General Bounds on the Mutual Information Between a Parameter and n Conditionally Independent Observations

David Haussler[*]
UC Santa Cruz

Manfred Opper[†]
Universität Würzburg

Abstract

Each parameter θ in an abstract parameter space Θ is associated with a different probability distribution on a set Y. A parameter θ is chosen at random from Θ according to some *a priori* distribution on Θ, and n conditionally independent random variables $Y^n = Y_1, \ldots Y_n$ are observed with common distribution determined by θ. We obtain bounds on the mutual information between the random variable Θ, giving the choice of parameter, and the random variable Y^n, giving the sequence of observations. We also bound the supremum of the mutual information, over choices of the prior distribution on Θ. These quantities have applications in density estimation, computational learning theory, universal coding, hypothesis testing, and portfolio selection theory. The bounds are given in terms of the metric and information dimensions of the parameter space Θ with respect to the Hellinger distance.

1 Introduction

Let us assume that the state of nature is defined by a parameter θ in an abstract parameter space Θ. We are not allowed to observe the state of nature directly, but rather must make inferences about it indirectly by observing a sequence of observations $y^n = y_1, \ldots, y_n$. These observations are the values of n random variables $Y^n = Y_1, \ldots, Y_n$ that are conditionally independent given the true state of nature θ and have a common distribution determined by this true state of nature. This general setup is fundamental in statistics and related disciplines, including coding and data compression [12], computational learning theory [19, 22, 20, 1, 28, 41], and portfolio selection theory [5]. It is usually called *density estimation* or *parameter estimation* in statistics, depending on the nature of the parameter space Θ. In this paper we explore this setup from an information theoretic point of view.

We can measure our progress in learning about the true state of nature by measuring how well we are able to predict the observation y_{t+1} after seeing the previous observations $y_1, \ldots y_t$. Let us assume that Y is discrete. If, after seeing these previous observations, we produce the estimate $\hat{P}_t$ of the distribution on Y defined by the true state of nature, then let us define our *loss* in predicting the observation y_{t+1} to be $-\log \hat{P}_t(y_{t+1})$. If the logarithm base 2 is used, then this loss can be interpreted as the number of bits needed to encode y_{t+1}, making appropriate use of the estimated distribution $\hat{P}_t$. The *total loss* in sequentially predicting the observations $y_1, \ldots, y_n$ can be defined as $-\sum_{t=1}^{n} \log \hat{P}_{t-1}(y_t)$. This can be viewed as the number of bits need to encode $y_1, \ldots, y_n$ by an adaptive coding scheme that makes use of the estimated distributions $\hat{P}_0, \ldots, \hat{P}_{n-1}$. Analogous definitions can be made for continuous Y using estimated densities.

It is useful to compare the total loss incurred by using a particular method for obtaining the estimated distributions $\hat{P}_t$ to the total loss that would be obtained if we knew the true state of nature, and thus used the true distribution on Y to encode each y_t. The difference between these two losses is called the *regret* or *net loss*. The expected regret, with respect to the random choice of $y_1, \ldots, y_n$ according to a fixed state of nature, measures the average extra number of bits we need to encode $y_1, \ldots, y_n$ over and above the average number of bits required by the optimal encoding method, which would use the true distribution. This is called the *redundancy* in coding theory, and is generally called *risk* in statistics (with respect to arbitrary loss or regret functions).

The risk of any method depends on the true state of nature $\theta \in \Theta$. Taking a Bayesian approach, let us define an *a priori* distribution on the parameter space Θ so that Θ itself can be viewed as a random variable. We can then quantify the performance of any method that produces

[*]Supported by NSF grant IRI-9123692. Email addresses: haussler@cse.ucsc.edu

[†]Supported by Heisenberg fellowship of DFG Email addresses: opper@physik.uni-wuerzburg.de

estimated distributions $\hat{P}_t$ by its average risk when the true state of nature is drawn at random according to the prior distribution on Θ. In this case the posterior distribution $P(Y_{t+1}|y_1,\ldots,y_t)$ is the unique optimal choice for $\hat{P}_t$, and the average risk (redundancy) of this optimal method (the *Bayes method*) is called the *Bayes risk*. A simple calculation (see e.g. [19]) shows that the Bayes risk is equal to the mutual information $I(\Theta;Y^n)$ between the random variable Θ and the random variable $Y^n = Y_1,\ldots,Y_n$, which can be interpreted as the average amount of information contained in the observation sequence Y^n about the true state of nature θ. This leads to the other standard interpretations of the mutual information in terms of density estimation and coding theory. Clarke and Barron discuss further interpretations of this quantity in the context of hypothesis testing and portfolio selection theory [10]. The Bayes risk can also be used to obtain a lower bound on the *minimax risk*, which is the minimum over all methods of encoding (or predicting) the observations $y_1,\ldots,y_n$ of the maximum risk over all true states of nature $\theta \in \Theta$. This quantity is called *information capacity* or minimax redundancy in information theory. We denote it by $C(\Theta;Y^n)$.

Because of it's key role in several areas, the mutual information $I(\Theta;Y^n)$ has been investigated by numerous authors. Early work by Ibragimov and Hasminskii showed that $I(\Theta;Y^n) \approx (D/2)\log n$ when Y is real-valued and the conditional distributions on Y are a smooth family of densities indexed by real-valued parameter vectors θ of dimension D, and certain other conditions apply [24]. In this case they were even able to estimate the lower order additive terms in this approximation, which involve the Fisher information and the entropy of the prior. Further related results were given by Efroimovich [14] and Clarke [9]. Clarke and Barron gave a detailed analysis, with applications, of the risk (redundancy) of the Bayes method as a function of the true state of nature [10]. These results were extended to the mutual information and minimax risk in [11]. Related lower bounds, which are often quoted, were obtained by Rissanen [38], based on certain asymptotic normality assumptions. In this paper we extend work from [19, 4], obtaining bounds on $I(\Theta;Y^n)$ in more general settings.

The approach taken here is to relate the mutual information $I(\Theta;Y^n)$ directly to certain metric properties of the parameter space Θ. The distance between two states of nature θ and θ^* is measured in terms of the distance between the conditional distributions on the observation space Y that they define. This distance is defined either as the relative entropy (Kullback-Leibler divergence) between the distributions, or as the Hellinger distance between the distributions. The former distance is used in the upper bounds on the mutual information, and the latter is used in the lower bounds. The volume of a relative entropy (or Hellinger) neighborhood of radius r around the true state of nature θ^* is defined as the prior probability of all $\theta \in \Theta$ within distance r of θ^*. The rate at which this volume scales as a function of r is shown to be the key quantity that determines the mutual information for large n for many cases. The key inequality is given in Theorem 1 in Section 2 below, examples are given in Section 3, and final results are given in Theorems 2 and 3 in Section 4.

