Correspondence

A Note on the Absolute Epsilon Entropy

James A. Bucklew, *Member, IEEE*

***Abstract*—A new characterization of the absolute epsilon entropy of an ergodic information source is presented. This characterization allows one to derive a new lower bound to this quantity, which is shown to be tight in some interesting special cases.**

***Index Terms*—Epsilon entropy, quantization, rate-distortion function.**

I. Introduction

The concept of the absolute epsilon entropy (ϵ-entropy) was introduced and developed in a series of papers [1]–[5] to describe the problem of data storage or transmission with a maximum error constraint. Loosely speaking, we desire to compress a source to R bits per sample where it is required that we must be able to reconstruct the original source sequence with error on each sample that never exceeds ϵ. It is desired to make R as small as possible subject to this constraint. This subject is of special theoretical interest because this sort of distortion measure is not the sample average of a single letter distortion measure (the usual case in classical rate-distortion theory). Important generalizations of many of the key concepts of absolute ϵ-entropy were undertaken by the study of a very general class of fidelity criteria called zero-one distortion measures [6].

A major result of this subject is that the absolute ϵ-entropy can be expressed as an information theoretic minimization. In this correspondence we give an alternate characterization of the absolute ϵ-entropy and show that in some interesting special cases closed form expressions for it can be obtained.

We caution the reader that there are more than one usage of the term ϵ-entropy. In the setting of a compact metric space valued alphabet A, it has been defined as $n^{-1}\log[N_{n,\epsilon}]$ where $N_{n,\epsilon}$ is the number of balls of radius ϵ or less that is needed to cover $A^n \equiv \times_{i=1}^n A$. Note there is no probability measure present or needed for this definition. This ϵ-entropy is discussed briefly in references [3], [8]. Another use of the word ϵ-entropy is an alternative naming of the unnormalized rate-distortion function. A recent reference making use of this terminology is [9].

II. Preliminaries

Let us now give some definitions and present some needed preliminaries. Let $(A, \mathcal{F})$ be a measurable space. A will denote the source alphabet. $(A^n, \mathcal{F}_n)$ will denote the usual product measurable space whose elements are n-tuples. $(A^\infty, \mathcal{F}_\infty)$ will denote the space consisting of A^∞, the set of all infinite sequences $(x_1, x_2, \cdots)$ from A, and the infinite product σ-algebra $\mathcal{F}_\infty$. Let T: $A^\infty \to A^\infty$ be the shift operator $T(x_1, x_2, \cdots) = (x_2, x_3, \cdots)$. We define our source μ to be a probability measure on A^∞, which is stationary and ergodic with respect to T.

Manuscript received July 13, 1990.

The author is with the Department of Electrical and Computer Engineering, 1415 Johnson Dr., University of Wisconsin–Madison, Madison, WI 53706.

IEEE Log Number 9039283.

Let d be a single letter distortion measure on A that is a measurable mapping from $A \times A$ into $\mathbb{R}^+$. Let d_n be the distortion measure on A^n defined as

$$d_n((x_1, x_2, \cdots, x_n), (y_1, y_2, \cdots, y_n)) = \max\{d(x_i, y_i): 1 \le i \le n\}.$$

We may define the ball of radius ϵ and center x in A^n by $S_x(\epsilon) \equiv \{y \in A^n: d_n(x, y) < \epsilon\}$. We say that a set $E \subset A^n$ is ϵ-admissible if $E \subset S_x(\epsilon)$ for some $x \in A^n$. By a partition of A^n, we will mean a countable partition of A^n into measurable sets. A partition is ϵ-admissible if every set in the partition is ϵ-admissible.

Let X^n: $A^\infty \to A^n$ be the projection $X^n(x_1, x_2, \cdots) = (x_1, x_2, \cdots, x_n)$. Let μ_n be the measure induced on A^n by μ. Define the n-dimensional ϵ-entropy H_n as

$$H_n = \inf\left\{ -\sum_{E \in \mathcal{P}} \mu_n(E)\log \mu_n(E): \mathcal{P} \text{ is an } \epsilon\text{-admissible partition}\right\}.$$

We may now define the absolute ϵ-entropy as

$$I_\epsilon \equiv \lim_{n\to\infty} \frac{1}{n} H_n,$$

where the limit is known to exist [6]. One of the major results of this theory is to show that the absolute ϵ-entropy can be computed as some sort of information theoretic minimization, rather than a minimization over all ϵ-admissible partitions as the definition suggests. To state this result we need a few more definitions.

