hypothesis, has a certain probability of yielding an outcome. We observe that Bayesian active learning is a special case of active hypothesis testing, where the hypotheses map to functions, actions map to samples, and the outcomes map to noisy observation of labels. This view of the problem allows for a natural extension of the model of [1] to the non-binary Bayesian active learning setting, where the label noise might be label dependent and asymmetric. Relying on this connection, we derive a universal lower bound on the *expected* number of samples

required to identify the true hypothesis among M with reliability ϵ as a function of noise model parameters. Our lower bound generalizes that of [2]. This lower bound, when specialized for the noisy generalized binary search suggests that the proposed schemes of [1] are suboptimal in general. The next contribution of this work is to propose and analyze an alternative strategy for sample collection.

To find an alternative strategy, we again take advantage of the connection between Bayesian learning and active sequential hypothesis testing. In [3], the authors introduced the notion of Extrinsic Jensen–Shannon (EJS) divergence, and proposed an active sequential hypothesis test that, at each step, selects the action that maximizes the EJS divergence. In this paper, we apply the corresponding sampling strategy to Bayesian active learning, and characterize the performance of this strategy. Our analysis improves on the analysis of [3]. Our bounds show that the number of label queries required by our algorithm is $O\left(\frac{\log M}{\alpha} + \frac{\log \frac{1}{\epsilon}}{\beta}\right)$, where M is the number of functions and α and β are terms, different from λ , that depend on the structure of the function class, the sample space, and the noise model.

To illustrate our bounds, in Section V, we focus on generalized binary search studied in [1] and consider the class of 1-neighborly functions and its three specific subclasses — intervals on the line, thresholds on the line, and a set of rich function classes. We show that the upper bounds on the number of labels required by the EJS policy are superior to those of [1] for all three subclasses for the asymptotic values of ϵ and M . In addition, we show through numerical simulations that our policy has better performance than the algorithms of [1] also in non-asymptotic regimes of practical interest.

There has been a large explosion of recent work on the theory of active learning [4]–[13] but despite the similarity of the titles, the models and the assumptions vary drastically with at times contradictory conclusions. Here we attempt to detail specific attributes of these papers and the connection/disconnect between our work and this literature. Early work on active learning [4], [5], [7] considered the realizable case where the binary labels are produced by a function in a given function class and are observed noise-free. Here, the function class is either finite, like our setting, or, unlike our setting, infinite but equipped with a fixed structure, such as the class of thresholds on a line, or the class of linear classifiers. In contrast with our work, however, the learner is only allowed to query the labels of samples among an unlabeled set of points which are drawn from the unlabeled data distribution. Also unlike ours, the goal here is to find a

function which has low *prediction error with respect to the data distribution*. Thus the challenge is to identify a function in the function class where the disagreement with the true labeling function is less than the required accuracy, and the prediction error occurs due to infiniteness of the function class or due to the indistinguishability of the functions with respect to the data distribution as opposed to noisy observations of the labels.

Since the realizability assumption can hardly ever be justified in practice, more recent literature [6]–[8], [10]–[12] has considered active learning in the non-realizable case. A line of work [7]–[11] considers active learning in the agnostic setting, where the binary labels are not necessarily generated by a function in a given function class, and the goal is to find a function in the function class which has low prediction error with respect to the labeled data distribution. Most of this work employs a *disagreement-based* strategy for label queries; the algorithm maintains a candidate set of functions that is guaranteed to contain the best function in the class with high probability, and queries the label of a sample only when there are two functions in the candidate set that disagree on its label. An important special case of the non-realizable setting relevant to our work is the bounded rate class noise of [6] in which labels are produced by a member of a given function class but are subjected to an exogenous (and non-persistent) observation noise. In such a setting, [6], [14] show that repeat queries can be effectively utilized to mitigate the effect of noise. In [12], the authors perform an information theoretic analysis of active learning in the agnostic setting and provide lower bounds on its sample complexity.

Finally, [13] considers the same setting as our work. Unlike us, they do not provide absolute upper and lower bounds on the query complexity. Instead, they consider sampling strategies that select the sample that maximizes the information gain based on a certain measure of information, and show that if the measure of information in question is adaptively submodular, then this strategy is competitive with the optimal strategy according to the same information measure.

In summary, our work differs from the previous work on active learning in three important ways. First, we are interested in a generalized learning setup where labels can be non-binary and observation noise can have a general non-symmetric and non-discrete nature. Second, we are interested in a sequential learning setting where the learner is allowed not only to query individual samples (hence, rendering the data distribution irrelevant), but also to determine the number of queries in an online fashion as a function of observations so far. Third, by considering

the simpler setup of a finite function class as well as an exogenous and non-persistent observation noise, we provide sharp lower and upper bounds on the query complexity. Our lower bound is purely information theoretic and is only a function of the observation noise which is the only inevitable source of inaccuracy in our model. Our upper bound, in contrast, is obtained via the analysis of an achievable scheme and sheds light on how the structure of the function class impacts the overall performance of our proposed scheme. Perhaps, most significantly, we show that the number of label queries required by the proposed scheme matches the lower bound asymptotically when the function/sample space is sufficiently “rich.”

The remainder of this paper is organized as follows. In Section II, we formulate the problem of Bayesian active learning. In Section III, we propose our heuristic policy for selecting samples. Section IV provides the main results of the paper. As a special case, noisy generalized binary search is discussed in Section V and a comparison to some of the known results is provided. Finally, we conclude the paper and discuss future work in Section VI.

Notation: Let $[x]^+ = \max\{x, 0\}$. For any set \mathcal{S} , $|\mathcal{S}|$ denotes the cardinality of \mathcal{S} . The space of all probability distributions on set \mathcal{A} is denoted by $\mathbb{P}(\mathcal{A})$. All logarithms are in base 2. The entropy function on a vector $\boldsymbol{\rho} = [\rho_1, \rho_2, \dots, \rho_M] \in [0, 1]^M$ is defined as $H(\boldsymbol{\rho}) = \sum_{i=1}^M \rho_i \log \frac{1}{\rho_i}$, with the convention that $0 \log \frac{1}{0} = 0$. Finally, the Kullback–Leibler (KL) divergence between two probability density functions $q(\cdot)$ and $q'(\cdot)$ on space \mathcal{Y} is defined as $D(q\|q') = \int_{\mathcal{Y}} q(y) \log \frac{q(y)}{q'(y)} dy$, with the convention $0 \log \frac{a}{0} = 0$ and $b \log \frac{b}{0} = \infty$ for $a, b \in [0, 1]$ with $b \neq 0$.

II. BAYESIAN ACTIVE LEARNING

In this section, we provide the mathematical description of the problem of Bayesian active learning.

Problem (P) [Bayesian Active Learning]

In the Bayesian active learning problem, we are given a *sample space* \mathcal{X} , a finite label set \mathcal{L} , and an *observation space* \mathcal{Y} . We are also given a set $\mathcal{H} = \{h_1, h_2, \dots, h_M\}$ of M distinct functions, where each $h_i : \mathcal{X} \rightarrow \mathcal{L}$ maps elements in the sample space \mathcal{X} to the label set \mathcal{L} . We assume that one of the functions in \mathcal{H} , denoted by h_θ , produces the correct labeling on \mathcal{X} .

The decision maker is allowed to *query* samples from \mathcal{X} . Querying a sample x generates an observation in $y \in \mathcal{Y}$ whose distribution is a given function of the true label as determined by

the function h_θ . More specifically, if h_θ is the true underlying function and hence $l = h_\theta(x)$ is the true label of sample x , then the result of a query on x is a \mathcal{Y} -valued random variable with probability density $f_l(\cdot)$. We assume that the observation densities $\{f_l(\cdot)\}_{l \in \mathcal{L}}$ are fixed and known, and observations are conditionally independent over time.

The goal of the decision maker is to determine the identity of the function in \mathcal{H} that generates the true labels by an adaptive sequential query of a small number of samples. We assume that the decision maker does not have any extra prior knowledge on the identity of the true function; in other words, it begins with a uniform prior over \mathcal{H} . Let τ be the stopping time at which the decision maker retires and declares the label generating function $h_{\hat{\theta}}$. Furthermore, let $\text{Pe} = P(\hat{\theta} \neq \theta)$ where θ is the index of the true function. In Bayesian active learning, the objective is to design a strategy for the decision maker for querying samples in \mathcal{X} such that, for any given $\epsilon > 0$, we have

$$\text{minimize } \mathbb{E}[\tau] \text{ subject to } \text{Pe} \leq \epsilon. \quad (1)$$

Here the minimization is taken over the choice of the stopping time τ and the learning strategy and the expectation is taken with respect to the observation distribution as well as the Bayesian uniform prior on the true function in \mathcal{H} .

Note that Bayesian learning strategy is more than a single sample query but instead is an adaptive and sequential rule that dictates the causal choice of (random) sample queries depending on the past observations and past queries prior to the stopping time. In this paper, we refer to this adaptive and sequential rule as a query scheme, \mathfrak{c} , which together with the particular realization of outputs $Y_{\mathfrak{c}}(0), Y_{\mathfrak{c}}(1), \dots, Y_{\mathfrak{c}}(\tau - 2)$, dictates the sample queries $X_{\mathfrak{c}}(1), X_{\mathfrak{c}}(2), \dots, X_{\mathfrak{c}}(\tau - 1)$.

Before we end this section, and in face of the difficulty in fully characterizing the optimal learning strategy in general we define weaker notions of optimality.

A. Asymptotic and Order Optimality

Definition 1. Let $\mathbb{E}[\tau_\epsilon^{\mathfrak{c}}]$ denote the expected number of samples required by query scheme \mathfrak{c} to achieve $\text{Pe} \leq \epsilon$. Furthermore, let $\mathbb{E}[\tau_\epsilon^*]$ be the minimum expected number of samples required to achieve $\text{Pe} \leq \epsilon$, where the minimum is taken over all possible strategies. Query scheme \mathfrak{c} is referred to as *asymptotically optimal* in ϵ (and M) if

$$\left(\lim_{M \rightarrow \infty} \right) \lim_{\epsilon \rightarrow 0} \frac{\mathbb{E}[\tau_\epsilon^{\mathfrak{c}}] - \mathbb{E}[\tau_\epsilon^*]}{\mathbb{E}[\tau_\epsilon^{\mathfrak{c}}]} = 0.$$

Query scheme \mathfrak{c} is referred to as *order optimal* in ϵ (and M) if

$$\left(\lim_{M \rightarrow \infty} \right) \lim_{\epsilon \rightarrow 0} \frac{\mathbb{E}[\tau_\epsilon^{\mathfrak{c}}] - \mathbb{E}[\tau_\epsilon^*]}{\mathbb{E}[\tau_\epsilon^{\mathfrak{c}}]} < 1.$$

It is clear from the definitions above that order optimality is weaker than asymptotic optimality. If a scheme \mathfrak{c} is asymptotically optimal in ϵ (and M), then $\mathbb{E}[\tau_\epsilon^{\mathfrak{c}}]$ and $\mathbb{E}[\tau_\epsilon^*]$ will have the same dominating terms in ϵ (and M); while order optimality of scheme \mathfrak{c} only implies that dominating terms in $\mathbb{E}[\tau_\epsilon^{\mathfrak{c}}]$ and $\mathbb{E}[\tau_\epsilon^*]$ are similar up to a constant factor.

III. PRELIMINARIES AND PROPOSED HEURISTIC

After providing some preliminary results and notations, including the definition of Extrinsic Jensen–Shannon (EJS) divergence, in this section we propose our EJS-based heuristic.

Let $\Omega = \{1, 2, \dots, M\}$. Recall that $\theta \in \Omega$ is the random variable that indicates the index of the true function and τ is the stopping time at which the decision maker retires and guesses the true index.

Casting the problem as a decision theoretic problem allows for the structural characterization of the information state, also known as sufficient statistics. Let the decision maker’s posterior belief about each possible function index $i \in \Omega$, updated after each sample query and observation for $t = 0, 1, \dots, \tau - 1$, be

$$\rho_i(t) := P(\{\theta = i\} | X^{t-1}, Y^{t-1}). \quad (2)$$

The decision maker’s posteriors about the true label generating function collectively,

$$\boldsymbol{\rho}(t) := [\rho_1(t), \rho_2(t), \dots, \rho_M(t)], \quad (3)$$

form a sufficient statistics for our Bayesian decision maker. In other words, the selection of sample query as a function of this posterior does not incur any loss of optimality [15]. In particular, the optimal decision maker guesses the function with the highest posterior at time τ to be the label generating function, i.e.,

$$\hat{\theta} = \arg \max_{i \in \Omega} \rho_i(\tau). \quad (4)$$

We also note that the dynamics of the information state, i.e., the posterior, follows Bayes’ rule. But before we make this more precise, let us consider an alternative representation of querying a sample $x \in \mathcal{X}$:

Definition 2. A sample $x \in \mathcal{X}$ generates a $|\mathcal{L}|$ -partition $\Xi^x := \{H_l^x\}_{l \in \mathcal{L}}$ of the function class, i.e., if $H_l^x = \{h \in \mathcal{H} : h(x) = l\}$, then $\mathcal{H} = \cup_{l \in \mathcal{L}} H_l^x$.

This view allows us to characterize the observation density given the belief vector $\boldsymbol{\rho}$ and queried sample x as

$$f_x^\rho(y) := \sum_{i \in \Omega} \rho_i f_{h_i(x)}(y) = \sum_{l \in \mathcal{L}} f_l(y) \sum_{i: h_i \in \mathcal{H}_l^x} \rho_i. \quad (5)$$

Therefore, given the belief vector $\boldsymbol{\rho}(t)$, querying sample x and observing (noisy) label y results in a refinement of the posterior according to the Bayes' rule, i.e.,

$$\boldsymbol{\rho}(t+1) = \Phi^x(\boldsymbol{\rho}(t), y) \quad (6)$$

where

$$\Phi^x(\boldsymbol{\rho}, y) := \left[\rho_1 \frac{f_{h_1(x)}(y)}{f_x^\rho(y)}, \rho_2 \frac{f_{h_2(x)}(y)}{f_x^\rho(y)}, \dots, \rho_M \frac{f_{h_M(x)}(y)}{f_x^\rho(y)} \right]. \quad (7)$$

Many of our results in the paper are obtained as a consequence of a connection between Bayesian active learning and the more general problem of Information Acquisition which has been discussed in full generality in [16]. In particular, taking cue from the seminal work of DeGroot on statistical decision theory [17], and our own prior work on active hypothesis testing [3], given a belief vector $\boldsymbol{\rho} \in \mathbb{P}(\Omega)$, the expected utility of the sample query $x \in \mathcal{X}$, or equivalently its corresponding $|\mathcal{L}|$ -partition $\Xi^x = \{H_l^x\}_{l \in \mathcal{L}}$, can be characterized by its Extrinsic Jensen–Shannon divergence [3]:

$$EJS(\boldsymbol{\rho}, x) := \sum_{l \in \mathcal{L}} \sum_{i: h_i \in \mathcal{H}_l^x} \rho_i D \left(f_l \parallel \frac{f_x^\rho - \rho_i f_l}{1 - \rho_i} \right). \quad (8)$$

We use this to construct our proposed heuristic deterministic Markov sample query strategy.

