EPSILON ENTROPY OF PROBABILITY DISTRIBUTIONS

EDWARD C. POSNER and EUGENE R. RODEMICH
JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY

1. Introduction

This paper summarizes recent work on the theory of epsilon entropy for probability distributions on complete separable metric spaces. The theory was conceived [3] in order to have a framework for discussing the quality of data storage and transmission systems.

The concept of data source was defined in [4] as a probabilistic metric space: a complete separable metric space together with a probability distribution under which open sets are measurable, so that the Borel sets are measurable. An ε partition of such a space is a partition by measurable ε sets, which, depending on context, can be sets of diameter at most ε or sets of radius at most $\frac{1}{2}\varepsilon$, that is, sets contained in spheres of radius $\frac{1}{2}\varepsilon$. The entropy $H(U)$ of a partition U is the Shannon entropy of the probability of the distribution consisting of the measures of the sets of the partition. The (one shot) epsilon entropy of X with distribution μ, $H_{\varepsilon;\mu}(X)$, is defined by

$$H_{\varepsilon;\mu}(X) = \inf_{U} \{H(U);\ U \text{ an } \varepsilon \text{ partition}\} \tag{1.1}$$

and, except for roundoff in the entropy function, a term less than 1, $H_{\varepsilon;\mu}(X)$ is the minimum expected number of bits necessary to describe X to within ε when storage is not allowed. The inf in (1.1) was shown to be a min in [4].

For X a compact metric space, Kolmogorov's epsilon entropy $H_\varepsilon(X)$ is defined as

$$H_\varepsilon(X) = \min_{U} \{\log \text{ card } (U);\ U \text{ an } \varepsilon \text{ partition}\} \tag{1.2}$$

and, except for roundoff in the logarithm, is the minimum number of bits necessary to describe X to within ε when words of fixed length are used.

Suppose one does experiments from X independently and then attempts storage or transmission. That is, take a cartesian product $X^{(n)}$ of X, with product measure $\mu^{(n)}$ and supremum metric. Thus, ε sets in the product are the subsets

This paper presents the results of one phase of research carried out under Contract NAS 7-100, sponsored by the National Aeronautics and Space Administration.

of products of ε sets. This is the notion that insures that knowledge of outcomes to within ε in $X^{(n)}$ forces knowledge to within ε in each of the factors. Then the following limit can be proved to exist:

$$I_{\varepsilon;\mu}(X) = \lim_{n\to\infty} \frac{1}{n} H_{\varepsilon;\mu^{(n)}}(X^{(n)}), \tag{1.3}$$

and is called the absolute epsilon entropy of X. It represents the minimum expected number of bits per sample needed to describe a sequence of independent outcomes of X when arbitrary storage between experiments can be used. Similarly, define the absolute epsilon entropy of the compact metric space X as

$$I_{\varepsilon}(X) = \lim_{n\to\infty} \frac{1}{n} H_{\varepsilon}(X^{(n)}), \tag{1.4}$$

with the same definition for the metric on $X^{(n)}$.

2. Relations with channel coding

It was shown in [3] that

$$I_{\varepsilon;\mu}(X) = \inf_{\rho} \{I(\rho);\ \rho \in R_{\varepsilon}(X)\}, \tag{2.1}$$

where $R_{\varepsilon}(X)$, in the case of the radius definition, is the class of probability distributions on $X \times X$ which are supported within $\frac{1}{2}\varepsilon$ of the diagonal, with a more complicated definition for the diameter case. Here $I(\rho)$ stands for mutual information. We do not know if the inf need be attained. However, (2.1), coupled with the continuity of $I_{\varepsilon;\mu}(X)$ from above in ε, proved in [3], allows us to prove a strong channel coding theorem and its converse [3]:

"If K is a memoryless channel with capacity Γ less than $I_{\varepsilon;\mu}(X)$ (of finite capacity if $I_{\varepsilon;\mu}(X)$ is infinite), then it is not possible to transmit outcomes of X over K such that, with probability approaching 1, an arbitrarily large fraction of a long sequence of outcomes are known to within an error not much more than ε. But if Γ is greater than $I_{\varepsilon;\mu}(X)$ (assuming $I_{\varepsilon;\mu}(X)$ is finite), then it is possible to transmit outcomes of X over the channel such that, with probability approaching 1, all the outcomes are known to within ε."

