Bayesian Active Learning Using Arbitrary Binary Valued Queries

Liu Yang[1], Steve Hanneke[2], and Jaime Carbonell[3]

[1] Machine Learning Department
Carnegie Mellon University
liuy@cs.cmu.edu
[2] Department of Statistics
Carnegie Mellon University
shanneke@stat.cmu.edu
[3] Language Technologies Institute
Carnegie Mellon University
jgc@cs.cmu.edu

Abstract. We explore a general Bayesian active learning setting, in which the learner can ask arbitrary yes/no questions. We derive upper and lower bounds on the expected number of queries required to achieve a specified expected risk.

Key words: Active Learning, Bayesian Learning, Sample Complexity, Information Theory

1 Introduction

In this work, we study the fundamental complexity of Bayesian active learning by examining the basic problem of learning from binary-valued queries. We are particularly interested in identifying a key quantity that characterizes the number of queries required to learn to a given accuracy, given knowledge of the prior distribution from which the target is sampled. This topic is interesting both in itself, and also as a general setting in which to derive lower bounds, which apply broadly to any active learning scenario in which binary-valued queries are employed, such as the popular setting of active learning with label requests (membership queries). The analysis of the Bayesian variant of this setting is important for at least two reasons: first, for practical reasons, as minimax analyses tend to emphasize scenarios much more difficult to learn from than what the world often offers us, so that the smoothed or average-case analysis offered by a Bayesian setting can often be an informative alternative, and second, for philosophical reasons, owing to the decision-theoretic interpretation of rational inference, which is typically formulated in a Bayesian setting.

There is much related work on active learning with binary-valued queries. However, perhaps the most relevant for us is the result of (Kulkarni et al., 1993). In this classic work, they allow a learning algorithm to ask *any* question with a yes/no answer, and derive a precise characterization of the number of these binary-valued queries necessary and sufficient for learning a target classifier to a prescribed accuracy, in a PAC-like framework. In particular, they find this quantity is essentially characterized by $\log \mathcal{M}(\epsilon)$, where $1 - \epsilon$ is the desired accuracy, and $\mathcal{M}(\epsilon)$ is the size of a maximal ϵ-packing of the concept space.

In addition to being quite interesting in their own right, these results have played a significant role in the recent developments in active learning with "label request" queries for binary classification (Hanneke, 2007b; Hanneke, 2007a; Dasgupta, 2005). Specifically, since label requests can be viewed as a type of binary-valued query, the number of label requests necessary for learning is naturally lower bounded by the number of arbitrary binary-valued queries necessary for learning. We therefore always expect to see some term relating to $\log \mathcal{M}(\epsilon)$ in our sample complexity bounds for active learning with label requests (though this factor is typically represented by its upper bound: $\propto VC(\mathbb{C}) \log(1/\epsilon)$, where $VC(\mathbb{C})$ is the VC dimension).

Also related is a certain thread of the literature on sample complexity bounds for Bayesian learning. In particular, (Haussler et al., 1994a) study the passive learning problem in a Bayesian setting, and study the effect of the information made available via access to the prior. In many cases, the learning problem is made significantly easier than the worst-case scenarios of the PAC model. In particular, (building from the work of (Haussler et al., 1994b)) they find that $VC(\mathbb{C})/\epsilon$ random labeled examples are sufficient to achieve expected error rate at most ϵ using the Bayes classifier.

Allowing somewhat more general types of queries than (Haussler et al., 1994a), a paper by (Freund et al., 1997; Seung et al., 1992) studied an algorithm known as Query by Committee (QBC). Specifically, QBC is allowed to sequentially decide which points to select, observing each response before selecting the next data point to observe. They found this additional flexibility can sometimes pay off significantly, reducing the expected number of queries needed exponentially to only $O(\log(1/\epsilon))$. However, these results only seem to apply to a very narrow family of problems, where a certain expected information gain quantity is lower bounded by a constant, a situation which seems fairly uncommon among the types of learning problems we are typically most interested in (informative priors, or clustered data). Thus, to our knowledge, the general questions, such as how much advantage we actually get from having access to the prior π, and what fundamental quantities describe the intrinsic complexity of the learning problem, remain virtually untouched in the published literature.