A. Proposed Heuristic

In this work, we focus on the following (possibly suboptimal) stopping rule. For any given query scheme \mathfrak{c} , querying samples is only stopped when one of the posteriors becomes larger than $1 - \epsilon$, where $\epsilon > 0$ is the desired probability of error:

$$\tilde{\tau}_\epsilon := \min\{t : \max_{i \in \Omega} \rho_i(t) \geq 1 - \epsilon\}. \quad (9)$$

Let $\mathbb{E}[\tau_\epsilon^*]$ and $\mathbb{E}[\tilde{\tau}_\epsilon^*]$ denote the optimal expected number of queries in (1) and the optimal expected number of queries with the (possibly suboptimal) stopping rule as given in (9), respectively. The following fact bounds these quantities both from above and below, and hence will be used in Section IV in bounding the loss of optimality in restricting attention to the above possibly suboptimal stopping rule.

Lemma 1. *Consider stopping times defined earlier with scalars $\iota \geq \epsilon > 0$. We have*

$$\mathbb{E}[\tilde{\tau}_\iota^*] \left(1 - \frac{\epsilon}{\iota}\right) \leq \mathbb{E}[\tau_\epsilon^*] \leq \mathbb{E}[\tilde{\tau}_\epsilon^*]. \quad (10)$$

The proof of Lemma 1 is similar to that of Lemma 3 in [18] and is given in Appendix IV.

We are now ready to fully describe our proposed heuristic.

Definition 3. Policy \mathbf{c}_{EJS} is a stationary deterministic Markov policy with a suboptimal stopping rule defined in (9) which at a given prior belief $\boldsymbol{\rho}$ queries sample $X_{\mathbf{c}_{EJS}} \in \arg \max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x)$.¹

IV. MAIN RESULTS

We now provide the main results – lower and upper bounds on the optimal number of queries to identify the true function with high accuracy. Note that we expect the query complexity of our problem to depend on the characterizations of the discrete memoryless communication channel (DMC) which corrupts the true label’s observations. This is a DMC with input alphabet set \mathcal{L} , output alphabet set \mathcal{Y} , and a collection of conditional probabilities $f_l(\cdot)$, $l \in \mathcal{L}$. We begin with a few assumptions on this channel.

Assumption 1. $C := \min_{g \in \mathbb{P}(\mathcal{Y})} \max_{l \in \mathcal{L}} D(f_l \| g) > 0$.

Assumption 2. $C_1 := \max_{k, l \in \mathcal{L}} D(f_k \| f_l) < \infty$.

Assumption 3. $C_2 := \max_{k, l \in \mathcal{L}} \sup_{y \in \mathcal{Y}} \frac{f_k(y)}{f_l(y)} < \infty$.

Note that C defined above is nothing but the Shannon capacity of the DMC with the collection of conditional probabilities $P(Y = y | L = l) = f_l(y)$, $l \in \mathcal{L}$ (See [19, Theorem 13.1.1]). In

¹Let \mathcal{A} denote the smallest partition of sample space \mathcal{X} , i.e., $\mathcal{X} = \cup_{A \in \mathcal{A}} A$, such that for every $A \in \mathcal{A}$ and $h \in \mathcal{H}$, the value of $h(x)$ remains constant for all $x \in A$. By definition, $EJS(\boldsymbol{\rho}, x) = EJS(\boldsymbol{\rho}, x')$ for every $x, x' \in A$, $A \in \mathcal{A}$. We have $|\mathcal{A}| \leq |\mathcal{L}|^M$, and hence, $\arg \max_{x \in \mathcal{X}}$ is a valid operation in $\arg \max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x)$.

particular, the minimum is achieved by g^* , a convex combination of $\{f_l\}_{l \in \mathcal{L}}$, i.e., $g^* = \sum_{l \in \mathcal{L}} \pi_l^* f_l$ where $\{\pi_l^*\}_{l \in \mathcal{L}}$ is referred to as the *capacity-achieving input distribution* and has the property that for each $k \in \mathcal{L}$, if $\pi_k^* > 0$, then $D(f_k \| g^*) = C$ (See [20, Theorem 4.5.1]). If Assumption 1 does not hold, that is if $C = 0$, the label queries will be completely noisy and no information can be retrieved from the label queries regarding the true function. In this sense, Assumption 1 is a necessary condition that ensures Problem (P) has a meaningful solution.

Parameter C_1 emerges as an important quantity in the problem of variable-length coding with feedback: It denotes the maximum exponential decay rate of the error probability [2]. It is straight forward to show that $C \leq C_1$ and hence, Assumptions 1 and 2 imply that also $C_1 > 0$ and $C < \infty$.

Since, in general, $C_1 \leq \log C_2$, Assumption 2 is redundant with respect to Assumption 3. For observation densities with finite support, i.e., when $|\mathcal{Y}| < \infty$, Assumption 3 ensures that the conditional distributions f_l , $l \in \mathcal{L}$, are absolutely continuous with respect to each other. Thus for observation densities with finite support, Assumption 3 is a necessary and sufficient condition to ensure Assumption 2. On the other hand, for observation kernels with unbounded support, Assumption 3, which is stronger than Assumption 2, is a technical assumption made for notational convenience, and will help us construct strong non-asymptotic bounds in closed form.

While the (non-asymptotic) bounds and analysis in this paper are all obtained under Assumptions 1 and 3, we have chosen to separately state Assumptions 2 and 3 in order to point out that it is possible to relax Assumption 3. More specifically, it is shown in [16] that at the cost of increasing notation, more complicated analysis, and loosening the non-asymptotic bounds, it is possible to relax Assumption 3 and obtain similar asymptotic characterizations only under Assumption 1 and a slightly stronger variant of Assumption 2.

A. Main Results: Lower Bound

In this subsection, we show the following lower bound on the minimum expected number of samples required to achieve $\text{Pe} \leq \epsilon$.

Theorem 1. Consider Problem (P) under Assumptions 1 and 3.

$$\mathbb{E}[\tau_\epsilon^*] \geq \left[\frac{(1 - \frac{3}{\log \frac{4}{\epsilon}} - \frac{\epsilon}{2} \log \frac{1}{\epsilon}) \log M - 2}{C} + \frac{\log \frac{1-\epsilon}{\epsilon} - 2 \log \log \frac{4}{\epsilon} - \log C_2 - 4}{C_1} \right]^+. \quad (11)$$

Theorem 1 is proved in Appendix I using results in dynamic programming. Our lower bound is similar to [21, Theorem 1], [22, Theorem 1], and [23, Theorem 6].

Next we provide upper bounds on the optimal expected sample size of Bayesian active learning.

B. Main Results: Upper Bounds

In this subsection, we characterize upper bounds on the expected number of sample queries in terms of the corresponding Extrinsic Jensen–Shannon (EJS) divergence obtained at each time.

In our presentation of these results, we will need the following notation:

$$\mathbb{P}_\epsilon^M(\Omega) = \left\{ \boldsymbol{\rho} \in \mathbb{P}(\Omega) : \max_{j \in \Omega} \rho_j \geq \tilde{\rho} \right\}, \quad (12)$$

where

$$\tilde{\rho} = 1 - \frac{1}{1 + \max\{\log M, \log \frac{1}{\epsilon}\}}. \quad (13)$$

Theorem 2. Consider Problem (P) under Assumptions 1 and 3. If there exists a positive value α such that at any given belief vector $\boldsymbol{\rho} \in \mathbb{P}(\Omega)$, it is possible to find a sample $x \in \mathcal{X}$ satisfying $EJS(\boldsymbol{\rho}, x) \geq \alpha$, then

$$\mathbb{E}[\tau_\epsilon^*] \leq \frac{\log M + \max\{\log \log M, \log \frac{1}{\epsilon}\} + 4C_2}{\alpha}. \quad (14)$$

Furthermore, if there exists a positive value $\beta > \alpha$ such that for all belief vectors $\boldsymbol{\rho} \in \mathbb{P}_\epsilon^M(\Omega)$, it is possible to find a sample $x \in \mathcal{X}$ satisfying $EJS(\boldsymbol{\rho}, x) \geq \beta$, then the following bound is obtained

$$\mathbb{E}[\tau_\epsilon^*] \leq \frac{\log M + \max\{\log \log M, \log \log \frac{1}{\epsilon}\}}{\alpha} + \frac{\log \frac{1}{\epsilon}}{\beta} + \frac{3(4C_2)^2}{\alpha\beta}. \quad (15)$$

The proof of the above theorem is constructive and is provided in Appendix II. In other words, the policy which selects and queries the label of the sample x for which $EJS(\boldsymbol{\rho}, x) \geq \alpha$, ensures an expected sample size which is smaller than or equal to the right hand side of (14). Now, by construction, policy \mathfrak{c}_{EJS} is such a policy. A similar statement holds for (15).

We remark that as β is the minimum value of $EJS(\boldsymbol{\rho}, x)$ over a subset of belief vectors $\boldsymbol{\rho} \in \mathbb{P}_\epsilon^M(\Omega)$, and α is the minimum value over all belief vectors, $\beta \geq \alpha$, (15) illustrates that we can get significantly better bounds when β is much greater than α .

C. Main Results: Asymptotic and Order Optimality

Note that the lower and upper bounds provided by Theorems 1 and 2 are non-asymptotic and hold for all values of M and ϵ . Nonetheless, they can be applied to establish the asymptotic and order optimality of \mathfrak{c}_{EJS} as defined in Section II-A:

Corollary 1. *The proposed Markov deterministic heuristic policy which maximizes the EJS divergence is order optimal in ϵ and M if there exists scalar $\alpha > 0$ satisfying the first condition of Theorem 2 such that $\alpha \not\rightarrow 0$ as $M \rightarrow \infty$ or $\epsilon \rightarrow 0$. Furthermore, it is asymptotically optimal in ϵ (and M) if β can be selected to be as large as C_1 (and α as large as C).*

However, the above results depend on characterizing non-zero values, if not sufficiently large values, for quantities α and β , which in turn depend on the function class \mathcal{H} and the set of samples that we are allowed to pick from. In the next subsection, we specialize the above results to several function classes in order to concretely illustrate the asymptotic performance of \mathfrak{c}_{EJS} .

D. Applications and Consequences

So far, we have only characterized the performance of \mathfrak{c}_{EJS} in terms of strictly positive scalars α and β , assuming they do exist. An important question remains as whether one can always find such scalars. In this section, we specifically look at an important function class example and provide nontrivial characterization of α and β , hence, demonstrating the relative looseness/tightness of the upper bounds. Furthermore, we discuss the asymptotic and order optimality of these bounds.

We begin with the following definitions which will allow us to generalize the notion of 1-neighborly, first suggested by [1]; then for this general class, we will obtain non-trivial scalars α and β satisfying the conditions of Theorem 2.

Consider the representation of a pair of samples x and x' in terms of their partitioning of the functions:

Definition 4. A pair of samples $x, x' \in \mathcal{X}$ partition the function class \mathcal{H} in an agreement set $A_{x,x'} := \{h \in \mathcal{H} : h(x) = h(x')\}$ and a disagreement set $\Delta_{x,x'} := \{h \in \mathcal{H} : h(x) \neq h(x')\}$.

Definition 5. A class of functions \mathcal{H} is referred to as locally identifiable if for any $h_i \in \mathcal{H}$, there exist samples $x, x' \in \mathcal{X}$ and labels $l, l' \in \mathcal{L}$ such that either of the following be true

- (i) $h_i \in \Delta_{x,x'} \cap H_l^x \cap H_{l'}^{x'}$ and $\mathcal{H} - \{h_i\} = A_{x,x'} \cup (H_l^x \cap H_{l'}^{x'})$, or
(ii) $\{h_i\} = A_{x,x'} \cap H_l^x$ and for all $k \neq l, l'$, $H_k^x \cup H_k^{x'} = \emptyset$.

In essence, the locally identifiable condition implies that for any function $h_i \in \mathcal{H}$, there are (at least) two samples x and x' in \mathcal{X} and two labels l and l' using which h_i can be distinguished from all other functions. As we will see in Section V, local identifiability is a fairly mild condition that is satisfied by a number of natural function classes.

The performance of c_{EJS} when the labeling function class is locally identifiable is characterized by the capacity of the (sub)channel with two inputs $l, l' \in \mathcal{L}$ denoted by $C_{ll'}$, i.e.,

$$C_{ll'} := \min_{g \in \mathbb{P}(\mathcal{Y})} \max\{D(f_l \| g), D(f_{l'} \| g)\}, \quad (16)$$

and consequently

$$C_{\min} := \min_{l, l' \in \mathcal{L}, l \neq l'} \min \left\{ C_{ll'}, D \left(f_{l'} \left\| \frac{1}{2} f_l + \frac{1}{2} f_{l'} \right. \right) \right\}. \quad (17)$$

Proposition 1. *When function class \mathcal{H} is locally identifiable, $\alpha \geq \frac{1}{M} C_{\min}$ and $\beta \geq \tilde{\rho} C_{\min}$. More precisely, for every belief vector ρ , there exists an $x \in \mathcal{X}$ such that*

$$EJS(\rho, x) \geq \begin{cases} \frac{1}{M} C_{\min} & \text{if } \rho \notin \mathbb{P}_\epsilon^M(\Omega) \\ \tilde{\rho} C_{\min} & \text{otherwise} \end{cases}. \quad (18)$$

Proof: To prove Proposition 1, it suffices to show that

$$\max_{x \in \mathcal{X}} EJS(\rho, x) \geq \max_{i \in \Omega} \rho_i C_{\min}.$$

Let $\hat{i} = \arg \max_{i \in \Omega} \rho_i$. By definition of the locally identifiable class, there exist $x_{\hat{i}}, x'_{\hat{i}} \in \mathcal{X}$ and $l, l' \in \mathcal{L}$ such that one of the following conditions holds

$$[h_{\hat{i}}(x_{\hat{i}}), h_{\hat{i}}(x'_{\hat{i}})] = [l, l'] \quad \text{and} \quad [h_j(x_{\hat{i}}), h_j(x'_{\hat{i}})] \in \bigcup_{k \in \mathcal{L}} \{[k, k]\} \cup \{[l', l]\}, \quad \forall j \neq \hat{i}, \quad (19)$$

$$[h_{\hat{i}}(x_{\hat{i}}), h_{\hat{i}}(x'_{\hat{i}})] = [l, l] \quad \text{and} \quad [h_j(x_{\hat{i}}), h_j(x'_{\hat{i}})] \in \{[l, l'], [l', l], [l', l']\}, \quad \forall j \neq \hat{i}. \quad (20)$$

For any $k, k' \in \mathcal{L}$, let

$$\pi_{kk'} := \sum_{j \in \Omega: [h_j(x_{\hat{i}}), h_j(x'_{\hat{i}})] = [k, k']} \frac{\rho_j}{1 - \rho_{\hat{i}}}.$$

Suppose (19) holds. Then

$$\max_{x \in \mathcal{X}} EJS(\rho, x)$$

$$\begin{aligned}
&\geq \max \{EJS(\boldsymbol{\rho}, x_i), EJS(\boldsymbol{\rho}, x'_i)\} \\
&\geq \rho_i \max \left\{ D\left(f_{h_i(x_i)} \parallel \sum_{j \neq i} \frac{\rho_j}{1 - \rho_i} f_{h_j(x_i)}\right), D\left(f_{h_i(x'_i)} \parallel \sum_{j \neq i} \frac{\rho_j}{1 - \rho_i} f_{h_j(x'_i)}\right) \right\} \\
&= \rho_i \max \left\{ D\left(f_l \parallel \sum_{k \in \mathcal{L}} \pi_{kk} f_k + \pi_{l'l} f_{l'}\right), D\left(f_{l'} \parallel \sum_{k \in \mathcal{L}} \pi_{kk} f_k + \pi_{l'l} f_l\right) \right\} \\
&\stackrel{(a)}{\geq} \rho_i \max \left\{ D\left(f_l \parallel \frac{\sum_{k \in \mathcal{L}} \pi_{kk} f_k + \pi_{l'l} f_{l'} + \pi_{l'l} f_l}{1 + \pi_{l'l}}\right), D\left(f_{l'} \parallel \frac{\sum_{k \in \mathcal{L}} \pi_{kk} f_k + \pi_{l'l} f_l + \pi_{l'l} f_{l'}}{1 + \pi_{l'l}}\right) \right\} \\
&\geq \rho_i \min_g \max \{D(f_l \parallel g), D(f_{l'} \parallel g)\} \\
&= \rho_i C_W \\
&\geq \max_{i \in \Omega} \rho_i C_{\min}, \tag{21}
\end{aligned}$$

where (a) follows by Fact 3 in Appendix IV.

