Characterizing Optimal Rates for Lossy Coding with Finite-Dimensional Metrics

Liu Yang, Steve Hanneke, and Jaime Carbonell

***Abstract*—We investigate the minimum expected number of bits sufficient to encode a random variable X while still being able to recover an approximation of X with expected distance from X at most D: that is, the optimal rate at distortion D, in a one-shot coding setting. We find this quantity is related to the entropy of a Voronoi partition of the values of X based on a maximal D-packing.**

***Index Terms*—Quantization, Lossy Coding, Binary Codes, Bayesian Learning, Active Learning**

I. Introduction

IN this work, we study the fundamental complexity of lossy coding. We are particularly interested in identifying a key quantity that characterizes the expected number of bits (called the *rate*) required to encode a random variable so that we may recover an approximation within expected distance D (called the *distortion*). This topic is a generalization of the well-known analysis of exact coding by Shannon [1], where it is known that the optimal expected number of bits is precisely characterized by the entropy. There are many problems in which exact coding is not practical or not possible, so that lossy coding becomes necessary: particularly for random variables taking values in uncountably infinite spaces. The topic of code lengths for lossy coding is interesting, both for its direct applications to compression, and also as a general setting in which to derive lower bounds for specializations of the setting.

There is much existing work on lossy binary codes. In the present work, we are interested in a "one-shot" analysis of lossy coding [2], in which we wish to encode a single random variable, in contrast to the analysis of "asymptotic" source coding [3], in which one wishes to simultaneously encode a sequence of random variables. Of particular relevance to the one-shot coding problem is the analysis of *quantization* methods that balance *distortion* with *entropy* [2], [4], [5]. In particular, it is now well-known that this approach can yield codes that respect a distortion contraint while nearly minimizing the rate, so that there are near-optimal codes of this type [2]. Thus, we have an alternative way to think of the optimal rate, in terms of the rate of the best distortion-constrained quantization method. While this is interesting, in that it allows us to restrict our focus in the design of effective coding techniques, it is not as directly helpful if we wish to understand the behavior of the optimal rate itself. That is, since we do not have an explicit description of the optimal quantizer, it may often be difficult to study the behavior of its rate under various interesting conditions. There exist classic results lower bounding the achievable rates, most notably the famous Shannon lower bound [6], which under certain restrictions on the source and the distortion metric, is known to be fairly tight in the *asymptotic* analysis of source coding [7]. However, there are few general results explicitly and tightly characterizing the (non-asymptotic) optimal rates for one-shot coding. In particular, to our knowledge, only a few special-case calculations of the exact value of this optimal rate have been explicitly carried out, such as vectors of independent Bernoulli or Gaussian random variables [3].

Below, we discuss a particular distortion-constrained quantizer, based on a Voronoi partition induced by a maximal packing. We are interested in the *entropy* of this quantizer, as a quantity used to characterize the optimal rate for codes of a given distortion. While it is clear that this entropy upper bounds the optimal rate, as this is the case for *any* distortion-constrained quantizer [2], the novelty of our analysis lies in noting the remarkable fact that the entropy of any quantizer constructed in this way also *lower bounds* the optimal rate. In particular, this provides a method for approximately calculating the optimal rate without the need to optimize over all possible quantizers. Our result is general, in that it applies to an arbitrary distribution and an arbitrary distortion measure from a general class of finite-dimensional pseudo-metrics. This generality is noteworthy, as it leads to interesting applications in statistical learning theory, which we describe below.

Our analysis is closely related to various notions that arise in the study of ϵ-entropy [8], [9], in that we are concerned with the entropy of a Voronoi partition induced by an ϵ-cover. The notion of ϵ-entropy has been related to the optimal rates for a given distortion (under a slightly different model than studied here) [8], [9]. However, there are some important distinctions, perhaps the most significant of which is that calculating the ϵ-entropy requires a prohibitive optimization of the entropy over all ϵ-covers; in contrast, the entropy term in our analysis can be calculated based on *any* maximal ϵ-packing (which is a particular type of ϵ-cover). Maximal ϵ-packings are easy to construct by greedily adding arbitrary new elements to the packing that are ϵ-far from all elements already added; thus, there is always a straightforward algorithmic approach to applying our results.