The "label request" query discussed in these Bayesian analyses represents a type of binary-valued query, though quite restricted compared to the powerful queries analyzed in the present work. As a first step toward a more complete understanding of the Bayesian active learning problem, we propose to return to the basic question of how many binary-valued queries are necessary and sufficient in general; but unlike the (Kulkarni et al., 1993) analysis, we adopt the Bayesian perspective of (Haussler et al., 1994a) and (Freund et al., 1997), so that the algorithms in question will directly depend on the prior π. In fact, we investigate the problem in a somewhat more general form, where reference to the underlying data distribution is replaced by direct reference to the induced pseudo-metric between elements of the concept space. As we point out below, this general problem has deep connections to many problems commonly studied in information theory (e.g., the analysis of lossy compression); for instance, one might view the well-known asymptotic results of rate distortion theory as a massively multitask variant of this problem. However, to our knowledge, the basic question of the number of binary queries necessary to approximate a single random target h^* to a given accuracy, given access to the distribution π of h^*, has not previously been addressed in generality.

Below, we are able to derive upper and lower bounds on the query complexity based on a natural analogue of the bounds of (Kulkarni et al., 1993). Specifically, we find that in this Bayesian setting, under an assumption of bounded doubling dimension, the query complexity is controlled by the entropy of a partition induced by a maximal ϵ-packing (specifically, the natural Voronoi partition); in particular, the worst-case value of this entropy is the $\log \mathcal{M}(\epsilon)$ bound of (Kulkarni et al., 1993), which represents a uniform prior over the regions of the partition. The upper bound is straightforward to derive, but nice to have; but our main contribution is the lower bound, the proof of which is somewhat more involved.

The rest of this paper is organized as follows. In Section 2, we introduce a few important quantities used in the statement of the main theorem. Following this, Section 3 contains a statement of our main result, along with some explanation. Section 4 contains the proof of our result, followed by Section 5, which states a few of the many remaining open questions about Bayesian active learning.

2 Definitions and Notation

We will formalize our discussion in somewhat more abstract terms.

Formally, throughout this discussion, we will suppose $\mathbb{C}^*$ is an arbitrary (nonempty) collection of objects, equipped with a separable pseudo-metric $\rho : \mathbb{C}^* \times \mathbb{C}^* \to [0, \infty)$. [4] We suppose $\mathbb{C}^*$ is equipped with its Borel σ-algebra induced by ρ. There is additionally a (nonempty, measurable) set $\mathbb{C} \subseteq \mathbb{C}^*$, and we denote by $\bar{\rho} = \sup_{h_1, h_2 \in \mathbb{C}} \rho(h_1, h_2)$. Finally, there is a probability measure π with $\pi(\mathbb{C}) = 1$, known as the "prior," and a $\mathbb{C}$-valued random variable h^* with distribution π, known as the "target." As the prior is essentially arbitrary, the results below will hold for *any* prior π.

As an example, in the special case of the binary classifier learning problem studied by (Haussler et al., 1994a) and (Freund et al., 1997), $\mathbb{C}^*$ is the set of all measurable classifiers $h : \mathcal{X} \to \{-1, +1\}$, $\mathbb{C}$ is the "concept space," h^* is the "target function," and $\rho(h_1, h_2) = \mathbb{P}_{X \sim \mathcal{D}}(h_1(X) \neq h_2(X))$, where $\mathcal{D}$ is the distribution of the (unlabeled) data; in particular, $\rho(h, h^*) = er(h)$ is the "error rate" of h.