On the other hand, if (20) holds, then

$$\begin{aligned}
&\max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x) \\
&\geq \rho_i \max \left\{ D(f_l \parallel \pi_{l'l} f_l + (\pi_{l'l} + \pi_{l'l'}) f_{l'}), D(f_l \parallel \pi_{l'l} f_l + (\pi_{l'l} + \pi_{l'l'}) f_{l'}) \right\} \\
&\stackrel{(a)}{\geq} \rho_i D\left(f_l \parallel \frac{1}{2} f_l + \frac{1}{2} f_{l'}\right) \\
&\geq \max_{i \in \Omega} \rho_i C_{\min}, \tag{22}
\end{aligned}$$

where (a) follows by Fact 3 in Appendix IV and since $\min\{\pi_{l'l}, \pi_{l'l'}\} \leq \frac{1}{2}$.

Combining (21) and (22), we have the assertion of the proposition. \blacksquare

The following corollary provides an upper bound on the expected number of sample queries.

Corollary 2. *Consider Problem (P) under Assumptions 1 and 3. If the function class \mathcal{H} is locally identifiable, then*

$$\mathbb{E}[\tau_\epsilon^*] \leq \frac{M(\log M + \max\{\log \log M, \log \log \frac{1}{\epsilon}\})}{C_{\min}} + \frac{\log \frac{1}{\epsilon}}{\tilde{\rho} C_{\min}} + \frac{3M(4C_2)^2}{\tilde{\rho} C_{\min}^2}. \tag{23}$$

Next, we define a subclass of the locally identifiable function class, and show that for this function class, α and β can be selected to match the denominators in the lower bound in (11).

Hence, the policy \mathfrak{c}_{EJS} is provably asymptotically optimal in ϵ and M .

Definition 6. We call the function class \mathcal{H} $\mathcal{R}(\mathcal{H})$ -sample-rich for $\mathcal{R}(\mathcal{H}) = \cup_{x \in \mathcal{X}} \Xi^x$. In the special case where $\mathcal{R}(\mathcal{H})$ includes all $(|\mathcal{L}|^M - |\mathcal{L}|)$ non-trivial $|\mathcal{L}|$ -partitions of \mathcal{H} , we simply refer to \mathcal{H} as sample-rich.

Proposition 2. When function class \mathcal{H} is sample-rich, $\alpha \geq C$ and $\beta \geq \tilde{\rho}C_1$.

Proof: To prove Proposition 2, we will show that for all belief vectors $\boldsymbol{\rho}$,

$$\max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x) \geq C,$$

and furthermore,

$$\max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x) \geq \max_{i \in \Omega} \rho_i C_1.$$

Recall from Section IV that

$$C = \min_{g \in \mathbb{P}(\mathcal{Y})} \max_{l \in \mathcal{L}} D(f_l \| g), \quad (24)$$

and the minimum is achieved by $g^* = \sum_{l \in \mathcal{L}} \pi_l^* f_l$ where π^* is the capacity achieving input distribution, i.e.,

$$D\left(f_k \left\| \sum_{l \in \mathcal{L}} \pi_l^* f_l\right.\right) = C \quad \text{for any } k \in \mathcal{L} \text{ such that } \pi_k^* > 0. \quad (25)$$

By definition of the sample-rich function class, for each $\mathbf{v} := [v_1, \dots, v_M] \in \mathcal{L}^M$, there exists a sample in \mathcal{X} , say $x_{\mathbf{v}}$, that satisfies $\mathbf{h}(x_{\mathbf{v}}) = \mathbf{v}$, where $\mathbf{h}(x) := [h_1(x), h_2(x), \dots, h_M(x)]$. Let

$$\lambda_{\mathbf{v}}^* = \prod_{i=1}^M \pi_{v_i}^*.$$

Note that $\sum_{\mathbf{v} \in \mathcal{L}^M} \lambda_{\mathbf{v}}^* = 1$. Moreover, for any $i, j \in \Omega$, $i \neq j$,

$$\sum_{\mathbf{v} \in \mathcal{L}^M : v_i = k} \lambda_{\mathbf{v}}^* = \pi_k^*, \quad \sum_{\mathbf{v} \in \mathcal{L}^M : v_i = k, v_j = l} \lambda_{\mathbf{v}}^* = \pi_k^* \pi_l^*.$$

Using weights $\{\lambda_{\mathbf{v}}^*\}_{\mathbf{v} \in \mathcal{L}^M}$ and taking average over all $\mathbf{v} \in \mathcal{L}^M$, we obtain

$$\begin{aligned} \max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x) &\geq \sum_{\mathbf{v}} \lambda_{\mathbf{v}}^* EJS(\boldsymbol{\rho}, x_{\mathbf{v}}) \\ &= \sum_{\mathbf{v}} \lambda_{\mathbf{v}}^* \sum_{i=1}^M \rho_i D\left(f_{h_i(x_{\mathbf{v}})} \left\| \sum_{j \neq i} \frac{\rho_j}{1 - \rho_i} f_{h_j(x_{\mathbf{v}})}\right.\right) \\ &= \sum_{i=1}^M \rho_i \sum_{k \in \mathcal{L}} \pi_k^* \sum_{\mathbf{v} : v_i = k} \frac{\lambda_{\mathbf{v}}^*}{\pi_k^*} D\left(f_k \left\| \sum_{j \neq i} \frac{\rho_j}{1 - \rho_i} f_{v_j}\right.\right) \end{aligned}$$

$$\begin{aligned}
&\stackrel{(a)}{\geq} \sum_{i=1}^M \rho_i \sum_{k \in \mathcal{L}} \pi_k^* D\left(f_k \parallel \sum_{j \neq i} \frac{\rho_j}{1 - \rho_i} \sum_{\mathbf{v}: v_i=k} \frac{\lambda_{\mathbf{v}}^*}{\pi_k^*} f_{v_j}\right) \\
&= \sum_{i=1}^M \rho_i \sum_{k \in \mathcal{L}} \pi_k^* D\left(f_k \parallel \sum_{j \neq i} \frac{\rho_j}{1 - \rho_i} \sum_{l \in \mathcal{L}} \sum_{\mathbf{v}: v_i=k, v_j=l} \frac{\lambda_{\mathbf{v}}^*}{\pi_k^*} f_l\right) \\
&= \sum_{i=1}^M \rho_i \sum_{k \in \mathcal{L}} \pi_k^* D\left(f_k \parallel \sum_{l \in \mathcal{L}} \pi_l^* f_l\right) \\
&\stackrel{(b)}{=} \sum_{i=1}^M \rho_i C \\
&= C,
\end{aligned}$$

where (a) follows from Jensen's inequality and (b) follows from (25).

Let $\hat{i} = \arg \max_{i \in \Omega} \rho_i$. Let $k, l \in \mathcal{L}$ be the labels satisfying $D(f_k \parallel f_l) = C_1$. By definition of the sample-rich function class, there exists a sample $x_{\hat{i}} \in \mathcal{X}$ that satisfies $h_{\hat{i}}(x_{\hat{i}}) = k$ and $h_j(x_{\hat{i}}) = l$ for all $j \neq \hat{i}$. We have

$$\max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x) \geq EJS(\boldsymbol{\rho}, x_{\hat{i}}) \geq \rho_{\hat{i}} D\left(f_{h_{\hat{i}}(x_{\hat{i}})} \parallel \sum_{j \neq \hat{i}} \frac{\rho_j}{1 - \rho_{\hat{i}}} f_{h_j(x_{\hat{i}})}\right) = \max_{i \in \Omega} \rho_i C_1.$$

■

As a simple corollary,

Corollary 3. *Consider Problem (P) under Assumptions 1 and 3. If the function class \mathcal{H} is sample-rich,*

$$\mathbb{E}[\tau_{\epsilon}^*] \leq \frac{\log M + \max\{\log \log M, \log \log \frac{1}{\epsilon}\}}{C} + \frac{\log \frac{1}{\epsilon}}{\tilde{\rho} C_1} + \frac{48C_2^2}{\tilde{\rho} C C_1}. \quad (26)$$

The above results show that for sample-rich function classes, \mathfrak{c}_{EJS} is asymptotically optimal in both ϵ and M .

The above results generalize the finding of [1] to a multi-label Bayesian learning with non-binary and asymmetric noise case. However, to make this comparison precise, we will dedicate the next section to specialize our general results above to the noisy generalized binary search of [1].

V. SPECIAL CASE: NOISY GENERALIZED BINARY SEARCH

We next compare our work with existing results. Since the only study of similar nature is that of noisy generalized binary search [1], we consider an application of our main results to noisy

generalized binary search among 1-neighborly functions, first introduced in [1]. This is a special case of our problem where functions are binary-valued, i.e., $\mathcal{L} = \{-1, +1\}$, the observation space $\mathcal{Y} = \{-1, +1\}$, and observation densities are of the following form:

$$f_l(y) = \begin{cases} 1-p & \text{if } y = l \\ p & \text{if } y = -l \end{cases},$$

for some $p \in (0, 1/2)$. In other words, for any sample x , if h_i is the true function, then the label $h_i(x)$ is observed through a binary symmetric channel with crossover probability p .

For the case of noisy generalized binary search, C , C_1 , and C_2 defined in Section IV can be further simplified to

$$\begin{aligned} C &:= 1 + p \log p + (1-p) \log(1-p), \\ C_1 &:= p \log \frac{p}{1-p} + (1-p) \log \frac{1-p}{p}, \\ C_2 &:= \frac{1-p}{p}. \end{aligned}$$

In order to emphasize the dependence of C , C_1 , and C_2 on the Bernoulli parameter p (corresponding to the observation noise), we denote them by $C(p)$, $C_1(p)$, and $C_2(p)$ respectively. Note that from Jensen's inequality, $C_1(p) \geq 2C(p)$.

Next we define a class of 1-neighborly functions first defined in [1, Definition 2].

Definition 7. A class of binary-valued functions \mathcal{H} is referred to as 1-neighborly if for any $h_i \in \mathcal{H}$, there exist $x, x' \in \mathcal{X}$ such that

$$\begin{cases} h_i(x) \neq h_i(x') \\ h_j(x) = h_j(x') \quad \text{if } j \neq i \text{ and } h_j(\cdot) \neq -h_i(\cdot) \end{cases}.$$

It is simple to see that the class of 1-neighborly functions is a subset of binary-valued locally identifiable function class. This implies the following baseline bound:

Corollary 4. When function class \mathcal{H} is 1-neighborly, we have $\alpha \geq \frac{1}{M}C(p)$ and $\beta \geq \tilde{\rho}C(p)$.

In comparison, [1] provides two sample query strategies, NGBS and MSGBS, whose performance (upper bound) depends strongly on the properties of the function class at hand.

Let n_0 denote the number of queries made by GBS to determine h_θ in the noiseless setting. The number of queries required by NGBS to attain $\text{Pe} \leq \epsilon$ is upper bounded by

$$\frac{n_0(\log n_0 + \log \frac{1}{\epsilon})}{(\frac{1}{2} - p)^2}. \quad (27)$$

Let \mathcal{A} denote the smallest partition of sample space \mathcal{X} , i.e., $\mathcal{X} = \cup_{A \in \mathcal{A}} A$, such that for every $A \in \mathcal{A}$ and $h \in \mathcal{H}$, the value of $h(x)$ is constant for all $x \in A$; and denote this value by $h(A)$. Furthermore, let

$$c^* := \min_{P \in \mathbb{P}(\mathcal{A})} \max_{h \in \mathcal{H}} \left| \sum_{A \in \mathcal{A}} h(A) P(A) \right|. \quad (28)$$

Under MSGBS, the number of queries required to ensure that $\text{Pe} \leq \epsilon$ is upper bounded by

$$\frac{\log M + \log \frac{1}{\epsilon}}{\min\{2(1 - c^*), 1\} \lambda(p)}, \quad (29)$$

where

$$\lambda(p) := \max_{p' \in (p, 1/2)} \frac{1}{4} \left(1 - \frac{p'(1-p)}{1-p'} - \frac{(1-p')p}{p'} \right). \quad (30)$$

Note that c^* (as well as n_0) in general depends on the function class \mathcal{H} . Since this dependence is implicit and hard to characterize in closed form for general function class \mathcal{H} , a direct comparison between (29) (or (27)) and (23) is not possible. As a result, next we focus on special cases of function classes studied in [1] for which a precise characterization of the achievable upper bound is available. Consequently, we next define two important subclasses of 1-neighborly binary-valued functions: 1) Disjoint class \mathcal{H}_D ; 2) Threshold class \mathcal{H}_T . We further specialize the choices of α and β for these classes.

Definition 8. Let \mathbf{e}_i , $i \in \Omega$, represent a vector of size M whose i^{th} element is $+1$ and all other elements are -1 . A collection of functions \mathcal{H} is referred to as *disjoint interval class* if $\cup_{x \in \mathcal{X}} \{\mathbf{h}(x)\} = \cup_{i \in \Omega} \{\mathbf{e}_i\} \subset \{-1, +1\}^M$, where $\mathbf{h}(x) := [h_1(x), h_2(x), \dots, h_M(x)]$. In other words, for any sample $x \in \mathcal{X}$, only one function in \mathcal{H} takes value $+1$ and all other functions take value -1 .