Let $q(\cdot|\cdot)$ be a transition probability on A^n. That is, for each $x \in A^n$, $q(\cdot|x)$ is a probability measure on $(A^n, \mathcal{F}_n)$, and for each $E \in \mathcal{F}_n$, $q(E|\cdot)$ is an $\mathcal{F}_n$ measurable function. Let $\mathcal{Q}_n$ be the collection of all transition probabilities q on A^n such that $q(S_x(\epsilon)|x) = 1$, $x \in A^n$. Note that $\mathcal{Q}_n$ depends on ϵ. If $q \in \mathcal{Q}_n$, let μq denote the probability measure on $(A^n \times A^n, \mathcal{F}_n \times \mathcal{F}_n)$ such that $\mu q(E \times F) = \int_E q(F|x)\, d\mu_n(x)$, $E, F \in \mathcal{F}_n$. Let ν_n denote the probability measure induced on $(A^n, \mathcal{F}_n)$ by taking $\nu_n(F) \equiv \mu q(A^n \times F)$. If U_n, V_n are, respectively, the projections $(x, y) \to x$ and $(x, y) \to y$ from $A^n \times A^n$ to A^n, let $I^*_{n,\epsilon}(\mu q)$ be the mutual information of the pair (U_n, V_n) calculated with respect to μq (i.e., $I^*_{n,\epsilon}(\mu q) \equiv \int \log(d\mu q / d\mu_n \times \nu_n)\, d\mu q$). We now define

$$I^*_\epsilon \equiv \lim_{n\to\infty} \inf_{q \in \mathcal{Q}_n} \frac{1}{n} I^*_{n,\epsilon}(\mu q),$$

where the limit is known to exist [6]. We may now state a theorem.

Theorem 1 (Posner and Rodemich [3], Kieffer [6]): $I^*_\epsilon = I_\epsilon$.

The coding theorems of ϵ-entropy are of greatest interest. In [6] the key construction is concerned with zero-one distortion measures. In our setting we define a sequence of distortion measures (indexed by block length for $x, y \in A^n$) as $\rho_n(x, y) = 0$ if $d_n(x, y) < \epsilon$, $\rho_n(x, y) = 1$ otherwise. A block code on sequences of length n is a finite set $B \subset A^n$. The rate of the code

is $(1/n)\log|B|$. For $x \in A^n$, define $\rho_n(x|B) \equiv \min\{\rho_n(x,y)\colon y \in B\}$. Each $x \in A^n$ is coded into a $y \in B$ such that $\rho_n(x,y) = \rho_n(x|B)$. The average distortion $\bar{\rho}(B)$, resulting from using the block code B to code the source, is $\bar{\rho}(B) = \int \rho_n(x|B)\, d\mu_n$. The coding theorem and its converse can now be stated.

Theorem 2 (Kieffer [6]): Suppose that there is an ϵ-admissible partition $\mathscr{P}$ of A such that $-\sum_{E \in \mathscr{P}} \mu_1(E)\log \mu_1(E) < \infty$. Then for each n, there exists a block code B_n on sequences of length n such that $\lim_{n\to\infty}(1/n)\log|B_n| = I_\epsilon$ and $\lim_{n\to\infty} \bar{\rho}_n(B_n) = 0$.

Theorem 3 (Kieffer [6]): Let $R < I_\epsilon$. Then $\inf\{\bar{\rho}_n(B)\colon B \subset A^n, (1/n)\log|B| \le R, n = 1,2,\cdots\} > 0$.

III. Development

The problem with the absolute ϵ-entropy is that, even with the characterization of Theorem 1, it is still difficult to compute. Our main result in this correspondence is to give an alternative characterization that will lead to some closed form expressions for I_ϵ and a useful lower bound. Let us first define some quantities. Let $\mathscr{M}_n$ denote the collection of all probability measures on A^n. Then let

$$R_{n,\epsilon} \equiv \inf_{Q \in \mathscr{M}_n} -\int \log\left[Q(S_x(\epsilon))\right] d\mu_n(x).$$

We may now define

$$R_\epsilon \equiv \varliminf_{n\to\infty} \frac{1}{n} R_{n,\epsilon}.$$

Lemma 1: $R_\epsilon = \lim_{n\to\infty}(1/n)R_{n,\epsilon}$.