To discuss the fundamental limits of learning with binary-valued queries, we define the quantity $\mathrm{QueryComplexity}(\epsilon)$, for $\epsilon > 0$, as the minimum possible expected number of binary queries for any learning algorithm guaranteed to return $\hat{h}$ with $\mathbb{E}[\rho(\hat{h}, h^*)] \leq \epsilon$, where the only random variable in the expectation is $h^* \sim \pi$ (and $\hat{h}$, which is itself determined by h^* and the sequence of queries). For simplicity, we restrict ourselves to deterministic algorithms in this paper, so that the only source of randomness is h^*.

Alternatively, there is a particularly simple interpretation of the notion of an algorithm based on arbitrary binary-valued queries, which leads to an equivalent definition

[4] The set $\mathbb{C}^*$ will not play any significant role in the analysis, except to allow for improper learning scenarios to be a special case of our setting.

of $\mathrm{QueryComplexity}(\epsilon)$: namely, a prefix-free code. That is, any deterministic algorithm that asks a sequence of yes/no questions before terminating and returning some $\hat{h} \in \mathbb{C}^*$ can be thought of as a binary decision tree (no = left, yes = right), with the return $\hat{h}$ values stored in the leaf nodes. Transforming each root-to-leaf path in the decision tree into a codeword (left = 0, right = 1), we see that the algorithm corresponds to a prefix-free binary code. Conversely, given any prefix-free binary code, we can construct an algorithm based on sequentially asking queries of the form "what is the first bit in the codeword $C(h^*)$ for h^*?", "what is the second bit in the codeword $C(h^*)$ for h^*?", etc., until we obtain a complete codeword, at which point we return the value that codeword decodes to. From this perspective, we can state an equivalent definition of $\mathrm{QueryComplexity}(\epsilon)$ in the language of lossy codes.

Formally, a *code* is a pair of (measurable) functions (C, D). The *encoder*, C, maps any element $h \in \mathbb{C}$ to a binary sequence $C(h) \in \bigcup_{q=0}^{\infty}\{0,1\}^q$ (the *codeword*). The *decoder*, D, maps any element $c \in \bigcup_{q=0}^{\infty}\{0,1\}^q$ to an element $D(c) \in \mathbb{C}^*$. For any $q \in \{0, 1, \ldots\}$ and $c \in \{0,1\}^q$, let $|c| = q$ denote the *length* of c. A *prefix-free* code is any code (C, D) such that no $h_1, h_2 \in \mathbb{C}$ have $c^{(1)} = C(h_1)$ and $c^{(2)} = C(h_2)$ with $c^{(1)} \neq c^{(2)}$ but $\forall i \leq |c^{(1)}|$, $c_i^{(2)} = c_i^{(1)}$: that is, no codeword is a prefix of another (longer) codeword.

Here, we consider a setting where the code (C, D) may be *lossy*, in the sense that for some values of $h \in \mathbb{C}$, $\rho(D(C(h)), h) > 0$. Our objective is to design the code to have small expected loss (in the ρ sense), while maintaining as small of an expected codeword length as possible, where expectations are over the target h^*, which is also the element of $\mathbb{C}$ we encode. The following defines the optimal such length.

Definition 1. *For any $\epsilon > 0$, define the* query complexity *as*

$$\begin{aligned}&\mathrm{QueryComplexity}(\epsilon)\\&= \inf\left\{\mathbb{E}\Big[|C(h^*)|\Big] : (C,D) \textit{ is a prefix-free code with } \mathbb{E}\left[\rho\Big(D(C(h^*)), h^*\Big)\right] \leq \epsilon\right\},\end{aligned}$$

where the random variable in both expectations is $h^* \sim \pi$.

Recalling the equivalence between prefix-free binary codes and deterministic learning algorithms making arbitrary binary-valued queries, note that this definition is equivalent to the earlier definition.

Returning to the specialized setting of binary classification for a moment, we see that this corresponds to the minimum possible expected number of binary queries for a learning algorithm guaranteed to have expected error rate at most ϵ.