Liu Yang is with the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA, email: `liuy@cs.cmu.edu`.

Steve Hanneke is with the Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 USA, email: `shanneke@stat.cmu.edu`.

Jaime Carbonell is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA, email: `jgc@cs.cmu.edu`.

Manuscript recieved SomeMonth xx, 2010; revised SomeMonth xx, 2010.

II. Definitions

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

A *code* is a pair of (measurable) functions (ϕ, ψ). The *encoder*, ϕ, maps any element $x \in \mathcal{X}$ to a binary sequence $\phi(x) \in \bigcup_{q=0}^{\infty}\{0,1\}^q$ (the *codeword*). The *decoder*, ψ, maps any element $c \in \bigcup_{q=0}^{\infty}\{0,1\}^q$ to an element $\psi(c) \in \mathcal{X}^*$. 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 (ϕ, ψ) such that no $x_1, x_2 \in \mathcal{X}$ have $c^{(1)} = \phi(x_1)$ and $c^{(2)} = \phi(x_2)$ 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. Let PF denote the set of all prefix-free binary codes.

Here, we consider a setting where the code (ϕ, ψ) may be *lossy*, in the sense that for some values of $x \in \mathcal{X}$, $\rho(\psi(\phi(x)), x) > 0$. 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. Formally, we have the following definition, which essentially describes a notion of optimality for a lossy code.

Definition 1. *For any* $\mathrm{D} > 0$*, define the optimal* rate *at distortion* D

$$\mathrm{R}(\mathrm{D}) = \inf\Big\{\mathbb{E}\Big[|\phi(X)|\Big] : (\phi,\psi) \in \mathrm{PF} \textit{ with } \mathbb{E}\left[\rho\Big(\psi(\phi(X)), X\Big)\right] \leq \mathrm{D}\Big\},$$

where the random variable in both expectations is $X \sim \pi$.

For our analysis, we will require a notion of dimensionality for the pseudo-metric ρ. For this, we adopt the well-known *doubling dimension* [10].

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

That is, for any $x \in \mathcal{X}$ *and* $\epsilon > 0$*, there exists a set* $\{x_i\}_{i=1}^{2^d}$ *of* 2^d *elements of* $\mathcal{X}$ *such that*

$$\{x' \in \mathcal{X} : \rho(x', x) \leq \epsilon\} \subseteq \bigcup_{i=1}^{2^d}\{x' \in \mathcal{X} : \rho(x', x_i) \leq \epsilon/2\}.$$

Note that, as defined here, d is a constant (i.e., has no dependence on the x or ϵ in its definition). In the analysis below, we will always assume $d < \infty$. The doubling dimension has been studied for a variety of spaces, originally by Gupta, Krauthgamer, & Lee [10], and subsequently by many others. In particular, Bshouty, Li, & Long [11] discuss the doubling dimension of spaces $\mathcal{X}$ of binary classifiers, in the context of statistical learning theory.

A. Definition of Packing Entropy

Our main result concerns the relation between the optimal rate at a given distortion with the entropy of a certain quantizer. We now turn to defining this latter quantity.

Definition 3. *For any* $\mathrm{D} > 0$*, define* $\mathcal{Y}(\mathrm{D}) \subseteq \mathcal{X}$ *as a maximal* D*-packing of* $\mathcal{X}$*. That is,* $\forall x_1, x_2 \in \mathcal{Y}(\mathrm{D})$*,* $\rho(x_1, x_2) \geq \mathrm{D}$*, and* $\forall x \in \mathcal{X} \setminus \mathcal{Y}(\mathrm{D})$*,* $\min_{x' \in \mathcal{Y}(\mathrm{D})} \rho(x, x') < \mathrm{D}$.