Proof of Lemma 1: We show that $R_{n,\epsilon}$ is a subadditive sequence (a sequence $\{z_n\}$ is subadditive if $z_{n+m} \le z_n + z_m$, $m,n = 1,2,\cdots$). The lemma will then follow from the fact (see [7, p. 618]) that if $\{z_n\}$ is a nonnegative subadditive sequence, then $\lim_{n\to\infty}(1/n)z_n$ exists and equals $\inf_n (1/n)z_n$.

$$\begin{aligned}
&R_{n+m,\epsilon} \\
&\quad = \inf_{Q \in \mathscr{M}_{n+m}} -\int \log\left[Q(S_x(\epsilon))\right] d\mu_{n+m}(x) \\
&\quad = \inf_{Q \in \mathscr{M}_{n+m}} -\int \log\left[Q(S_{x_1}(\epsilon)) \cap Q(S_{x_2}(\epsilon))\right] d\mu_{n+m}(x) \\
&\qquad\qquad \text{(where } x = (x_1, x_2) \in A^n \times A^m\text{)} \\
&\quad \le \inf_{Q \in \mathscr{M}_{n+m}\colon Q = Q_1 \times Q_2,\, Q_1 \in \mathscr{M}_n,\, Q_2 \in \mathscr{M}_m} \\
&\qquad -\int \log\left[Q(S_{x_1}(\epsilon)) \cap Q(S_{x_2}(\epsilon))\right] d\mu_{n+m}(x) \\
&\quad = \inf_{Q \in \mathscr{M}_{n+m}\colon Q = Q_1 \times Q_2,\, Q_1 \in \mathscr{M}_n,\, Q_2 \in \mathscr{M}_m} \\
&\qquad \cdot \left[-\int \log\left[Q(S_{x_1}(\epsilon))\right] d\mu_n(x) \right. \\
&\qquad\quad \left. -\int \log\left[Q(S_{x_2}(\epsilon))\right] d\mu_m(x) \right] \\
&\quad = \inf_{Q \in \mathscr{M}_n} -\int \log\left[Q(S_{x_1}(\epsilon))\right] d\mu_n(x) \\
&\qquad + \inf_{Q \in \mathscr{M}_m} -\int \log\left[Q(S_{x_2}(\epsilon))\right] d\mu_m(x) \\
&\quad = R_{n,\epsilon} + R_{m,\epsilon}.
\end{aligned}$$

□

Theorem 4: $R_\epsilon = I_\epsilon$.

Proof of Theorem 4: For every $q \in \mathscr{D}_n$, it is known ([3, Lemma 13]) that $I^*_{n,\epsilon}(\mu q) \ge -\int \log[\nu_n(S_x(\epsilon))]\, d\mu_n(x)$. Hence we can take the infimum over all $q \in \mathscr{D}_n$ on both sides of the inequality. We can make the right-hand side smaller if we extend the infimum to include all $Q \in \mathscr{M}_n$ instead of just those Q that correspond to induced measures ν_n. Hence, after scaling by $1/n$ and taking the limit as $n \to \infty$, we have $I^*_\epsilon \ge R_\epsilon$. Hence by Theorem 1, $I_\epsilon \ge R_\epsilon$.

For the reverse inequality, we consider a random coding scheme. We use the zero-one distortion measure as in the setting of Theorems 2 and 3. Suppose $I_\epsilon > R > R_\epsilon$ for some real number R. We wish to show a contradiction.

Fix a $\delta > 0$ so that $R > R_\epsilon + \delta$. Choose an integer n and a $Q^*_n \in \mathscr{M}_n$ such that $R_\epsilon + \delta > (1/n)\int \log[Q^*_n(S_x(\epsilon))]\, d\mu_n(x)$. We then choose $[\exp(mnR)]$ independent mn-dimensional code words for our block code B_m, each consisting of m independent n-tuples from the distribution Q^*_n. We now wish to bound the average distortion produced by such a procedure.

Fix some $x = (x_1, x_2, \cdots, x_m)$ where each component $x_i \in A^n$, $1 \le i \le m$. First consider the probability that one codeword $y = (y_1, y_2, \cdots, y_m)$ is within ϵ of x:

$$\begin{aligned}
&Q^*_n(\max\{d_n(x_1,y), d_n(x_2,y), \cdots, d_n(x_m,y)\} \le \epsilon) \\
&\qquad = \prod_{i=1}^m Q^*_n(S_{x_i}(\epsilon)).
\end{aligned}$$