Given this coding perspective, we should not be surprised to see an entropy quantity appear in the results of the next section. Specifically, define the following quantities.

Definition 2. *For any $\epsilon > 0$, define $\mathcal{Y}(\epsilon) \subseteq \mathbb{C}$ as a maximal ϵ-packing of $\mathbb{C}$. That is, $\forall h_1, h_2 \in \mathcal{Y}(\epsilon)$, $\rho(h_1, h_2) \geq \epsilon$, and $\forall h \in \mathbb{C} \setminus \mathcal{Y}(\epsilon)$, the set $\mathcal{Y}(\epsilon) \cup \{h\}$ does not satisfy this property.*

For our purposes, if multiple maximal ϵ-packings are possible, we can choose to define $\mathcal{Y}(\epsilon)$ arbitrarily from among these; the results below hold for any such choice.

Recall that any maximal ϵ-packing of $\mathbb{C}$ is also an ϵ-cover of $\mathbb{C}$, since otherwise we would be able to add to $\mathcal{Y}(\epsilon)$ the $h \in \mathbb{C}$ that escapes the cover.

Next we define a complexity measure, a type of entropy, which serves as our primary quantity of interest in the analysis of $\mathrm{QueryComplexity}(\epsilon)$. It is specified in terms of a partition induced by $\mathcal{Y}(\epsilon)$, defined as follows.

Definition 3. *For any $\epsilon > 0$, define*

$$\mathcal{P}(\epsilon) = \left\{ \left\{ h \in \mathbb{C} : f = \operatorname*{argmin}_{g \in \mathcal{Y}(\epsilon)} \rho(h, g) \right\} : f \in \mathcal{Y}(\epsilon) \right\},$$

where we break ties in the argmin *arbitrarily but consistently (e.g., based on a predefined preference ordering of $\mathcal{Y}(\epsilon)$). If the* argmin *is not defined (i.e., the* min *is not realized), take any $f \in \mathcal{Y}(\epsilon)$ with $\rho(f, h) \leq \epsilon$ (one must exist by maximality of $\mathcal{Y}(\epsilon)$).*

Definition 4. *For any finite (or countable) partition $\mathcal{S}$ of $\mathbb{C}$ into measurable regions (subsets), define the* entropy *of $\mathcal{S}$*

$$\mathcal{H}(\mathcal{S}) = -\sum_{S \in \mathcal{S}} \pi(S) \log_2 \pi(S).$$

In particular, we will be interested in the quantity $\mathcal{H}(\mathcal{P}(\epsilon))$ in the analysis below.

Finally, we will require a notion of dimensionality for the pseudo-metric ρ. For this, we adopt the well-known *doubling dimension* (Gupta et al., 2003).

Definition 5. *Define the* doubling dimension *d as the smallest value d such that, for any $h \in \mathbb{C}$, and any $\epsilon > 0$, the size of the minimal $\epsilon/2$-cover of the ϵ-radius ball around h is at most 2^d.*

That is, for any $h \in \mathbb{C}$ and $\epsilon > 0$, there exists a set $\{h_i\}_{i=1}^{2^d}$ of 2^d elements of $\mathbb{C}$ such that

$$\{h' \in \mathbb{C} : \rho(h', h) \leq \epsilon\} \subseteq \bigcup_{i=1}^{2^d} \{h' \in \mathbb{C} : \rho(h', h_i) \leq \epsilon/2\}.$$

Note that, as defined here, d is a constant (i.e., has no dependence on h or ϵ). See (Bshouty et al., 2009) for a discussion of the doubling dimension of spaces $\mathbb{C}$ of binary classifiers, in the context of learning theory.

3 Main Result

Our main result can be summarized as follows. Note that, since we took the prior to be *arbitrary* in the above definitions, this result holds for *any* prior π.