3. A useful inequality

An important and useful inequality relating $H_{\varepsilon;\mu}$ and $I_{\varepsilon;\mu}$ was proved in [3]:

$$H_{\varepsilon;\mu}(X) \leqq I_{\varepsilon;\mu}(X) + \log^{+} I_{\varepsilon;\mu}(X) + C, \tag{3.1}$$

where C is a universal constant. In other words, it doesn't help much to store independent experiments if the entropy is large. This result, coupled with (2.1), makes it easy to obtain asymptotic bounds on $H_{\varepsilon;\mu}(X)$, which we shall do in subsequent sections.

4. Relation between $I_{\varepsilon;\mu}$ and I_ε

Here is a surprising relation between $I_{\varepsilon;\mu}(X)$ and $I_\varepsilon(X)$ for X compact [1]:

$$I_\varepsilon(X) = \max_\mu I_{\varepsilon;\mu}(X) \tag{4.1}$$

for all but countably many ε, a condition which we now know cannot be removed. What this result means is that Nature can choose a μ on X so bad that nothing can be saved by using variable length coding. The proof uses von Neumann's minimax theorem to prove, as an intermediate step, that

$$\max_\mu I_{\varepsilon;\mu}(X) = \log \frac{1}{v(\varepsilon)}, \tag{4.2}$$

where $v(\varepsilon)$ is the value of the zero sum two person game, which, in the radius case, has as its space of pure strategies the points of X, with payoff 0 or 1 to the first player; the payoff is 1 if and only if the points chosen by the two players have distance at most $\frac{1}{2}\varepsilon$.

5. Finite dimensional spaces

The differential entropy of a density function p an Euclidean n space is defined, when it exists, as

$$H(p) = \int p \log \frac{1}{p}\, dm(x), \tag{5.1}$$

where $m(x)$ is Lebesgue measure. The relation between differential entropy and epsilon entropy was considered in [2]. The metric can be any norm $\|\cdots\|_S$ on n space, where S, of Lebesgue measure v_1, is a compact symmetric convex set in E^n, and is the unit sphere under $\|\cdots\|_S$. Let p be a sufficiently nice density function on E^n, so that E^n is a probabilistic metric space with probability μ given by $\mu(A) = \int_A p\, dm$. Then as $\varepsilon \to 0$,

$$H_{\varepsilon;\mu}(E^n) = n \log \frac{2}{\varepsilon} + \log \frac{1}{v_1} + H(p) + C(S) + o(1) \tag{5.2}$$

for a constant $C(S)$ called the entropic packing constant. Furthermore, $C(S)$ is between 0 and 1, and is 0 if and only if translates of S fill E^n. Moreover, $C(S)$ as a function of S is continuous in the Hausdorff metric, in which the distance between two compact sets is the maximum over the two sets of the distance of a point in one of the two sets from the other set. A somewhat analogous result holds for H_ε of the unit n cube, but the analogous $C'(S)$, the deterministic packing constant, is not bounded by 1 but rather can be as large as $(1 - o(1))$ $\log n$ as $n \to \infty$.

If a Borel probability μ on E^n with Euclidean distance has mean 0 and a second moment $\sigma^2 = E\|x\|^2$, $H_{\varepsilon;\mu}(E^n)$ is finite, even though (5.2) does not hold [4]; in fact, for small ε,

$$H_{\varepsilon;\mu}(E^n) \leqq n \log \frac{1}{\varepsilon} + \frac{n}{2} \log n + n \log \sigma + 1 + \log (2 + \sqrt{\pi}). \tag{5.3}$$

The normal distribution comes under either of those two cases. When $n = 1$, the unique minimizing partition is known [6]. It is the partition by consecutive intervals of length ε such that the mean is in the center of one of the intervals. The proof is hard; a simpler one would be nice. For $n > 1$, minimizing partitions are not known, even for the independent normal of equal variances.

