Fixed-Rate Universal Lossy Source Coding and Rates of Convergence for Memoryless Sources

Tamás Linder, Gábor Lugosi, and Kenneth Zeger, *Senior Member, IEEE*

Abstract— **A fixed-rate universal lossy coding scheme is introduced for independent and identically distributed (i.i.d.) sources. It is shown for finite alphabet sources and arbitrary single letter distortion measures that as the sample size n grows the expected distortion obtained using this universal scheme converges to Shannon's distortion rate function $D(R)$ at a rate $O(\log n/n)$. The scheme can be extended to universal quantization of real i.i.d sources subject to a squared error criterion. It is shown in this case that the per-letter distortion converges to $D(R)$ at a rate $O(\sqrt{\log n/n})$ both in expectation and almost surely for any real-valued bounded i.i.d. source.**

Index Terms—**Universal source coding, convergence rates, vector quantization.**

I. Introduction

A UNIVERSAL lossy source code (or universal source code subject to a fidelity criterion) is a code whose performance for any source in a given family approaches the performance of the optimal code for that particular source as the length n of the encoded source sequence increases. Different types of universality for lossy coding are defined in [15] following similar concepts for noiseless coding given by Davisson [7]. The type of universality to be dealt with in this paper is weak minimax universality, where the universal code is assumed to perform in an asymptotically (i.e., for large block length) optimal manner for each source in a given class, however not necessarily uniformly over the class for a finite block length n.

The first result demonstrating the existence of a weakly minimax fixed-rate universal lossy source code for the class of stationary sources was given by Ziv [26] under certain assumptions on the source and reproduction alphabets and on the distortion measure. The conditions of Ziv's result are satisfied, for example, for the mean-squared distortion and the class of real-valued stationary sources whose marginals have finite second moment. An extension of this result and various new results dealing with different types of universality in fixed-rate coding of stationary sources are proved in [15]. Both necessary and sufficient conditions are given by Kieffer [9] for the existence of weak minimax universal variable rate noiseless and fixed-rate lossy codes for the class of stationary and ergodic sources.

The above results deal with the general problem of universal coding of stationary, or stationary and ergodic sources, and thus inherently do not provide convergence rates for the average distortion (fixed-rate case) or the average rate (variable-rate case) to the optimum value as the block length increases. Indeed, it was proved by Shields [18] that there does not exist any universal rate of convergence for the rate redundancy in universal noiseless coding of ergodic sources. However, for certain smaller classes of sources, very sharp results are known for universal noiseless coding. For example, for discrete memoryless sources with alphabet size L it is known that the minimax lower bound to the per letter average redundancy is $[(L-1)/2]n^{-1}\log n$, asymptotically for large n, and this convergence rate is achievable pointwise (see Krichevsky and Trofimov [10] and the references therein). Rissanen [17] provides a $(d/2)n^{-1}\log n$ lower bound for classes of "smoothly" parametrized sources, where d is the dimension of the parameter space. This result includes in particular the class of memoryless sources and the class of Markov sources of any given order.

Far fewer and less general results are available for universal lossy coding. For discrete memoryless sources Yu and Speed [22] have demonstrated the existence of a variable-length universal scheme where the average code length of the code approaches the rate distortion function $R(D)$ at a rate $O(n^{-1}\log n)$, while the per-letter distortion is pointwise upper-bounded by D. Their result applies to classes of independent and identically distributed (i.i.d.) sources over a finite alphabet such that, at a given rate, the second-order partial derivatives of the rate distortion function with respect to the source probabilities are bounded uniformly for each source. Unfortunately, they do not give an easier characterization of these source classes. Linder, Lugosi, and Zeger [11] analyzed Ziv's [26] fixed-rate universal scheme and established a $O(\sqrt{\log\log n/\log n})$ rate of convergence of the mean-square error on the class of real-valued bounded i.i.d. sources. The slower convergence rate apparently results from the fact that this scheme is universal over the larger class of all stationary and ergodic sources.

Ziv's [26] scheme was also analyzed in [24] and [2] by assuming both high resolution (i.e., high rate) and a fixed vector dimension. However, no rigorous rates of convergence have been derived from [24] and [2] for universal lossy coding. What is missing to make that connection is a rigorous

Manuscript received January 14, 1994; revised June 2, 1994. This research was supported in part by the National Science Foundation and the Joint Services Electronics Program.

T. Linder is with the Department of Telecommunications, Technical University of Budapest, Budapest, Hungary.

G. Lugosi is with the Department of Mathematics, Faculty of Electrical Engineering, Technical University of Budapest, Budapest, Hungary.

K. Zeger is with the Coordinated Science Laboratory, Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, IL 61801 USA.

IEEE Log Number 9410409.

theory of high-resolution quantization as the vector dimension increases. This problem was circumvented in [3], where a variety of two-stage coding schemes (including Ziv's method) were rigorously analized.

The present paper introduces a fixed-rate universal lossy coding scheme for memoryless sources. We show in Section II, Theorem 1 that for all finite alphabet memoryless sources the distortion of the scheme approaches the distortion rate function at least as fast as $cn^{-1}\log n$ with an increasing block length n, where the constant c is explicitly identified. The idea behind the construction is an extension of the "enumerative" (Lynch [12], Davisson [6], and Cover [4]) or "combinatorial" (Shtarkov and Babkin [19]) method of universal noiseless coding. First, the type (i.e., empirical distribution or composition) of the source sequence of length n is encoded, then the sequence itself is encoded using a fixed-rate source code which is optimal with respect to the uniform distribution on the type class (the set of sequences of length n with the same empirical distribution). A key tool used in our proof of the convergence rate is a result of Pilc [16], who established a $O(n^{-1}\log n)$ convergence rate in Shannon's source coding theorem for discrete memoryless sources on finite alphabets. Our scheme can also be extended to universal lossy coding of Markov sources, giving an $O(n^{-1}\log n)$ rate redundancy as the "price of universality." The code constructions given may not be practical, but are used to demonstrate the existence of universal codes with the claimed properties.

In Section III our construction in Section II for finite alphabets is extended to infinite alphabets where we obtain convergence rates for universal coding of real memoryless sources subject to a squared error criterion. We show in Theorem 2 that the per-letter mean-square distortion of the resulting universal quantization scheme converges to the distortion rate function at a rate $O(\sqrt{n^{-1}\log n})$ in expectation for all bounded memoryless sources. A more involved analysis reveals (Theorem 3) that the same convergence rate also holds for the per-letter sample distortion almost surely. We also demonstrate that a slight modification of the scheme results in a code weakly minimax universal over the class of memoryless sources with finite second moment. Then we investigate the "price of universality" (the rate and distortion redundancy of the scheme when compared with the nth-order operational distortion rate function of each source evaluated at rate R) for this modified scheme for unbounded memoryless sources. We show that the price of universality is on the order of $(n^{-1}\log n)^{1/2-\epsilon}$, where ϵ can be made arbitrarily small by assuming that the source has a finite kth moment for k large enough.

In Section IV we compare our result for finite alphabets with the result of Yu and Speed, which is the only other presently known result of this type. The two results provide the same speed of convergence, though Yu and Speed's code provides $O(n^{-1}\log n)$ convergence of the expected code length, with the sample distortion almost surely upper-bounded by a constant, while our code has constant transmission rate and a $O(n^{-1}\log n)$ convergence of the expected distortion. A recent work by Zhang *et al.* [25] reports $n^{-1}\log n$ type lower bounds on the distortion (resp., rate) redundancy of the best n-length fixed-rate (resp., variable-rate) lossy source code designed for a memoryless source. This means that both Yu and Speed's and our convergence rate is optimal, although the best constants are currently unknown. As seen in Section III, our scheme is easily modified for real sources, while there seem to exist technical difficulties to extending Yu and Speed's universal code. We know of no universal quantization results comparable to our Theorems 2 and 3.

II. The Finite Alphabet Case

Let $\mathcal{A}$ and $\mathcal{B}$ be two *finite* sets called the source and reproduction alphabets, respectively, and let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent identically distributed random variables taking values in $\mathcal{A}$. Suppose we are given a single letter distortion measure $d : \mathcal{A} \times \mathcal{B} \to [0, \infty)$ with

$$\min_{y\in\mathcal{B}} d(x, y) = 0$$

for all $x \in \mathcal{A}$. Let $\mathbf{E}$ denote the expectation operator. The (average) distortion between $x^n = (x_1, \cdots, x_n) \in \mathcal{A}^n$ and $y^n = (y_1, \cdots, y_n) \in \mathcal{B}^n$ is given as

$$d_n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i).$$

The *distortion rate function* $D(R)$ (see Berger [1]) of $\{X_n\}$ is defined for any $R \geq 0$ as

$$D(R) = \min_{I(X;Y)\leq R} \mathbf{E}d(X, Y),$$

where the minimum is taken over all pairs of random variables (X, Y) taking values in $\mathcal{A} \times \mathcal{B}$ such that X has the common distribution P of the X_n, and the mutual information of X and Y is less than or equal to R.