For our purposes, if multiple maximal D-packings are possible, we can choose to define $\mathcal{Y}(\mathrm{D})$ arbitrarily from among these; the results below hold for any such choice. Recall that any maximal D-packing of $\mathcal{X}$ is also a D-cover of $\mathcal{X}$, since otherwise we would be able to add to $\mathcal{Y}(\mathrm{D})$ the $x \in \mathcal{X}$ that escapes the cover. That is, $\forall x \in \mathcal{X}$, $\exists y \in \mathcal{Y}(\mathrm{D})$ s.t. $\rho(x, y) < \mathrm{D}$.

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

Definition 4. *For any* $\mathrm{D} > 0$*, define*

$$\mathcal{Q}(\mathrm{D}) = \left\{\left\{x \in \mathcal{X} : z = \operatorname*{argmin}_{y \in \mathcal{Y}(\mathrm{D})} \rho(x, y)\right\} : z \in \mathcal{Y}(\mathrm{D})\right\},$$

where we break ties in the argmin *arbitrarily but consistently (e.g., based on a predefined preference ordering of* $\mathcal{Y}(\mathrm{D})$*).*

Definition 5. *For any finite (or countable) partition* $\mathcal{S}$ *of* $\mathcal{X}$ *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 $\mathcal{H}(\mathcal{Q}(\mathrm{D}))$ in the analysis below.

III. Main Result

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

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

$$\mathcal{H}\left(\mathcal{Q}\left(\mathrm{D}\log_2(\bar{\rho}/\mathrm{D})\right)\right) - c \leq \mathrm{R}(\mathrm{D}) \leq \mathcal{H}\left(\mathcal{Q}\left(\mathrm{D}\right)\right) + 1.$$

It should not be surprising that entropy terms play a key role in this result, as the entropy is essential to the analysis of exact coding [1]. Furthermore, $\mathrm{R}(\mathrm{D})$ is tightly characterized by the minimum achievable entropy among all quantizers of distortion at most D [2]. The interesting aspect of Theorem 1 is that we can explicitly describe a particular quantizer with near-optimal rate, and its entropy can be explicitly calculated for a variety of scenarios $(\mathcal{X}, \rho, \pi)$. As for the behavior of $\mathrm{R}(\mathrm{D})$ within the range between the upper and lower bounds of Theorem 1, we should expect the upper bound to be tight when high-probability subsets of the regions in $\mathcal{Q}(\mathrm{D})$ are point-wise

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

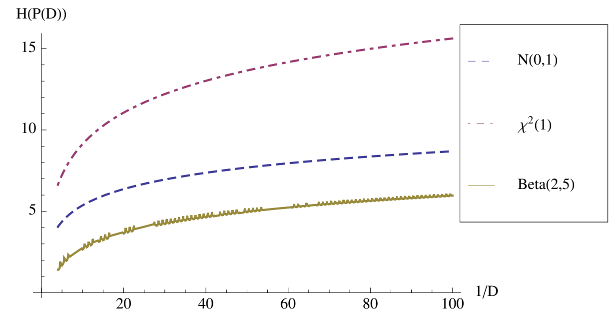

Fig. 1. Plots of $\mathcal{H}(\mathcal{Q}(\mathrm{D}))$ as a function of $1/\mathrm{D}$, for various distributions π on $\mathcal{X} = \mathbb{R}$.

well-separated, while $\mathrm{R}(\mathrm{D})$ may be much smaller (perhaps closer to the lower bound) when this is violated to a large degree, for reasons described in the proof below.