In Section 4 we obtain new bounds on the mutual information and minimax risk in terms of various abstract notions of the dimension and capacity of the parameter space Θ. These include definitions of dimension via the metric entropy as introduced by Kolmogorov and Tikhomirov in [26] and commonly used in the theory of empirical processes (see e.g. [13, 31, 16, 8]), as well as Renyi's information dimension [36, 15, 25] and the associated Posner-Rodemich ϵ-entropy [32, 33, 34, 35]. Metric entropies defined in terms of the relative entropy and Hellinger distances have been used in the context of density estimation by LeCam [27], Birgé [6, 7], Hasminskii and Ibragimov [17], and van de Geer [40], the latter with an explicit goal of applying methods from empirical processes to this problem. We conclude in Section 5 with a brief discussion of the problem of obtaining bounds on the average instantaneous regret for the encoding of the tth observation as a function of t, or equivalently the average additional information gained about the true state of nature from the tth observation, which is related to loss measures more commonly examined in computational learning theory (see e.g. [20, 1]). Further applications of the results given here to this and related problems are described in [29, 23].

2 Basic Definitions and Main Result

We use the following notational conventions. For a random variable X, P_X denotes the distribution function for X, and if X has a density then it is denoted p_X. For countable X, $P(x)$ denotes the probability that $X = x$ and for continuous X, $p(x)$ is shorthand for $p_X(x)$. For a (measurable) function $f(x)$, $\mathbb{E}_X f(x)$ denotes the expectation of f. Given random variables X and Y, $P_{X,Y}$ denotes their joint distribution, $\mathbb{E}_{X,Y} f(x,y)$ the expectation of f with respect to this joint distribution, etc. The conditional distribution of Y given that $X = x$ is denoted $P_{Y|x}$, and $p_{Y|x}$ denotes the conditional density. $P(y|x)$ and $p(y|x)$ denote the probability and conditional density of y given $X = x$, resp. The conditional expectation of f given that $X = x$ is denoted $\mathbb{E}_{Y|x} f(x,y)$. Similar notation is used to condition on an event. Finally, the marginal distribution of Y is denoted P_Y, a specific marginal probability is denoted $P(y) = \mathbb{E}_X P(y|x)$, the marginal density is denoted p_Y, and a specific density value is denoted $p(y) = \mathbb{E}_X p(y|x)$. Throughout the paper we employ the convention that $0 \ln 0 = 0 \ln \infty = 0 \ln \frac{0}{0} = 0$.

Let $\Theta, Y_1, Y_2, \ldots$ be random variables such that $Y_1, Y_2, \ldots$ are conditionally independent and identically distributed given Θ, their common distribution denoted by $P_{Y|\theta}$ for each $\theta \in \Theta$. The random variable Θ takes values in an arbitrary set. For simplicity we assume that either Y

takes values in a countable set, or Y is continuous[1] with conditional densities $p_{Y|\theta}$ defined for each $\theta \in \Theta$. Our default assumption will be that Y is countable, and all our results will be stated explicitly for this case, with comments on how analogous results can be obtained for continuous Y. All functions in our results are assumed to be measurable without explicit mention.

For each $n \geq 1$, let $Y^n = Y_1, \ldots, Y_n$ and let y^n denote a typical value of Y^n. We give upper and lower bounds on the mutual information between Θ and Y^n, defined for countable Y by

$$I(\Theta; Y^n) = \mathbb{E}_{\Theta, Y^n} \ln \frac{P(y^n|\theta)}{P(y^n)},$$

and similarly for continuous Y, using the corresponding densities denoted with lower-case p.

In obtaining these bounds, we use two notions of "distance" between probability distributions. Let $P = \{p_i\}$ and $Q = \{q_i\}$ be two probability mass functions on a countable set. The *Kullback-Leibler (KL) divergence* (or *relative entropy distance*) between P and Q is defined by

$$D_K(P||Q) = \sum_i p_i \ln \frac{p_i}{q_i}$$

and the *(squared) Hellinger distance* between P and Q is defined by

$$D_H(P, Q) = 2 \sum_i (\sqrt{p_i} - \sqrt{q_i})^2.$$

For densities $p(x)$ and $q(x)$, the analogous definitions are $D_K(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} dx$ and $D_H(p, q) = 2 \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$.

Note that

$$D_H(P, Q) = 4(1 - \sum_i \sqrt{p_i q_i}), \tag{1}$$

($D_H(p, q) = 4(1 - \int \sqrt{p(x)q(x)} dx)$ in the continuous case) and therefore $D_H(P, Q) \leq 4$, in contrast to the KL divergence, which can be arbitrarily large.

We employ these notions of distance between distributions to define two corresponding notions of "distance" between the points in Θ. For each $\theta^*, \theta \in \Theta$, let

$$\Delta_K(\theta^*, \theta) = D_K(P_{Y|\theta^*}||P_{Y|\theta})$$

and

$$\Delta_H(\theta^*, \theta) = D_H(P_{Y|\theta^*}, P_{Y|\theta}).$$

Our main theorem gives bounds on $I(\Theta; Y^n)$ in terms of the expected logarithms of two Laplace transforms, one of the KL divergence and one of the Hellinger distance.

[1] Specifically, we assume that conditional densities $p_{Y|\theta}$ are defined with respect to a common sigma-finite measure on a suitable probability space, such as k-dimensional real space with the Lebesgue measure.

Theorem 1 *For every $n \geq 1$,*

$$\begin{aligned} -I\!E_{\Theta^*} \ln I\!E_\Theta e^{-\frac{n}{4}\Delta_H(\theta^*, \theta)} &\leq -I\!E_{\Theta^*} \ln I\!E_\Theta \left(1 - \frac{1}{4}\Delta_H(\theta, \theta^*)\right)^n \\ &\leq I(\Theta; Y^n) \\ &\leq -I\!E_{\Theta^*} \ln I\!E_\Theta e^{-n\Delta_K(\theta^*, \theta)}. \end{aligned}$$

The upper bound follows from results given in [3] and is mentioned there, however we give a simple and direct proof. To the best of our knowledge, the lower bound is new. The proof is given in a series of lemmas and calculations. We begin with the upper bound. This requires the following lemma which has been previously utilized in the framework of Statistical Physics [39].

Lemma 1 *For any random variables W and V and real-valued function $u(w, v)$,*

$$-I\!E_V \ln I\!E_W e^{u(w,v)} \leq -\ln I\!E_W e^{I\!E_V u(w,v)}.$$

Proof: First note that by Hölder's inequality, for any real-valued functions u_1 and u_2,

$$\begin{aligned} &\mathbb{E}_W e^{\alpha u_1(w) + (1-\alpha)u_2(w)} = \mathbb{E}_W (e^{u_1(w)})^\alpha (e^{u_2(w)})^{(1-\alpha)} \\ &\leq \left(\mathbb{E}_W e^{u_1(w)}\right)^\alpha \left(\mathbb{E}_W e^{u_2(w)}\right)^{(1-\alpha)} \end{aligned}$$

Taking logs, this shows that $\ln \mathbb{E}_W e^{u(w,v)}$ is convex in u. The result then follows by applying Jensen's inequality. □