An n-length block code is a mapping $g : \mathcal{A}^n \to \mathcal{B}^n$. Let $|g| = |g(\mathcal{A}^n)|$ denote the cardinality of the range of g. The *rate* of g is defined as $R = n^{-1}\log|g|$. Here and throughout this paper we use base–2 logarithm. The (expected) distortion $\Delta(g)$ of the code g is given by

$$\Delta(g) = \mathbf{E}d_n(X^n, g(X^n))$$

where $X^n = (X_1, \cdots, X_n)$. Let $D_n(R)$ denote the minimum distortion which can be achieved by encoding $\{X_n\}$ with an n-length block code of rate not exceeding R, i.e.

$$D_n(R) = \min_{g:n^{-1}\log|g|\leq R} \Delta(g)$$

where the minimum is taken over all $g : \mathcal{A}^n \to \mathcal{B}^n$ satisfying the rate constraint. Shannon's source coding theorem states that $D_n(R) \geq D(R)$ for all n, and $D_n(R) \to D(R)$ as $n \to \infty$. The following upper bound on the rate of this convergence is given by Pilc:

Propostion 1 (Pilc [16]): If $R > 0$ and $D(R) > 0$, then the minimum expected distortion in fixed-rate coding an n-block from a finite alphabet memoryless source is upper-bounded as

$$D_n(R) - D(R) \leq \left(\frac{|D'(R)|}{2} + o(1)\right)\left(\frac{\log n}{n}\right),$$

where $D'(R)$ is the derivative of $D(\cdot)$ at R.

We remark that a careful inspection of the proof of Pilc's result reveals that the $o(1)$ term above converges to zero uniformly as $n \to \infty$ in a small enough neighborhood of R. Pilc's result asserts that for a *particular given* discrete memoryless source there exists a sequence of such fixed-rate codes which produce distortion only $O(n^{-1}\log n)$ higher than the ultimate lower bound $D(R)$. This sequence of codes may be different for each different source. We next demonstrate the existence of a *fixed* sequence of codes $g_n : \mathcal{A}^n \to \mathcal{B}^n$, $n = 1, 2, \cdots$ of rate not exceeding R that achieves this $O(n^{-1}\log n)$convergence rate *universally* for *any* memoryless source over $\mathcal{A}$ whose distortion rate function evaluated at R is positive.

Theorem 1: For all $R > 0$ there exists a sequence of fixed-rate codes $\{g_n\}_{n=1}^{\infty}$ with rate $n^{-1}\log|g_n| \leq R$, such that for any i.i.d source $\{X_n\}$ on a finite alphabet $\mathcal{A}$ with $D(R) > 0$ the difference between the expected squared error of the length n code g_n and the distortion rate function is bounded as

$$\Delta(g_n) - D(R) \leq |D'(R)|(|\mathcal{A}| + 1/2 + o(1))\left(\frac{\log n}{n}\right).$$

The construction of the code will be an extension of a method of noiseless universal coding [19]. The general idea is most clearly explained in [7, Theorem 6]. In our case, the procedure consists of partitioning $\mathcal{A}^n$ into $k(n)$ sets $A_1, \cdots, A_{k(n)}$, where each A_i has the property that X^n is uniformly distributed over A_i given $X^n \in A_i$. If $X^n \in A_i$, the index i is encoded using at most $\log k(n) + 1$ bits. Then a source code with minimum expected distortion with respect to the uniform distribution on A_i is designed, and X^n is encoded using this code. The receiver can decode this code since the index i was transmitted. In our case, the A_i's will be the n-length sequences with the same empirical distribution.

In the spirit of Ziv's pioneering paper in universal lossy coding [26] and Wyner's results in the analysis of convergence rates for quantization schemes [20], [21], we have chosen the parameters determining the overall rate of the universal code so that the rate redundancy is zero, i.e., $n^{-1}\log|g_n| \leq R$.

Proof: The *type* P_{x^n} of a sequence $x^n \in \mathcal{A}^n$ is its empirical distribution, defined as

$$P_{x^n}(a) = \frac{1}{n}|\{i : x_i = a, 1 \leq i \leq n\}|, \qquad a \in \mathcal{A}.$$

For a given probability distribution $\hat{P}$ on $\mathcal{A}$ let $T_{\hat{P}}^n = \{x^n \in \mathcal{A}^n : P_{x^n} = \hat{P}\}$ be the set of all x^n having the same empirical distribution as $\hat{P}$, a possibly empty set. $T_{\hat{P}}^n$ is called the *type class* of x^n, if $\hat{P} = P_{x^n}$. Note that the number of different types (hence the number of type classes) is upper-bounded by $(n+1)^{|\mathcal{A}|}$. The terminology above is that of Csiszár and Körner [5], although we will not need to fully exploit their powerful "method of types."

For each probability distribution Q on $\mathcal{A}^n$, let $g_{Q,R}$ denote an n-length source code of rate R which is optimal with respect to Q, i.e., $|g_{Q,R}| \leq 2^{nR}$, and for any other n-length source code g with $|g| \leq 2^{nR}$

$$\mathbf{E}_Q[d_n(X^n, g_{Q,R}(X^n))] \leq \mathbf{E}_Q[d_n(X^n, g(X^n))] \tag{1}$$

where we used the notation

$$\mathbf{E}_Q[f(X^n)] = \sum_{x^n \in \mathcal{A}^n} Q(x^n) f(x^n) \tag{2}$$

for the expectation with respect to Q of a real-valued function f on $\mathcal{A}^n$.

Our code is described as follows. For a given source sequence x^n with type $\hat{P} = P_{x^n}$ let Q denote the uniform distribution concentrated on the type class $T_{\hat{P}}^n$ of x^n. Set

$$R_n = R - (|\mathcal{A}|/n)\log(n+1) \tag{3}$$

which is a positive number for n large enough. Now we define our code g_n by

$$g_n(x^n) = g_{Q,R_n}(x^n)$$

that is, g_n maps x^n to the codeword that the code of rate R_n, which is optimal with respect to Q, assigns to x^n. Since there can be more than one optimal code g_{Q,R_n}, the encoding rule is made unique by specifying a rule for choosing g_{Q,R_n} from among the possible candidates. Now $|g_{Q,R_n}| \leq 2^{nR_n}$ for each Q, and the number of different Q's is the same as the number of different type classes, which is upper-bounded by $(n+1)^{|\mathcal{A}|}$. Thus we conclude that $|g_n| \leq (n+1)^{|\mathcal{A}|}2^{nR_n} = 2^{nR}$, that is, g_n satisfies the rate constraint R.

We will demonstrate the simple but crucial observation that for any i.i.d. source $\{X_n\}$, the code g_n does as well as the optimal code of rate R_n matched to the source. Let P be the common distribution of the X_i's, and with some abuse of notation denote the distribution of the sequence X^n also by P. Then since the conditional distribution of X^n given the type $\hat{P}$ is uniform over $T_{\hat{P}}^n$, we have

$$\begin{aligned}\mathbf{E}[d_n(X^n, g_n(X^n))|P_{X^n} = \hat{P}] &= \mathbf{E}_Q[d_n(X^n, g_{Q,R_n}(X^n))] \\ &\leq \mathbf{E}_Q[d_n(X^n, g_{P,R_n}(X^n))] \\ &= \mathbf{E}[d_n(X^n, g_{P,R_n}(X^n))|P_{X^n} = \hat{P}]\end{aligned} \tag{4}$$

where the inequality follows from the defining optimality (1) of g_{Q,R_n}. Taking expectations we obtain

$$\begin{aligned}\Delta(g_n) &= \mathbf{E}d_n(X^n, g_n(X^n)) \leq \mathbf{E}d_n(X^n, g_{P,R_n}(X^n)) \\ &= D_n(R_n).\end{aligned} \tag{5}$$

Therefore

$$\begin{aligned}\Delta(g_n) - D(R) &\leq D_n(R_n) - D(R) \\ &= D_n(R_n) - D(R_n) \qquad (6) \\ &\quad + D(R_n) - D(R). \qquad (7)\end{aligned}$$

We can upper bound (6) by using Pilc's result (Proposition 1) as follows. As $n \to \infty$ we have

$$D_n(R_n) - D(R_n) \le \left(\frac{|D'(R_n)|}{2} + o(1)\right)\left(\frac{\log n}{n}\right) = \left(\frac{|D'(R)|}{2} + o(1)\right)\left(\frac{\log n}{n}\right) \quad (8)$$

since $R_n < R$, $R_n \to R$ as $n \to \infty$, and by the continuity of $D'(R)$ in R. The continuity of $D'(R)$ at any $R > 0$ for which $D(R) > 0$ follows from the existence of $D'(R)$ for such R and from the convexity of $D(R)$ in R (see e.g., [1]). An upper bound on (7) is given by using a first-order Taylor expansion of $D(\cdot)$ at R,

$$D(R_n) - D(R) \le |D'(R)|(R - R_n)(1 + o(1)) = (|D'(R)| \cdot |\mathcal{A}| + o(1))\frac{\log(n+1)}{n}. \quad (9)$$

From (8) and (9) we conclude that

$$\Delta(g_n) - D(R) \le |D'(R)|(|\mathcal{A}| + 1/2 + o(1))\left(\frac{\log n}{n}\right)$$

and the proof of the theorem is complete. ■

Remark: One may be interested in distortion criteria other than the expected distortion. We can consider, for example, the probability that the average sample distortion exceeds a certain distortion level D. Marton [13] investigated the asymptotic behavior of the sequence

$$-\frac{1}{n}\min_{g_n}\{\log \Pr\{d_n(X^n, g_n(X^n)) > D(R)\}\}$$

where the minimum is taken over all codes g_n with $n^{-1}\log|g_n| \le R'$ for a given $R' > R$. She identified the positive liminf and limsup of the above sequence as $n \to \infty$, thus establishing the error exponent in fixed-rate lossy source coding. In fact, the exponential upper bound on the minimum of $\Pr\{d_n(X^n, g_n(X^n)) > D(R)\}$ taken over all codes with $n^{-1}\log|g_n| \le R'$ is given by a code construction independent of the source, and is thus universally achievable.