Although this result is stated for bounded psuedo-metrics 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 $\mathcal{X}$ with constant probability and calculating the lower bound for that region. For instance, to get a lower bound for π as a Gaussian distribution on $\mathbb{R}$, we might note that $\pi([-1/2, 1/2])$ times the expected loss under the *conditional* $\pi(\cdot|[-1/2, 1/2])$ lower bounds the total expected loss. Thus, calculating the lower bound of Theorem 1 under the conditional $\pi(\cdot|[-1/2, 1/2])$ while replacing D with $\mathrm{D}/\pi([-1/2, 1/2])$ provides a lower bound on $\mathrm{R}(\mathrm{D})$.

To get a feel for the behavior of $\mathcal{H}\left(\mathcal{Q}\left(\mathrm{D}\right)\right)$, we have plotted it as a function of $1/\mathrm{D}$ for several distributions, in Figure 1.

IV. Proof of Theorem 1

We first state a lemma, due to Gupta, Krauthgamer, & Lee [10], which will be useful in the proof of Theorem 1.

Lemma 1. *[10] For any $\gamma \in (0, \infty)$, $\delta \in [\gamma, \infty)$, and $x \in \mathcal{X}$,*

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

In particular, note that this lemma implies that the minimum of $\rho(x, y)$ over $y \in \mathcal{Y}(\mathrm{D})$ is always *achieved* in Definition 4, so that $\mathcal{Q}(\mathrm{D})$ is well-defined.

We are now ready for the proof of Theorem 1.

Proof of Theorem 1: Throughout the proof, we will consider a set-valued random quantity $Q_{\mathrm{D}}(X)$ with value equal to the set in $\mathcal{Q}(\mathrm{D})$ containing X, and a corresponding $\mathcal{X}$-valued random quantity $Y_{\mathrm{D}}(X)$ with value equal the sole point in $Q_{\mathrm{D}}(X) \cap \mathcal{Y}(\mathrm{D})$: that is, the target's nearest representative in the D-packing. Note that, by Lemma 1, $|\mathcal{Y}(\mathrm{D})| < \infty$ for all $\mathrm{D} \in (0, 1)$. We will also adopt the usual notation for entropy (e.g., $\mathcal{H}(Q_{\mathrm{D}}(X))$) and conditional entropy (e.g., $\mathcal{H}(Q_{\mathrm{D}}(X)|Z)$) [3], both in base 2.

To establish the upper bound, we simply take ϕ as the Huffman code for the random quantity $Q_{\mathrm{D}}(X)$ [3], [12]. It is well-known that the expected length of a Huffman code for $Q_{\mathrm{D}}(X)$ is at most $\mathcal{H}(Q_{\mathrm{D}}(X))+1$ (in fact, is equal $\mathcal{H}(Q_{\mathrm{D}}(X))$ when the probabilities are powers of 2) [3], [12], and each possible value of $Q_{\mathrm{D}}(X)$ is assigned a unique codeword so that we can perfectly recover $Q_{\mathrm{D}}(X)$ (and thus also $Y_{\mathrm{D}}(X)$) based on $\phi(X)$. In particular, define $\psi(\phi(X)) = Y_{\mathrm{D}}(X)$. Finally, recall that any maximal D-packing is also a D-*cover*. Thus, since every element of the set $Q_{\mathrm{D}}(X)$ has $Y_{\mathrm{D}}(X)$ as its closest representative in $\mathcal{Y}(\mathrm{D})$, we must have $\rho(X, \psi(\phi(X))) = \rho(X, Y_{\mathrm{D}}(X)) < \mathrm{D}$. In fact, as this proof never relies on $\bar{\rho} < \infty$, this establishes the upper bound even in the case $\bar{\rho} = \infty$.

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 $\mathcal{Q}(\mathrm{D}\log_2(\bar{\rho}/\mathrm{D}))$ were separated by some distance, since we could make an argument based on Fano's inequality to say that since any $\hat{X} = \psi(\phi(X))$ is "close" to at most one region, the expected distance from X is at least as large as half this inter-region distance times a quantity proportional to the conditional entropy $\mathcal{H}(Q_{\mathrm{D}}(X)|\phi(X))$, so that $\mathcal{H}(\phi(X))$ can be related to $\mathcal{H}(Q_{\mathrm{D}}(X))$.