Using this lemma, we now get the following chain of inequalities

$$\begin{aligned} I(\Theta; Y^n) &= \mathbb{E}_{\Theta^*} \mathbb{E}_{Y^n|\theta^*} \ln \frac{P(y^n|\theta^*)}{P(y^n)} \\ &= \mathbb{E}_{\Theta^*} \mathbb{E}_{Y^n|\theta^*} \ln \frac{P(y^n|\theta^*)}{\mathbb{E}_\Theta P(y^n|\theta)} \\ &= -\mathbb{E}_{\Theta^*} \mathbb{E}_{Y^n|\theta^*} \ln \mathbb{E}_\Theta \frac{P(y^n|\theta)}{P(y^n|\theta^*)} \\ &= -\mathbb{E}_{\Theta^*} \mathbb{E}_{Y^n|\theta^*} \ln \mathbb{E}_\Theta e^{\ln \frac{P(y^n|\theta)}{P(y^n|\theta^*)}} \\ &\leq -\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta e^{\mathbb{E}_{Y^n|\theta^*} \ln \frac{P(y^n|\theta)}{P(y^n|\theta^*)}} \\ &= -\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta e^{-D_K(P_{y^n|\theta^*}||P_{y^n|\theta})} \\ &= -\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta e^{-n\Delta_K(\theta^*, \theta)}, \end{aligned}$$

where the last equality follows from the fact that the KL divergence is additive over the product of independent distributions (see e.g. [12]). Note that by our convention that $0 \ln 0 = 0$, for each θ^*, any y^n such that $P(y^n|\theta^*) = 0$ can simply be removed in the first equality above and then reintroduced in the exponent of the second to the last inequality, thus avoiding any division by zero. Similar inequalities hold in the case of continuous Y using densities in place of probabilities. This establishes the upper bound of Theorem 1.

Turning to the lower bounds, the first inequality follows from the fact that $1 - x \leq e^{-x}$. To establish the tighter

lower bound (the second inequality), we use the following lemma, which is new, as far as we can tell. For any random variables W and V, where V is countable and any real λ, define

$$I_\lambda(W;V) = -\mathbb{E}_{W^*,V} \ln \mathbb{E}_W \left(\frac{P(v|w)}{P(v|w^*)} \right)^\lambda .$$

Similarly, if V is continuous and the appropriate densities exist, $I_\lambda(W;V) = -\mathbb{E}_{W^*,V} \ln \mathbb{E}_W \left(\frac{p(v|w)}{p(v|w^*)} \right)^\lambda$. Note that, using the formula for mutual information given in the third equality in the chain above (and substituting $\Theta = W$, $Y^n = V$), it is clear that $I_1(W;V) = I(W;V)$, the mutual information between W and V.

Lemma 2 *For any random variables W and V, $I_\lambda(W;V)$ is maximal at $\lambda = 1$.*

Proof: Assume V is countable. We show that the difference $I(W;V) - I_\lambda(W;V)$ is nonnegative. This difference can be written as

$$\begin{aligned}
& -\mathbb{E}_{W^*,V} \ln \frac{P^{\lambda-1}(v|w^*)\mathbb{E}_W P(v|w)}{\mathbb{E}_W P^\lambda(v|w)} \\
\geq & -\ln \mathbb{E}_{W^*,V} \frac{P^{\lambda-1}(v|w^*)\mathbb{E}_W P(v|w)}{\mathbb{E}_W P^\lambda(v|w)} \\
= & -\ln \sum_v \mathbb{E}_{W^*} P(v|w^*) \frac{P^{\lambda-1}(v|w^*)\mathbb{E}_W P(v|w)}{\mathbb{E}_W P^\lambda(v|w)} \\
= & -\ln \sum_v \frac{\mathbb{E}_{W^*} P^\lambda(v|w^*)\mathbb{E}_W P(v|w)}{\mathbb{E}_W P^\lambda(v|w)} \\
= & -\ln \sum_v \mathbb{E}_W P(v|w) \\
= & 0,
\end{aligned}$$

where in the first line we have used Jensen's inequality. A similar argument can be used when V is continuous. □

Using this lemma, along with Jensen's inequality and Equation (1), we have

$$\begin{aligned}
I(\Theta;Y^n) &= -\mathbb{E}_{\Theta^*}\mathbb{E}_{Y^n|\theta^*} \ln \mathbb{E}_\Theta \frac{P(y^n|\theta)}{P(y^n|\theta^*)} \\
&\geq -\mathbb{E}_{\Theta^*}\mathbb{E}_{Y^n|\theta^*} \ln \mathbb{E}_\Theta \sqrt{\frac{P(y^n|\theta)}{P(y^n|\theta^*)}} \\
&\geq -\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta \mathbb{E}_{Y^n|\theta^*} \sqrt{\frac{P(y^n|\theta)}{P(y^n|\theta^*)}} \\
&= -\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta \sum_{Y^n} \sqrt{P(y^n|\theta)P(y^n|\theta^*)} \\
&= -\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta \sum_{Y^n} \prod_{t=1}^n \sqrt{P(y_t|\theta)P(y_t|\theta^*)} \\
&= -\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta \prod_{t=1}^n \sum_{y_t \in Y_t} \sqrt{P(y_t|\theta)P(y_t|\theta^*)} \\
&= -\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta (1 - \frac{1}{4} D_H(P_{Y|\theta}, P_{Y|\theta^*}))^n
\end{aligned}$$

As in the proof of the upper bound, to avoid division by zero, we can remove y^n such that $P(y^n|\theta^*) = 0$ from the first line, and reintroduce them in the fourth line. Similar inequalities hold in the case of continuous Y using densities in place of probabilities. This establishes the lower bounds, and completes the proof of Theorem 1. □

3 Examples

We now illustrate Theorem 1 by using it to characterize the asymptotic behavior of the mutual information $I(\Theta;Y^n)$ for large n in some typical cases. The simplest case is the case of countable Θ. Here and in the sequel, we will assume for simplicity that for all distinct $\theta, \theta^* \in \Theta$, the conditional distributions $P_{Y|\theta}$ and $P_{Y|\theta^*}$ (or densities $p_{Y|\theta}$ and $p_{Y|\theta^*}$) differ on a set of positive measure, and hence $\Delta_H(\theta,\theta^*) > 0$. Otherwise we can replace Θ by a set of equivalence classes with the property that $\theta \equiv \theta^*$ iff $p_{Y|\theta} = p_{Y|\theta^*}$ (except on a set of measure zero) in a natural way, without changing $I(\Theta;Y^n)$.

Suppose Θ is countable, say $\Theta = \{\theta_i\}$. Let $H(\Theta) = -\sum_i P(\theta_i) \ln P(\theta_i)$ denote the entropy of Θ, which may be infinite. Then

Corollary 1

$$\lim_{n\to\infty} I(\Theta;Y^n) = H(\Theta).$$

Proof: If $H(\Theta)$ is infinite then clearly

$$\limsup_{n\to\infty} I(\Theta;Y^n) \leq H(\Theta).$$

Assume $H(\Theta)$ is finite. Let

$$H(\Theta|Y^n) = -\mathbb{E}_{Y^n} \sum_i P(\theta_i|y^n) \ln P(\theta_i|y^n),$$

the conditional entropy of Θ given Y^n. Note that this quantity is nonnegative. When $H(\Theta)$ is finite it is easily verified that

$$I(\Theta;Y^n) = H(\Theta) - H(\Theta|Y^n)$$

(see e.g. [12]), and thus $\limsup_{n\to\infty} I(\Theta;Y^n) \leq H(\Theta)$ in this case as well.