Markov Sources

The construction of Theorem 1 easily extends to Markov sources just as in the case of combinatorial codes [19], [7], without affecting the $O(n^{-1}\log n)$ rate of convergence. For Markov sources, however, we do not have a counterpart of Pilc's theorem, and therefore we are forced to compare the scheme's distortion to $D_n(R)$ instead of $D(R)$. Thus the code will operate at a distortion not exceeding $D_n(R)$, and at a rate $r_n > R$, and we will investigate the price of universality, $r_n - R$. Note that in the previous theorem, by changing R_n to R, we have $\Delta(g_n) \le D_n(R)$, and the rate redundancy (in this case the price of universality) becomes $n^{-1}|\mathcal{A}|\log(n+1)$. In this case the distortion redundancy $\Delta(g_n) - D(R)$ is caused only by the "price of finite block length," $D_n(R) - D(R)$, which can be upper-bounded using Pilc's theorem.

To obtain a scheme for finite-alphabet kth-order stationary Markov sources first encode $X^k = (X_1, \cdots, X_k)$, $k < n$, using a fixed-rate code with at most $k\log L + 1$ bits (L is the alphabet size $|\mathcal{A}|$). Then the number of occurrences of each $x^{k+1} \in \mathcal{A}^{k+1}$ is counted on X^n

$$n(x^{k+1}) = \left|\{i : (X_i, \cdots, X_{i+k}) = x^{k+1}, 1 \le i \le n-k\}\right|. \quad (10)$$

The $n(x^{k+1})$'s, the so-called $(k+1)$-grams, can be noiselessly encoded in at most $L^{k+1}\log n + 1$ bits. By the kth-order Markov property the conditional distribution of X^n given the first k letters and the $(k+1)$-grams is *uniform* over the set of all $x^n \in \mathcal{A}^n$ having the given X^k prefix and the given $(k+1)$-grams. For this uniform distribution, the optimal (minimum-distortion) source code of rate R is designed and X^n is encoded using this code. By the same argument as in (4) and (5) (only R_n is replaced by R), we obtain

$$\Delta(g_n) \le D_n(R).$$

The overall rate r_n of g_n is

$$r_n \le R + \frac{1}{n}\left(k\log L + L^{k+1}\log n + 2\right).$$

Thus the price of universality is still $O(n^{-1}\log n)$, as in the memoryless case.

III. The Real Alphabet Case

Let the source and reproduction alphabets be the real line (i.e., $\mathcal{A} = \mathcal{B} = \mathcal{R}$), and let the distortion between sequences $x^n, y^n \in \mathcal{R}^n$ (n-dimensional Euclidean space) be measured by the average squared error

$$d_n(x^n, y^n) = \frac{1}{n}\|x^n - y^n\|^2 = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|^2.$$

A *memoryless stationary source* is a sequence of real-valued i.i.d. random variables $\{X_n\}_{n=1}^{\infty}$. For notational simplicity we introduce an auxiliary random variable X which has the common distribution of the X_i's. As in the finite-alphabet case, its distortion rate function $D(R)$ with respect to the mean squared error is defined for $R \ge 0$ as

$$D(R) = \inf_{I(X;Y)\le R}\mathbf{E}|X - Y|^2$$

and the infimum is taken over all pairs of real random variables (X, Y) such that X has the common distribution of the X_i's, and the mutual information between X and Y is at most R. The definition of $I(X;Y)$ for this general case is given, e.g., in [1, ch. 7]. An *n-length block code* is a mapping of $\mathcal{R}^n$ into a finite subset of $\mathcal{R}^n$. If the block code $g : \mathcal{R}^n \to \{y_1, \cdots, y_N\} \subset \mathcal{R}^n$ is onto, then we write $|g| = N$, and the *rate* of g is defined as $n^{-1}\log|g|$. Given the source $\{X_n\}_{n=1}^{\infty}$, the *expected distortion* of g is

$$\Delta(g) = \mathbf{E}d_n(X^n, g(X^n)) = \frac{1}{n}\mathbf{E}\|X^n - g(X^n)\|^2.$$

As before, let $D_n(R)$ denote the minimum distortion which can be achieved using an n-length block code of rate not exceeding R, i.e.

$$D_n(R) = \inf_{g: n^{-1}\log|g|\le R}\Delta(g).$$

By Shannon's source coding theorem (see, e.g., [1]) $D_n(R) \geq D(R)$ for all n, and when $\mathbf{E}|X|^2 < \infty$, we have $\lim_n D_n(R) = D(R)$, just as in the finite-alphabet case. Unfortunately, there are no known results for wide classes of real sources as strong as Pilc's finite-alphabet convergence rates. One exception is that Wyner [20] proved for memoryless Gaussian sources, that $D_n(R) - D(R) = O(n^{-1}\log n)$, and he later showed in [21] that the weaker bound $D_n(R) - D(R) = O(\sqrt{n^{-1}\log n})$ holds for any stationary Gaussian source with a spectral density having a Lipschitz-continuous derivative. Zamir and Feder [23] have shown in a recent work that an $O(n^{-1}\log n)$ rate redundancy is achievable for stationary Gaussian sources with nice spectral densities by means of a variable-rate scheme using subtractive dither. There are presently no known results providing a $O(n^{-1}\log n)$ rate of convergence for non-Gaussian real i.i.d. sources. For our purposes, however, an earlier result of the present authors will suffice.

Propostion 2 (Linder, Lugosi, Zeger [11]): Let $\{X_n\}$ be an i.i.d. real source of bounded support, i.e., $\Pr\{|X_1| \leq B\} = 1$ for some $B > 0$. Then for all $R > 0$ such that $D(R) > 0$ there exists a constant $c(R)$ such that

$$D_n(R) - D(R) \leq (c(R) + o(1))\sqrt{\frac{\log n}{n}}.$$

$c(R)$ is continuous in R, and the term $o(1)$ converges to zero uniformly in a small enough neighborhood of R.

The following result states that there exists a *fixed* sequence of block codes of length n and of rate not larger than R that achieves this $O(\sqrt{n^{-1}\log n})$ rate of convergence *universally* over the class of i.i.d. sources of bounded support.

Theorem 2: For any $R > 0$ there exists a sequence of fixed-rate codes $\{g_n\}_{n=1}^{\infty}$ with rate $n^{-1}\log|g_n| \leq R$, such that for any i.i.d. real source $\{X_n\}$ of bounded support with $D(R) > 0$, the difference between the expected squared error of the length n code g_n and the distortion rate function is bounded as

$$\Delta(g_n) - D(R) = O\left(\sqrt{\frac{\log n}{n}}\right).$$

Given a sequence of real random variables $Y_1, Y_2, \cdots$ and a sequence of real numbers $a_1, a_2, \cdots$, we say that $Y_n = O(a_n)$ almost surely (a.s.), if

$$\text{limsup}_{n\to\infty}(Y_n/a_n) < +\infty$$

with probability 1. The following theorem says that an $O(\sqrt{n^{-1}\log n})$ rate of convergence also holds for the sample distortion with probability 1. The theorem is proved in Appendix I.

Theorem 3: For any $R > 0$ there exists a sequence of fixed-rate codes $\{g_n\}_{n=1}^{\infty}$ with rate $n^{-1}\log|g_n| \leq R$, such that for any i.i.d. real source $\{X_n\}$ of bounded support with $D(R) > 0$, the difference between the squared error of the length n code g_n and the distortion rate function satisfies (with $X^n = (X_1, \cdots, X_n)$)

$$\frac{1}{n}\|X^n - g_n(X^n)\|^2 - D(R) = O\left(\sqrt{\frac{\log n}{n}}\right) \quad \text{a.s.}$$

Proof of Theorem 2: The construction is a generalization of the finite alphabet case.

Definitions for the code construction: Since the source has bounded support, there exists a positive integer M such that $\Pr\{|X| \leq M\} = 1$. For each $x^n = (x_1, \cdots, x_n) \in \mathcal{R}^n$ define the integer M_n by

$$M_n(x^n) = \left\lceil \max_{1\leq i\leq n} |x_i| \right\rceil. \tag{11}$$

Then we have $M_n \leq M$ a.s. For each positive integer L, let $U_{n,L}$ denote the L-level uniform scalar quantizer on the interval $[-M_n, M_n]$, i.e., for $j = 0, \cdots, L-1$

$$U_{n,L}(x) = -M_n + \frac{(2j+1)M_n}{L}$$

whenever

$$-M_n + \frac{2jM_n}{L} \leq x < -M_n + \frac{2(j+1)M_n}{L}$$

and also $U_{n,L}(M_n) = M_n(L-1)/L$. Note that both the domain $[-M_n, M_n]$ and the range

$$\left\{-M_n + \frac{(2j+1)M_n}{L} : j = 0, \cdots, L-1\right\} \tag{12}$$

of $U_{n,L}$ depend on the maximum of the absolute values of the X_i's, for $i = 1, \cdots, n$.