However, the general case is not always so simple, as the regions can generally be quite close to each other (even adjacent), so that it is possible for $\hat{X}$ to be close to multiple regions. Thus, the proof will first "color" the regions of $\mathcal{Q}(\mathrm{D}\log_2(\bar{\rho}/\mathrm{D}))$ in a way that guarantees no two regions of the same color are within distance $\mathrm{D}\log_2(\bar{\rho}/\mathrm{D})$ of each other. Then we apply the above simple argument for each color separately (i.e., lower bounding the expected distance from X under the conditional given the color of $Q_{\mathrm{D}\log_2(\bar{\rho}/\mathrm{D})}(X)$ by a function of the conditional entropy under the conditional), and average over the colors to get a global lower bound. The details follow.

Fix any $\mathrm{D} \in (0, \bar{\rho}/2)$, and for brevity let $\alpha = \mathrm{D}\log_2(\bar{\rho}/\mathrm{D})$. We suppose (ϕ, ψ) is some prefix-free binary code.

Define a function $\mathcal{K} : \mathcal{Q}(\alpha) \rightarrow \mathbb{N}$ such that $\forall Q_1, Q_2 \in \mathcal{Q}(\alpha)$,

$$\mathcal{K}(Q_1) = \mathcal{K}(Q_2) \implies \inf_{x_1 \in Q_1, x_2 \in Q_2} \rho(x_1, x_2) \geq \alpha, \qquad (1)$$

and suppose $\mathcal{K}$ has minimum $\mathcal{H}(\mathcal{K}(Q_\alpha(X)))$ subject to (1). We will refer to $\mathcal{K}(Q)$ as the *color* of Q.

Now we are ready to bound the expected distance from X. Let $\hat{X} = \psi(\phi(X))$, and let $Q_\alpha(\hat{X}; \mathcal{K})$ denote the set $Q \in \mathcal{Q}(\alpha)$ having $\mathcal{K}(Q) = \mathcal{K}$ with smallest $\inf_{x \in Q} \rho(x, \hat{X})$ (breaking ties arbitrarily). We know

$$\mathbb{E}[\rho(\hat{X}, X)] = \mathbb{E}\left[\mathbb{E}[\rho(\hat{X}, X)|\mathcal{K}(Q_\alpha(X))]\right]. \qquad (2)$$

Furthermore, by (1) and a triangle inequality, we know no $\hat{X}$ can be closer than $\alpha/2$ to more than one $Q \in \mathcal{Q}(\alpha)$ of a given color. Therefore,

$$\begin{aligned}&\mathbb{E}[\rho(\hat{X}, X)|\mathcal{K}(Q_\alpha(X))]\\ &\quad \geq \frac{\alpha}{2}\mathbb{P}(Q_\alpha(\hat{X}; \mathcal{K}(Q_\alpha(X))) \neq Q_\alpha(X)|\mathcal{K}(Q_\alpha(X))). \qquad (3)\end{aligned}$$

By Fano's inequality, we have

$$\mathbb{E}\left[\mathbb{P}(Q_\alpha(\hat{X};\mathcal{K}(Q_\alpha(X))) \neq Q_\alpha(X)|\mathcal{K}(Q_\alpha(X)))\right] \geq \frac{\mathcal{H}(Q_\alpha(X)|\phi(X),\mathcal{K}(Q_\alpha(X))) - 1}{\log_2|\mathcal{Y}(\alpha)|}. \quad (4)$$