Let us define $p_n(x)$ as the probability that no codeword is within ϵ of x. We then have

$$\begin{aligned}
p_n(x) &= \left[1 - \prod_{i=1}^m Q^*_n(S_{x_i}(\epsilon))\right]^{[\exp(mnR)]} \\
&\le \exp\left(-[\exp(mnR)] \cdot \prod_{i=1}^m Q^*_n(S_{x_i}(\epsilon)) \right).
\end{aligned}$$

We may upper bound the ensemble average zero-one distortion of this random coding procedure by

$$\begin{aligned}
E\{\bar{\rho}(B_m)\} &= \int p_n(x)\, d\mu_n(x) \\
&\le \int \exp\left(-[\exp(mnR)] \cdot \prod_{i=1}^m Q^*_n(S_{x_i}(\epsilon)) \right) d\mu_n(x)
\end{aligned}$$

Now note that by the mean ergodic theorem (if μ is n-ergodic for all positive integers n, otherwise utilize an ergodic mode argument as in [8, pp. 278–281]) we have

$$\begin{aligned}
\lim_{m\to\infty} \frac{1}{m} \log \prod_{i=1}^m Q^*_n(S_{x_i}(\epsilon)) &= \int \log\left[Q^*_n(S_x(\epsilon))\right] d\mu_n(x) \\
&> -n[R_\epsilon + \delta] > -nR.
\end{aligned}$$

Hence

$$\lim_{m\to\infty} [\exp(mnR)] \cdot \prod_{i=1}^m Q^*_n(S_{x_i}(\epsilon)) = \infty.$$

By the Lebesgue dominated convergence theorem, we have

$$\begin{aligned}
&\lim_{m\to\infty} E\{\bar{\rho}(B_m)\} \\
&\quad \le \lim_{m\to\infty} \int \exp\left(-[\exp(mnR)] \cdot \prod_{i=1}^m Q^*_n(S_{x_i}(\epsilon)) \right) d\mu_n(x) = 0.
\end{aligned}$$

Since this is the ensemble behavior, there must exist at least one block code with this distortion performance. However, this is a contradiction to Theorem 3 and the converse inequality must be true. □

Corollary 1: Suppose that $\{x_n\}$ is a sequence of real valued random variables with ergodic probability measure μ. Further, suppose that μ_n has a probability density $m_n(\cdot)$ for all $n < \infty$ and $d(x, y) = |x - y|$ for $x, y \in \mathbb{R}$. Then $I_\epsilon \geq H(X) + \log[1/2\epsilon]$, where $H(X)$ is the differential entropy of the random variable sequence.

Proof of Corollary 1: Note that $\int_{S_x(\epsilon)} dx = (2\epsilon)^n$. Consider

$$R_{n,\epsilon} = \inf_{Q \in \mathcal{M}_n} -\int \log\left[Q(S_x(\epsilon))\right] d\mu_n(x)$$

$$= \inf_{Q \in \mathcal{M}_n} -\int \log\left[Q(S_x(\epsilon))\right] m_n(x)\, dx$$

$$= -n \log[2\epsilon]$$

$$+ \inf_{Q \in \mathcal{M}_n} -\int \log\left[\int_{S_x(\epsilon)} dQ/(2\epsilon)^n\right] m_n(x)\, dx$$

$$\geq -n \log[2\epsilon]$$

$$+ -\int \log\left[m_n(x)\right] m_n(x)\, dx$$

$$= -n \log[2\epsilon] + H(x_1, x_2, \cdots, x_n)$$

where $H(x_1, x_2, \cdots, x_n)$ is the differential entropy of $(x_1, x_2, \cdots, x_n)$. We now scale by $1/n$ and take the limit as $n \to \infty$ to obtain the corollary statement. □

Corollary 2: Let $\{x_n\}$ be an i.i.d. sequence of absolutely continuous real valued random variables. Let $d(x, y) = |x - y|$ and denote the probability density of x_1 as $m(\cdot)$. Let $U[-\epsilon, \epsilon]$ denote the uniform distribution on the interval $[-\epsilon, \epsilon]$. If there exists some probability density $q(\cdot)$ on R such that $m(x) = q(x)^* U[-\epsilon, \epsilon]$, then $I_\epsilon = -\log[2\epsilon] + H(X)$.