Let $\hat{P}_n$ denote the empirical distribution of $X^n = (X_1, \cdots, X_n)$, i.e., for each measurable set $A \subset \mathcal{R}$

$$\hat{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in A\}}$$

where I_B is the indicator of the event B. Define $\hat{P}_{n,L}$ as the empirical distribution of the quantized vector

$$U_{n,L}(X^n) = (U_{n,L}(X_1), \cdots, U_{n,L}(X_n)). \tag{13}$$

Let $T^n_{\hat{P}_n}$ be the set of all $x^n \in \mathcal{R}^n$ having "type" $\hat{P}_n$, and let $T^n_{\hat{P}_{n,L}}$ be the set of n-vectors, having type $\hat{P}_{n,L}$ (note that $T^n_{\hat{P}_{n,L}}$ is in the range of $U_{n,L}$). Finally, let Q_n and $Q_{n,L}$ be the uniform distributions on $T^n_{\hat{P}_n}$ and $T^n_{\hat{P}_{n,L}}$, respectively.

For any probability measure μ on $\mathcal{R}^n$ with a finite second moment, define $g_{\mu,R}$ as an n-length nearest neighbor code of rate R, which is optimal with respect to μ, i.e.

$$\mathbf{E}_\mu\|x^n - g_{\mu,R}(x^n)\|^2 \leq \mathbf{E}_\mu\|x^n - g(x^n)\|^2$$

for any code g of rate less than or equal to R. Such optimal codes do exist for distributions with finite second moments, although they are not necessarily unique. Note that we are always able to choose an optimal code $g_{\mu,R}$ that uses the nearest neighbor encoding rule, i.e., if the range of $g_{\mu,R}$ is $\{y_1, \cdots, y_N\}$, then $\forall x \in \mathcal{R}^n$

$$g_{\mu,R}(x) = y_i \text{ if } \|x - y_i\| \leq \|x - y_j\|, \quad \forall j \neq i$$

where "ties" are broken according to some rule. Also note that if $\mu([-c,c]^n) = 1$ for some $c > 0$, then $g_{\mu,R}$ can have no codevector outside $[-c,c]^n$.

Construction of the code: The main idea behind the construction of our code is as follows. First, we observe n data points $x_1, x_2, \cdots, x_n$. We transmit the smallest integer M_n which is greater than the largest (in magnitude) data point among $x_1, x_2, \cdots, x_n$. We use $2\lceil \log M_n \rceil + 2$ bits for this. Then we uniformly quantize (with L levels) the points $x_1, x_2, \cdots, x_n$ to obtain the quantized points $U_{n,L}(x_1), U_{n,L}(x_2), \cdots, U_{n,L}(x_n)$. The type of this quantized n-vector is then transmitted using $\lceil L \log (n+1) \rceil$ bits. Then, an optimal quantizer is designed by both the encoder and decoder for a training set consisting of all n-vectors with the same type as $(U_{n,L}(x_1), U_{n,L}(x_2), \cdots, U_{n,L}(x_n))$. This quantizer is then used to quantize the vector $x_1, x_2, \cdots, x_n$ itself, and the remaining bits are transmitted to the decoder to specify the index of the codevector closest to $x_1, x_2, \cdots, x_n$. The decoder receives this index and produces the appropriate code vector as an approximation to the input sequence $x_1, x_2, \cdots, x_n$. By choosing the quantity L to increase with n at a rate proportional to $\sqrt{n/\log n}$, we get the desired rate of convergence of the distortion to $D(R)$.

Formally, we define our code g_n as

$$g_n(x^n) = g_{Q_{n,L}, \bar{R}_n}(x^n)$$

where

$$\bar{R}_n = R - \frac{1}{n}(\lceil L \log (n+1) \rceil + 2\lceil \log M_n \rceil + 2). \tag{14}$$

When $\bar{R}_n \le 0$, set $g_n(x^n) = 0$, the all-zero vector. Thus first the optimal code of rate $\bar{R}_n$ for the uniform distribution over $T^n_{\hat{P}_{n,L}}$ is determined (a function of x^n), and then we set $g_n(x^n)$ as the codevector this code assigns to x^n. Since the distribution $Q_{n,L}$ is supported by a finite set, there are only a finite number of possible optimal quantizers from which to choose $g_{Q_{n,L}, \bar{R}_n}$ according to some rule ensuring uniqueness. Since $M_n \le M$ a.s., it can be seen that $\bar{R}_n > 0$ for n large enough, if L does not grow too fast with n. Note that g_n *does not* necessarily use nearest neighbor encoding.

To see that the rate of g_n is at most R, we use the simple fact, that $n^{-1} \log |g_n| \le R$ if there exists a one-to-one mapping of the range of g into $\{0,1\}^{\lceil nR \rceil}$, i.e., the output of g_n can be encoded losslessly using binary sequences of fixed length $\lceil nR \rceil$. The integer M_n can always be transmitted in $2\lceil \log M_n \rceil + 2$ bits by repeating twice in a row each bit in the binary representation of M_n and then sending the string "01" to terminate the transmission. It is possible to reduce this number to $\lceil \log M_n \rceil + 2\lceil \log \lceil \log M_n \rceil \rceil + 2$ bits using Elias' prefix code [8] for the positive integers, though our slightly less efficient scheme will suffice here. The type $\hat{P}_{n,L}$ can be described using $\lceil L \log (n+1) \rceil$ bits, and given M_n and the type $\hat{P}_{n,L}$, the output of $g_{\hat{P}_{n,L}, \bar{R}_n}$ requires $n\bar{R}_n$ bits. These add up to nR bits altogether.

The difference $\Delta(g_n) - D(R)$ can be upper-bounded as follows. Let us introduce the code $\bar{g}_n$ defined as

$$\bar{g}_n(x^n) = g_{Q_n, \bar{R}_n}(x^n)$$

that is, if the input x^n has empirical distribution $\hat{P}_n$, then $\bar{g}_n$ is the optimal source code of rate $\bar{R}_n$ for the uniform distribution over $T^n_{\hat{P}_n}$. Let

$$R_n = R - \frac{1}{n}(\lceil L \log (n+1) \rceil + 2\lceil \log M \rceil + 2).$$

Then $R_n \le \bar{R}_n$ a.s., and R_n depends only on n whereas $\bar{R}_n$ is also a function of the input x^n. If P denotes the distribution of X^n, then g_{P,R_n} is the optimal source code of rate R_n for X^n. Note that by the previous definitions $n^{-1}\mathbf{E}[\|X^n - g_{P,R_n}(X^n)\|^2] = D_n(R_n)$. We use the following decomposition:

$$\Delta(g_n) - D(R) = \mathbf{E}(A_1) + \mathbf{E}(A_2) + A_3 \tag{15}$$

where

$$\begin{aligned}
A_1 &= \frac{1}{n}\mathbf{E}\left[\|X^n - g_n(X^n)\|^2 | \hat{P}_n\right] \\
&\quad - \frac{1}{n}\mathbf{E}\left[\|X^n - \bar{g}_n(X^n)\|^2 | \hat{P}_n\right] \\
A_2 &= \frac{1}{n}\mathbf{E}\left[\|X^n - \bar{g}_n(X^n)\|^2 | \hat{P}_n\right] \\
&\quad - \frac{1}{n}\mathbf{E}\left[\|X^n - g_{P,R_n}(X^n)\|^2 | \hat{P}_n\right] \\
A_3 &= D_n(R_n) - D(R).
\end{aligned}$$

We present a chain of inequalities to give a uniform almost sure upper bound on A_1.

$$\begin{aligned}
nA_1 &= \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_{n,L}, \bar{R}_n}(x^n)\|^2\right] \\
&\quad - \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_n, \bar{R}_n}(x^n)\|^2\right] \\
&\le \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_{n,L}, \bar{R}_n}(U_{n,L}(x^n))\|^2\right] \\
&\quad - \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_n, \bar{R}_n}(x^n)\|^2\right] && (16) \\
&\le \mathbf{E}_{Q_n}\left[\|U_{n,L}(x^n) - g_{Q_{n,L}, \bar{R}_n}(U_{n,L}(x^n))\|^2\right] \\
&\quad - \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_n, \bar{R}_n}(x^n)\|^2\right] + \frac{4nM_n^2}{L} && (17) \\
&= \mathbf{E}_{Q_{n,L}}\left[\|x^n - g_{Q_{n,L}, \bar{R}_n}(x^n)\|^2\right] \\
&\quad - \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_n, \bar{R}_n}(x^n)\|^2\right] + \frac{4nM_n^2}{L} && (18) \\
&\le \mathbf{E}_{Q_{n,L}}\left[\|x^n - g_{Q_n, \bar{R}_n}(x^n)\|^2\right] \\
&\quad - \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_n, \bar{R}_n}(x^n)\|^2\right] + \frac{4nM_n^2}{L} && (19) \\
&= \mathbf{E}_{Q_n}\left[\|U_{n,L}(x^n) - g_{Q_n, \bar{R}_n}(U_{n,L}(x^n))\|^2\right] \\
&\quad - \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_n, \bar{R}_n}(x^n)\|^2\right] + \frac{4nM_n^2}{L} && (20) \\
&\le \mathbf{E}_{Q_n}\left[\|U_{n,L}(x^n) - g_{Q_n, \bar{R}_n}(x^n)\|^2\right] \\
&\quad - \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_n, \bar{R}_n}(x^n)\|^2\right] + \frac{4nM_n^2}{L} && (21) \\
&\le \frac{8nM_n^2}{L} \\
&\le \frac{8nM^2}{L} \quad \text{a.s.} && (22)
\end{aligned}$$