It is generally true that, for a prefix-free binary code $\phi(X)$, $\phi(X)$ 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 [1], [3] implies $\mathcal{H}(\phi(X)) \leq \mathbb{E}[|\phi(X)|]$. Also, recalling that $\mathcal{Y}(\alpha)$ is maximal, and therefore also an α-cover, we have that any $Q_1, Q_2 \in \mathcal{Q}(\alpha)$ with $\inf_{x_1\in Q_1, x_2\in Q_2} \rho(x_1,x_2) \leq \alpha$ have $\rho(Y_\alpha(x_1), Y_\alpha(x_2)) \leq 3\alpha$ (by a triangle inequality). Therefore, Lemma 1 implies that, for any given $Q_1 \in \mathcal{Q}(\alpha)$, there are at most 12^d sets $Q_2 \in \mathcal{Q}(\alpha)$ with $\inf_{x_1\in Q_1, x_2\in Q_2} \rho(x_1,x_2) \leq \alpha$. We therefore know there exists a function $\mathcal{K}' : \mathcal{Q}(\alpha) \to \mathbb{N}$ satisfying (1) such that $\max_{Q\in\mathcal{Q}(\alpha)} \mathcal{K}'(Q) \leq 12^d$ (i.e., we need at most 12^d colors to satisfy (1)). That is, if we consider coloring the sets $Q \in \mathcal{Q}(\alpha)$ sequentially, for any given Q_1 not yet colored, there are $< 12^d$ sets $Q_2 \in \mathcal{Q}(\alpha) \setminus \{Q_1\}$ 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 $\mathcal{K}'(Q_1)$. In particular, by our choice of $\mathcal{K}$ to minimize $\mathcal{H}(\mathcal{K}(Q_\alpha(X)))$ subject to (1), this implies

$$\mathcal{H}(\mathcal{K}(Q_\alpha(X))) \leq \mathcal{H}(\mathcal{K}'(Q_\alpha(X))) \leq \log_2(12^d) \leq 4d.$$

Thus,

$$\begin{aligned}
&\mathcal{H}(Q_\alpha(X)|\phi(X),\mathcal{K}(Q_\alpha(X))) \\
&= \mathcal{H}(Q_\alpha(X),\phi(X),\mathcal{K}(Q_\alpha(X))) \\
&\qquad - \mathcal{H}(\phi(X)) - \mathcal{H}(\mathcal{K}(Q_\alpha(X))|\phi(X)) \\
&\geq \mathcal{H}(Q_\alpha(X)) - \mathcal{H}(\phi(X)) - \mathcal{H}(\mathcal{K}(Q_\alpha(X))) \\
&\geq \mathcal{H}(Q_\alpha(X)) - \mathbb{E}\left[|\phi(X)|\right] - 4d \\
&= \mathcal{H}(\mathcal{Q}(\alpha)) - \mathbb{E}\left[|\phi(X)|\right] - 4d. \qquad (5)
\end{aligned}$$

Thus, combining (2), (3), (4), and (5), we have

$$\begin{aligned}
\mathbb{E}[\rho(\hat{X},X)] &\geq \frac{\alpha}{2}\frac{\mathcal{H}(\mathcal{Q}(\alpha)) - \mathbb{E}\left[|\phi(X)|\right] - 4d - 1}{\log_2|\mathcal{Y}(\alpha)|} \\
&\geq \frac{\alpha}{2}\frac{\mathcal{H}(\mathcal{Q}(\alpha)) - \mathbb{E}\left[|\phi(X)|\right] - 4d - 1}{d\log_2(4\bar{\rho}/\alpha)},
\end{aligned}$$

where the last inequality follows from Lemma 1.

