Rates of Convergence in the Source Coding Theorem, in Empirical Quantizer Design, and in Universal Lossy Source Coding

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

Abstract—Rate of convergence results are established for vector quantization. Convergence rates are given for an increasing vector dimension and/or an increasing training set size. In particular, the following results are shown for memoryless real-valued sources with bounded support at transmission rate R: (1) If a vector quantizer with fixed dimension k is designed to minimize the empirical mean-square error (MSE) with respect to m training vectors, then its MSE for the true source converges in expectation and almost surely to the minimum possible MSE as $O(\sqrt{\log m / m})$; (2) The MSE of an optimal k-dimensional vector quantizer for the true source converges, as the dimension grows, to the distortion-rate function $D(R)$ as $O(\sqrt{\log k / k})$; (3) There exists a fixed-rate universal lossy source coding scheme whose per-letter MSE on n real-valued source samples converges in expectation and almost surely to the distortion-rate function $D(R)$ as $O(\sqrt{\log \log n / \log n})$; (4) Consider a training set of n real-valued source samples blocked into vectors of dimension k, and a k-dimension vector quantizer designed to minimize the empirical MSE with respect to the $m = \lfloor n / k \rfloor$ training vectors. Then the per-letter MSE of this quantizer for the true source converges in expectation and almost surely to the distortion-rate function $D(R)$ as $O(\sqrt{\log \log n / \log n})$, if one chooses $k = \lfloor (1 / R)(1 - \epsilon)\log n \rfloor$ for any $\epsilon \in (0, 1)$.

Index Terms—Universal lossy source coding, empirical vector quantizer design, convergence rates, large deviation theory.

I. INTRODUCTION

CONSIDER an independent and identically distributed (i.i.d.) sequence of real-valued random variables $X_1, X_2, \cdots$ that we wish to quantize and transmit. While many techniques exist for designing good quantizers from either a mathematical description of the source distribution or from training data, little is known theoretically about how quickly these algorithms converge for

arbitrary sources. Similarly, there are no known results for real alphabet sources that indicate how quickly the distortion of fixed-rate universal lossy source coders converges to Shannon's distortion-rate function as the input size grows. In addition, many of the existing proofs of quantizer distortion convergence do not give almost sure convergence, but rather convergence in expectation or in probability. In this paper we provide rate of convergence results to answer a number of these open questions.

In particular, we establish convergence rates for four important source coding quantities. These include convergence of the mean-square distortion of: (1) empirically designed quantizers of a fixed vector dimension to the best quantizer of the same dimension for the true source, as the training set size grows; (2) optimal vector quantizers for the true source to the distortion-rate function, as the vector dimension grows; (3) fixed-rate universal lossy source coding (where the encoding rate is measured by the logarithm of the number of codewords normalized by the encoded block-length) to the distortion-rate function, as the number of input samples grows; (4) empirically designed quantizers to the distortion-rate function, as the number of input samples grows. All of the probabilistic convergences are given in expectation and almost surely. The four main results stated above are given respectively in our Corollary 1 and Theorems 2, 3, and 4.

Our first result (Theorem 1 and more specifically Corollary 1) concerns designing quantizers from training data, one of the central problems of data compression. The problem of vector quantization is the following: We are to design a vector quantizer of a given dimension with a given number of code vectors for a source (modeled by a sequence of real random variables). The information about the source to carry out this design is given in the form of a *training sequence*, a sequence of samples drawn according to the source distribution. The performance of the designed quantizer is measured by its sample distortion over a sequence that is independent of the training sequence and is distributed as the source, called the *test sequence*. The basic assumption is that the test sequence is much longer than the training sequence, therefore this sample distortion (assuming some ergodic properties on the source) is very close to the expected distortion of the quantizer. In fact we assume, as is customary, that the test sequence is long enough to measure the performance of

the designed quantizer by the expected distortion (see, e.g., [1]–[4]). The expectation can be taken with respect to both the test and the training sequences, or only with respect to the test sequence, in which case the distortion is a function of the training sequence. In both cases, the main parameter affecting the distortion is the size of the training sequence, which can be viewed as the number of examples the vector quantizer must be trained on.

A natural strategy to obtain a good quantizer for a given source is to try to minimize the empirical distortion, i.e., the performance of the quantizer measured on the training data. Many successful techniques have been proposed in the literature for minimizing the empirical distortion, and we refer the reader to [2] for a good summary of such algorithms. One of the most widely used of these techniques is the generalized Lloyd–Max algorithm (also known as the Linde–Buzo–Gray algorithm), which produces good, but not necessarily empirically optimal, quantizers from training sets. Sabin and Gray [4] demonstrated that this algorithm performs eventually as well as if the true distribution of the training data were known. In [5], the behavior of the distortion of empirically designed quantizers as the training set size increases was investigated. The asymptotics of this distortion of the form $A + m^{-\alpha}$ were experimentally verified, where A is a constant, m is the number of training vectors, and $\alpha > 0$. An experimental approach for image coding is given in [6], where the decay rate (in terms of training set size) for a special Hamming-type distortion was obtained from a Vapnik–Chervonenkis argument and then experimentally fit to match the data.