Thus $A_1 \le 8M^2/L$ with probability 1. Inequalities (16) and (21) follow from the fact that our optimal source codes use nearest neighbor encoding. Inequalities (17) and (22) follow

from the observation that for any $x^n, y^n \in [-M_n, M_n]^n$, we have

$$\left|\|x^n-y^n\|^2-\|U_{n,L}(x^n)-y^n\|^2\right| \le 4M_n \sum_{j=1}^{n} |x_j-U_{n,L}(x_j)| \le \frac{4nM_n^2}{L}. \tag{23}$$

Equations (18) and (20) follow from the fact that

$$\mathbf{E}_{Q_n}[f(U_{n,L}(x^n))] = \mathbf{E}_{Q_{n,L}}[f(x^n)]$$

for any function f since

$$T^n_{\hat{P}_{n,L}} = \{U_{n,L}(x^n) \ : \ x^n \in T^n_{\hat{P}_n}\}.$$

Inequality (19) follows since $g_{Q_{n,L},\bar{R}_n}$ is optimal with respect to the probability measure $Q_{n,L}$.

Next we show that $A_2 \le 0$ almost surely. This follows since X^n is uniformly distributed over $T^n_{\hat{P}_n}$ when conditioned on $\hat{P}_n$, so that

$$\begin{aligned}\mathbf{E}\left[\|X^n - \bar{g}_n(X^n)\|^2|\hat{P}_n\right] &= \mathbf{E}_{Q_n}\left[\|x^n - g_{Q_n,\bar{R}_n}(x^n)\|^2\right] \\ &\le \mathbf{E}_{Q_n}\left[\|x^n - g_{P,R_n}(x^n)\|^2\right] \quad \text{a.s.} \\ &= \mathbf{E}\left[\|X^n - g_{P,R_n}(X^n)\|^2|\hat{P}_n\right] \quad \text{a.s.}\end{aligned} \tag{24}$$

where the inequality holds because $R_n \le \bar{R}_n$ a.s., and $g_{Q_n,\bar{R}_n}$ is matched to the distribution Q_n.

Finally, A_3 is upper-bounded using Proposition 2 and a first-order Taylor expansion of $D(\cdot)$ around R.

$$\begin{aligned}A_3 &= [D_n(R_n) - D(R_n)] + [D(R_n) - D(R)] \\ &\le (c(R_n) + o(1))\sqrt{\frac{\log n}{n}} \\ &\quad + |D'(R)|(R - R_n)(1 + o(1)) \qquad (25) \\ &= (c(R_n) + o(1))\sqrt{\frac{\log n}{n}} \\ &\quad + |D'(R)|(1 + o(1))\frac{1}{n}(\lceil L\log(n+1)\rceil \\ &\quad + 2\lceil \log M\rceil + 2) \\ &= (c(R) + o(1))\sqrt{\frac{\log n}{n}} \\ &\quad + |D'(R)|(1 + o(1))L\left(\frac{\log n}{n}\right) \qquad (26)\end{aligned}$$

where we assumed in (25) that $n^{-1}L\log(n+1) \to 0$ as $n \to \infty$, so that $R_n \to R$. In (26) we used the fact that $c(R_n) = c(R) + o(n)$, by the continuity of $c(\cdot)$. Combining the bounds in (22), (24), and (26), we obtain

$$\Delta(g_n) - D(R) \le \frac{8M^2}{L} + (c(R) + o(1))\sqrt{\frac{\log n}{n}} + |D'(R)|(1 + o(1))L\left(\frac{\log n}{n}\right).$$

We can now choose $L \sim \sqrt{n/\log n}$ as a function of n, which gives

$$\Delta(g_n) - D(R) \le (8M^2 + c(R) + |D'(R)| + o(1))\sqrt{\frac{\log n}{n}}.$$

This completes the proof of the theorem. ■

Remark: It is not hard to see that Theorem 2 can be extended by considering more general distortion measures. For example, we could use distortion measures in the form

$$d_n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} f(|x_i - y_i|)$$

where $f : \mathcal{R} \to [0, \infty)$ is a nonnegative, nondecreasing, and continuous function. In this case the rate of convergence is controlled by the rate at which $f(x + a) \to f(a)$ as $x \downarrow 0$.

Universality of code for unbounded sources: In what follows we show that a slightly modified version of our scheme is weakly minimax universal over the class of memoryless sources with finite second moment, i.e., for each such source $\Delta(g_n) - D(R) \to 0$ as $n \to \infty$. Then we will consider the price of universality for memoryless sources having finite higher moments.

Sources with finite second moment: Let $f_n : \mathcal{R} \to \mathcal{R}$ be the limiter function

$$f_n(x) = xI_{\{|x|\le a_n\}} + a_n I_{\{x>a_n\}} - a_n I_{\{x<-a_n\}}$$

where $\{a_n\}$ is a sequence of positive numbers to be specified later. The modified scheme simply encodes (using the scheme of Theorem 2) $\tilde{X}^n = (f_n(X_1), \cdots, f_n(X_n))$ instead of X^n. The only change is that we set $M_n = a_n$, instead of using (11). Since $|f_n(X_i)| \le a_n$ a.s., by putting $R_n = \bar{R}_n$ we obtain as in (15)–(21), and (24)

$$\frac{1}{n}\mathbf{E}\|g_n(\tilde{X}^n) - \tilde{X}^n\|^2 - \tilde{D}_n(\bar{R}_n) \le \frac{8a_n^2}{L} \tag{27}$$

where $\tilde{D}(R)$ is the distortion of the optimal quantizer of rate R designed for $\tilde{X}^n$. Now by the triangle inequality and the definition of $\tilde{X}^n$ we have

$$\begin{aligned}&\mathbf{E}\|g_n(\tilde{X}^n) - X^n\|^2 \\ &\quad \le \left(\sqrt{\mathbf{E}\|g_n(\tilde{X}^n) - \tilde{X}^n\|^2} + \sqrt{\mathbf{E}\|\tilde{X}^n - X^n\|^2}\ \right)^2 \\ &\quad = \left(\sqrt{\mathbf{E}\|g_n(\tilde{X}^n) - \tilde{X}^n\|^2} + \sqrt{n\mathbf{E}[|X_1|^2 I_{\{|X_1|>a_n\}}]}\ \right)^2.\end{aligned} \tag{28}$$

Thus when $\mathbf{E}|X_1|^2 < \infty$ and $a_n \to \infty$ as $n \to \infty$, it follows from (28) that the distortion $\Delta(g_n) = n^{-1}\mathbf{E}\|g_n(\tilde{X}^n) - X^n\|^2$ is upper-bounded as

$$\Delta(g_n) \le \frac{1}{n}\mathbf{E}\|g_n(\tilde{X}^n) - \tilde{X}^n\|^2 + o(1), \quad \text{as } n \to \infty. \tag{29}$$

By an argument very similar to (16)–(21) (shown more generally in (35)–(38)) we have that

$$\tilde{D}_n(\bar{R}_n) \le D_n(\bar{R}_n) + o(1), \quad \text{as } n \to \infty. \tag{30}$$

Choosing L so that $\bar{R}_n \to R$ as $n \to \infty$ we obtain $D_n(\bar{R}_n) \to D(R)$. If $a_n \to \infty$ is chosen so that $a_n = o(\sqrt{L})$, then (27), (29), and (30) imply that $\Delta(g_n) - D(R) \to 0$ as $n \to \infty$ and the code is shown to be weakly minimax universal.