Thus, for any code with

$$\mathbb{E}\left[|\phi(X)|\right] < \mathcal{H}(\mathcal{Q}(\alpha)) - 4d - 1 - 2d\frac{\log_2(4\bar{\rho}/\mathrm{D})}{\log_2(\bar{\rho}/\mathrm{D})},$$

we have $\mathbb{E}[\rho(\hat{X},X)] > \mathrm{D}$, which implies

$$\mathrm{R}(\mathrm{D}) \geq \mathcal{H}(\mathcal{Q}(\alpha)) - 4d - 1 - 2d\frac{\log_2(4\bar{\rho}/\mathrm{D})}{\log_2(\bar{\rho}/\mathrm{D})}.$$

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

$$\mathrm{R}(\mathrm{D}) \geq \mathcal{H}(\mathcal{Q}(\alpha)) - O(d).$$

■

V. Application to Bayesian Active Learning

As an example, in the special case of the problem of learning a binary classifier, as studied by [13] and [14], $\mathcal{X}^*$ is the set of all measurable classifiers $h : \mathcal{Z} \to \{-1,+1\}$, $\mathcal{X}$ is called the "concept space," X is called the "target function," and $\rho(X_1,X_2) = \mathbb{P}(X_1(Z) \neq X_2(Z))$, where Z is some $\mathcal{Z}$-valued random variable. In particular, $\rho(X_1,X)$ is called the "error rate" of X_1.

We may then discuss a *learning protocol* based on binary-valued queries. That is, we suppose some learning machine is able to pose yes/no questions to an oracle, and based on the responses it proposes a *hypothesis* $\hat{X}$. We may ask how many such yes/no questions must the learning machine pose (in expectation) before being able to produce a hypothesis $\hat{X} \in \mathcal{X}^*$ with $\mathbb{E}[\rho(\hat{X},X)] \leq \epsilon$, known as the *query complexity*.

If the learning machine is allowed to pose *arbitrary* binary-valued queries, then this setting is precisely a special case of the general lossy coding problem studied above. That is, any learning machine that asks a sequence of yes/no questions before terminating and returning some $\hat{X} \in \mathcal{X}^*$ can be thought of as a binary decision tree (no = left, yes = right), with the return $\hat{X}$ 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 $\phi(X)$ for X?", "what is the second bit in the codeword $\phi(X)$ for X?", etc., until we obtain a complete codeword, at which point we return the value that codeword decodes to. From this perspective, the query complexity is precisely $\mathrm{R}(\epsilon)$.

This general problem of learning with arbitrary binary-valued queries was studied previously by Kulkarni, Mitter, & Tsitsiklis [15], in a *minimax* analysis (studying the worst-case value of X). In particular, they find that for a given distribution for Z, the worst-case query complexity is essentially characterized by $\log|\mathcal{Y}(\epsilon)|$. The techniques employed are actually far more general than the classifier-learning problem, and actually apply to any pseudo-metric space. Thus, we can abstractly think of their work as a minimax analysis of lossy coding.

In addition to being quite interesting in their own right, the results of Kulkarni, Mitter, & Tsitsiklis [15] have played a significant role in the recent developments in active learning with *label request* queries for binary classification [16]–[18], in which the learning machine may only ask questions of the form, "What is the value $X(z)$?" for certain values $z \in \mathcal{Z}$. 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 $\log|\mathcal{Y}(\epsilon)|$ in any minimax query complexity results for active learning with label requests (though this factor is typically represented by its upper bound: $\propto V \cdot \log(1/\epsilon)$, where V is the VC dimension).

Similarly to how the work of Kulkarni, Mitter, & Tsitsiklis [15] can be used to argue that $\log|\mathcal{Y}(\epsilon)|$ is a lower bound on the minimax query complexity of active learning with label requests, Theorem 1 can be used to argue that $\mathcal{H}(\mathcal{Q}(\epsilon \log_2(1/\epsilon))) - O(d)$ is a lower bound on the query complexity of learning relative to a given distribution for X (called a *prior*, in the language of Bayesian statistics), rather than the worst-case value of X. Furthermore, as with [15], this lower bound remains valid for learning with label requests, since label requests are a type of binary-valued query. Thus, we should expect a term related to $\mathcal{H}(\mathcal{Q}(\epsilon))$ or $\mathcal{H}(\mathcal{Q}(\epsilon \log_2(1/\epsilon)))$ to appear in any tight analysis of the query complexity of Bayesian learning with label requests.