Definition 9. Let \mathbf{u}_i , $i \in \Omega$, represent a vector of size M whose first i elements are -1 and all other elements are $+1$. A collection of functions \mathcal{H} is referred to as *threshold class* if $\cup_{x \in \mathcal{X}} \{\mathbf{h}(x)\} = \cup_{i \in \Omega} \{\mathbf{u}_i\} \subset \{-1, +1\}^M$.

Fact 1 (see [24]). *For the disjoint interval class \mathcal{H}_D , $n_0 \leq M$ and $c^* = 1 - \frac{2}{M}$. For the threshold function class \mathcal{H}_T , $n_0 \leq \log M$ and $c^* = 0$. For the sample-rich function class \mathcal{H}_R , $n_0 \leq \log M$ and $c^* = 0$.*

We are now ready to contrast these results with our findings. In particular, we have

Proposition 3. *For the disjoint interval class \mathcal{H}_D , $\alpha \geq \frac{1}{M}C_1(p)$ and $\beta \geq \tilde{\rho}C_1(p)$. For the threshold function class \mathcal{H}_T , $\alpha \geq C(p)$ and $\beta \geq C(p)$. For the sample-rich function class \mathcal{H}_R , $\alpha \geq C(p)$ and $\beta \geq \tilde{\rho}C_1(p)$.*

The proof of Proposition 3 is provided in Appendix III-A.

Table I summarizes our results and specializes the upper bounds in [24] and lists the number of samples required by the policies NGBS, MSGBS, and \mathfrak{c}_{EJS} to attain $\text{Pe} \leq \epsilon$. Furthermore, these bounds together with (52) establish asymptotic and order optimality of \mathfrak{c}_{EJS} .²

Recall that policies NGBS and MSGBS are non-sequential in the sense that they stop after a fixed number of samples, regardless of the probability of error. The numbers shown in Table I are the number of samples that these policies require to achieve $\text{Pe} \leq \epsilon$. Policy \mathfrak{c}_{EJS} is sequential and Table I shows the expected number of samples required by this policy to achieve $\text{Pe} \leq \epsilon$.

TABLE I
PERFORMANCE COMPARISON OF NGBS, MSGBS, AND \mathfrak{c}_{EJS} ON DIFFERENT FUNCTION CLASSES.

Function class	NGBS	MSGBS	\mathfrak{c}_{EJS}
Disjoint \mathcal{H}_D	$\frac{M(\log M + \log \frac{1}{\epsilon})}{(\frac{1}{2}-p)^2}$	$\frac{M(\log M + \log \frac{1}{\epsilon})}{4\lambda(p)}$	$\left(\frac{M \log M}{C_1(p)} + \frac{\log \frac{1}{\epsilon}}{C_1(p)}\right) (1 + o(1))$
	order optimal in ϵ	order optimal in ϵ	asymptotic optimal in ϵ
Threshold \mathcal{H}_T	$\frac{\log M(\log \log M + \log \frac{1}{\epsilon})}{(\frac{1}{2}-p)^2}$	$\frac{\log M + \log \frac{1}{\epsilon}}{\lambda(p)}$	$\left(\frac{\log M}{C(p)} + \frac{\log \frac{1}{\epsilon}}{C(p)}\right) (1 + o(1))$
	order optimal in ϵ	order optimal in ϵ, M	order optimal in ϵ, M
Sample-rich \mathcal{H}_R	$\frac{\log M(\log \log M + \log \frac{1}{\epsilon})}{(\frac{1}{2}-p)^2}$	$\frac{\log M + \log \frac{1}{\epsilon}}{\lambda(p)}$	$\left(\frac{\log M}{C(p)} + \frac{\log \frac{1}{\epsilon}}{C_1(p)}\right) (1 + o(1))$
	order optimal in ϵ	order optimal in ϵ, M	asymptotic optimal in ϵ, M

To provide a comparison between the obtained bounds, in asymptotic regime, Fig. 1 compares the denominators of the upper bounds given in Table I. Note that our upper bound provides

²The term $o(1)$ goes to zero as $\epsilon \rightarrow 0$ or $M \rightarrow \infty$. See Appendix III-B for more details.

improvement over those corresponding to NGBS and MSGBS. Particularly, the gap between the bounds is very significant for small values of the Bernoulli parameter p and for large values of $\frac{1}{\epsilon}$ and M .

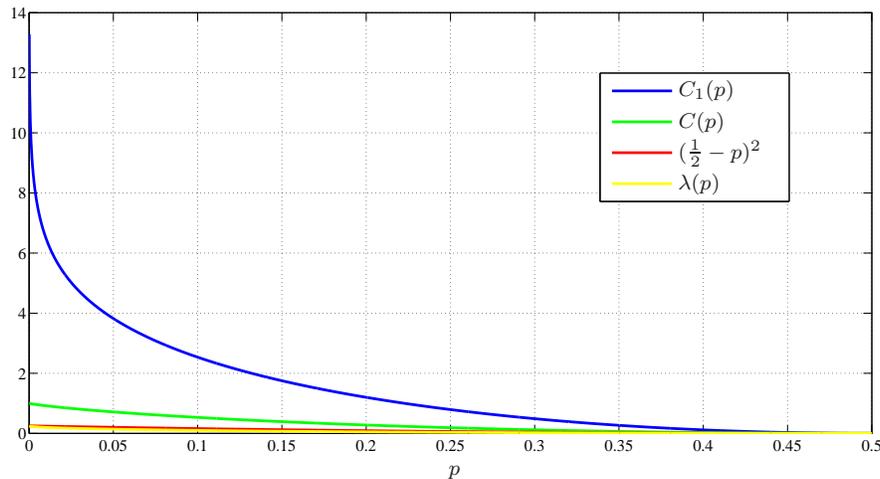


Fig. 1. Comparison of $C(p)$, $C_1(p)$, $(\frac{1}{2} - p)^2$, and $\lambda(p)$, for $p \in (0, 1/2)$.

Remark. With no tight lower bound on the performance of NGBS and MSGBS, the above comparison must not be confused with a comparative analysis between c_{EJS} versus NGBS and MSGBS. In fact, the gap between the above upper bounds could potentially be due to the analysis limitation in [24] of these algorithms rather than their performance.

Next, policies c_{EJS} and MSGBS are compared numerically for the problem of noisy generalized binary search with parameter p and a rich function class of size M (we do not consider NGBS since it is outperformed by MSGBS). This numerical study not only sheds light on non-asymptotic performance of both policies but also provides a direct comparison between the performance of these policies (as opposed to a comparison between the upper bounds on the performance of these policies given in Table I).

In order to have a fair comparison, the candidate policies are compared in both sequential and non-sequential scenarios. In the sequential scenario, the policies stop as soon as the belief about one of the functions passes a threshold $1 - \epsilon$, and the expected number of queries is considered as a measure of performance; while in the non-sequential scenario, the policies are

compared based on their average probability of making a wrong declaration after N number of label queries.

Figs. 2 and 3 show the performance of \mathfrak{c}_{EJS} and MSGBS for the sequential scenario while Figs. 4 and 5 compare their performance for the non-sequential scenario.

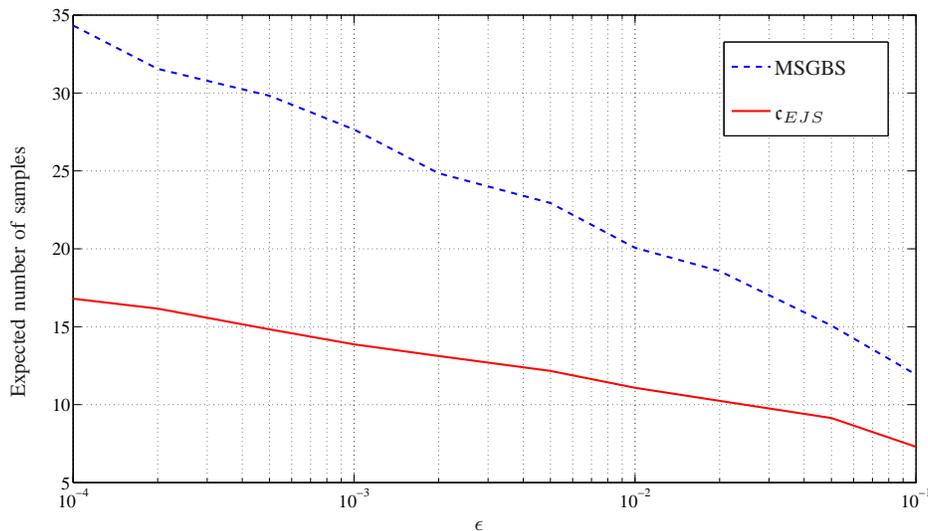


Fig. 2. Sequential noisy generalized binary search with parameter $p = 0.2$, desired probability of error ϵ , and a rich function class of size $M = 5$. The expected number of samples is plotted as ϵ varies.

The figures show the superior performance of \mathfrak{c}_{EJS} over MSGBS in both scenarios and for different values of ϵ , N , and M .

VI. DISCUSSION AND FUTURE WORK

In this paper, we consider the problem of noisy Bayesian active learning. In this setting, we propose a heuristic policy for querying the labels of samples using Extrinsic Jensen–Shannon divergence, and provide upper bounds on its performance. In addition, we provide information-theoretic lower bounds on the query complexity of any sampling strategy. Comparison to the state-of-the-art [24] shows that our sampling strategy achieves superior performance for several natural function classes.

Our lower and upper bounds reveal that Bayesian active learning in the presence of noise is a two-phase problem, where the lengths of the phases correspond to the two terms in Theorems 1

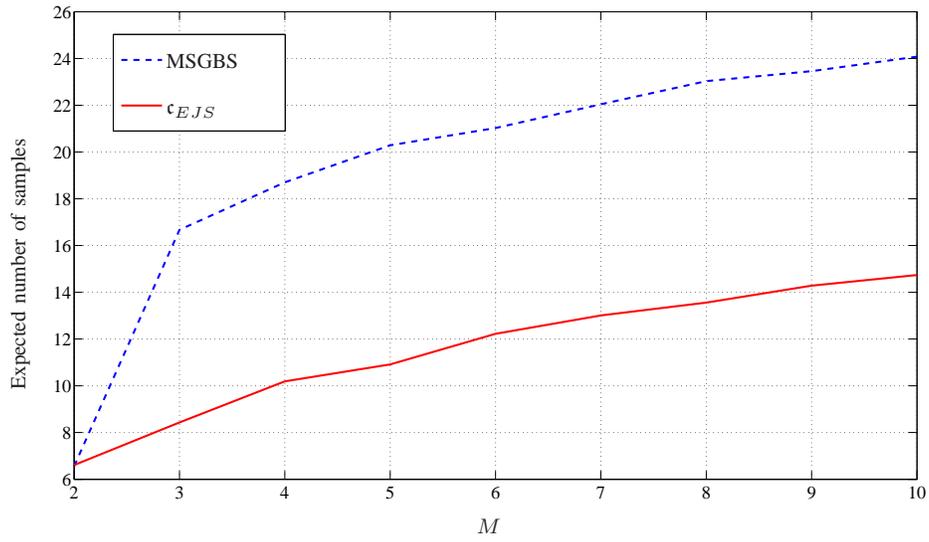


Fig. 3. Sequential noisy generalized binary search with parameter $p = 0.2$, desired probability of error $\epsilon = 0.01$, and a rich function class of size M . The expected number of samples is plotted as M varies.

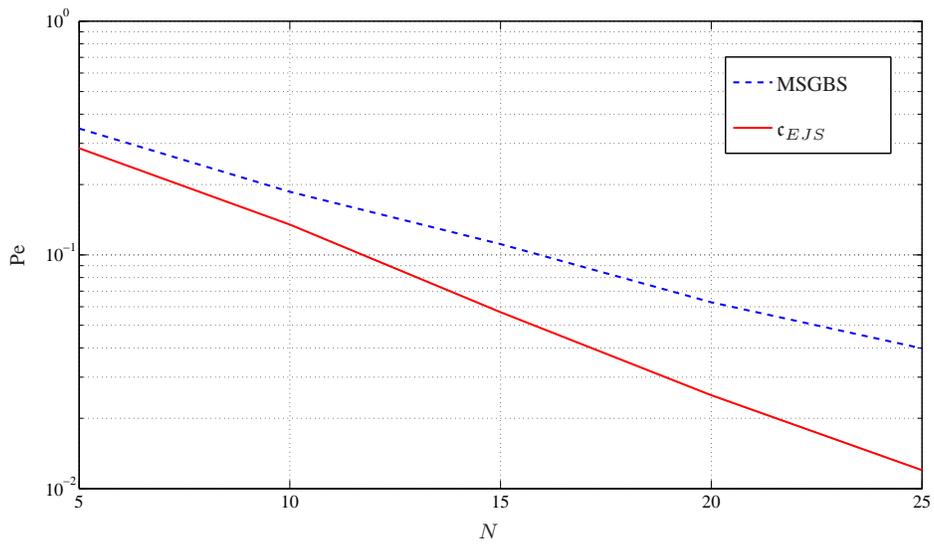


Fig. 4. Non-sequential noisy generalized binary search with parameter $p = 0.2$, total number of label queries N , and a rich function class of size $M = 5$. The average probability of error is plotted as N varies.

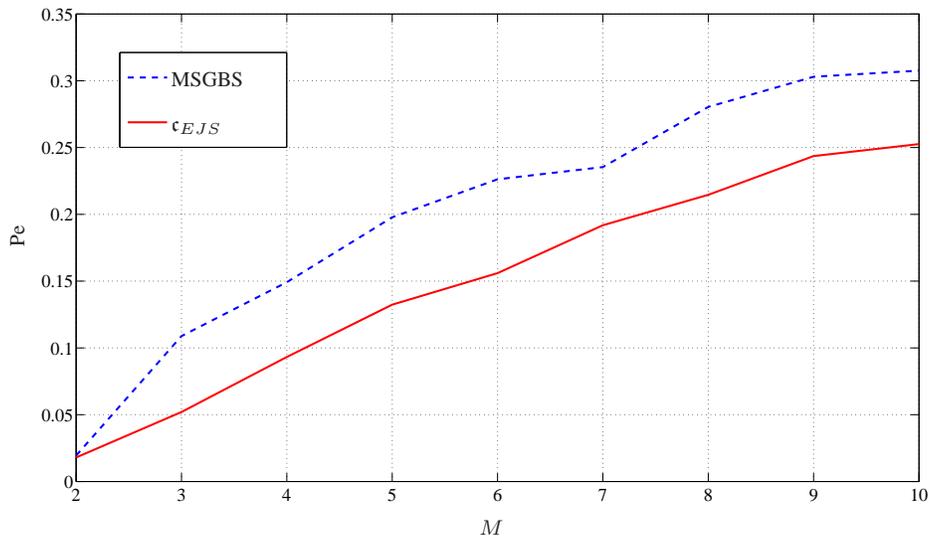


Fig. 5. Non-sequential noisy generalized binary search with parameter $p = 0.2$, total number of label queries $N = 10$, and a rich function class of size M . The average probability of error is plotted as M varies.

and 2. The first phase corresponds to a *search* among the M functions in the class, and the second phase corresponds to a testing phase where we seek to increase our confidence in the result. An important direction of future research is to extend our algorithms to more general function classes such as linear classifiers and to establish its connection to other notions used to measure the query complexity of active learning such as Alexander's capacity [7], [9], [12] and the splitting index [5].