Theorem 1. *If $d < \infty$ and $\bar{\rho} < \infty$, then there is a constant $c = O(d)$ such that $\forall \epsilon \in (0, \bar{\rho}/2)$,*

$$\mathcal{H}\left(\mathcal{P}\left(\epsilon \log_2(\bar{\rho}/\epsilon)\right)\right) - c \leq \mathrm{QueryComplexity}(\epsilon) \leq \mathcal{H}\left(\mathcal{P}\left(\epsilon\right)\right) + 1.$$

Due to the deep connections of this problem to information theory, it should not be surprising that entropy terms play a key role in this result. Indeed, this type of entropy seems to give a good characterization of the asymptotic behavior of the query complexity in this setting. We should expect the upper bound to be tight when the regions in $\mathcal{P}(\epsilon)$ are point-wise well-separated. However, it may be looser when this is not the case, for reasons discussed in the next section.

Although this result is stated for bounded psuedometrics ρ, it also has implications for unbounded ρ. In particular, the proof of the upper bound holds as-is for unbounded ρ. Furthermore, we can always use this lower bound to construct a lower bound for unbounded ρ, simply restricting to a bounded subset of $\mathbb{C}$ with constant probability and calculating the lower bound for that region. For instance, to get a lower bound for π being a Gaussian distribution on $\mathbb{R}$, we might note that $\pi([-1/2, 1/2])$ times the expected error rate under the *conditional* $\pi(\cdot|[-1/2, 1/2])$ lower bounds the total expected error rate. Thus, calculating the lower bound of Theorem 1 under the conditional $\pi(\cdot|[-1/2, 1/2])$ while replacing ϵ with $\epsilon/\pi([-1/2, 1/2])$ provides a lower bound on $\mathrm{QueryComplexity}(\epsilon)$.

4 Proof of Theorem 1

We first state a lemma that will be useful in the proof.

Lemma 1. *(Gupta et al., 2003) For any $\gamma \in (0, \infty)$, $\delta \in [\gamma, \infty)$, and $h \in \mathbb{C}$, we have*

$$|\{h' \in \mathcal{Y}(\gamma) : \rho(h', h) \leq \delta\}| \leq \left(\frac{4\delta}{\gamma}\right)^d.$$

Proof. See (Gupta et al., 2003). □

Proof (of Theorem 1). Throughout the proof, we will consider a set-valued random quantity $P_\epsilon(h^*)$ with value equal to the set in $\mathcal{P}(\epsilon)$ containing h^*, and a corresponding $\mathbb{C}$-valued random quantity $Y_\epsilon(h^*)$ with value equal the sole point in $P_\epsilon(h^*) \cap \mathcal{Y}(\epsilon)$: that is, the target's nearest representative in the ϵ-packing. Note that, by Lemma 1, $|\mathcal{Y}(\epsilon)| < \infty$ for all $\epsilon \in (0, 1)$. We will also adopt the usual notation for entropy (e.g., $\mathcal{H}(P_\epsilon(h^*))$) and conditional entropy (e.g., $\mathcal{H}(P_\epsilon(h^*)|X)$), both in base 2; see (Cover & Thomas, 2006) for definitions.

To establish the upper bound, we simply take C as the Huffman code for the random quantity $P_\epsilon(h^*)$ (Cover & Thomas, 2006). It is well-known that the expected length of a Huffman code for $P_\epsilon(h^*)$ is at most $\mathcal{H}(P_\epsilon(h^*)) + 1$ (in fact, is equal $\mathcal{H}(P_\epsilon(h^*))$ when the probabilities are powers of 2) (Cover & Thomas, 2006), and each possible value of $P_\epsilon(h^*)$ is assigned a unique codeword so that we can perfectly recover $P_\epsilon(h^*)$ (and thus also $Y_\epsilon(h^*)$) based on $C(h^*)$. In particular, define $D(C(h^*)) = Y_\epsilon(h^*)$. Finally, recall that any maximum ϵ-packing is also an ϵ-*cover*; that is, for every $h \in \mathbb{C}$, there is at least one $h' \in \mathcal{Y}(\epsilon)$ with $\rho(h, h') \leq \epsilon$ (otherwise, we could add h to the packing, contradicting its maximality). Thus, since every element of the set $P_\epsilon(h^*)$ has $Y_\epsilon(h^*)$ as its closest representative in $\mathcal{Y}(\epsilon)$, we must have $\rho(h^*, D(C(h^*))) = \rho(h^*, Y_\epsilon(h^*)) \leq \epsilon$.