For the lower bound, using Theorem 1 and Fatou's lemma

$$\begin{aligned}
\liminf_{n\to\infty} I(\Theta;Y^n) &\geq \liminf_{n\to\infty} -\sum_i P(\theta_i) \ln \sum_j P(\theta_j) e^{-\frac{n}{4}\Delta_H(\theta_i,\theta_j)} \\
&\geq -\sum_i P(\theta_i) \liminf_{n\to\infty} \ln \sum_j P(\theta_j) e^{-\frac{n}{4}\Delta_H(\theta_i,\theta_j)} \\
&= -\sum_i P(\theta_i) \ln P(\theta_i) \\
&= H(\Theta).
\end{aligned}$$

□

This result generalizes the similar result in [11] (Corollary 1) by removing the additional conditions assumed there. In the case that Θ is finite, results of Renyi [37] show further that the difference $I(\Theta;Y^n) - H(\Theta)$ converges to zero exponentially fast in n. More general results, including the above corollary, follow from results in Pinsker's book [30] (see also [2]).

For uncountable Θ, $I(\Theta;Y^n)$ is typically unbounded as n grows. To illustrate Theorem 1 in this case, we calculate the bounds given there for a simple Gaussian case, where an exact formula for the mutual information can be obtained. Let Y be a Gaussian distributed D-dimensional random vector with mean θ and spherical density

$$P(y|\theta) = (2\pi)^{-D/2} e^{-\frac{||y-\theta||^2}{2}} \tag{2}$$

In this case Hellinger and KL distances are easily calculated to be

$$\Delta_H(\theta^*,\theta) = 4(1 - e^{-\frac{||\theta^*-\theta||^2}{8}}) \tag{3}$$

and

$$\Delta_K(\theta^*,\theta) = \frac{1}{2}||\theta^* - \theta||^2$$

Note that for $||\theta^* - \theta||$ small, both distances become asymptotically equal. In general, however $\Delta_K(\theta^*,\theta)$ can become arbitrary large. Let us next assume that the prior density of the mean θ is also a spherical Gaussian with density

$$p(\theta) = (2\pi)^{-D/2} e^{-\frac{||\theta||^2}{2}}. \tag{4}$$

Then, after some straightforward integrations one obtains the simple result for the mutual information

$$I(\Theta;Y^n) = \frac{D}{2}\ln(n+1). \tag{5}$$

The upper bound from Theorem 1 also reduces to standard Gaussian integrals and gives

$$-\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta e^{-n\Delta_K(\theta^*,\theta)} = \frac{D}{2}\ln(n+1) + \frac{Dn}{2(n+1)}. \tag{6}$$

While the weaker lower bound has no nice closed form, the tighter lower bound in Theorem 1 has the same form as the upper bound in this Gaussian case, and gives

$$\begin{aligned} &-\mathbb{E}_{\Theta^*} \ln \mathbb{E}_\Theta \left(1 - \frac{1}{4}\Delta_H(\theta,\theta^*)\right)^n \\ &= \frac{D}{2}\ln(\frac{n}{4}+1) + \frac{Dn}{2n+4}. \end{aligned}$$

Thus the upper and lower bounds differ by an additive term which is about $\frac{D}{2}\ln 4$. Using Laplace's method to estimate the weaker lower bound shows that it is very close to the tighter lower bound, asymptotically differing from the upper bound by the same additive term. Numerical evaluation for the $D = 1$ case shows that all bounds are quite good for $n \geq 5$.

4 Mutual information, entropy and dimension

In this section we will explore the relationship between the properties of the metric space $(\Theta, \Delta_H^{1/2})$ and the mutual information $I(\Theta;Y^n)$. This is a metric space because $\Delta_H^{1/2}(\theta^*,\theta)$ is an l_1 distance for countable Y and an L_1 distance for continuous Y, and we are assuming that $p_{Y|\theta}$ and $p_{Y|\theta^*}$ differ on a set of positive measure for distinct θ and θ^*, and hence $\Delta_H^{1/2}(\theta,\theta^*) > 0$.

There are several notions of the entropy, capacity and dimension of a metric space equipped with a measure that are useful here. For the following definitions, let (X,ρ) be a complete separable metric space and μ be a measure on X defined on the Borel subsets of X.

Definition 1 *(Metric entropy, also called Kolmogorov ϵ-entropy [26]) A partition Π of X is a collection $\{\pi_i\}$ of Borel subsets of X that are pairwise disjoint and whose union is X. The diameter of a set $A \subseteq X$ is given by $diam(A) = \sup_{x,y\in A}\rho(x,y)$. The diameter of a partition is the supremum of the diameters of the sets in the partition. For $\epsilon > 0$, let $D_\epsilon(X,\rho)$ denote the cardinality of the smallest finite partition of X of diameter at most ϵ, or ∞ if no such finite partition exists. The metric entropy of (X,ρ) is defined by*

$$K_\epsilon(X) = K_\epsilon(X,\rho) = \ln D_\epsilon(X,\rho).$$

(The metric ρ is omitted when understood from the context here and below.)

Definition 2 *(Posner-Rodemich ϵ-entropy [35]) Let $\Pi = \{\pi_i\}$ be a countable partition of X. The entropy of Π is defined as $H(\Pi) = -\sum_i \mu(\pi_i)\ln\mu(\pi_i)$. Let $\mathcal{P}_\epsilon$ denote the set of all countable partitions of X of diameter at most ϵ. The Posner-Rodemich ϵ-entropy of (X,ρ,μ) is defined by*

$$H_{\epsilon,\mu}(X) = \inf_{\Pi\in\mathcal{P}_\epsilon} H(\Pi).$$

Definition 3 *(volume-scaling entropy) For any $x \in X$, the ball of radius ϵ centered at x is defined by $b_\epsilon(x) = \{y : \rho(x,y) \leq \epsilon\}$. The volume-scaling entropy of (X,ρ,μ) is given by*

$$V_{\epsilon,\mu}(X) = -I\!E_X \ln \mu(b_\epsilon(x)).$$

Definition 4 *(Laplace transform entropy) The Laplace transform entropy of (X,ρ,μ) is given by*

$$L_{\epsilon,\mu}(X) = -I\!E_{X^*} \ln I\!E_X e^{-(\rho(x^*,x)/\epsilon)^2}.$$

Note that the metric entropy is independent of the measure μ, whereas the other entropies depend on μ. Measure independent versions of these can be defined as follows.

Definition 5

$$H_\epsilon(X) = \sup_\mu H_{\epsilon,\mu}(X),$$

where the supremum is over all measures μ defined on the Borel subsets of X. We call $H_\epsilon(X)$ the Posner-Rodemich ϵ-capacity of X. The volume-scaling and Laplace transform capacities $V_\epsilon(X)$ and $L_\epsilon(X)$ are defined analogously.

For each of these entropies and capacities, there is a corresponding notion of dimension.