APPENDIX I

PROOF OF THEOREM 1

From Lemma 1, we have

$$\mathbb{E}[\tau_\epsilon^*] \geq \mathbb{E}[\tilde{\tau}_\iota^*] \left(1 - \frac{\epsilon}{\iota}\right). \quad (31)$$

Let $V_\iota^* : \mathbb{P}(\Omega) \rightarrow \mathbb{R}_+$ be the solution to the following fixed point equation:

$$V_\iota(\boldsymbol{\rho}) = \begin{cases} 0 & \text{if } \max_{j \in \Omega} \rho_j \geq 1 - \iota \\ 1 + \min_{x \in \mathcal{X}} \mathbb{E}[V_\iota(\Phi^x(\boldsymbol{\rho}, Y))], & \text{otherwise} \end{cases}$$

where Φ^x , $x \in \mathcal{X}$, is the Bayes operator defined in (7).

It follows from Propositions 9.8 and 9.10 in [25] that

$$\mathbb{E}[\tilde{\tau}_\iota^*] = V_\iota^*([1/M, \dots, 1/M]). \quad (32)$$

The assertion of the Theorem follows from (31), (32), and Lemma 2 at the end of this section, and by setting $\iota = \frac{\epsilon}{2} \log \frac{4}{\epsilon}$ and $\delta = \frac{1}{\log \frac{4}{\epsilon}}$, as shown below.

$$\begin{aligned} \mathbb{E}[\mathcal{T}_\epsilon^*] &\geq \left(1 - \frac{2}{\log \frac{4}{\epsilon}}\right) \left[\frac{\left(1 - \frac{1}{\log \frac{4}{\epsilon}} - \frac{\epsilon}{2} \log \frac{4}{\epsilon}\right) \log M - 2}{C} + \frac{\log \frac{1 - \frac{\epsilon}{2} \log \frac{4}{\epsilon}}{\frac{\epsilon}{2} \log \frac{4}{\epsilon}} - \log \log \frac{2}{\epsilon} - \log C_2 - 1}{C_1} \right]^+ \\ &\geq \left[\frac{\left(1 - \frac{2}{\log \frac{4}{\epsilon}}\right) \left(1 - \frac{\epsilon}{2} \log \frac{4}{\epsilon}\right) \log M - \frac{\log M}{\log \frac{4}{\epsilon}} - 2}{C} \right. \\ &\quad \left. + \frac{\left(1 - \frac{2}{\log \frac{4}{\epsilon}}\right) \log \frac{1}{\frac{\epsilon}{2} \log \frac{4}{\epsilon}} - \log \frac{1}{1 - \frac{\epsilon}{2} \log \frac{4}{\epsilon}} - \log \log \frac{2}{\epsilon} - \log C_2 - 1}{C_1} \right]^+ \\ &\geq \left[\frac{\left(1 - \frac{2}{\log \frac{4}{\epsilon}} - \frac{\epsilon}{2} \log \frac{1}{\epsilon}\right) \log M - \frac{\log M}{\log \frac{4}{\epsilon}} - 2}{C} \right. \\ &\quad \left. + \frac{\log \frac{1-\epsilon}{\epsilon} - \log \log \frac{4}{\epsilon} - 1 - \log \frac{1-\epsilon}{1 - \frac{\epsilon}{2} \log \frac{4}{\epsilon}} - \log \log \frac{2}{\epsilon} - \log C_2 - 1}{C_1} \right]^+ \\ &\geq \left[\frac{\left(1 - \frac{3}{\log \frac{4}{\epsilon}} - \frac{\epsilon}{2} \log \frac{1}{\epsilon}\right) \log M - 2}{C} + \frac{\log \frac{1-\epsilon}{\epsilon} - 2 \log \log \frac{4}{\epsilon} - \log C_2 - 4}{C_1} \right]^+. \end{aligned} \quad (33)$$

Lemma 2. *At any information state $\rho \in \mathbb{P}(\Omega)$ and for any $\iota \in (0, 1)$ and $\delta \in (0, 1/2)$,*

$$V_\iota^*(\rho) \geq \left[\frac{H(\rho) - F_M(\delta) - F_M(\iota)}{C} + \frac{\log \frac{1-\iota}{\iota} - \log \frac{1-\delta}{\delta} - \log C_2 - 1}{C_1} \mathbf{1}_{\{\max_{i \in \Omega} \rho_i \leq 1-\delta\}} \right]^+ \quad (34)$$

where $F_M(z) := H([z, 1-z]) + z \log(M-1)$ for $0 \leq z \leq 1$.

Proof: The proof of Lemma 2 follows closely the proof of Lemma 1 and Theorem 2 in [16] and is provided next.

First we will use the following technical lemma, proved in Appendix IV.

Lemma 3. *Any functional $V : \mathbb{P}(\Omega) \rightarrow \mathbb{R}_+$ that satisfies the following:*

$$V(\rho) \leq \begin{cases} 0 & \text{if } \max_{j \in \Omega} \rho_j \geq 1 - \iota \\ 1 + \min_{x \in \mathcal{X}} \mathbb{E}[V(\Phi^x(\rho, Y))] & \text{otherwise} \end{cases}, \quad (35)$$

provides a uniform lower bound for the optimal value function V_ι^* .

Next we define $J(\boldsymbol{\rho}) = \max\{J'(\boldsymbol{\rho}), J''(\boldsymbol{\rho})\}$ where

$$J'(\boldsymbol{\rho}) := \left[\frac{-F_M(\iota)}{C} + \sum_{i=1}^M \rho_i \frac{\log \frac{1-\iota}{\iota} - \log \frac{\rho_i}{1-\rho_i} - 1}{C_1} \right]^+, \quad (36)$$

and J'' is the right-hand side of (34), i.e.,

$$J''(\boldsymbol{\rho}) := \left[\frac{H(\boldsymbol{\rho}) - F_M(\delta) - F_M(\iota)}{C} + \frac{\log \frac{1-\iota}{\iota} - \log \frac{1-\delta}{\delta} - \log C_2 - 1}{C_1} \mathbf{1}_{\{\max_{i \in \Omega} \rho_i \leq 1-\delta\}} \right]^+.$$

We show that J satisfies (35) and hence, $V_\iota^* \geq J = \max\{J', J''\} \geq J''$.

We use Jensen's inequality to show that

$$J'(\boldsymbol{\rho}) \leq 1 + \min_{x \in \mathcal{X}} \mathbb{E}[J'(\Phi^x(\boldsymbol{\rho}, Y))], \quad \forall \boldsymbol{\rho} \in \mathbb{P}(\Omega). \quad (37)$$

For any $\boldsymbol{\rho}$ such that $J'(\boldsymbol{\rho}) = 0$, inequality (37) holds trivially. For any $\boldsymbol{\rho}$ such that $J'(\boldsymbol{\rho}) > 0$ and for any $x \in \mathcal{X}$, we have

$$\begin{aligned} \mathbb{E}[J'(\Phi^x(\boldsymbol{\rho}, Y))] &\geq \frac{-F_M(\iota)}{C} + \sum_{i=1}^M \int \rho_i f_{h_i(x)}(y) \frac{\log \frac{1-\iota}{\iota} - \log \frac{\rho_i f_{h_i(x)}(y)}{\sum_{j \neq i} \rho_j f_{h_j(x)}(y)} - 1}{C_1} dy \\ &= J'(\boldsymbol{\rho}) - \sum_{i=1}^M \rho_i \frac{\int f_{h_i(x)}(y) \log \frac{f_{h_i(x)}(y)}{\sum_{j \neq i} \frac{\rho_j}{1-\rho_j} f_{h_j(x)}(y)} dy}{C_1} \\ &\geq J'(\boldsymbol{\rho}) - \sum_{i=1}^M \rho_i \frac{\sum_{j \neq i} \frac{\rho_j}{1-\rho_j} D(f_{h_i(x)} \| f_{h_j(x)})}{C_1} \\ &\geq J'(\boldsymbol{\rho}) - 1. \end{aligned}$$

For all $\boldsymbol{\rho}$ satisfying $\max_{i \in \Omega} \rho_i > 1 - \delta$,

$$H(\boldsymbol{\rho}) < (1 - \delta) \log \frac{1}{1 - \delta} + (M - 1) \times \frac{\delta}{M - 1} \log \frac{1}{\delta / (M - 1)} = F_M(\delta),$$

hence, $J'' = 0$. In other words, $J(\boldsymbol{\rho}) = J'(\boldsymbol{\rho}) > 0$ implies that $\max_{i \in \Omega} \rho_i \leq 1 - \delta$.

Let $\hat{\boldsymbol{\rho}} = \Phi^x(\boldsymbol{\rho}, y)$. If $\max_{i \in \Omega} \hat{\rho}_i \leq 1 - \delta$, then

$$J(\hat{\boldsymbol{\rho}}) \geq J''(\hat{\boldsymbol{\rho}}) \geq \frac{H(\hat{\boldsymbol{\rho}}) - F_M(\delta) - F_M(\iota)}{C} + \frac{\log \frac{1-\iota}{\iota} - \log \frac{1-\delta}{\delta} - \log C_2 - 1}{C_1}. \quad (38)$$

On the other hand, if $\max_{i \in \Omega} \hat{\rho}_i > 1 - \delta$, we get

$$J(\hat{\boldsymbol{\rho}}) = J'(\hat{\boldsymbol{\rho}})$$

$$\begin{aligned}
&= \left[\frac{-F_M(\iota)}{C} + \sum_{i=1}^M \hat{\rho}_i \frac{\log \frac{1-\iota}{\iota} - \log \frac{\hat{\rho}_i}{1-\hat{\rho}_i} - 1}{C_1} \right]^+ \\
&\stackrel{(a)}{\geq} \left[\frac{-F_M(\iota)}{C} + \sum_{i=1}^M \hat{\rho}_i \frac{\log \frac{1-\iota}{\iota} - \log \frac{1-\delta}{\delta} - \log C_2 - 1}{C_1} \right]^+ \\
&\geq \frac{-F_M(\iota)}{C} + \frac{\log \frac{1-\iota}{\iota} - \log \frac{1-\delta}{\delta} - \log C_2 - 1}{C_1}, \tag{39}
\end{aligned}$$

where (a) follows from the fact that under Assumption 3 and for all $i \in \Omega$,

$$\begin{aligned}
\log \frac{\hat{\rho}_i}{1-\hat{\rho}_i} &\leq \left| \log \frac{\hat{\rho}_i}{1-\hat{\rho}_i} - \log \frac{\rho_i}{1-\rho_i} \right| + \left| \log \frac{\rho_i}{1-\rho_i} \right| \\
&\leq \left| \log \frac{\rho_i f_{h_i(x)}(y)}{\sum_{j \neq i} \rho_j f_{h_j(x)}(y)} - \log \frac{\rho_i}{1-\rho_i} \right| + \log \frac{1-\delta}{\delta} \\
&= \left| \log \frac{f_{h_i(x)}(y)}{\sum_{j \neq i} \frac{\rho_j}{1-\rho_i} f_{h_j(x)}(y)} \right| + \log \frac{1-\delta}{\delta} \\
&\leq \log C_2 + \log \frac{1-\delta}{\delta}.
\end{aligned}$$

From the above facts, we obtain:

- **Case 1:** For all $\boldsymbol{\rho}$ such that $J(\boldsymbol{\rho}) = 0$ or $J(\boldsymbol{\rho}) = J'(\boldsymbol{\rho})$, it is trivial from (37) that

$$J(\boldsymbol{\rho}) = J'(\boldsymbol{\rho}) \leq 1 + \min_{x \in \mathcal{X}} \mathbb{E}[J'(\Phi^x(\boldsymbol{\rho}, Y))] \leq 1 + \min_{x \in \mathcal{X}} \mathbb{E}[J(\Phi^x(\boldsymbol{\rho}, Y))]. \tag{40}$$

- **Case 2:** For all $\boldsymbol{\rho}$ such that $J(\boldsymbol{\rho}) = J''(\boldsymbol{\rho}) > 0$, and for any $x \in \mathcal{X}$, we have

$$\begin{aligned}
\mathbb{E}[J(\Phi^x(\boldsymbol{\rho}, Y))] &= \int J(\Phi^x(\boldsymbol{\rho}, y)) f_x^\rho(y) dy \\
&\stackrel{(a)}{\geq} \frac{\int H(\Phi^x(\boldsymbol{\rho}, y)) f_x^\rho(y) dy - F_M(\delta) - F_M(\iota)}{C} \\
&\quad + \frac{\log \frac{1-\iota}{\iota} - \log \frac{1-\delta}{\delta} - \log C_2 - 1}{C_1} \mathbf{1}_{\{\max_{i \in \Omega} \rho_i \leq 1-\delta\}} \\
&= J''(\boldsymbol{\rho}) - \frac{I(\boldsymbol{\rho}; f_x^\rho)}{C} \\
&\geq J''(\boldsymbol{\rho}) - 1 \\
&\stackrel{(b)}{=} J(\boldsymbol{\rho}) - 1, \tag{41}
\end{aligned}$$

where (a) follows from (38) and (39), and (b) holds since $\boldsymbol{\rho}$ is such that $J(\boldsymbol{\rho}) = J''(\boldsymbol{\rho})$.

Combining (40) and (41), we have that

$$J(\boldsymbol{\rho}) \leq 1 + \min_{x \in \mathcal{X}} \mathbb{E}[J(\Phi^x(\boldsymbol{\rho}, Y))]. \tag{42}$$

What remains is to show that $J(\boldsymbol{\rho}) = 0$ for all $\boldsymbol{\rho} \in \mathbb{P}(\Omega)$ such that $\max_{i \in \Omega} \rho_i \geq 1 - \iota$.