In fact, as this proof never relies on $d < \infty$ or $\bar{\rho} < \infty$, this establishes the upper bound even in the case $d = \infty$ or $\bar{\rho} = \infty$.

The proof of the lower bound is somewhat more involved, though the overall idea is simple enough. Essentially, the lower bound would be straightforward if the regions of $\mathcal{P}(\epsilon \log_2(\bar{\rho}/\epsilon))$ were separated by some distance, since we could make an argument based on Fano's inequality to say that since any $\hat{h}$ is "close" to at most one region, the expected distance from h^* is at least as large as half this inter-region distance times a quantity proportional to the entropy. However, it is not always so simple, as the regions can generally be quite close to each other (even adjacent), so that it is possible for $\hat{h}$ to be close to multiple regions. Thus, the proof will first "color" the regions of $\mathcal{P}(\epsilon \log_2(\bar{\rho}/\epsilon))$ in a way that guarantees no two regions of the same color are within distance $\epsilon \log_2(\bar{\rho}/\epsilon)$ of each other. Then we apply the above simple argument for each color separately (i.e., lower bounding the expected distance from h^* under the conditional given the color of $P_{\epsilon \log_2(\bar{\rho}/\epsilon)}(h^*)$ by a function of the entropy under the conditional), and average over the colors to get a global lower bound. The details follow.

Fix any $\epsilon \in (0, \bar{\rho}/2)$, and for brevity let $\alpha = \epsilon \log_2(\bar{\rho}/\epsilon)$. We suppose (C, D) is some prefix-free binary code (representing the learning algorithm's queries and return policy).

Define a function $\mathcal{K} : \mathcal{P}(\alpha) \to \mathbb{N}$ such that $\forall P_1, P_2 \in \mathcal{P}(\alpha)$,

$$\mathcal{K}(P_1) = \mathcal{K}(P_2) \implies \inf_{h_1 \in P_1, h_2 \in P_2} \rho(h_1, h_2) \geq \alpha, \tag{1}$$

and suppose $\mathcal{K}$ has minimum $\mathcal{H}(\mathcal{K}(P_\alpha(h^*)))$ subject to (1). We will refer to $\mathcal{K}(P)$ as the *color* of P.

Now we are ready to bound the expected distance from h^*. Let $\hat{h} = D(C(h^*))$ denote the element returned by the algorithm (decoder), and let $P_\alpha(\hat{h}; \mathcal{K})$ denote the set $P \in \mathcal{P}(\alpha)$ having $\mathcal{K}(P) = \mathcal{K}$ with smallest $\inf_{h \in P} \rho(h, \hat{h})$ (breaking ties arbitrarily). We know

$$\mathbb{E}[\rho(\hat{h}, h^*)] = \mathbb{E}\left[\mathbb{E}[\rho(\hat{h}, h^*) | \mathcal{K}(P_\alpha(h^*))]\right]. \tag{2}$$

Furthermore, by (1) and a triangle inequality, we know no $\hat{h}$ can be $\alpha/3$-close to more than one $P \in \mathcal{P}(\alpha)$ of a given color. Therefore,

$$\mathbb{E}[\rho(\hat{h}, h^*) | \mathcal{K}(P_\alpha(h^*))] \geq \frac{\alpha}{3} \mathbb{P}(P_\alpha(\hat{h}; \mathcal{K}(P_\alpha(h^*))) \neq P_\alpha(h^*) | \mathcal{K}(P_\alpha(h^*))). \tag{3}$$