6. Epsilon entropy in $L_2[0, 1]$

The bound of (5.3) depends on n, and, in fact, examples can be given that show this dependence can actually occur. It is not surprising then, that if $L_2[0, 1]$ is made into a probabilistic metric space by the measure induced by a mean continuous separable stochastic process on the unit interval, then the epsilon entropy can be infinite, even though the expectation of $\|x\|^2$ is always finite for such a process. In fact, [4] proved that a given convergent sequence $\{\lambda_n\}$ of nonnegative numbers written in nonincreasing order is the set of eigenvalues of some mean continuous stochastic process on the unit interval of infinite epsilon entropy for some $\varepsilon > 0$ if and only if

$$\sum n\lambda_n = \infty. \tag{6.1}$$

Conversely, if $\Sigma\, n\lambda_n = \infty$, there is a process with infinite epsilon entropy for every $\varepsilon > 0$. Thus, a slightly stronger condition than finite second moment is necessary to insure finite epsilon entropy in the infinite dimensional case.

7. Product entropy of Gaussian processes

In this section, we shall consider the definition of *product entropy* if $X = L_2[0, 1]$, but only for mean continuous Gaussian processes. We shall defer the case of the epsilon entropy of Gaussian processes until Section 9. Product entropy $J_\varepsilon(X)$ is defined as the minimum entropy over all *product* epsilon partitions of $L_2[0, 1]$. A product epsilon partition is a (countable) epsilon partition of all of X except a set of probability zero by sets which are hyperrectangles (with respect to eigenfunction coordinates) of diagonal epsilon. Thus,

$$J_\varepsilon(X) \geqq H_\varepsilon(X). \tag{7.1}$$

Surprisingly, $J_\varepsilon(X)$ is infinite for one ε if and only if it is infinite for all ε, and is finite if and only if the "entropy of the eigenvalues" $\Sigma\, \lambda_n \log 1/\lambda_n$ is finite, where the eigenvalues are written in nonincreasing order [5]. We have no good explanation of why the entropy of the eigenvalues occurs as the condition for

finite epsilon entropy. Furthermore, $\Sigma \lambda_n \log 1/\lambda_n$ is necessary and sufficient in order that there be a product epsilon partition, and is also the condition that there be a hyperrectangle of positive probability and finite diameter. Incidentally, $J_\varepsilon(X)$ depends only on the eigenvalues of the Gaussian process, as do $H_{\varepsilon;\mu}(X)$ and $I_{\varepsilon;\mu}(X)$, where μ is the measure on $L_2[0, 1]$ induced by the mean continuous Gaussian process. In [6], product entropy is estimated rather precisely in terms of the eigenvalues and the optimum product partitions found by a variational argument. By remarks of Section 5, the optimal partitions are products of centered partitions on each coordinate axis.

The interpretation of product entropy is as follows. One wishes to transmit sample functions of the process so that one knows outcomes to within ε in the L_2 norm, but only wishes to consider methods which involve correlating the sample function with the eigenfunction and then sending a quantized version of these correlations. Since the diagonal of the product sets has diameter ε, the method guarantees knowledge of the sample function to within ε. Unfortunately, as we shall see, this method is not very good compared to the optimal compression schemes which are not restricted to product partitions but can use arbitrary ε partitions.