For $\boldsymbol{\rho} \in \mathbb{P}(\Omega)$ such that $\max_{i \in \Omega} \rho_i \geq 1 - \iota$, we have:

$$\begin{aligned}
J'(\boldsymbol{\rho}) &= \left[\sum_{i=1}^M \rho_i \frac{\log \frac{1-\iota}{\iota} - \log \frac{\rho_i}{1-\rho_i} - 1}{C} - \frac{F_M(\iota)}{C} \right]^+ \\
&\leq \left[\sum_{\{i \in \Omega: \rho_i < 1-\iota\}} \rho_i \frac{\log \frac{1}{\iota} + \log \frac{1}{\rho_i} - 1}{C_1} - \frac{F_M(\iota)}{C} \right]^+ \\
&\stackrel{(a)}{\leq} \left[\left(\sum_{\{i \in \Omega: \rho_i < 1-\iota\}} \rho_i \right) \frac{\log \frac{1}{\iota} + \log \frac{|\{i \in \Omega: \rho_i < 1-\iota\}|}{\sum_{\{i \in \Omega: \rho_i < 1-\iota\}} \rho_i} - 1}{C_1} - \frac{F_M(\iota)}{C} \right]^+ \\
&\stackrel{(b)}{\leq} \left[\frac{\iota \log \frac{1}{\iota} + \iota \log(M-1)}{C_1} - \frac{F_M(\iota)}{C} \right]^+ \\
&\stackrel{(c)}{=} 0,
\end{aligned} \tag{43}$$

where (a) follows by Jensen's inequality; (b) follows from the facts that $\sum_{\{i \in \Omega: \rho_i < 1-\iota\}} \rho_i \leq \iota < 1$ for any $\boldsymbol{\rho} \in \mathbb{P}(\Omega)$ that satisfies $\max_{i \in \Omega} \rho_i \geq 1 - \iota$, and $x \log \frac{1}{x} \leq 1$ for $x \in [0, 1]$; and (c) holds since $\iota \log \frac{1}{\iota} \leq H([\iota, 1 - \iota])$ and $C \leq C_1$.

On the other hand, for J'' and any $\boldsymbol{\rho} \in \mathbb{P}(\Omega)$ such that $\max_{i \in \Omega} \rho_i \geq 1 - \iota$, we have:

$$\begin{aligned}
J''(\boldsymbol{\rho}) &\leq \left[\frac{H(\boldsymbol{\rho}) - F_M(\iota)}{C} + \frac{\log \frac{1-\iota}{\iota} - \log \frac{1-\delta}{\delta}}{C_1} \mathbf{1}_{\{\max_{i \in \Omega} \rho_i \leq 1-\delta\}} \right]^+ \\
&\stackrel{(a)}{\leq} \left[\frac{\log \frac{1-\iota}{\iota} - \log \frac{1-\delta}{\delta}}{C_1} \mathbf{1}_{\{\delta \leq \iota, \max_{i \in \Omega} \rho_i \leq 1-\delta\}} \right]^+ \\
&= 0,
\end{aligned} \tag{44}$$

where (a) follows from concavity of the entropy function.

Combining (43) and (44), we have that

$$J(\boldsymbol{\rho}) = 0 \quad \text{if } \max_{i \in \Omega} \rho_i \geq 1 - \iota. \tag{45}$$

It is implied from (42) and (45) that J satisfies (35) and hence, $V_\iota^* \geq J = \max\{J', J''\} \geq J''$.

This is a slightly stronger result than (34). ■

APPENDIX II

PROOF OF THEOREM 2

First let us consider inequality (14) in Theorem 2.

Notice that for all $i \in \Omega$, upon selecting $X(t) = x$ and observing $Y(t) = y$, the belief state evolves as

$$\rho_i(t+1) = \rho_i(t) \frac{f_{h_i(x)}(y)}{f_x^{\boldsymbol{\rho}(t)}(y)}.$$

Let $U(\cdot)$ be the average log-likelihood function defined as

$$U(\boldsymbol{\rho}) := \sum_{i=1}^M \rho_i \log \frac{1 - \rho_i}{\rho_i}, \quad (46)$$

and let $\mathcal{F}(t) = \sigma\{X(0), Y(0), \dots, X(t-1), Y(t-1)\}$ denote the history of samples and observations up to time t . We have

$$\begin{aligned} & \mathbb{E}[U(\boldsymbol{\rho}(t+1)) | \mathcal{F}(t)] \\ &= \sum_{x \in \mathcal{X}} P(X(t) = x) \mathbb{E} \left[\sum_{i=1}^M \rho_i(t+1) \log \frac{1 - \rho_i(t+1)}{\rho_i(t+1)} \middle| \mathcal{F}(t), X(t) = x \right] \\ &= \sum_{x \in \mathcal{X}} P(X(t) = x) \int_{\mathcal{Y}} \sum_{i=1}^M \rho_i(t) f_{h_i(x)}(y) \log \frac{\sum_{j \neq i} \rho_j(t) f_{h_j(x)}(y)}{\rho_i(t) f_{h_i(x)}(y)} dy \\ &= \sum_{i=1}^M \rho_i(t) \log \frac{1 - \rho_i(t)}{\rho_i(t)} + \sum_{x \in \mathcal{X}} P(X(t) = x) \sum_{i=1}^M \int_{\mathcal{Y}} \rho_i(t) f_{h_i(x)}(y) \log \frac{\sum_{j \neq i} \frac{\rho_j(t)}{1 - \rho_i(t)} f_{h_j(x)}(y)}{f_{h_i(x)}(y)} dy \\ &= U(\boldsymbol{\rho}(t)) - \sum_{x \in \mathcal{X}} P(X(t) = x) \sum_{i=1}^M \rho_i(t) D(f_{h_i(x)} \| \sum_{j \neq i} \frac{\rho_j(t)}{1 - \rho_i(t)} f_{h_j(x)}) \\ &= U(\boldsymbol{\rho}(t)) - \sum_{x \in \mathcal{X}} P(X(t) = x) EJS(\boldsymbol{\rho}(t), x). \end{aligned}$$

Remember that \mathbf{c}_{EJS} , at any time $t < \tau$, selects a sample that maximizes the EJS divergence, i.e., $X(t) = \arg \max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}(t), x)$. Thus, under \mathbf{c}_{EJS} , the sequence $\{U(\boldsymbol{\rho}(t))\}$ satisfies

$$\begin{aligned} \mathbb{E}[U(\boldsymbol{\rho}(t+1)) | \mathcal{F}(t)] &= U(\boldsymbol{\rho}(t)) - \max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}(t), x) \\ &\stackrel{(a)}{\leq} U(\boldsymbol{\rho}(t)) - \alpha, \end{aligned} \quad (47)$$

where (a) follows from the assumption of Theorem 2. In other words, the sequence $\{-\frac{U(\boldsymbol{\rho}(t))}{\alpha} - t\}$ forms a submartingale with respect to the filtration $\{\mathcal{F}(t)\}$. Let us define a stopping time

$$v := \min \left\{ t : \max_{i \in \Omega} \rho_i(t) \geq 1 - \min \left\{ \frac{1}{\log 2M}, \epsilon \right\} \right\}.$$

It is clear that $\tilde{\tau}_\epsilon \leq v$ and hence, $\mathbb{E}[\tilde{\tau}_\epsilon] \leq \mathbb{E}[v]$ under any query scheme. By Doob's Stopping Theorem,

$$\frac{-U(\boldsymbol{\rho}(0))}{\alpha} \leq \mathbb{E} \left[\frac{-U(\boldsymbol{\rho}(v))}{\alpha} - v \right].$$

Rearranging the terms, we obtain

$$\begin{aligned} \mathbb{E}[v] &\leq \frac{U(\boldsymbol{\rho}(0))}{\alpha} + \mathbb{E} \left[\frac{-U(\boldsymbol{\rho}(v))}{\alpha} \right] \\ &\stackrel{(a)}{\leq} \frac{\log M + \mathbb{E}[-U(\boldsymbol{\rho}(v-1)) + U(\boldsymbol{\rho}(v-1)) - U(\boldsymbol{\rho}(v))]}{\alpha} \\ &\stackrel{(b)}{\leq} \frac{\log M + \max\{\log \log M, \log \frac{1}{\epsilon}\} + \mathbb{E}[U(\boldsymbol{\rho}(v-1)) - U(\boldsymbol{\rho}(v))]}{\alpha} \\ &\stackrel{(c)}{\leq} \frac{\log M + \max\{\log \log M, \log \frac{1}{\epsilon}\} + C_2 \left(3 + \frac{1}{\log 2M} \log(M-1)\right)}{\alpha} \\ &\leq \frac{\log M + \max\{\log \log M, \log \frac{1}{\epsilon}\} + 4C_2}{\alpha}, \end{aligned} \quad (48)$$

where (a) follows from the fact that initially the functions are equiprobable, i.e., $\boldsymbol{\rho}(0) = [1/M, \dots, 1/M]$ and hence $U(\boldsymbol{\rho}(0)) = \log(M-1)$, (b) holds since $\rho_i(v-1) < 1 - \min\{\frac{1}{\log 2M}, \epsilon\}$ for all $i \in \Omega$ and hence,

$$-U(\boldsymbol{\rho}(v-1)) = \sum_{i=1}^M \rho_i(v-1) \log \frac{\rho_i(v-1)}{1 - \rho_i(v-1)} < \log \frac{1 - \min\{\frac{1}{\log 2M}, \epsilon\}}{\min\{\frac{1}{\log 2M}, \epsilon\}} < \max\{\log \log M, \log \frac{1}{\epsilon}\},$$

and (c) follows from Lemma 6 in Appendix IV.

The proof of Inequality (15) in Theorem 2 follows similar lines. Recall from (13) that $\tilde{\rho} = 1 - \frac{1}{1 + \max\{\log M, \log \frac{1}{\epsilon}\}}$. Notice that if $\rho_i(t) < \tilde{\rho}$ for all $i \in \Omega$, then

$$U(\boldsymbol{\rho}(t)) = \sum_{i=1}^M \rho_i(t) \log \frac{1 - \rho_i(t)}{\rho_i(t)} > \sum_{i=1}^M \rho_i(t) \log \frac{1 - \tilde{\rho}}{\tilde{\rho}} = \log \frac{1 - \tilde{\rho}}{\tilde{\rho}}.$$

Similar to (47), we can show that

$$\mathbb{E}[U(\boldsymbol{\rho}(t+1)) | \mathcal{F}(t)] \leq \begin{cases} U(\boldsymbol{\rho}(t)) - \alpha & \text{if } U(\boldsymbol{\rho}(t)) > \log \frac{1-\tilde{\rho}}{\tilde{\rho}} \\ U(\boldsymbol{\rho}(t)) - \beta & \text{if } U(\boldsymbol{\rho}(t)) \leq \log \frac{1-\tilde{\rho}}{\tilde{\rho}} \end{cases}. \quad (49)$$

Furthermore, from Lemma 6 in Appendix IV, we know that if $\max_{i \in \Omega} \rho_i(t) \geq \tilde{\rho}$, then

$$|U(\boldsymbol{\rho}(t)) - U(\boldsymbol{\rho}(t-1))| \leq C_2 (3 + (1 - \tilde{\rho}) \log(M-1)) \leq 4C_2. \quad (50)$$

The rest of the proof follows directly from (49) and (50) and Fact 2 in Appendix IV.

III NOISY GENERALIZED BINARY SEARCH

Let $g_p(\cdot)$ and $\bar{g}_p(\cdot)$ be probability density functions on \mathcal{Y} defined as follows:

$$g_p(y) = \begin{cases} p & \text{if } y = -1 \\ 1 - p & \text{if } y = +1 \end{cases}, \quad \bar{g}_p(y) = g_p(-y). \quad (51)$$

It can be easily shown that:

$$C(p) = D(g_p \| \frac{g_p + \bar{g}_p}{2}) = D(\bar{g}_p \| \frac{g_p + \bar{g}_p}{2}) \quad \text{and} \quad C_1(p) = D(g_p \| \bar{g}_p) = D(\bar{g}_p \| g_p).$$

A. Proof of Proposition 3

The result for the sample-rich class follows from Proposition 2. Next we provide the proof for the class of disjoint interval functions and threshold functions.

1) Disjoint Class:

To prove this case, we will show that

$$\max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x) \geq \max_{i \in \Omega} \rho_i C_1(p).$$

Let $\hat{i} = \arg \max_{i \in \Omega} \rho_i$. By definition of the class of disjoint interval functions, there exists a sample $x_{\hat{i}} \in \mathcal{X}$ that satisfies $\mathbf{h}(x_{\hat{i}}) = \mathbf{e}_{\hat{i}}$. We have

$$EJS(\boldsymbol{\rho}, x_{\hat{i}}) \geq \rho_{\hat{i}} D\left(f_{h_{\hat{i}}(x_{\hat{i}})} \| \sum_{j \neq \hat{i}} \frac{\rho_j}{1 - \rho_{\hat{i}}} f_{h_j(x_{\hat{i}})}\right) = \rho_{\hat{i}} D(g_p \| \bar{g}_p) = \rho_{\hat{i}} C_1(p).$$

2) Threshold Class:

We will prove that

$$\max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x) \geq C(p).$$

At any belief vector $\boldsymbol{\rho} \in \mathbb{P}(\Omega)$, there exists $k, k \in \Omega$, such that $\sum_{j=1}^k \rho_j \leq \frac{1}{2}$ and $\sum_{j=1}^{k+1} \rho_j > \frac{1}{2}$. Let x_k and x_{k+1} be samples in \mathcal{X} that satisfy $\mathbf{h}(x_k) = \mathbf{u}_k$ and $\mathbf{h}(x_{k+1}) = \mathbf{u}_{k+1}$ respectively. Let $\delta_1 = \frac{1}{2} - \sum_{j=1}^k \rho_j$ and $\delta_2 = \sum_{j=1}^{k+1} \rho_j - \frac{1}{2}$. Notice that $\rho_{k+1} = \delta_1 + \delta_2$.

There are two cases:

- Case 1: $\delta_1 \leq \delta_2$. We have

$$EJS(\boldsymbol{\rho}, x_k) = \sum_{i=1}^M \rho_i D\left(f_{h_i(x_k)} \| \sum_{j \neq i} \frac{\rho_j}{1 - \rho_i} f_{h_j(x_k)}\right)$$

$$\begin{aligned}
&= \sum_{i=1}^k \rho_i D\left(\bar{g}_p \parallel \frac{1/2 - \delta_1 - \rho_i}{1 - \rho_i} \bar{g}_p + \frac{1/2 + \delta_1}{1 - \rho_i} g_p\right) \\
&\quad + \rho_{k+1} D\left(g_p \parallel \frac{1/2 - \delta_1}{1 - \rho_{k+1}} \bar{g}_p + \frac{1/2 - \delta_2}{1 - \rho_{k+1}} g_p\right) \\
&\quad + \sum_{i=k+2}^M \rho_i D\left(g_p \parallel \frac{1/2 - \delta_1}{1 - \rho_i} \bar{g}_p + \frac{1/2 + \delta_1 - \rho_i}{1 - \rho_i} g_p\right) \\
&\stackrel{(a)}{\geq} (1/2 - \delta_1) D\left(g_p \parallel (1/2 + \delta_1) \bar{g}_p + (1/2 - \delta_1) g_p\right) \\
&\quad + (\delta_1 + \delta_2) D\left(g_p \parallel \frac{1}{2} \bar{g}_p + \frac{1}{2} g_p\right) \\
&\quad + (1/2 - \delta_2) D\left(g_p \parallel (1/2 - \delta_1) \bar{g}_p + (1/2 + \delta_1) g_p\right) \\
&\stackrel{(b)}{\geq} D\left(g_p \parallel (1 - \gamma) \bar{g}_p + \gamma g_p\right) \\
&\stackrel{(c)}{\geq} D\left(g_p \parallel \frac{1}{2} \bar{g}_p + \frac{1}{2} g_p\right) \\
&= C(p),
\end{aligned}$$

where

$$\gamma = (1/2 - \delta_1)^2 + \frac{1}{2}(\delta_1 + \delta_2) + (1/2 - \delta_2)(1/2 + \delta_1),$$

inequality (a) follows from Fact 3 in Appendix IV and (51), (b) holds since KL divergence is convex, and (c) follows from the fact that $\gamma = \frac{1}{2} + \delta_1(\delta_1 - \delta_2) \leq \frac{1}{2}$ and by Fact 3.