Definition 6 *The upper and lower metric dimensions [26] are defined by*

$$\overline{\mathbf{dim}}_K(X) = \limsup_{\epsilon \to 0} \frac{K_\epsilon(X)}{-\ln \epsilon}$$

and

$$\underline{\mathbf{dim}}_K(X) = \liminf_{\epsilon \to 0} \frac{K_\epsilon(X)}{-\ln \epsilon},$$

respectively. When $\overline{\mathbf{dim}}_K(X) = \underline{\mathbf{dim}}_K(X)$, then this value is denoted $\mathbf{dim}_K(X)$ and called the metric dimension of X. The information dimension [36, 15, 25], volume-scaling dimension, and Laplace transform dimension are defined analogously by substituting $H_{\epsilon,\mu}$, $V_{\epsilon,\mu}$ and $L_{\epsilon,\mu}$ respectively for K_ϵ in the above definition, and are denoted $\mathbf{dim}_{H,\mu}(X)$, $\mathbf{dim}_{V,\mu}(X)$ and $\mathbf{dim}_{L,\mu}(X)$, respectively. Measure independent forms of these dimensions are defined using the capacities H_ϵ, V_ϵ and L_ϵ, and are denoted $\mathbf{dim}_H(X)$, $\mathbf{dim}_V(X)$ and $\mathbf{dim}_L(X)$, respectively.

The following lemma describes some of the relationships between these notions of entropy, capacity and dimension.

Lemma 3 *1. $L_{\epsilon,\mu}(X) - 1 \le V_{\epsilon,\mu}(X) \le H_{\epsilon,\mu}(X) \le K_\epsilon(X)$.*

2. ([34]) $V_{\epsilon,\mu}(X) \ge H_{2\epsilon,\mu}(X)$.

3. If $K_\epsilon(X)$ is finite then

$$L_{\epsilon,\mu}(X) \ge V_{\sqrt{2\epsilon^2 K_\epsilon(X)},\mu}(X) - (\ln 2 + 2/e).$$

4. $V_\epsilon(X) \ge K_{2\epsilon}(X)$.

5. If $\overline{\mathbf{dim}}_K(X)$ is finite then

$$\overline{\mathbf{dim}}_{L,\mu}(X) = \overline{\mathbf{dim}}_{V,\mu}(X) = \overline{\mathbf{dim}}_{H,\mu}(X) \le \overline{\mathbf{dim}}_K(X)$$

and

$$\overline{\mathbf{dim}}_L(X) = \overline{\mathbf{dim}}_V(X) = \overline{\mathbf{dim}}_H(X) = \overline{\mathbf{dim}}_K(X),$$

and similarly for $\underline{\mathbf{dim}}$ and $\mathbf{dim}$, when the latter exists.

Proof: We begin with part (1). To verify the first inequality, note that for any nonnegative random variable Z,

$$\begin{aligned} -\ln \mathbb{E}_Z e^{-(z/\epsilon)^2} &\le -\ln\left(P(Z \le \epsilon)\mathbb{E}_{Z|Z\le\epsilon} e^{-(z/\epsilon)^2}\right) \\ &\le -\ln P(Z \le \epsilon) + 1. \end{aligned}$$

For fixed x^*, let $Z = \rho(x^*, x)$. It follows that

$$L_{\epsilon,\mu}(X) \le -\mathbb{E}_{X^*} \ln \mu(\rho(x^*, x) \le \epsilon) + 1 = V_{\epsilon,\mu}(X) + 1.$$

To verify the second inequality, let Π be any partition of diameter at most ϵ. For any $x \in X$, let $[x]_\Pi$ denote the set in Π containing x. Then $H(\Pi) = -\mathbb{E}_X \ln \mu([x]_\Pi)$. Since $\mathrm{diam}([x]_\Pi) \le \epsilon$, $[x]_\Pi \subseteq b_\epsilon(x)$. Therefore

$$-\mathbb{E}_X \ln \mu(b_x(\epsilon)) \le H(\Pi)$$

for all partitions Π of diameter at most ϵ. The second inequality follows. Finally, the last inequality follows from the fact that the entropy of a partition is always at most the logarithm of the cardinality of the partition.

In contrast to part (1), the inequality in part (2) is a quite nontrivial result. The proof is given in [34] (see also [35], inequality (9.7).)

Part (3) is verified as follows. Note that for any nonnegative random variable Z and $r > 0$,

$$\begin{aligned} -\ln \mathbb{E}_Z e^{-(z/\epsilon)^2} &= -\ln(P(Z \le r)\mathbb{E}_{Z|Z\le r} e^{-(z/\epsilon)^2} \\ &\quad + P(Z > r)\mathbb{E}_{Z|Z>r} e^{-(z/\epsilon)^2}) \\ &\ge -\ln\left(P(Z \le r) + \mathbb{E}_{Z|Z>r} e^{-(z/\epsilon)^2}\right) \\ &\ge -\ln\left(P(Z \le r) + e^{-(r/\epsilon)^2}\right) \\ &\ge -\ln\left(2 * \max\{P(Z \le r), e^{-(r/\epsilon)^2}\}\right) \\ &= -\ln 2 + \min\{-\ln P(Z \le r), (r/\epsilon)^2\}. \end{aligned}$$

As above, for fixed x^*, let $Z = \rho(x^*, x)$. Let $A_\epsilon(r) = \{x^* : -\ln \mu(b_r(x^*)) > (r/\epsilon)^2\}$. Then for all $r > 0$,

$$\begin{aligned} L_{\epsilon,\mu}(X) &\ge -\ln 2 + \mathbb{E}_{X^*} \min\{-\ln \mu(b_r(x^*)), (r/\epsilon)^2\} \\ &\ge -\ln 2 - \mathbb{E}_{X^*} \ln \mu(b_r(x^*)) \\ &\quad + \int_{A_\epsilon(r)} \ln \mu(b_r(x^*)) d\mu(x^*) \\ &= -\ln 2 + V_{r,\mu}(X) + \int_{A_\epsilon(r)} \ln \mu(b_r(x^*)) d\mu(x^*) \end{aligned}$$

Let $r = \sqrt{2\epsilon^2 K_\epsilon(X)}$. Let Π be a minimum cardinality partition of X of diameter at most ϵ, so that $\ln |\Pi| = K_\epsilon(X)$. Note that for all x^* in $A_\epsilon(r)$,

$$\mu([x^*]_\Pi) \le \mu(b_r(x^*)) < e^{-2K_\epsilon(X)}$$

by definition of r and $A_\epsilon(r)$. From this, the fact that $-x \ln x$ is an increasing function for $0 < x < 1/e$, and the fact that $xe^{-x} \le 1/e$, we have

$$\begin{aligned} -\int_{A_\epsilon(r)} \ln \mu(b_r(x^*)) d\mu(x^*) &\le -\int_{A_\epsilon(r)} \ln \mu([x^*]_\Pi) d\mu(x^*) \\ &\le -\sum_{\pi \in \Pi : \pi \cap A_\epsilon(r) \ne \emptyset} \mu(\pi) \ln \mu(\pi) \\ &\le |\Pi| 2K_\epsilon(X) e^{-2K_\epsilon(X)} \\ &= 2K_\epsilon(X) e^{-K_\epsilon(X)} \\ &\le 2/e. \end{aligned}$$

The inequality of part (3) follows.