VI. Open Problems

In our present context, there are several interesting questions, such as whether the $\log(\bar{\rho}/\mathrm{D})$ 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., in the statistical learning special case, by making a VC class assumption instead).

Acknowledgments

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

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.

[2] J. C. Kieffer, "A survey of the theory of source coding with a fidelity criterion," *IEEE Transactions on Information Theory*, vol. 39, no. 5, pp. 1473–1490, 1993.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

[4] P. L. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 139–149, 1982.

[5] A. Gersho, "Asymptotically optimal block quantization," *IEEE Transactions on Information Theory*, vol. 25, no. 4, pp. 373–380, 1979.

[6] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Rec., Part 4*, pp. 142–163, 1959.

[7] T. Linder and R. Zamir, "One the asymptotic tightness of the shannon lower bound," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 2026–2031, 1994.

[8] E. C. Posner, E. R. Rodemich, and H. Rumsey, Jr., "Epsilon entropy of stochastic processes," *The Annals of Mathematical Statistics*, vol. 38, no. 4, pp. 1000–1020, 1967.

[9] E. C. Posner and E. R. Rodemich, "Epsilon entropy and data compression," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 2079–2125, 1971.

[10] A. Gupta, R. Krauthgamer, and J. R. Lee, "Bounded geometries, fractals, and low-distortion embeddings," in *In Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003.

[11] N. H. Bshouty, Y. Li, and P. M. Long, "Using the doubling dimension to analyze the generalization of learning algorithms," *Journal of Computer and System Sciences*, vol. 75, no. 6, pp. 323–335, 2009.

[12] D. A. Huffman, "A method for the construction of minimum-redundancy codes," in *Proceedings of the I.R.E.*, 1952, pp. 1098–1102.

[13] D. Haussler, M. Kearns, and R. Schapire, "Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension," *Machine Learning*, vol. 14, no. 1, pp. 83–113, 1994.

[14] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, no. 2, pp. 133–168, 1997.

[15] S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis, "Active learning using arbitrary binary valued queries," *Machine Learning*, vol. 11, no. 1, pp. 23–35, 1993.

[16] S. Hanneke, "Teaching dimension and the complexity of active learning," in *In Proceedings of the 20th Annual Conference on Learning Theory*, 2007.

[17] ——, "A bound on the label complexity of agnostic active learning," in *In Proceedings of the 24th International Conference on Machine Learning*, 2007.

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

Liu Yang is currently pursuing a Ph.D. degree in Machine Learning at Carnegie Mellon University. Her main research interests are in computational and statistical learning theory. Her recent focus has been the theoretical analysis of Bayesian active learning. She received a B.S. degree in Electronics and Information Engineering in 2005 from the Hua Zhong University of Science and Technology, and received the M.S. degree in Machine Learning in 2010 from Carnegie Mellon University.
Homepage: http://www.cs.cmu.edu/∼liuy

Steve Hanneke is a Visiting Assistant Professor in the Department of Statistics at Carnegie Mellon University. His research focuses on statistical learning theory, with an emphasis on the theoretical analysis of active learning. He received a B.S. in Computer Science in 2005 from the Univeristy of Illinois at Urbana-Champaign, and received his Ph.D. in Machine Learning in 2009 from Carnegie Mellon University.
Homepage: http://www.stat.cmu.edu/∼shanneke

Jaime Carbonell is the Allen Newell Professor of Computer Science at Carnegie Mellon University and the director of the Language Technologies Institute. His research interests span several areas of artificial intelligence, including machine learning, machine translation, information retrieval, and automated text summarization. He received his Ph.D. in Computer Science from Yale University.
Homepage: http://www.cs.cmu.edu/∼jgc