- Case 2: $\delta_1 > \delta_2$. We have

$$\begin{aligned}
EJS(\boldsymbol{\rho}, x_{k+1}) &= \sum_{i=1}^k \rho_i D\left(\bar{g}_p \parallel \frac{1/2 + \delta_2 - \rho_i}{1 - \rho_i} \bar{g}_p + \frac{1/2 - \delta_2}{1 - \rho_i} g_p\right) \\
&\quad + \rho_{k+1} D\left(\bar{g}_p \parallel \frac{1/2 - \delta_1}{1 - \rho_{k+1}} \bar{g}_p + \frac{1/2 - \delta_2}{1 - \rho_{k+1}} g_p\right) \\
&\quad + \sum_{i=k+2}^M \rho_i D\left(g_p \parallel \frac{1/2 + \delta_2}{1 - \rho_i} \bar{g}_p + \frac{1/2 - \delta_2 - \rho_i}{1 - \rho_i} g_p\right) \\
&\stackrel{(a)}{\geq} (1/2 - \delta_1) D\left(g_p \parallel (1/2 - \delta_2) \bar{g}_p + (1/2 + \delta_2) g_p\right) \\
&\quad + (\delta_1 + \delta_2) D\left(g_p \parallel \frac{1}{2} \bar{g}_p + \frac{1}{2} g_p\right) \\
&\quad + (1/2 - \delta_2) D\left(g_p \parallel (1/2 + \delta_2) \bar{g}_p + (1/2 - \delta_2) g_p\right)
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{\geq} D\left(g_p \parallel (1 - \gamma')\bar{g}_p + \gamma'g_p\right) \\
&\stackrel{(c)}{\geq} D\left(g_p \parallel \frac{1}{2}\bar{g}_p + \frac{1}{2}g_p\right) \\
&= C(p),
\end{aligned}$$

where

$$\gamma' = (1/2 - \delta_1)(1/2 + \delta_2) + \frac{1}{2}(\delta_1 + \delta_2) + (1/2 - \delta_2)^2,$$

inequality (a) follows from Fact 3 in Appendix IV and (51), (b) holds since KL divergence is convex, and (c) follows from the fact that $\gamma' = \frac{1}{2} + \delta_2(\delta_2 - \delta_1) < \frac{1}{2}$ and by Fact 3.

Therefore,

$$\max_{x \in \mathcal{X}} EJS(\boldsymbol{\rho}, x) \geq \max\{EJS(\boldsymbol{\rho}, x_k), EJS(\boldsymbol{\rho}, x_{k+1})\} \geq C(p).$$

B. Noisy Generalized Binary Search: Asymptotic Analysis

For disjoint function class \mathcal{H}_D and from Theorem 2 and Proposition 3,

$$\begin{aligned}
\mathbb{E}[\tau_\epsilon^*] &\leq \frac{\log M + \max\{\log \log M, \log \log \frac{1}{\epsilon}\}}{\frac{1}{M}C_1(p)} + \frac{\log \frac{1}{\epsilon}}{\tilde{\rho}C_1(p)} + \frac{3(4C_2(p))^2}{\frac{1}{M}C_1(p)\tilde{\rho}C_1(p)} \\
&\stackrel{(a)}{\leq} \frac{M \log M + M \log \log \frac{M}{\epsilon}}{C_1(p)} + \frac{\log \frac{1}{\epsilon} + 1}{C_1(p)} + \frac{6M(4C_2(p))^2}{(C_1(p))^2} \\
&\leq \left(\frac{M \log M}{C_1(p)} + \frac{\log \frac{1}{\epsilon}}{C_1(p)}\right) \times \left(1 + \frac{M \log \log \frac{M}{\epsilon} + 1 + 6M(4C_2(p))^2/C_1(p)}{M \log M + \log \frac{1}{\epsilon}}\right) \\
&= \left(\frac{M \log M}{C_1(p)} + \frac{\log \frac{1}{\epsilon}}{C_1(p)}\right)(1 + o(1)),
\end{aligned}$$

where $o(1) \rightarrow 0$ as $\epsilon \rightarrow 0$ or $M \rightarrow \infty$ and (a) holds since $\frac{1}{\tilde{\rho}} = 1 + \frac{1}{\max\{\log M, \log \frac{1}{\epsilon}\}} \leq 2$.

For threshold function class \mathcal{H}_T and from Theorem 2 and Proposition 3,

$$\begin{aligned}
\mathbb{E}[\tau_\epsilon^*] &\leq \frac{\log M + \max\{\log \log M, \log \frac{1}{\epsilon}\} + 4C_2(p)}{C(p)} \\
&\leq \left(\frac{\log M}{C(p)} + \frac{\log \frac{1}{\epsilon}}{C(p)}\right) \times \left(1 + \frac{\log \log M + 4C_2(p)}{\log \frac{M}{\epsilon}}\right) \\
&= \left(\frac{\log M}{C(p)} + \frac{\log \frac{1}{\epsilon}}{C(p)}\right)(1 + o(1)),
\end{aligned}$$

where $o(1) \rightarrow 0$ as $\epsilon \rightarrow 0$ or $M \rightarrow \infty$.

For rich function class \mathcal{H}_R and from Theorem 2 and Proposition 3,

$$\begin{aligned}
\mathbb{E}[\tau_\epsilon^*] &\leq \frac{\log M + \max\{\log \log M, \log \log \frac{1}{\epsilon}\}}{C(p)} + \frac{\log \frac{1}{\epsilon}}{\tilde{\rho}C_1(p)} + \frac{3(4C_2(p))^2}{C(p)\tilde{\rho}C_1(p)} \\
&\stackrel{(a)}{\leq} \frac{\log M + \log \log \frac{M}{\epsilon}}{C(p)} + \frac{\log \frac{1}{\epsilon} + 1}{C_1(p)} + \frac{6(4C_2(p))^2}{C(p)C_1(p)} \\
&\leq \left(\frac{\log M}{C(p)} + \frac{\log \frac{1}{\epsilon}}{C_1(p)} \right) \times \left(1 + \frac{C_1(p) \log \log \frac{M}{\epsilon} + C(p) + 6(4C_2(p))^2}{C(p) \log \frac{M}{\epsilon}} \right) \\
&= \left(\frac{\log M}{C(p)} + \frac{\log \frac{1}{\epsilon}}{C_1(p)} \right) (1 + o(1)),
\end{aligned}$$

where $o(1) \rightarrow 0$ as $\epsilon \rightarrow 0$ or $M \rightarrow \infty$ and (a) holds since $\frac{1}{\tilde{\rho}} = 1 + \frac{1}{\max\{\log M, \log \frac{1}{\epsilon}\}} \leq 2$.

It follows from Proposition 1 that

$$\begin{aligned}
\mathbb{E}[\tau_\epsilon^*] &\geq \frac{\log M}{C(p)} \left(1 - \frac{2}{\log \frac{4}{\epsilon}} - \epsilon \log \frac{1}{\epsilon} \right) - \frac{2}{C(p)} + \frac{\log \frac{1}{\epsilon}}{C_1(p)} \left(1 - \frac{2 \log \log \frac{2}{\epsilon} + \log C_2(p) + 4}{\log \frac{1}{\epsilon}} \right) \\
&\geq \left(\frac{\log M}{C(p)} + \frac{\log \frac{1}{\epsilon}}{C_1(p)} \right) \times \left(1 - \epsilon \log \frac{1}{\epsilon} - \frac{2 \log \log \frac{2}{\epsilon} + \log C_2(p) + 4 + 2C_1(p)/C(p)}{\log \frac{1}{\epsilon}} \right) \\
&= \left(\frac{\log M}{C(p)} + \frac{\log \frac{1}{\epsilon}}{C_1(p)} \right) (1 - o(1)), \tag{52}
\end{aligned}$$

where $o(1) \rightarrow 0$ as $\epsilon \rightarrow 0$.

IV TECHNICAL LEMMAS

In this appendix, we provide some preliminary lemmas and facts. These lemmas are technical and only helpful in proving the main results of the paper.

Lemma 1. *Consider stopping times defined earlier with scalars $\iota \geq \epsilon > 0$. We have*

$$\mathbb{E}[\tilde{\tau}_\iota^*] \left(1 - \frac{\epsilon}{\iota} \right) \leq \mathbb{E}[\tau_\epsilon^*] \leq \mathbb{E}[\tilde{\tau}_\epsilon^*].$$

Proof: Under any query scheme with the stopping rule (9):

$$\text{Pe} = \mathbb{E}[1 - \max_{i \in \Omega} \rho_i(\tilde{\tau}_\epsilon)] \leq \epsilon,$$

hence, by construction,

$$\mathbb{E}[\tau_\epsilon^*] \leq \mathbb{E}[\tilde{\tau}_\epsilon^*]. \tag{53}$$

On the other hand, let us consider $\mathbb{E}[\tilde{\tau}_\iota^*]$ for any $\iota > \epsilon$. Let τ_ϵ be a stopping time at which the probability of error satisfies $\text{Pe} \leq \epsilon$. Under any query scheme,

$$\begin{aligned}
\mathbb{E}[\tau_\epsilon] &\geq \mathbb{E}[\tau_\epsilon | \max_{j \in \Omega} \rho_j(\tau_\epsilon) \geq 1 - \iota] P(\max_{j \in \Omega} \rho_j(\tau_\epsilon) \geq 1 - \iota) \\
&\stackrel{(a)}{\geq} \mathbb{E}[\tau_\epsilon | \max_{j \in \Omega} \rho_j(\tau_\epsilon) \geq 1 - \iota] \left(1 - \iota^{-1} \mathbb{E}[1 - \max_{j \in \Omega} \rho_j(\tau_\epsilon)]\right) \\
&\stackrel{(b)}{\geq} \mathbb{E}[\tau_\epsilon | \max_{j \in \Omega} \rho_j(\tau_\epsilon) \geq 1 - \iota] \left(1 - \frac{\epsilon}{\iota}\right) \\
&\geq \mathbb{E}[\tilde{\tau}_\iota^*] \left(1 - \frac{\epsilon}{\iota}\right)
\end{aligned} \tag{54}$$

where (a) follows from Markov inequality and (b) follows from the definition of τ_ϵ which implies that $\text{Pe} = \mathbb{E}[1 - \max_{j \in \Omega} \rho_j(\tau_\epsilon)] \leq \epsilon$. From (54),

$$\mathbb{E}[\tilde{\tau}_\iota^*] \left(1 - \frac{\epsilon}{\iota}\right) \leq \mathbb{E}[\tau_\epsilon^*]. \tag{55}$$

■

Lemma 3. Any functional $V : \mathbb{P}(\Omega) \rightarrow \mathbb{R}_+$ that satisfies the following:

$$V(\boldsymbol{\rho}) \leq \begin{cases} 0 & \text{if } \max_{j \in \Omega} \rho_j \geq 1 - \iota \\ 1 + \min_{x \in \mathcal{X}} \mathbb{E}[V(\Phi^x(\boldsymbol{\rho}, Y))] & \text{otherwise} \end{cases},$$

provides a uniform lower bound for the optimal value function V_ι^* .

Proof: To prove the above fact, we have to slightly modify the state space and introduce new notations. We assume that after taking the retire-declare action, the system goes to the termination state, denoted by F , and remains in that state for the rest of the time. The state space is modified to $\mathcal{S} = \mathbb{P}(\Omega) \cup \{F\}$ to include the termination state. For $x \in \mathcal{X} \cup \{d_1, d_2, \dots, d_M\}$, $s \in \mathcal{S}$, let

$$c^x(s) = \begin{cases} 1 & \text{if } s = \boldsymbol{\rho} \in \mathbb{P}(\Omega), x \in \mathcal{X} \\ \infty & \text{if } s = \boldsymbol{\rho} \in \mathbb{P}(\Omega), \max_{j \in \Omega} \rho_j < 1 - \iota, x \in \{d_1, \dots, d_M\} \\ 0 & \text{if } s = \boldsymbol{\rho} \in \mathbb{P}(\Omega), \max_{j \in \Omega} \rho_j \geq 1 - \iota, x \in \{d_1, \dots, d_M\} \\ 0 & \text{if } s = F \end{cases}.$$

The Bayes operator is modified as follows:

$$\Phi^x(s, y) = \begin{cases} \Phi^x(\boldsymbol{\rho}, y) & \text{if } s = \boldsymbol{\rho} \in \mathbb{P}(\Omega), x \in \mathcal{X} \\ F & \text{if } s = \boldsymbol{\rho} \in \mathbb{P}(\Omega), x \in \{d_1, \dots, d_M\} \\ F & \text{if } s = F \end{cases}.$$

Using the notations above, condition (35) is rewritten as

$$\begin{aligned} V(F) &= 0, \\ V(s) &\leq \min_{x \in \mathcal{X} \cup \{d_1, \dots, d_M\}} \{c^x(s) + \mathbb{E}[V(\Phi^x(s, Y))]\}, \quad \forall s \in \mathcal{S} - \{F\}. \end{aligned} \quad (56)$$

Let S_0, S_1, S_2, \dots be a sequence of random variables denoting the belief states at times $t = 0, 1, 2, \dots$ starting from belief state s , i.e.,

$$\begin{aligned} S_0 &= s, \\ S_n &= \Phi^{X(n-1)}(S_{n-1}, Y), \quad \forall n, n > 0. \end{aligned}$$

Using (56) iteratively for N times, we obtain

$$\begin{aligned} V(s) &\leq \mathbb{E}_{\pi^*}[c^{X(0)}(s)] + \mathbb{E}_{\pi^*}[V(\Phi^{X(0)}(s, Y))] \\ &= \mathbb{E}_{\pi^*}[c^{X(0)}(S_0)] + \mathbb{E}_{\pi^*}[V(S_1)] \\ &\leq \mathbb{E}_{\pi^*}\left[\sum_{n=0}^1 c^{X(n)}(S_n)\right] + \mathbb{E}_{\pi^*}[V(S_2)] \\ &\leq \mathbb{E}_{\pi^*}\left[\sum_{n=0}^{N-1} c^{X(n)}(S_n)\right] + \mathbb{E}_{\pi^*}[V(S_N)], \end{aligned}$$

where subscript π^* implies that actions are selected according to an optimal policy π^* .³ Taking the limit as $N \rightarrow \infty$, we obtain

$$\begin{aligned} V(s) &\stackrel{(a)}{\leq} \mathbb{E}_{\pi^*}\left[\sum_{n=0}^{\infty} c^{X(n)}(S_n)\right] + \lim_{N \rightarrow \infty} \mathbb{E}_{\pi^*}[V(S_N)] \\ &\stackrel{(b)}{=} V_\iota^*(s) + \lim_{N \rightarrow \infty} \mathbb{E}_{\pi^*}[V(S_N)] \\ &= V_\iota^*(s) + \lim_{N \rightarrow \infty} \mathbb{E}_{\pi^*}[V(F)\mathbf{1}_{\{S_N=F\}} + V(S_N)\mathbf{1}_{\{S_N \neq F\}}] \\ &= V_\iota^*(s) + \lim_{N \rightarrow \infty} \mathbb{E}_{\pi^*}[V(S_N)\mathbf{1}_{\{S_N \neq F\}}] \\ &= V_\iota^*(s), \end{aligned}$$

where (a) follows from the monotone convergence theorem and (b) follows from the definition of V_ι^* . ■

³The existence of an optimal policy follows from [25, Corollary 9.12.1] and since $|\mathcal{L}| < \infty$.