Sources with finite higher order moments: We will assume that $\mathbf{E}|X_1|^k < \infty$ for some $k > 8$, and investigate the price of universality $\Delta(g_n) - D_n(R)$ for the previous scheme. We will not force the rate redundancy to be zero, instead we choose the overall rate of g_n to be

$$r_n = R + \frac{1}{n}\lceil L \log n \rceil. \tag{31}$$

Thus the only change is that we replace $\bar{R}_n$ by the constant rate R. Since the sequence $\{a_n\}$ is deterministic, the rate redundancy $r_n - R$ is caused only by the fixed-rate encoding of the quantized type $\hat{P}_{n,L}$. We are going to show that for any $\epsilon > 2/(k-4)$ the sequence $\{a_n\}$ can be chosen such that

$$\Delta(g_n) - D_n(R) = O\left(\left(\frac{\log n}{n}\right)^{1/2-\epsilon}\right) \tag{32}$$

and

$$r_n - R = O\left(\left(\frac{\log n}{n}\right)^{1/2-\epsilon}\right). \tag{33}$$

To prove these we only have to examine (29) and (30) more carefully. Defining $h(a) = \mathbf{E}[|X_1|^2 I_{\{|X|>a\}}]$ we obtain from (29) that

$$\Delta(g_n) \le \frac{1}{n}\mathbf{E}\|g_n(\tilde{X}^n) - \tilde{X}^n\|^2 + O\left(\sqrt{h(a_n)}\right). \tag{34}$$

Let $\tilde{P}$ denote the distribution of $\tilde{X}^n$, and let $g_{\tilde{P},R}$ and $g_{P,R}$ be two R-rate quantizers which are optimal for $\tilde{X}^n$ and X^n, respectively. Then

$$\begin{aligned}
\tilde{D}_n(R) &= \frac{1}{n}\mathbf{E}\|g_{\tilde{P},R}(\tilde{X}^n) - \tilde{X}^n\|^2 && (35)\\
&\le \frac{1}{n}\mathbf{E}\|g_{P,R}(\tilde{X}^n) - \tilde{X}^n\|^2 && (36)\\
&\le \frac{1}{n}\left(\sqrt{\mathbf{E}\|g_{P,R}(\tilde{X}^n - X^n)\|^2} + \sqrt{\mathbf{E}\|X^n - \tilde{X}^n\|^2}\right)^2 && (37)\\
&\le \frac{1}{n}\left(\sqrt{\mathbf{E}\|g_{P,R}(X^n) - X^n\|^2} + \sqrt{nh(a_n)}\right)^2 && (38)\\
&= D_n(R) + O\left(\sqrt{h(a_n)}\right), \quad \text{as} \quad a_n \to \infty
\end{aligned}$$

where (38) follows because $g_{P,R}$ is a nearest neighbor quantizer. Combining this with (27) and (34) implies

$$\Delta(g_n) - D_n(R) = O\left(\frac{a_n^2}{L}\right) + O\left(\sqrt{h(a_n)}\right). \tag{39}$$

Now by the Schwarz and Chebyshev inequalities for all $a > 0$ we have

$$\begin{aligned}
\mathbf{E}[|X_1|^2 I_{|X_1|>a}] &\le \sqrt{\mathbf{E}|X_1|^4}\sqrt{\mathbf{P}\{|X_1| > a\}}\\
&\le \sqrt{\mathbf{E}|X_1|^4}\frac{\sqrt{\mathbf{E}|X_1|^k}}{a^{k/2}}
\end{aligned}$$

giving $h(a) = O(a^{-k/2})$ as $a \to \infty$. To achieve (33) we choose $L \sim (n/\log n)^{1/2+\epsilon}$. Since $k > 4 + 2/\epsilon$ we can choose $\rho > 0$ such that $\rho < \epsilon$ and $k > 4 + 2/\rho$. Putting $a_n = n^\rho$, these two conditions for ρ ensure that (39) implies (32).

IV. Discussion

One can ask whether the $O(n^{-1}\log n)$ convergence rate for the finite-alphabet case (Theorem 1) is optimal. To be more precise, the question is whether there exists a fixed-rate scheme $\{g_n\}$ of rate R with distortion

$$\Delta(g_n) = D(R) + o\left(\frac{\log n}{n}\right) \tag{40}$$

for all i.i.d. sources on alphabet $\mathcal{A}$ with $D(R) > 0$. One could use a result of Pilc [16], which asserts that the following lower bound holds:

$$D_n(R) - D(R) \ge |D'(R)|(1 + o(1))\frac{\log n}{2n}$$

for any finite alphabet i.i.d. source. This, of course, contradicts (40), thus proving the optimality of the $O(n^{-1}\log n)$ convergence rate in fixed-rate universal lossy coding of finite alphabet i.i.d. sources. However, as has been pointed out by several researchers recently, it appears that Pilc's proof of the above lower bound is flawed. Specifically, in [16, eq. (14)] an unsubstantiated approximation is used. Recently, however, Zhang *et al.* [25] have reported a rederivation of Pilc's lower bound. This would imply that (40) cannot hold, and in this sense, our convergence rate would be optimal, although the optimal constant has yet to be determined.

The optimality of the $O(\sqrt{n^{-1}\log n})$ covergence rate in the real alphabet case (Theorem 2) is an open problem. Here it might be possible to obtain an improved result by tightening some upper bounds in the analysis of our scheme. Indeed, any simultaneous improvement of Proposition 2 and the upper bound in (22) would result in an improved upper bound in Theorem 2.

The only available result analogous to the flavor of our Theorem 1 seems to be Yu and Speed's variable-length universal code [22] for finite alphabet memoryless sources. They proved the existence of a sequence of codes $\{g_n\}$ such that

$$d_n(x^n, g_n(x^n)) \le D, \quad \forall x^n \in \mathcal{A}^n \tag{41}$$

and

$$\begin{aligned}
\frac{1}{n}\mathbf{E}[l(g_n(X^n))] \le R(P,D) + (|\mathcal{A}|\cdot|\mathcal{B}| + |\mathcal{A}| + 4)\left(\frac{\log n}{n}\right)\\
+ O(n^{-1}).
\end{aligned} \tag{42}$$

where $l(\cdot)$ denotes the code length of a binary prefix code for encoding the range of g_n, and $R(P,D)$ is the rate distortion function of the i.i.d. source $\{X_n\}$ with generic distribution P. Thus the distortion of this code is *pointwise* upper bounded by D, while its expected average code length is $R(D,P) + O(n^{-1}\log n)$.

There is an apparent duality between the above result and our code in Theorem 1. We proved the existence of a sequence of fixed-rate codes $\{g_n\}$ such that

$$\begin{aligned}
\mathbf{E}[d_n(X^n, g_n(X^n))] \le D(P,R) + D'(P,R)\\
\times(|\mathcal{A}| + 1/2 + o(1))\left(\frac{\log n}{n}\right)
\end{aligned} \tag{43}$$

and

$$\frac{1}{n}\log|g_n| \le R \tag{44}$$

for all i.i.d. sources $\{X_n\}$ over $\mathcal{A}$ with generic distribution P such that $D(P,R) > 0$. We imposed no other condition on P. The bound (42) proved to hold only for sources with distribution P such that $R(P,D)$ has bounded second-order partial derivatives with respect to the source letter probabilities in the neighborhood of P. This condition is satisfied, for example, for binary sources with the Hamming distortion when $R(P,D) > 0$. A result in [25] also implies that the rate $n^{-1}\log n$ in (42) can not be improved, although the constant $(|\mathcal{A}|\cdot|\mathcal{B}| + |\mathcal{A}| + 4)$ may not be optimal.

We know of no result on universal quantization of bounded real sources comparable to our Theorem 2. One way to obtain a similar result for variable-length quantization (or "entropy coded quantization") might be an extension of the Yu and Speed result using a similar fine uniform quantization argument as in the proof of Theorem 2. However, source and reproduction alphabets of size L give a constant $|\mathcal{A}|\cdot|\mathcal{B}| = L^2$ in (42). Thus to achieve a $O(\sqrt{n^{-1}\log n})$ convergence rate of the expected code lengt to the rate distortion function of the source one has to set $L \sim (n/\log n)^{1/4}$, and then upper-bound the difference between the rate distortion function of the true source and that of the uniformly quantized source by $O(L^{-2})$. It is not clear that this can be done, although an $O(L^{-1})$ upper bound for this difference is easy to obtain. Assuming all this is done, one still has to show that the differentiability conditions on $R(P,D)$ are satisfied for the squared distortion.

It is interesting to note the difference between the two stage codes used for stationary sources [26], [11], [3] and our method. In the former case, the entire codebook is transmitted as overhead resulting in an $O(\sqrt{\log\log n/\log n})$ redundancy, while in our case (for i.i.d. and Markov sources) it is more efficient to transmit the approximate empirical statistics to obtain the much faster $O(\sqrt{n^{-1}\log n})$ convergence rate.

Finally, one might ask whether the $O(n^{-1}\log n)$ convergence rate in Theorem 1 for the expected distortion redundancy with finite alphabets also holds for almost sure type convergence. The best almost sure convergence rate that we have so far been able to obtain for the finite alphabet case is the much weaker $O(\sqrt{n^{-1}\log n})$, which can be derived using the same technique as in the real alphabet case.

Appendix I

Proof of Theorem 3: We will prove that the code constructed in the proof of Theorem 2 satisfies the claim. Using the notations introduced there we can write the difference

$$\begin{aligned}\frac{1}{n}\|X^n - g_n(X^n)\|^2 - D(R) &= \frac{1}{n}\|X^n - g_n(X^n)\|^2 \\ &\quad - \frac{1}{n}\mathbf{E}\left[\|X^n - g_n(X^n)\|^2|\hat{P}_n\right] \qquad (45)\\ &\quad + A_1 + A_2 + A_3 \\ &\quad + \frac{1}{n}\mathbf{E}\left[\|X^n - g_{P,R_n}(X^n)\|^2|\hat{P}_n\right] \\ &\quad - \frac{1}{n}\mathbf{E}\left[\|X^n - g_{P,R_n}(X^n)\|^2\right] \qquad (46)\end{aligned}$$

where A_1, A_2, and A_3 are defined in (15). We have already shown in the previous proof that $A_1 \le 8M^2/L = O(\sqrt{n^{-1}\log n})$ a.s., $A_2 \le 0$ a.s., and $A_3 = O(\sqrt{n^{-1}\log n})$. Therefore, it suffices to show that (45) and (46) are $O(\sqrt{n^{-1}\log n})$ a.s.