In [5], conditions are given which guarantee either

$$J_\varepsilon(X) = O(H_{\varepsilon;\mu}(X)) \tag{7.2}$$

or

$$J_\varepsilon(X) \sim H_{\varepsilon;\mu}(X) \tag{7.3}$$

as $\varepsilon \to 0$ for a mean continuous Gaussian process. The condition for (7.2) is that the sum of the eigenvalues beyond the nth (in nonincreasing order) be $O(n\lambda_n)$. For (7.3), the condition is that the sum be $o(n\lambda_n)$. In the first case, in fact,

$$J_\varepsilon(X) = O(L_\varepsilon(X)) \tag{7.4}$$

and in the second

$$J_\varepsilon(X) \sim L_\varepsilon(X), \tag{7.5}$$

where $L_\varepsilon(X)$ is a general lower bound for the epsilon entropy of a mean continuous Gaussian process to be discussed later.

For a stationary band limited Gaussian process on the unit interval with continuous spectral density,

$$\lambda_n \sim n^{-1}(Cn)^{-2n}, \tag{7.6}$$

C constant, as $n \to \infty$. Thus, the $o(n\lambda_n)$ condition is satisfied and $L_\varepsilon(X)$ can be evaluated. The final result is

$$J_\varepsilon(X) \sim \frac{\frac{1}{2}\left(\log \frac{1}{\varepsilon}\right)^2}{\log \log \frac{1}{\varepsilon}} \tag{7.7}$$

as $\varepsilon \to 0$. Now

$$J_\varepsilon(X) \sim n \log \frac{1}{\varepsilon} \tag{7.8}$$

as $\varepsilon \to 0$ if the process has only finitely many nonzero eigenvalues, n in number. So in the case at hand where there are infinitely many nonzero eigenvalues, the growth of $J_\varepsilon(X)$ had to be faster than any constant times $\log(1/\varepsilon)$. The rate of growth given by (7.8) is not much faster, however. This is an expression of the fact that the sample functions from such a process are entire functions, hence not very random, and they should be easy to approximate.

8. Entropy in $C[0, 1]$

The case of $C[0, 1]$ is much more difficult than $L_2[0, 1]$, partly because it is hard to determine whether a given process has continuous paths. However, in [7] it is proved that if the mean continuous separable stochastic process on $[0, 1]$ satisfies,

$$\begin{aligned} E(x(0))^2 &\leqq A, \\ E(x(s) - x(t))^2 &\leqq A|s - t|^a, \qquad s, t \in [0, 1], \end{aligned} \tag{8.1}$$

for some $A \geqq 0$, $1 < a \leqq 2$, then the paths are continuous with probability 1, a known result, and, if μ is the measure induced by the process on $C[0, 1]$, then

$$H_{\varepsilon;\mu} \leqq C(a) A^{1/a} \varepsilon^{-2/a} \tag{8.2}$$

for $\varepsilon < \sqrt{A}$, where $C(a)$ depends only on a. The proof was achieved by constructing an ε partition of $C[0, 1]$ using uniformly spread points on $[0, 1]$ and facts about the modulus of continuity of the process that are forced by the given conditions on the covariance function.

Conditions for finite entropy in function spaces other than $L_2[0, 1]$ and $C[0, 1]$ have not been looked for or found, even in the case of Gaussian processes. In that case, a bound like (8.2) is found for $C[0, 1]$, which is valid under weaker conditions on the covariance function of the process. Multidimensional time processes have not been considered at all.

9. Bounds for L_2 Gaussian processes

Various lower bounds have been found for L_2 Gaussian processes. The first is the bound L_ε of [6]. This bound is defined as follows. Let $\{\lambda_n\}$ in nonincreasing order be the eigenvalues of the process. Define $b = b(\varepsilon)$ for $\varepsilon > 0$ by

$$\begin{aligned} \sum \frac{\lambda_n}{1 + b(\varepsilon)\lambda_n} &= \varepsilon^2, & \varepsilon^2 &< \sum \lambda_n, \\ b(\varepsilon) &= 0, & \varepsilon^2 &\geqq \sum \lambda_n. \end{aligned} \tag{9.1}$$