Proof of Corollary 2: The proof follows immediately by noting that

$$\int_{S_x(\epsilon)} dQ/(2\epsilon)^n = Q^* \underset{i=1}{\overset{n}{\times}} U[-\epsilon, \epsilon]$$

(the convolution of Q with the uniform distribution over the n-dimensional rectangle of side 2ϵ centered at the origin). The optimal choice for Q will then be that which makes this convolution equal to $m_n(x) = \prod_{i=1}^n m(x_i)$. □

Remark: Interestingly enough, if $m = q^* U[-\epsilon, \epsilon]$ for some q, then $M(\omega) = Q(\omega) \cdot [\sin(\omega\epsilon)/\omega\epsilon]$ where $M(\cdot)$ and $Q(\cdot)$ are the characteristic functions of the densities m and q, respectively. This condition implies that $M(\omega)$ necessarily must have zeros at least at the zeros of $[\sin(\omega\epsilon)/\omega\epsilon]$. Hence this is a fairly restrictive condition. As an aside, the condition that $M(\omega)$ have zeros at the zeros of $\sin(\omega\epsilon)/\omega\epsilon$ is precisely the necessary condition of the so-called Quantization Theorem. The Quantization Theorem states that if and only if $M(\cdot)$ has this zero set, then a uniform quantizer with step size 2ϵ will have a uniformly distributed quantization error [10]. For a variety of source distributions such as uniforms, triangulars, a uniform quantizer will attain the R_ϵ rate distortion performance bound.

IV. Discussion and Conclusion

In the study of block codes in references [3], [6], a failure to encode an n-tuple with maximum distortion less than ϵ is only accorded a distortion of one. If it is desired to absolutely make sure that the error would never exceed ϵ, then a more appropriate distortion for a block decoding failure would be to assign a distortion of ∞. It is easy to see that no block code with a finite number of codewords could have distortion less than ∞ for any sequence of i.i.d. random variables whose probability distribution has unbounded support. One would have to consider block codes with infinite numbers of codewords, and measure the rates by entropy instead of the logarithm of the number of codewords. Another possibility is to consider the following coding scheme.

Suppose that the source is ergodic. Imagine a coding system where the transmitter and receiver have an infinite tape of i.i.d. samples taken from the distribution $Q_n^* \in \mathcal{M}_n$. Fix $R, \delta > 0$, and m_0 to be a positive integer. Suppose we have a block of m ($\geq m_0$) n-dimensional source letters to be transmitted. Choose a code by selecting $\exp[nmR]$ m-tuples of samples from the tape. If the source sequence can be coded with maximum distortion less than $\epsilon + \delta$, then transmit the appropriate codeword and set $m = m_0$ and consider the next source n-tuple. If no codeword is close enough, then take the next source letter n-tuple and append it to the previous block to form an $m+1$ source letter n-tuple. Choose a new code of $\exp[(m+1)nR]$ $(m+1)$-tuples from the tape, and see if any codeword is within maximum distortion of less than $\epsilon + \delta$, etc. Since this is a variable block length coding scheme, there must be additional overhead of transmitting the block length. This adds an additional $\log[m]/m$ term to the needed transmission rate to send a block of length m. This overhead can be made as small as desired by taking m_0 large.

It is not *a priori* clear that anything would ever be transmitted. Using the techniques in the proof of Theorem 4, we can show that for μ-a.e. sequence, this process terminates with probability 1 as long as $R > R_{n,\epsilon}$. Furthermore, the expected amount of time to wait until a transmission is finite.

References

[1] E. Posner, E. Rodemich, and H. Rumsey Jr., "Epsilon entropy of stochastic processes," *Ann. Math. Stat.*, vol. 38, pp. 1000–1020, 1967.

[2] E. Posner and E. Rodemich, "Differential entropy and tiling," *J. Stat. Phys.*, vol. 1, no. 1, pp. 57–69, 1969.

[3] ____, "Epsilon entropy and data compression," *Ann. Math. Stat.*, vol. 42, no. 6, pp. 2079–2125, 1971.

[4] R. McEliece and E. Posner, "Hiding and covering in a compact metric space," *Ann. Stat.*, vol. 1, no. 4, pp. 729–739, 1973.

[5] E. Posner, "Random coding strategies for minimum entropy," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 4, pp. 389–391, July 1975.

[6] J. Kieffer, "Block coding for an ergodic source relative to a zero-one valued fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 4, pp. 432–437, July 1978.

[7] N. Dunford and J. Schwartz, *Linear Operators — Part I.* New York: Wiley Interscience, 1957.

[8] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[9] H. Rosenthal and J. Binia, "On the epsilon entropy of mixed random variables," *IEEE Trans. Inform. Theory*, vol. IT-34, no. 5, pp. 1110–1114, Sept. 1988.

[10] A Sripad and D. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-25, no. 3, pp. 442–448, Oct. 1977.