First we show that

$$\left|\frac{1}{n}\|X^n - g_{P,R_n}(X^n)\|^2 - \frac{1}{n}\mathbf{E}\left[\|X^n - g_{P,R_n}(X^n)\|^2\right]\right| = O\left(\sqrt{\frac{\log n}{n}}\right) \text{ a.s.} \tag{47}$$

and

$$\left|\frac{1}{n}\|X^n - g_{P,R_n}(X^n)\|^2 - \frac{1}{n}\mathbf{E}\left[\|X^n - g_{P,R_n}(X^n)\|^2|\hat{P}_n\right]\right| = O\left(\sqrt{\frac{\log n}{n}}\right) \text{ a.s.} \tag{48}$$

from which it follows that (46) is $O(\sqrt{n^{-1}\log n})$ a.s. To prove (47) we will use the following probability inequality by McDiarmid [14] for functions of independent random variables.

Lemma 1 (McDiarmid [14]): Let $X_1,\cdots,X_n$ be independent random variables taking values in a measurable space $(\mathcal{X},\mathcal{S})$, and let $h : \mathcal{X}^n \to \mathcal{R}$ be a bounded, measurable function such that

$$\sup_{x_i',x_1,\cdots,x_n\in A} |h(x_1,\cdots,x_n) - h(x_1,\cdots,x_{i-1},x_i',x_{i+1},\cdots,x_n)| \le c_i$$

for $1 \le i \le n$. That is, changing the ith variable, the value of h can change at most by c_i. Then for every n and every $t > 0$

$$\Pr\{|h(X_1,\cdots,X_n) - \mathbf{E}h(X_1,\cdots,X_n)| \ge t\} \le 2e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

Now, since g_{P,R_n} is a nearest neighbor encoder and its codevectors are contained in $[-M,M]^n$, if $x^n,\bar{x}^n \in [-M,M]^n$ such that they differ only in their ith coordinate, then

$$\begin{aligned}&\frac{1}{n}\|x^n - g_{P,R_n}(x^n)\|^2 - \frac{1}{n}\|\bar{x}^n - g_{P,R_n}(\bar{x}^n)\|^2 \\ &\quad\le \frac{1}{n}\|x^n - g_{P,R_n}(\bar{x}^n)\|^2 - \frac{1}{n}\|\bar{x}^n - g_{P,R_n}(\bar{x}^n)\|^2 \\ &\quad\le \frac{4M}{n}\sum_{j=1}^n |x_j - \bar{x}_j| \\ &\quad= \frac{4M}{n}|x_i - \bar{x}_i| \\ &\quad\le \frac{8M^2}{n}. \end{aligned} \tag{49}$$

Thus using Lemma 1 with $\mathcal{X} = [-M,M]$,

$$h(x^n) = n^{-1}\|x^n - g_{P,R_n}(x^n)\|^2$$

and $c_i = 8M^2/n$, gives

$$\Pr\left\{\left|\frac{1}{n}\|X^n - g_{P,R_n}(X^n)\|^2 - \frac{1}{n}\mathbf{E}[\|X^n - g_{P,R_n}(X^n)\|^2]\right| > t\right\} \le 2e^{-nt^2/(32M^4)}.$$

Choosing

$$t = c\sqrt{n^{-1}\log n}$$

with the constant $c > 4M^2\sqrt{2\ln 2}$, the right-hand side is summable in n, and (47) results by the Borel–Cantelli lemma.

The proof of (48) requires more effort. We will upper-bound the conditional probability

$$\Pr\left\{\left|\frac{1}{n}\|X^n - g_{P,R_n}(X^n)\|^2 - \frac{1}{n}\mathbf{E}\left[\|X^n - g_{P,R_n}(X^n)\|^2|\hat{P}_n\right]\right| > t\middle|\hat{P}_n\right\}$$

and then take expectations to obtain an upper bound on the unconditioned probability of the same event. Note that if $\hat{P}_n = P_{x^n}$, then the conditional distribution of X^n given $\hat{P}_n$ is the same as the distribution of the random vector $(x_{Z_1}, x_{Z_2}, \cdots, x_{Z_n})$, where $(Z_1, \cdots, Z_n)$ is a random permutation of $\{1, \cdots, n\}$. Thus we need a large deviation inequality similar to Lemma 1 for functions of random permutations. The following lemma is proved in Appendix II.

Lemma 2: Let $h : \{1, \ldots, n\}^n \to \mathcal{R}$ be a function such that

$$\max_{z_i', z_1, \cdots, z_n \in \{1,\cdots,n\}} |h(z_1, \cdots, z_n) - h(z_1, \cdots, z_{i-1}, z_i', z_{i+1}, \cdots, z_n)| \le c_i \quad (50)$$

for $i = 1, \cdots, n$. If $(Z_1, \cdots, Z_n)$ is a random permutation of $\{1, \cdots, n\}$, then for all $t > 0$

$$\Pr\{|h(Z_1^n) - \mathbf{E}h(Z_1^n)| \ge t\} \le 2e^{-t^2/2\sum_{i=1}^n c_i^2}.$$

For a given $x^n \in [-M, M]^n$ let $h : \{1, \cdots, n\}^n \to \mathcal{R}$ be defined by

$$h(i_1, \cdots, i_n) = \frac{1}{n}\|(x_{i_1}, \cdots, x_{i_n}) - g_{P,R_n}(x_{i_1}, \cdots, x_{i_n})\|^2$$

for $(i_1, \cdots, i_n) \in \{1, \cdots, n\}^n$. Then, since g_{P,R_n} has the nearest neighbor property, by (49) h can change at most by $8M^2/n$ if only one of its coordinates is changed. Furthermore, the distribution of $n^{-1}\|X^n - g_{P,R_n}(X^n)\|^2$ conditioned on $\hat{P}_n = P_{x^n}$ is the same as the distribution of $h(Z_1, \cdots, Z_n)$. Thus we can use Lemma 2 with $c_i = 8M^2/n$ to obtain (see first equation at bottom of this page) Taking the expectation of both sides we get (see the second equation at bottom of this page) from which (48) follows by the Borel–Cantelli principle. Thus we have shown that (46) is $O(\sqrt{n^{-1}\log n})$ a.s.

We only have left to show that (45) is $O(\sqrt{n^{-1}\log n})$ a.s. But this requires no extra proof because g_n is a nearest neighbor encoder for a given $\hat{P}_n = P_{x^n}$ (namely, $g_n(x^n) = g_{Q_n,\bar{R}_n}(x^n)$), thus we can define h as in the previous argument (with $g_{Q_n,\bar{R}_n}$ in place of g_{P,R_n}), obtaining

$$\Pr\left\{\left|\frac{1}{n}\|X^n - g_n(X^n)\|^2 - \frac{1}{n}\mathbf{E}[\|X^n - g_n(X^n)\|^2|\hat{P}_n]\right| > t\middle|\hat{P}_n\right\} = \Pr\{|h(Z_1, \cdots, Z_n) - \mathbf{E}h(Z_1, \cdots, Z_n)| > t\} \le 2e^{-nt^2/(32M^4)}.$$

This, of course, implies that (45) is $O(\sqrt{n^{-1}\log n})$ a.s., which completes the proof that

$$n^{-1}\|X^n - g_n(X^n)\|^2 - D(R) = O(\sqrt{n^{-1}\log n}). \quad \blacksquare$$

APPENDIX II

The next result is used to establish Lemma 2.

Lemma 3 (McDiarmid [14]): Let $Z_1^n = Z_1, \cdots, Z_n$ be random variables with Z_i taking values in a set A_i. Let $f : A_1 \times \cdots \times A_n \to \mathcal{R}$ be an appropriately measurable function. Suppose that there are constants $b_1, \cdots, b_n$ such that for each $1 \le i \le n$

$$|\mathbf{E}[f(Z_1^n)|Z_1 = z_1, \cdots, Z_{i-1} = z_{i-1}, Z_i = z_i] - \mathbf{E}[f(Z_1^n)|Z_1 = z_1, \cdots, Z_{i-1} = z_{i-1}, Z_i = z_i']| \le b_i \quad (51)$$

for each $z_j \in A_j$, $j = 1, \cdots, i-1$, and $z_i, z_i' \in A_i$. Then for any $t > 0$

$$\Pr\{|f(Z_1^n) - \mathbf{E}f(Z_1^n)| \ge t\} \le 2e^{-2t^2/\sum_{i=1}^n b_i^2}.$$

Proof of Lemma 2: We will show that h and $Z_1, \cdots, Z_n$ satisfy condition (51) of Lemma 3 with

$$\sum_{i=1}^n b_i^2 \le 4\sum_{i=1}^n c_i^2.$$

Fix $1 \le i \le n$ and let $z_1, \cdots, z_{i-1}, z_i$ and $z_1, \cdots, z_{i-1}, z_i'$ be two i-tuples of distinct positive integers not exceeding n, and put $C = \{1, \cdots, n\}$, $B = \{z_1, \cdots, z_{i-1}, z_i\}$, and