By Fano's inequality, we have

$$\mathbb{E}\Big[\mathbb{P}(P_\alpha(\hat{h}; \mathcal{K}(P_\alpha(h^*))) \neq P_\alpha(h^*) | \mathcal{K}(P_\alpha(h^*)))\Big] \geq \frac{\mathcal{H}(P_\alpha(h^*) | C(h^*), \mathcal{K}(P_\alpha(h^*))) - 1}{\log_2 |\mathcal{Y}(\alpha)|}. \tag{4}$$

It is generally true that, for a prefix-free binary code $C(h^*)$, $C(h^*)$ is a lossless prefix-free binary code for itself (i.e., with the identity decoder), so that the classic entropy lower bound on average code length (Cover & Thomas, 2006) implies $\mathcal{H}(C(h^*)) \leq \mathbb{E}[|C(h^*)|]$. Also, recalling that $\mathcal{Y}(\alpha)$ is maximal, and therefore also an α-cover, we have that any $P_1, P_2 \in \mathcal{P}(\alpha)$ with $\inf_{h_1 \in P_1, h_2 \in P_2} \rho(h_1, h_2) \leq \alpha$ have $\rho(Y_\alpha(h_1), Y_\alpha(h_2)) \leq 3\alpha$ (by a triangle inequality). Therefore, Lemma 1 implies that, for any given $P_1 \in$

$\mathcal{P}(\alpha)$, there are at most 12^d sets $P_2 \in \mathcal{P}(\alpha)$ with $\inf_{h_1 \in P_1, h_2 \in P_2} \rho(h_1, h_2) \leq \alpha$. We therefore know there exists a function $\mathcal{K}' : \mathcal{P}(\alpha) \to \mathbb{N}$ satisfying (1) such that $\max_{P \in \mathcal{P}(\alpha)} \mathcal{K}'(P) \leq 12^d$ (i.e., we need at most 12^d colors to satisfy (1)). That is, if we consider coloring the sets $P \in \mathcal{P}(\alpha)$ sequentially, for any given P_1 not yet colored, there are $< 12^d$ sets $P_2 \in \mathcal{P}(\alpha) \setminus \{P_1\}$ within α of it, so there must exist a color among $\{1, \ldots, 12^d\}$ not used by any of them, and we can choose that for $\mathcal{K}'(P_1)$. In particular, by our choice of $\mathcal{K}$ to minimize $\mathcal{H}(\mathcal{K}(P_\alpha(h^*)))$ subject to (1), this implies

$$\mathcal{H}(\mathcal{K}(P_\alpha(h^*))) \leq \mathcal{H}(\mathcal{K}'(P_\alpha(h^*))) \leq \log_2(12^d) \leq 4d.$$

Thus,

$$\begin{aligned}
&\mathcal{H}(P_\alpha(h^*)|C(h^*), \mathcal{K}(P_\alpha(h^*))) && (5)\\
&= \mathcal{H}(P_\alpha(h^*), C(h^*), \mathcal{K}(P_\alpha(h^*))) - \mathcal{H}(C(h^*)) - \mathcal{H}(\mathcal{K}(P_\alpha(h^*))|C(h^*)) && (6)\\
&\geq \mathcal{H}(P_\alpha(h^*)) - \mathcal{H}(C(h^*)) - \mathcal{H}(\mathcal{K}(P_\alpha(h^*))) \geq \mathcal{H}(P_\alpha(h^*)) - \mathbb{E}\left[|C(h^*)|\right] - 4d &&\\
&= \mathcal{H}(\mathcal{P}(\alpha)) - \mathbb{E}\left[|C(h^*)|\right] - 4d. && (7)
\end{aligned}$$