To verify part (4), let us say that $S \subseteq X$ is ϵ-separated if for all distinct $x, y \in S$, $\rho(x,y) > \epsilon$. Let $M_\epsilon(X)$ denote the cardinality of the largest finite ϵ-separated subset of X, or ∞ if arbitrarily large ϵ-separated subsets exist. It is readily verified that $M_\epsilon(X) \geq D_{2\epsilon}(X)$ (see e.g. [26]). Now let S be an ϵ-separated subset of X of maximal cardinality $M_\epsilon(X)$. If S is finite, then let μ be the uniform distribution on S. Then $V_{\epsilon,\mu}(X) = \ln M_\epsilon(X) \geq \ln D_{2\epsilon}(X) = K_{2\epsilon}(X)$. Hence $V_\epsilon(X) \geq K_{2\epsilon}(X)$. If S is infinite, then by taking uniform distributions on larger and larger subsets of S, we see that $V_\epsilon(X) = \infty$, so the inequality holds here as well.

Finally, part (5) follows easily from parts (1-4). Indeed, part (1) implies that

$$\overline{\mathbf{dim}}_{L,\mu}(X) \leq \overline{\mathbf{dim}}_{V,\mu}(X) \leq \overline{\mathbf{dim}}_{H,\mu}(X) \leq \overline{\mathbf{dim}}_K(X),$$

and similarly for $\underline{\mathbf{dim}}$ and $\mathbf{dim}$, when the latter exists. When $\overline{\mathbf{dim}}_K(X)$ is finite then $K_\epsilon(X)$ is $O(\ln(1/\epsilon))$. Thus, as is easily verified, parts (2) and (3) imply that $\overline{\mathbf{dim}}_{L,\mu}(X) \geq \overline{\mathbf{dim}}_{V,\mu}(X) \geq \overline{\mathbf{dim}}_{H,\mu}(X)$, and similarly for $\underline{\mathbf{dim}}$ and $\mathbf{dim}$, when the latter exists. Using part (4), we can also drop the μ in the subscript and extend this chain of inequalities to include $\overline{\mathbf{dim}}_K(X)$. □

We now apply the above lemma to the problem of estimating $I(\Theta; Y^n)$ and $C(\Theta; Y^n)$. For the remainder of the paper, we assume that Y is a complete separable metric space. Let $A(Y)$ denote the set of all probability measures defined on the Borel subsets of Y. We assume that the parameter set Θ is such that $\{P_{Y|\theta} : w \in \Theta\}$ is a Borel subset of $A(Y)$, under the topology generated by the Hellinger distance $\Delta_H^{1/2}$.

Recall now the discussion of the minimax risk $C(\Theta; Y^n)$ given in Section 1. A formal definition is

$$C(\Theta; Y^n) = \inf_Q \sup_{\theta \in \Theta} \mathbb{E}_{Y^n|\theta} \ln \frac{P(y^n|\theta)}{Q(y^n)},$$

where $\inf_Q$ ranges over all probability distributions Q defined on the Borel subsets of Y^n. This quantity is related to the mutual information as described in the following lemma.

Lemma 4 *([18])*

$$C(\Theta; Y^n) = \sup_{P_\Theta} I(\Theta; Y^n),$$

where the supremum is over all prior distributions P_Θ defined on the Borel subsets[2] of the parameter space Θ, holding fixed the family of conditional distributions $P_{Y|\theta}$, $\theta \in \Theta$.

□

Theorem 2 *If $\overline{\mathbf{dim}}_K(\Theta, \Delta_H^{1/2})$ is finite then*

1.

$$\liminf_{n\to\infty} \frac{I(\Theta; Y^n)}{\ln n} \geq \frac{\underline{\mathbf{dim}}_{H,P_\Theta}(\Theta, \Delta_H^{1/2})}{2}$$

and

2.

$$\liminf_{n\to\infty} \frac{C(\Theta; Y^n)}{\ln n} \geq \frac{\underline{\mathbf{dim}}_K(\Theta, \Delta_H^{1/2})}{2}.$$

Proof: It follows from the lower bound of Theorem 1 that $I(\Theta; Y^n) \geq L_{\sqrt{4/n}, P_\Theta}(\Theta, \Delta_H^{1/2})$. Hence

$$\sup_{P_\Theta} I(\Theta; Y^n) \geq L_{\sqrt{4/n}}(\Theta, \Delta_H^{1/2}).$$

Part (1) then follows using Lemma 3 part (5) with $\epsilon = \sqrt{1/n}$, and part (2) follows similarly, using also Lemma 4. □

To obtain analogous upper bounds, we will need to first investigate the relationship between the Hellinger and the relative entropy distances. The following is a useful lemma in this direction. See e.g. [6] for related results.

Lemma 5 *For any z, $0 \leq z \leq \infty$, define*

$$b(z) = \frac{z - \ln z - 1}{2(1 - \sqrt{z})^2} \in [1/2, \infty].$$

For all distributions $P = \{p_i\}$ and $Q = \{q_i\}$,

$$b(s) D_H(P,Q) \leq D_K(P||Q) \leq b(r) D_H(P,Q),$$

where $r = \inf_i \frac{q_i}{p_i}$ and $s = \sup_i \frac{q_i}{p_i}$. The analogous result holds for any densities p and q with $r = \inf_X \frac{q(x)}{p(x)}$ and $s = \sup_X \frac{q(x)}{p(x)}$.

Proof: We do the case of countable P and Q; the case of densities p and q is similar. For each i, let $r_i = q_i/p_i$. If there exists an i such that $q_i = 0$ and $p_i > 0$ then $r = \inf r_i = 0$ and both $D_K(P||Q)$ and $b(r)$ are infinite, so the result holds trivially. Hence we assume there is no i such that $q_i = 0$ and $p_i > 0$. Let $A = \{i : p_i = 0 \text{ and } q_i > 0\}$ and $\bar{A}$ be the complement of A. Then

$$\begin{aligned}
D_K(P||Q) &= \sum_{i \in \bar{A}} p_i \ln \frac{p_i}{q_i} \\
&= -\sum_{i \in \bar{A}} p_i \ln r_i \\
&= -\sum_{i \in \bar{A}} p_i \ln r_i + \sum_i q_i - \sum_i p_i \\
&= -\sum_{i \in \bar{A}} p_i \ln r_i + \sum_{i \in \bar{A}} p_i r_i + \sum_{i \in A} q_i - \sum_{i \in \bar{A}} p_i \\
&= \sum_{i \in \bar{A}} p_i (r_i - \ln r_i - 1) + \sum_{i \in A} q_i.
\end{aligned}$$

Also

$$\begin{aligned}
D_H(P,Q) &= 2 \sum_i (\sqrt{p_i} - \sqrt{q_i})^2 \\
&= 2 \sum_{i \in \bar{A}} p_i (1 - \sqrt{r_i})^2 + 2 \sum_{i \in A} q_i.
\end{aligned}$$

[2] Here we use the Borel subsets with respect to the topology of weak convergence of measures, as in [18].