Then

$$L_\varepsilon(X) = \frac{1}{2}\sum \log\,[1 + \lambda_n b(\varepsilon)] \tag{9.2}$$

is a lower bound for $H_{\varepsilon;\mu}(X)$. It was derived from the obvious inequality

$$H_{\varepsilon;\mu}(X) \geqq E \log \frac{1}{\mu[S_\varepsilon(x)]}, \tag{9.3}$$

and so is quite weak. (The term $S_\varepsilon(x)$ is the ball of center x and radius ε.) A stronger bound $M_\varepsilon(X)$, never any worse, is

$$M_\varepsilon(X) = L_{\varepsilon/2}(X) - \frac{1}{8}\varepsilon^2 b\left(\frac{1}{2}\varepsilon\right). \tag{9.4}$$

This bound is derived by bounding the probability density drop in a Gaussian distribution under translation, and using the fact, proved in [6], that the sphere of radius $\frac{1}{2}\varepsilon$ about the origin in L_2 under a Gaussian distribution is at least as probable as any set of diameter ε. The $M_\varepsilon(X)$ bound is the best asymptotic lower bound we have for arbitrary L_2 Gaussian processes, but, for special ones, the next section gives an improvement.

The $L_\varepsilon(X)$ lower bound was introduced chiefly because it is also proved in [5] that

$$H_\varepsilon(X) \lesssim L_{m\varepsilon}(X) \qquad \text{for any } m < \frac{1}{2}. \tag{9.5}$$

This difficult proof uses products of partitions of finite dimensional eigensubspaces of the process where the dimension increases without bound. In each finite dimensional subspace, "shell" partitions are used, partitions which are composed of partitions between regions of concentric spheres of properly varying radii. The deterministic epsilon entropy of the n sphere in Euclidean space needed to be estimated in the proof. This proves the finiteness of $H_\varepsilon(X)$ for X a mean continuous Gaussian process on $[0, 1]$.

From results of Section 3, we know $H_\varepsilon(X) \lesssim I_\varepsilon(X)$. The measure ρ on $X \times X$, which is defined by choosing a point of the first factor according to μ and then assigning probability in $S_{\varepsilon/2}(x)$ according to μ, has, as in [1], mutual information

$$I(\rho) \leqq E\left(\log \frac{1}{\mu[S_{\varepsilon/2}(x)]}\right). \tag{9.6}$$

Thus, for any probabilistic metric space,

$$H_{\varepsilon;x}(X) \lesssim E\left(\log \frac{1}{\mu[S_{\varepsilon/2}(x)]}\right). \tag{9.7}$$

This, coupled with (9.3), gives a pair of bounds valid in general, but not readily computable.

10. Examples of L_2 entropy bounds

The bounds of the previous section lead to some interesting asymptotic expressions which yield quick answers when the eigenvalues of the process are known asymptotically. For example, a Gaussian process arising as the solution of a linear differential equation with constant coefficients driven by white Gaussian noise can be handled. For such processes, [8] shows that

$$\lambda_n \sim An^{-p}, \qquad p > 1, A > 0, \tag{10.1}$$

and shows how to find A and p from the equation with simple calculations. The results of the previous section then yield

$$M_\varepsilon(X) \leqq H_\varepsilon(X) \lesssim L_{\varepsilon/2}(X), \tag{10.2}$$

$$M_\varepsilon(X) \sim (p-1)\left(\frac{\pi}{p \sin \pi/p}\right)^{p/(p-1)} 2^{2/(p-1)-1}\left(\frac{A}{\varepsilon^2}\right)^{1/(p-1)}, \tag{10.3}$$

$$L_\varepsilon(X) \sim \frac{p}{2}\left(\frac{\pi}{p \sin \pi/p}\right)^{p/(p-1)} \frac{A}{\varepsilon^2}\Bigg)^{1/(p-1)} \tag{10.4}$$