Thus, combining (2), (3), (4), and (7), we have

$$\begin{aligned}
\mathbb{E}[\rho(\hat{h}, h^*)] &\geq \frac{\alpha}{3} \frac{\mathcal{H}(\mathcal{P}(\alpha)) - \mathbb{E}\left[|C(h^*)|\right] - 4d - 1}{\log_2 |\mathcal{Y}(\alpha)|}\\
&\geq \frac{\alpha}{3} \frac{\mathcal{H}(\mathcal{P}(\alpha)) - \mathbb{E}\left[|C(h^*)|\right] - 4d - 1}{d \log_2(4\bar{\rho}/\alpha)},
\end{aligned}$$

where the last inequality follows from Lemma 1.

Thus, for any code with

$$\mathbb{E}\left[|C(h^*)|\right] < \mathcal{H}(\mathcal{P}(\alpha)) - 4d - 1 - 3d \frac{\log_2(4\bar{\rho}/\epsilon)}{\log_2(\bar{\rho}/\epsilon)},$$

we have $\mathbb{E}[\rho(\hat{h}, h^*)] > \epsilon$, which implies

$$\mathrm{QueryComplexity}(\epsilon) \geq \mathcal{H}(\mathcal{P}(\alpha)) - 4d - 1 - 3d \frac{\log_2(4\bar{\rho}/\epsilon)}{\log_2(\bar{\rho}/\epsilon)}.$$

Since $\log_2(4\bar{\rho}/\epsilon)/\log_2(\bar{\rho}/\epsilon) \leq 3$, we have

$$\mathrm{QueryComplexity}(\epsilon) = \mathcal{H}(\mathcal{P}(\alpha)) - O(d).$$

□

5 Open Problems

Generally, we feel this topic of Bayesian active learning is relatively unexplored, and as such there is an abundance of ripe open problems ready for solvers.

In our present context, there are several interesting questions, such as whether the $\log(\bar{\rho}/\epsilon)$ factor in the entropy argument of the lower bound can be removed, whether the additive constant in the lower bound might be improved, and in particular whether a similar result might be obtained without assuming $d < \infty$ (e.g., by making a VC class assumption instead).

Additionally, one can ask for necessary and sufficient conditions for this entropy lower bound to be achievable via a restricted type of query, such as label requests (membership queries).

Overall, the challenge here is to understand, to as large an extent as possible, how much benefit we get from having access to the prior, and what the general form of improvements we can expect in the query complexity given this information are.

Acknowledgments

Liu Yang would like to extend her sincere gratitude to Avrim Blum and Venkatesan Guruswami for several enlightening and highly stimulating discussions.

Bibliography

Bshouty, N. H., Li, Y., & Long, P. M. (2009). Using the doubling dimension to analyze the generalization of learning algorithms. *Journal of Computer and System Sciences*, *75*, 323–335.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons, Inc.

Dasgupta, S. (2005). Coarse sample complexity bounds for active learning. *In Advances in Neural Information Processing Systems 18*.

Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*, *28*, 133–168.

Gupta, A., Krauthgamer, R., & Lee, J. R. (2003). Bounded geometries, fractals, and low-distortion embeddings. *In Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*.

Hanneke, S. (2007a). A bound on the label complexity of agnostic active learning. *In Proceedings of the 24th International Conference on Machine Learning*.

Hanneke, S. (2007b). Teaching dimension and the complexity of active learning. *In Proceedings of the 20th Annual Conference on Learning Theory*.

Haussler, D., Kearns, M., & Schapire, R. (1994a). Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning*, *14*, 83–113.

Haussler, D., Littlestone, N., & Warmuth, M. (1994b). Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, *115*, 248–292.

Kulkarni, S. R., Mitter, S. K., & Tsitsiklis, J. N. (1993). Active learning using arbitrary binary valued queries. *Machine Learning*, *11*, 23–35.

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. *In Proceedings of the 5th Workshop on Computational Learning Theory*.