Therefore

$$\begin{aligned}\frac{D_K(P||Q)}{D_H(P,Q)} &= \frac{\sum_{i\in\bar{A}} p_i(r_i - \ln r_i - 1) + \sum_{i\in A} q_i}{2\sum_{i\in\bar{A}} p_i(1-\sqrt{r_i})^2 + 2\sum_{i\in A} q_i} \\ &\leq \max\left(\frac{1}{2}, \sup_{i\in\bar{A}} \frac{r_i - \ln r_i - 1}{2(1-\sqrt{r_i})^2}\right) \\ &= \max\left(\frac{1}{2}, \sup_{i\in\bar{A}} b(r_i)\right) \\ &= b(r)\end{aligned}$$

since b is a continuously decreasing function that is always at least $1/2$, $r_i = \infty$ for $i \in A$, and $r = \inf_i r_i$. This establishes the upper bound. By similar reasoning, if A is nonempty then

$$\frac{D_K(P||Q)}{D_H(P,Q)} \geq \min\left(\frac{1}{2}, \inf_{i\in\bar{A}} b(r_i)\right) = \frac{1}{2} = b(\infty) = b(s)$$

and if A is empty then

$$\frac{D_K(P||Q)}{D_H(P,Q)} \geq \inf_{i\in\bar{A}} b(r_i) = b(s).$$

Hence the lower bound follows in either case. □

It is easy to verify that the function $b(z)$ is decreasing and continuous, $b(0) = \infty$, $b(1) = 1$ and $b(\infty) = 1/2$. Therefore, the lemma shows that when the ratios q_i/p_i are near 1 then $D_K(P||Q) \approx D_H(P,Q)$.

Let $z = \inf_{\theta,\theta^*\in\Theta \text{ and } y\in Y} \frac{P(y|\theta)}{P(y|\theta^*)}$. By Lemma 5,

$$\frac{1}{2}\Delta_H(\theta^*,\theta) \leq \Delta_K(\theta^*,\theta) \leq b(z)\Delta_H(\theta^*,\theta)$$

for all $\theta, \theta^* \in \Theta$. Hence when z is positive, these distances always differ by at most a fixed constant factor. Upper bounds on $I(\Theta;Y^n)$ and $C(\Theta;Y^n)$ in the same form as the lower bounds given above are easily obtained in this case.

Theorem 3 *If there exists $b < \infty$ such that $\Delta_K(\theta^*,\theta) \leq b\Delta_H(\theta^*,\theta)$ for all $\theta^* \in \Theta$ and in particular, if*

$$\inf_{\theta,\theta^*\in\Theta \text{ and } y\in Y} \frac{P(y|\theta)}{P(y|\theta^*)} < \infty,$$

then

1. $$\limsup_{n\to\infty} \frac{I(\Theta;Y^n)}{\ln n} \leq \frac{\overline{\dim}_{H,P_\Theta}(\Theta,\Delta_H^{1/2})}{2}$$

   and

2. $$\limsup_{n\to\infty} \frac{C(\Theta;Y^n)}{\ln n} \leq \frac{\overline{\dim}_K(\Theta,\Delta_H^{1/2})}{2}.$$

Proof: The particular conditions for the existence of the finite b follow from Lemma 5, as mentioned above. When such a b exists, it follows from the upper bound of Theorem 1 that $I(\Theta;Y^n) \leq L_{\sqrt{1/(bn)},P_\Theta}(\Theta,\Delta_H^{1/2})$. Part (1) then follows using Lemma 3 part (1) with $\epsilon = \sqrt{1/n}$, and part (2) follows similarly, using also Lemma 4. □

Note that when the upper and lower metric and information dimensions used in Theorems 2 and 3 match, as is typically the case, then (modulo the other assumptions stated in the theorems) these theorems give an exact asymptotic characterization of the mutual information and minimax risk, showing each is proportional to $\ln n$, and giving the constant of proportionality.

5 Bounds on instantaneous information gain

Let us define the instantaneous information gain at time n by

$$\begin{aligned}I^\Delta(\Theta;Y^n) &= \mathbb{E}_{\Theta,Y^n} \ln \frac{P(y_n|y^{n-1},\theta)}{P(y_n|y^{n-1})} \\ &= \mathbb{E}_{\Theta,Y^n} \ln \frac{P(y_n|\theta)}{P(y_n|y^{n-1})}\end{aligned}$$

for countable Y and similarly using densities for continuous Y. (Here y^{n-1} denotes $y_1,\ldots,y_{n-1}$, as above.) The last equality follows from the assumed conditional independence of the Y_t. The quantity $I^\Delta(\Theta;Y^n)$ can be viewed as the average amount of information gained about Θ by observing Y_n, over and above the information gained from observing $Y_1,\ldots,Y_{n-1}$. It is easily verified that

$$\sum_{t=1}^n I^\Delta(\Theta;Y^t) = I(\Theta;Y^n),$$

so the total average instantaneous information gain is the same as the average total information gain, or mutual information, as expected (see e.g. [19]). As the instantaneous information gain is of interest in many areas, including computational learning theory [20, 1], we briefly discuss here how our results may be used to analyze it.

Lemma 6 *Let $a_1, a_2, \ldots$ be a sequence of real numbers such that $a = \lim_{n\to\infty} \frac{a_n}{\ln n}$ exists (possibly being infinite). Assume $b_1, b_2, \ldots$ is a sequence of real numbers such that $a_n = \sum_{t=1}^n b_t$ for all n. Then $\lim_{n\to\infty} nb_n = a$ if this limit exists.*

Proof: Suppose $\lim_{n\to\infty} nb_n = b < \infty$. Then for all $\epsilon > 0$ there exists an n_0 such that for all $t > n_0$, $|b_t - b/t| \leq \epsilon/t$. Since there is a constant c such that $|\sum_{t=1}^n 1/t - \ln n| \leq c$, we have for all $n \geq n_0$,

$$\begin{aligned}|a_n - b\ln n| &\leq bc + |\sum_{t=1}^n (b_t - b/t)| \\ &\leq bc + |\sum_{t=1}^{n_0} (b_t - b/t)| + \epsilon\ln n + \epsilon c.\end{aligned}$$

Hence $\lim_{n\to\infty} \frac{a_n}{\ln n} = b$. The argument for the case of infinite b is similar. □

It follows from this lemma that whenever

$$\lim_{n\to\infty} \frac{I(\Theta; Y^n)}{\ln n} = d \in [0, \infty],$$

for example, as in the cases described in the Gaussian example in Section 3 and in Section 4 of this paper, then either $\lim_{n\to\infty} nI^{\Delta}(\Theta; Y^n) = d$ or this limit does not exist. Thus whenever $\lim_{n\to\infty} nI^{\Delta}(\Theta; Y^n)$ exists and $\lim_{n\to\infty} \frac{I(\Theta;Y^n)}{\ln n} = d \in (0, \infty)$, then we get the typical growth rate for the average instantaneous information gain of

$$I^{\Delta}(\Theta; Y^n) \approx \frac{d}{n}.$$

In particular, we have the following

Theorem 4 *Assume that both* $\mathbf{dim}_{H,P_\Theta}(\Theta, \Delta_H^{1/2})$ *and* $\lim_{n\to\infty} nI^{\Delta}(\Theta; Y^n)$ *exist, the former is finite, and there exists* $b < \infty$ *such that* $\Delta_K(\theta^*, \theta) \le b\Delta_H(\theta^*, \theta)$ *for all* $\theta^* \in \Theta$. *Then*

$$\lim_{n\to\infty} nI^{\Delta}(\Theta; Y^n) = \frac{\mathbf{dim}_{H,P_\Theta}(\Theta, \Delta_H^{1/2})}{2}.$$

Proof: Follows directly from Theorems 2 and 3, and Lemma 5. □

Since $I^{\Delta}(\Theta; Y^n)$ is a lower bound on the average instantaneous log loss on the nth observation of any adaptive method of prediction, this gives a useful lower bound for computational learning theory applications.