For example, if X is the Wiener process

$$E[X(s)X(t)] = \min(s, t), \qquad s, t \in [0, 1], \tag{10.5}$$

then

$$\lambda_n = \frac{1}{\pi^2(n - 1/2)^2}, \qquad n \geqq 1, \tag{10.6}$$

and so $A = 1/\pi^2, p = 2$, and

$$\frac{1}{2\varepsilon^2} \lesssim H_\varepsilon(X) \lesssim \frac{1}{\varepsilon^2}. \tag{10.7}$$

A better lower bound valid in general for L_2 is given by

$$\begin{aligned} H_\varepsilon(X) &\geqq M_\varepsilon(X) + \frac{1}{2}\sum E\left\{\frac{\lambda_n x_n^2 q^2(x)}{[\varepsilon + \lambda_n q(x)]^2}\right\} \\ &= N_\varepsilon(X), \end{aligned} \tag{10.8}$$

say, where $q(x) = 0, \|x\| \leqq \varepsilon$, and

$$\sum \frac{x_n^2}{[\varepsilon + \lambda_n q(x)]^2} = 1, \qquad \|x\| > \varepsilon. \tag{10.9}$$

Here x_n denotes the component of x along the nth eigenfunction. Equation (10.8) is a refined version of (9.4), and is useful only when the eigenvalues decrease slowly enough so that $q(x)$ is almost deterministic. The condition turns out to be satisfied if (10.1) is, and allows $N_\varepsilon(X)$ to be found asymptotically as

$$N_\varepsilon(X) \sim (p-1)\left(\frac{\pi}{p \sin \pi/p}\right)^{p/(p-1)}\left[2^{2/(p-1)} + \frac{1}{2}p^{-p/(p-1)}\right]\left(\frac{A}{\varepsilon^2}\right)^{1/(p-1)} \tag{10.10}$$

Thus, for X the Wiener process, (10.7) can be improved to

$$\frac{17}{32\varepsilon^2} \lesssim H_\varepsilon(X) \lesssim \frac{1}{\varepsilon^2}. \tag{10.11}$$

The result given by (10.11) is all we know about the entropy of the Wiener process. Our lower bounds just are not good enough to prove our conjecture

$$H_\varepsilon(X) \sim \frac{1}{\varepsilon^2} \tag{10.12}$$

for X the Wiener process.

Notice, however, that for the Wiener process on $C[0, 1]$, where one might expect the entropy to be much larger because of the more stringent covering requirement, Section 8 yields

$$H_\varepsilon(X \text{ in } C[0, 1]) = O(H_\varepsilon(X \text{ in } L_2[0, 1])). \tag{10.13}$$

In fact, (10.13) holds for any Gaussian process satisfying the eigenvalue condition (10.1). With this surprising result we close the paper.

REFERENCES

[1] R. J. McEliece and E. C. Posner, "Hide and seek, data storage, and entropy," *Ann. Math. Statist.*, Vol. 2 (1971), pp. 1706–1716.

[2] E. C. Posner and E. R. Rodemich, "Differential entropy and tiling," *J. Statist. Phys.*, Vol. 1 (1969), pp. 57–69.

[3] ———, "Epsilon entropy and data compression," *Ann. Math. Statist.*, Vol. 42 (1971), pp. 2079–2125.

[4] E. C. Posner, E. R. Rodemich, and H. Rumsey, Jr., "Epsilon entropy of stochastic processes," *Ann. Math. Statist.*, Vol. 38 (1967), pp. 1000–1020.

[5] ———, "Product entropy of Gaussian distributions," *Ann. Math. Statist.*, Vol. 40 (1969), pp. 870–904.

[6] ———, "Epsilon entropy of Gaussian distributions," *Ann. Math. Statist.*, Vol. 40 (1969), pp. 1272–1296.

[7] E. R. Rodemich and E. C. Posner, "Epsilon entropy of stochastic processes with continuous paths," in preparation.

[8] H. Widom, "Asymptotic behavior of eigenvalues of certain integral operators," *Arch. Rational Mech. Anal.*, Vol. 17 (1967), pp. 215–229.