Lemma 4. For any $i \in \Omega$,

$$\left| \log \frac{\rho_i(t+1)}{1-\rho_i(t+1)} - \log \frac{\rho_i(t)}{1-\rho_i(t)} \right| \leq \log C_2.$$

Proof:

$$\begin{aligned} \left| \log \frac{\rho_i(t+1)}{1-\rho_i(t+1)} - \log \frac{\rho_i(t)}{1-\rho_i(t)} \right| &= \left| \log \frac{\rho_i(t) f_{h_i(X(t))}(Y(t))}{\sum_{j \neq i} \rho_j(t) f_{h_j(X(t))}(Y(t))} - \log \frac{\rho_i(t)}{1-\rho_i(t)} \right| \\ &= \left| \log \frac{f_{h_i(X(t))}(Y(t))}{\sum_{j \neq i} \frac{\rho_j(t)}{1-\rho_i(t)} f_{h_j(X(t))}(Y(t))} \right| \\ &\leq \max_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \log \frac{f_{h_i(x)}(y)}{\min_{j \neq i} f_{h_j(x)}(y)} \\ &\leq \log C_2. \end{aligned}$$

■

Lemma 5. For any $i \in \Omega$,

$$|\rho_i(t+1) - \rho_i(t)| \leq \rho_i(t)(1-\rho_i(t))(C_2 - 1).$$

Proof:

$$\begin{aligned} |\rho_i(t+1) - \rho_i(t)| &= \rho_i(t) \left| \frac{f_{h_i(X(t))}(Y(t))}{\sum_{j=1}^M \rho_j(t) f_{h_j(X(t))}(Y(t))} - 1 \right| \\ &= \rho_i(t) \left| \frac{(1-\rho_i(t)) f_{h_i(X(t))}(Y(t)) - \sum_{j \neq i} \rho_j(t) f_{h_j(X(t))}(Y(t))}{\sum_{j=1}^M \rho_j(t) f_{h_j(X(t))}(Y(t))} \right| \\ &= \rho_i(t)(1-\rho_i(t)) \left| \frac{f_{h_i(X(t))}(Y(t)) - \sum_{j \neq i} \frac{\rho_j(t)}{1-\rho_i(t)} f_{h_j(X(t))}(Y(t))}{\rho_i(t) f_{h_i(X(t))}(Y(t)) + (1-\rho_i(t)) \sum_{j \neq i} \frac{\rho_j(t)}{1-\rho_i(t)} f_{h_j(X(t))}(Y(t))} \right| \\ &\leq \rho_i(t)(1-\rho_i(t)) \left(\frac{\max \left\{ f_{h_i(X(t))}(Y(t)), \sum_{j \neq i} \frac{\rho_j(t)}{1-\rho_i(t)} f_{h_j(X(t))}(Y(t)) \right\}}{\min \left\{ f_{h_i(X(t))}(Y(t)), \sum_{j \neq i} \frac{\rho_j(t)}{1-\rho_i(t)} f_{h_j(X(t))}(Y(t)) \right\}} - 1 \right) \\ &\leq \rho_i(t)(1-\rho_i(t)) \left(\max_{k,l \in \mathcal{L}} \sup_{y \in \mathcal{Y}} \frac{f_k(y)}{f_l(y)} - 1 \right) \end{aligned}$$

$$= \rho_i(t)(1 - \rho_i(t))(C_2 - 1).$$

■

Lemma 6. For any $\delta \in (0, \frac{1}{2}]$, if $\max_{i \in \Omega} \rho_i(t) \geq 1 - \delta$, then

$$|U(\boldsymbol{\rho}(t)) - U(\boldsymbol{\rho}(t-1))| \leq C_2(3 + \delta \log(M-1)).$$

Proof: Without loss of generality assume $\rho_i(t) \geq 1 - \delta$. We obtain

$$\begin{aligned} & |-U(\boldsymbol{\rho}(t-1)) + U(\boldsymbol{\rho}(t))| \\ &= \left| \sum_{i=1}^M \rho_i(t-1) \log \frac{\rho_i(t-1)}{1 - \rho_i(t-1)} - \sum_{i=1}^M \rho_i(t) \log \frac{\rho_i(t)}{1 - \rho_i(t)} \right| \\ &= \left| \sum_{i=1}^M \rho_i(t-1) \left(\log \frac{\rho_i(t-1)}{1 - \rho_i(t-1)} - \log \frac{\rho_i(t)}{1 - \rho_i(t)} \right) + \sum_{i=1}^M (\rho_i(t-1) - \rho_i(t)) \log \frac{\rho_i(t)}{1 - \rho_i(t)} \right| \\ &\leq \max_{i \in \Omega} \left| \log \frac{\rho_i(t-1)}{1 - \rho_i(t-1)} - \log \frac{\rho_i(t)}{1 - \rho_i(t)} \right| + \left| \sum_{i=1}^M (\rho_i(t-1) - \rho_i(t)) \log \frac{\rho_i(t)}{1 - \rho_i(t)} \right| \\ &\stackrel{(a)}{\leq} \log C_2 + \sum_{i=1}^M |\rho_i(t-1) - \rho_i(t)| \cdot \left| \log \frac{\rho_i(t)}{1 - \rho_i(t)} \right| \\ &\stackrel{(b)}{\leq} \log C_2 + C_2 \sum_{i=1}^M \rho_i(t)(1 - \rho_i(t)) \left| \log \frac{\rho_i(t)}{1 - \rho_i(t)} \right| \\ &\leq \log C_2 + C_2 \rho_i(t)(1 - \rho_i(t)) \left| \log \frac{\rho_i(t)}{1 - \rho_i(t)} \right| + C_2 \sum_{i \neq \hat{i}} \rho_i(t) \log \frac{1}{\rho_i(t)} \\ &\stackrel{(c)}{\leq} \log C_2 + C_2 + C_2 \left(\sum_{i \neq \hat{i}} \rho_i(t) \right) \log \frac{M-1}{\sum_{i \neq \hat{i}} \rho_i(t)} \\ &\leq \log C_2 + C_2 + C_2(\delta \log(M-1) + 1) \\ &\stackrel{(d)}{\leq} C_2(3 + \delta \log(M-1)), \end{aligned}$$

where (a) and (b) follow respectively from Lemmas 4 and 5; and (c) follows from Jensen's inequality and the fact that

$$z(1-z) \left| \log \frac{z}{1-z} \right| \leq 1, \quad z \in [0, 1];$$

and (d) holds since $C_2 \geq 1$ and hence $\log C_2 \leq C_2$. ■

Fact 2 (Lemma 10 in [18]). *Assume that the sequence $\{\xi(t)\}$, $t = 0, 1, 2, \dots$ forms a submartingale with respect to a filtration $\{\mathcal{F}(t)\}$. Furthermore, assume there exist positive constants K_1 , K_2 , and K_3 such that*

$$\begin{aligned}\mathbb{E}[\xi(t+1)|\mathcal{F}(t)] &\geq \xi(t) + K_1 \text{ if } \xi(t) < 0, \\ \mathbb{E}[\xi(t+1)|\mathcal{F}(t)] &\geq \xi(t) + K_2 \text{ if } \xi(t) \geq 0, \\ |\xi(t+1) - \xi(t)| &\leq K_3 \text{ if } \max\{\xi(t+1), \xi(t)\} \geq 0.\end{aligned}$$

Consider the stopping time $\nu = \min\{t : \xi(t) \geq B\}$, $B > 0$. Then we have the inequality

$$\mathbb{E}[\nu] \leq \frac{B - \xi(0)}{K_2} + \xi(0)\mathbf{1}_{\{\xi(0) < 0\}} \left(\frac{1}{K_2} - \frac{1}{K_1} \right) + \frac{3K_3^2}{K_1K_2}.$$

Fact 3 (Lemma 1 in [18]). *For any two distributions P and Q on a set \mathcal{Y} and $\gamma \in [0, 1]$, $D(P||\gamma P + (1 - \gamma)Q)$ is decreasing in γ .*

REFERENCES

- [1] R. D. Nowak, “Noisy generalized binary search,” in *Neural Information Processing Systems (NIPS)*, 2009.
- [2] M. V. Burnashev, “Data transmission over a discrete channel with feedback. Random transmission time,” *Problemy Peredachi Informatsii*, vol. 12, no. 4, pp. 10–30, 1975.
- [3] M. Naghshvar and T. Javidi, “Extrinsic Jensen–Shannon divergence with application in active hypothesis testing,” in *IEEE International Symposium on Information Theory (ISIT)*, 2012.
- [4] S. Dasgupta, “Two faces of active learning,” *Theoretical Computer Science*, vol. 412, no. 19, pp. 1767–1781, Apr. 2011.
- [5] —, “Coarse sample complexity bounds for active learning,” in *Neural Information Processing Systems (NIPS)*, 2005.
- [6] M. Kääriäinen, “Active learning in the non-realizable case,” in *17th International Conference on Algorithmic Learning Theory (ALT)*, 2006, pp. 63–77.
- [7] S. Hanneke, “A bound on the label complexity of agnostic active learning,” in *24th Annual International Conference on Machine Learning (ICML)*, 2007, pp. 353–360.
- [8] M. F. Balcan, A. Beygelzimer, and J. Langford, “Agnostic active learning,” *Journal of Computer and System Sciences*, vol. 75, no. 1, pp. 78–89, 2009.
- [9] S. Dasgupta, D. Hsu, and C. Monteleoni, “A general agnostic active learning algorithm,” in *NIPS*, 2007.
- [10] A. Beygelzimer, S. Dasgupta, and J. Langford, “Importance weighted active learning,” in *26th Annual International Conference on Machine Learning (ICML)*, 2009, pp. 49–56.
- [11] A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang, “Agnostic active learning without constraints,” in *Advances in Neural Information Processing Systems (NIPS)*, 2010.
- [12] M. Raginsky and A. Rakhlin, “Lower bounds for passive and active learning,” in *Neural Information Processing Systems (NIPS)*, 2011.
- [13] D. Golovin, A. Krause, and D. Ray, “Near-optimal bayesian active learning with noisy observations,” in *NIPS*, 2010, pp. 766–774.

- [14] Y. Sakakibara, "On learning from queries and counterexamples in the presence of noise," *Information Processing Letters*, vol. 37, no. 5, pp. 279–284, Mar. 1991. [Online]. Available: [http://dx.doi.org/10.1016/0020-0190\(91\)90220-C](http://dx.doi.org/10.1016/0020-0190(91)90220-C)
- [15] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.
- [16] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," 2013, to appear in *the Annals of Statistics* (available on arXiv:1203.4626).
- [17] M. H. DeGroot, *Optimal statistical decisions*. New York, NY: McGraw-Hill Book Co., 1970.
- [18] M. Naghshvar, T. Javidi, and M. Wigger, "Extrinsic Jensen–Shannon divergence: Applications to variable-length coding," available on arXiv: 1307.0067.
- [19] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2006.
- [20] R. G. Gallager, *Information theory and reliable communication*. New York, NY: John Wiley & Sons, Inc., 1968.
- [21] M. V. Burnashev, "Sequential discrimination of hypotheses with control of observations," *Mathematics of the USSR-Izvestiya*, vol. 15, no. 3, pp. 419–440, 1980.
- [22] P. Berlin, B. Nakiboglu, B. Rimoldi, and E. Telatar, "A simple converse of Burnashev's reliability function," *IEEE Transactions on Information Theory*, vol. 55, pp. 3074–3080, 2009.
- [23] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4903–4925, August 2011.
- [24] R. D. Nowak, "The geometry of generalized binary search," *IEEE Transactions on Information Theory*, vol. 57, no. 12, pp. 7893–7906, December 2011.
- [25] D. P. Bertsekas and S. E. Shreve, *Stochastic optimal control: The discrete-time case*. Belmont, MA: Athena Scientific, 2007.

Mohammad Naghshvar (S'07) received the B.S. degree in electrical engineering from Sharif University of Technology in 2007. He obtained the M.Sc. degree and the Ph.D. degree in electrical engineering (communication theory and systems) both from University of California San Diego in 2009 and 2013, respectively. He is currently a senior R&D engineer at Qualcomm Technologies Inc., San Diego, CA. His research interests include active hypothesis testing and optimal experimental design, stochastic control and optimization, wireless communication and information theory, and routing and scheduling in wireless networks.

Tara Javidi (S'96-M'02) studied electrical engineering at the Sharif University of Technology from 1992 to 1996. She received her MS degrees in Electrical Engineering: Systems, and Applied Mathematics: Stochastics, from the University of Michigan, Ann Arbor, MI. She received her PhD in electrical engineering and computer science from the University of Michigan, Ann Arbor, in 2002. From 2002 to 2004, she was an assistant professor at the Electrical Engineering Department, University of Washington, Seattle. In 2005, she joined University of California, San Diego, where she is currently an associate professor of electrical and computer engineering. Her research interests are in communication networks, stochastic resource allocation, stochastic control theory, and wireless communications.

Kamalika Chaudhuri is an assistant professor in the Department of Computer Science and Engineering, University of California, San Diego. She received a bachelor of technology degree in computer science and engineering from the Indian Institute of Technology, Kanpur, in 2002, and a Ph.D. degree in computer science from the University of California at Berkeley in 2007. Her research focuses on the design and analysis of machine-learning algorithms and their applications. In particular, she is interested in privacy-preserving machine learning, where the goal is to develop machine-learning methods for sensitive data while still preserving the privacy of the individuals in the data set.