6 Conclusion

Via a new pair of inequalities (Theorem 1), we have given general asymptotic estimates of the mutual information $I(\Theta; Y^n)$ and minimax risk $C(\Theta; Y^n)$ in terms of the dimension of the metric space $(\Theta, \Delta_H^{(1/2)})$ and related quantities. Some assumptions were required in obtaining our results; it would be interesting to see how these can be weakened. Also, we have treated only the finite dimensional case, which leads to an $O(\ln n)$ scaling law for these quantities. It would be interesting to see what general scaling laws can be established in the infinite dimensional case using Theorem 1. The possibility of using Theorem 1 to obtain bounds in the thermodynamic limit of large sample size and large dimension (see e.g. [21]) should also be explored. Finally, in practice it is important to get accurate estimates of $I(\Theta; Y^n)$ for moderate sample size n. We believe that the bounds given in Theorem 1 will be quite close in most practical cases, but this remains to be verified.

Acknowledgements

We thank Andrew Barron and Tom Cover for helpful suggestions in this research.

References

[1] S. Amari and N. Murata. Statistical theory of learning curves under entropic loss. *Neural Computation*, 5:140–153, 1993.

[2] A. Barron. The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem. *The Annals of Probability*, 13:1292–1303, 1985.

[3] A. Barron. The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Dept. of Statistics, U. Ill. Urbana-Champaign, 1987.

[4] A. Barron, B. Clarke, and D. Haussler. Information bounds for the risk of bayesian predictions and the redundancy of universal codes. In *Proc. International Symposium on Information Theory.*

[5] A. Barron and T. Cover. A bound on the financial value of information. *IEEE Trans. on Information Theory*, 34:1097–1100, 1988.

[6] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift fuer Wahrscheinlichkeitstheorie und verwandte gebiete*, 65:181–237, 1983.

[7] L. Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probability theory and related fields*, 71:271–291, 1986.

[8] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

[9] B. Clarke. *Asymptotic cumulative risk and Bayes risk under entropy loss with applications.* PhD thesis, Dept. of Statistics, University of Ill., 1989.

[10] B. Clarke and A. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

[11] B. Clarke and A. Barron. Jefferys' prior is asymptotically least favorable under entropy risk. *J. Statistical Planning and Inference*, 41:37–60, 1994.

[12] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley, 1991.

[13] R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2–142, 1984.

[14] S. Y. Efroimovich. Information contained in a sequence of observations. *Problems in Information Transmission*, 15:178–189, 1980.

[15] J. D. Farmer, E. Ott, and J. A. Yorke. The dimension of chaotic attractors. *Physica D*, 7:153–180, 1983.

[16] E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12:929–989, 1984.

[17] R. Hasminskii and I. Ibragimov. On density estimation in the view of Kolmogorov's ideas in approximation theory. *Annals of statistics*, 18:999–1010, 1990.

[18] D. Haussler. A general minimax result for relative entropy. 1995.

[19] D. Haussler and A. Barron. How well do Bayes methods work for on-line prediction of $\{+1, -1\}$ values? In *Proceedings of the Third NEC Symposium on Computation and Cognition*. SIAM, 1992.

[20] D. Haussler, M. Kearns, and R. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning*, 14:84–114, 1994.

[21] D. Haussler, M. Kearns, H. S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. In *Proceedings of the Seventh Annual ACM Workshop on Computational Learning Theory*, 1994.

[22] D. Haussler, J. Kivinen, and M. Warmuth. Tight worst-case loss bounds for predicting with expert advice. Technical Report UCSC-CRL-94-36, University of California at Santa Cruz, Computer and Information Sciences, 1994.

[23] D. Haussler and M. Opper. Mutual information and Bayes methods for learning a distribution. In *Proc. Workshop on the Theory of Neural Networks: The Statistical Mechanics Perspective*. World Scientific, 1995. to appear.

[24] I. Ibragimov and R. Hasminskii. On the information in a sample about a parameter. In *Second Int. Symp. on Information Theory*, pages 295–309, 1972.

[25] T. Kawabata and A. Dembo. The rate-distortion dimension of sets and measures. *IEEE Trans. on Info. Th.*, 40:1564–1572, 1994.

[26] A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Translations (Ser. 2)*, 17:277–364, 1961.

[27] L. LeCam. *Asymptotic methods in statistical decision theory*. Springer, 1986.

[28] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

[29] M. Opper and D. Haussler. General bounds for predictive errors in supervised learning. In *Proc. Workshop on the Theory of Neural Networks: The Statistical Mechanics Perspective*. World Scientific, 1995. to appear.

[30] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes (Transl.)*. Holden Day, 1964.

[31] D. Pollard. *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Math. Stat. and Am. Stat. Assoc., 1990.

[32] E. Posner, E. Rodemich, and H. Rumsey. Epsilon-entropy of stochastic processes. *Ann. Math. Statist.*, 38:1000–1020, 1967.

[33] E. Posner, E. Rodemich, and H. Rumsey. Epsilon-entropy of gaussian processes. *Ann. Math. Statist.*, 40:1272–1296, 1969.

[34] E. Posner and R. Rodemich. Epsilon-entropy and data compression. *Ann. Math. Statist.*, 42:2079–2125, 1971.

[35] E. Posner and R. Rodemich. Epsilon-entropy of proability distributions. In L. C. et.al., editor, *Sixth Berkeley Sym. on Math., Stat. and Prob.*, volume 2, pages 699–707. 1972.

[36] A. Renyi. On the dimension and entropy of probability distributions. *Acta Math. Acad. Sci. Hung.*, 10:193–215, 1959.

[37] A. Renyi. On the amount of information concerning an unknown parameter in a sequence of observations. *Publ. Math. Inst. Hungar. Acad. Sci.*, 9:617–625, 1964.

[38] J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.

[39] K. Symanzik. Proof and refinements of an inequality of feynman. *J.Math. Phys.*, 6:1155–, 1965.

[40] S. van deGeer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Annals of Statistics*, 21:14–44, 1993.

[41] K. Yamanishi. A learning criterion for stochastic rules. *Machine Learning*, 1992. Special Issue on the Proceedings of the 3nd Workshop on Computational Learning Theory.