---

$$\Pr\left\{\left|\frac{1}{n}\|X^n - g_{P,R_n}(X^n)\|^2 - \frac{1}{n}\mathbf{E}\left[\|X^n - g_{P,R_n}(X^n)\|^2|\hat{P}_n\right]\right| > t\middle|\hat{P}_n\right\} = \Pr\{|h(Z_1, \cdots, Z_n) - \mathbf{E}h(Z_1, \cdots, Z_n)| > t\} \le 2e^{-nt^2/(128M^4)}.$$

$$\Pr\left\{\left|\frac{1}{n}\|X^n - g_{P,R_n}(X^n)\|^2 - \frac{1}{n}\mathbf{E}\left[\|X^n - g_{P,R_n}(X^n)\|^2|\hat{P}_n\right]\right| > t\right\} \le 2e^{-nt^2/(128M^4)}$$

$B' = \{z_1, \cdots, z_{i-1}, z_i'\}$. Note that $B - B' = \{z_i\}$ and $B' - B = \{z_i'\}$. Furthermore, for any finite ordered set S, let $\Pi(S)$ denote the set of all permutations of S, and for any $\pi \in \Pi(S)$ let $\pi(i)$ denote its ith coordinate. Then for each $i < n$

$$
\begin{aligned}
&|\mathbf{E}[h(Z_1^n)|Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1}, Z_i = z_i] \\
&\qquad -\mathbf{E}[h(Z_1^n)|Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1}, Z_i = z_i']| \\
&= \Bigg| \frac{1}{(n-i)!} \sum_{\pi \in \Pi(C-B)} h(z_1, \cdots, z_{i-1}, z_i, \pi(1), \cdots, \pi(n-i)) \\
&\quad - \frac{1}{(n-i)!} \sum_{\pi \in \Pi(C-B')} h(z_1, \cdots; z_{i-1}, z_i', \pi(1), \cdots, \pi(n-i)) \Bigg| \\
&= \Bigg| \frac{1}{(n-i)!} \sum_{j=1}^{n-i} \sum_{\substack{\pi \in \Pi(C-B) \\ \pi(j) = z_i'}} h(z_1, \cdots, z_{i-1}, z_i, \pi(1), \cdots, \pi(n-i)) \\
&\quad - \frac{1}{(n-i)!} \sum_{j=1}^{n-i} \sum_{\substack{\pi \in \Pi(C-B') \\ \pi(j) = z_i}} h(z_1, \cdots, z_{i-1}, z_i', \pi(1), \cdots, \pi(n-i)) \Bigg| \\
&\le \frac{1}{(n-i)!} \sum_{j=1}^{n-i} (n-i-1)!\,(c_i + c_{i+j})
\end{aligned}
$$

where the inequality holds by (50) since terms in the two summations can be paired so that the arguments of h differ exactly in the ith and the $(i+j)$th positions. Thus the constants b_i in Lemma 3 are given as

$$
b_i = \begin{cases} \frac{1}{(n-i)} \sum_{j=i+1}^{n} (c_i + c_j), & \text{if } i < n \\ c_i, & \text{if } i = n. \end{cases}
$$

By the convexity of $x \to x^2$ we have for each $i < n$

$$
\begin{aligned}
b_i^2 &= \left(\frac{1}{(n-i)} \sum_{j=i+1}^{n} (c_i + c_j) \right)^2 \\
&\le \frac{1}{(n-i)} \sum_{j=i+1}^{n} (c_i + c_j)^2 \\
&\le \frac{1}{(n-i)} \sum_{j=i+1}^{n} 2(c_i^2 + c_j^2) \\
&= 2c_i^2 + \frac{2}{(n-i)} \sum_{j=i+1}^{n} c_j^2.
\end{aligned}
$$

This gives

$$
\begin{aligned}
\sum_{i=1}^{n} b_i^2 &\le 2 \sum_{i=1}^{n} c_i^2 + 2 \sum_{i=1}^{n-1} \frac{1}{(n-i)} \sum_{j=i+1}^{n} c_j^2 \\
&= 2 \sum_{i=1}^{n} c_i^2 + 2 \sum_{j=2}^{n} a_j c_j^2 \qquad (52)
\end{aligned}
$$

where we changed the order of summation in the second term on the right-hand side of the inequality, and

$$
a_j = \sum_{i=1}^{j-1} \frac{1}{n-i}.
$$

Note that

$$
\sum_{j=2}^{n} a_j = n - 1
$$

and $0 < a_2 < \cdots < a_n$. For any fixed $\pi \in \Pi(1, \cdots, n)$ the random variables $h_\pi(Z_1, \cdots, Z_n) = h(Z_{\pi(1)}, \cdots, Z_{\pi(n)})$ and $h(Z_1, \cdots, Z_n)$ have the same distribution. Thus the probability $\Pr\{|h - \mathbf{E}h| \ge t\}$ does not change if we replace h with h_π. In particular, we can choose π so that $c_1 \ge \cdots \ge c_n$, which minimizes the sum

$$
\sum_{j=2}^{n} a_j c_j^2.
$$

Thus we get

$$
\begin{aligned}
\sum_{j=2}^{n} a_j c_j^2 &\le \frac{1}{(n-1)!} \sum_{\pi \in \Pi(2, \cdots, n)} \sum_{j=2}^{n} a_{\pi(j)} c_j^2 \\
&= \sum_{j=2}^{n} c_j^2 \frac{1}{(n-1)!} \sum_{\pi \in \Pi(2, \cdots, n)} a_{\pi(j)} \\
&= \sum_{j=2}^{n} c_j^2 \frac{1}{(n-1)} \sum_{l=2}^{n} a_l \\
&= \sum_{j=2}^{n} c_j^2.
\end{aligned}
$$

Combining this with (52) we obtain

$$
\sum_{i=1}^{n} b_i^2 \le 4 \sum_{i=1}^{n} c_i^2.
$$

This completes the proof of the lemma. ■

ACKNOWLEDGMENT

The authors wish to thank J. Ziv, A. Wyner, and P. Narayan for informing us that there was a flaw in Pilc's proof of the source coding theorem rate of convergence lower bound, and S. Khudanpur for pointing out the exact location of the flaw. Also, thanks are due to an anonymous reviewer for suggesting the modification of our scheme for unbounded sources.

REFERENCES

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

[2] P. A. Chou and M. Effros, "Rate and distortion redundancies for universal source coding with respect to a fidelty criterion," summary in *IEEE Int. Symp. on Information Theory* (San Antonio, TX, 1992).

[3] P. A. Chou, M. Effros, and R. M. Gray, "A vector quantization approach to universal noiseless coding and quantization," submitted to *IEEE Trans. Inform. Theory*, 1994.

[4] T. Cover, "Enumerative source coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 73–76, Jan. 1973.

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[6] L. D. Davisson, "Comments on 'Sequence time coding for data compression'," *Proc. IEEE*, vol. 54, p. 2010, Dec. 1966.
[7] ——, "Universal lossless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[8] P. Elias, "Universal codeword sets and representation of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 192–203, 1975.
[9] J. C. Kieffer, "A unified approach to week universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 674–682, Nov. 1978.
[10] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.
[11] T. Linder, G. Lugosi, and K. Zeger, "Rates of convergence in the source coding thorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1728–1740, Nov. 1994.
[12] T. J. Lynch, "Sequence time coding for data compression," *Proc. IEEE*, vol. 54, pp. 1490–1491, Oct. 1966.
[13] K. Marton, "Error exponent for source coding with a fidelty criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 179–199, Mar. 1974.
[14] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics 1989*. Cambridge, UK: Cambridge Univ. Press, 1989, pp. 148–188.
[15] D. L. Neuhoff, R. M. Gray, and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 511–523, Sept. 1975.
[16] R. Pilc, "The transmission distortion of a source as a function of the encoding block length," *Bell Syst. Tech. J.*, vol. 47, pp. 827–885, 1968.
[17] J. Rissanen. "Stochastic complexity and modeling," *Annals Statist.*, vol. 14, pp. 1080–1100, 1986.
[18] P. C. Shields, "Universal redundancy rates do not exist," *IEEE Trans. Inform. Theory*, vol. 39, pp. 520–524, Mar. 1993.
[19] Yu. M. Shtarkov and V. F. Babkin, "Combinatorial encoding for discrete stationary sources," in *Proc. 2nd Int. Symp. on Information Theory*, Akadémiai Kiadó, 1971, pp. 249–256.
[20] A. D. Wyner, "Communication of analog data from a Gaussian source over a noisy channel," *Bell Syst. Tech. J.*, pp. 801–812, May–June 1968.
[21] ——, "On the transmission of correlated Gaussian data over a noisy channel with finite encoding block length," *Inform. Contr.*, vol. 20, pp. 193–215, 1972.
[22] B. Yu and T. P. Speed, "A rate of convergence result for a universal D-semifaithful code," *IEEE Trans. Inform. Theory*, vol. 39, pp. 813–821, May 1993.
[23] R. Zamir and M. Feder, "Information rates of pre/post filtered dithered quantizers," submitted to *IEEE Trans. Inform. Theory*, 1993.
[24] K. Zeger, A. Bist, and T. Linder, "Universal source coding with codebook transmission," *IEEE Trans. Commun.*, vol. 42, pp. 336–346, Feb. 1994.
[25] Z. Zhang, E.-H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion," preprint, 1994.
[26] J. Ziv, "Coding of sources with unknown statistics—Part II: Distortion relative to a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 389–394, May 1972.