We focus on the performance of globally optimal quantizers. The first step in this direction was a result proved by Pollard [7], [3] (see also [1] and [8]), who showed that if the training data consist of independent copies of the random variable to be quantized, then as the size of the training data grows, the distortion of the empirically designed quantizer converges almost surely to the distortion of the optimal quantizer, for any distribution with finite second moment. In Section II, we derive an upper bound on the finite sample performance, i.e., the rate at which this distortion converges to the optimal one as the size of the training sequence grows.

Another related topic is universal lossy source coding. Loosely speaking, a sequence of quantizers at a fixed transmission rate and of increasing vector dimension is said to be a universal lossy source code over a particular class of sources if, for *any* source in the specified class, as the dimension increases, the distortions of the quantizers approach the distortion of the globally optimal quantizer for the source at the same transmission rate. The class of sources chosen and the method of measuring the rate and the distortion give different modes of universality. In Section IV, we examine a *fixed-rate* lossy coding scheme which is universal on the class of all stationary and ergodic sources of bounded support, where the distortion is measured by the squared error. We investigate the rate at which the distortion of the universal code decreases to its

lower bound, the distortion rate function, as the encoded block length increases, on the subclass of i.i.d. sources.

Another interesting type of vector quantizer design problem is the following: Given a finite set of training data, design a quantizer of a given rate with the lowest possible per-letter distortion. This problem is more complicated than ordinary vector quantizer design since in addition to designing the quantizer, one must determine the vector *dimension* of the quantizer. In particular, it is not the distortion of the best possible quantizer of a given dimension that one seeks to approach as the training set grows in size, but rather the distortion-rate function $D(R)$ of the source, since this is the lowest distortion achievable using quantizers of arbitrarily large dimension and with a rate constraint R.

Suppose we are given n scalar samples from an i.i.d. source, from which we are to design a quantizer of a given dimension with low distortion. Clearly, as Shannon's source coding theorem with respect to a fidelity criterion suggests, we should try to design a vector quantizer of large dimension since this improves the *potential* to get close to $D(R)$. At the same time, however, the *ability* to do so decreases, as it is harder to estimate the best quantizer in higher dimensions. This is because the codebook size of full-search fixed-rate vector quantizers increases exponentially with the quantizer dimension. Also, since the number of scalar training samples n is fixed, the number of available k-dimensional training vectors that can be formed from this set decreases as the dimension k grows. The empirically optimal quantizer would tend to overfit the training data with the increasing quantizer dimension. Therefore, a compromise in this trade-off must be determined. To handle the problem, it is not enough to know that the performance of the empirically optimal quantizer converges to the optimum; one should also know about the *finite* sample performance of such quantizers.

In Section II, exponential large deviation inequalities are derived for the empirical distortion (Lemma 1) and for the average distortion (Theorem 1), of quantizers with minimal empirical distortion. Lemma 1 is later used to obtain convergence rates in universal coding, while Theorem 1 is subsequently used for the problem of quantizer design. These nonasymptotic inequalities are valid for all memoryless sources of bounded support. One of the main tools we use to obtain these results is the elegant Vapnik–Chervonenkis inequality (Lemma 3). For the problem of vector quantizer design, Corollary 1, an immediate consequence of Theorem 1, shows that if a vector quantizer with fixed dimension k is designed to minimize the empirical distortion with respect to m training vectors, then its distortion for the true source converges in expectation and almost surely to the minimum possible distortion as $O(\sqrt{\log m / m})$. This is proved for memoryless real-valued sources with bounded support.

Section III concerns the other side of the trade-off, the difference between the distortion of the best k-dimensional quantizer of rate R and the distortion-rate function

$D(R)$, where $R = (\log N)/k$, where N is the codebook size of the quantizer. Pilc [9] proved that for memoryless sources with a finite alphabet, this difference is of order $\log k/k$. We show in Theorem 2 that for real memoryless sources (and real reproduction alphabets), this difference is at most of order $\sqrt{\log k/k}$.

Section IV investigates the problem of fixed-rate universal lossy coding, that is, transmitting messages of an unknown source with distortion under a certain rate R. This was first formulated by Ziv [10], who proved the existence of a fixed-rate scheme whose expected distortion converges to $D(R)$ for any stationary process. This scheme was analyzed using high-resolution quantization theory by Zeger, Bist, and Linder [11] and Chou and Effros [12]. Neuhoff, Gray, and Davisson [13] extended Ziv's results and provided a unified approach to fixed-rate universal lossy source coding. Pursley and Davisson [14] demonstrated universal variable-rate source codes. Recently, Yu and Speed [15] obtained bounds on the rate of convergence of variable-rate universal coding for finite alphabet memoryless sources.

In Section IV, the results of the previous sections are applied to the problem of fixed-rate universal lossy coding. It is shown in Theorem 3 that for real memoryless bounded sources, Ziv's scheme gives a $\sqrt{\log \log n/\log n}$ type convergence rate of the average distortion to $D(R)$, in both expectation and almost surely, where n is the input block length.

In Section V, the problem of choosing the dimension of an empirically designed quantizer of rate R (used on independent test data) is considered. Using the results of Sections II and III, it is shown in Theorem 4 that the quantizer dimension k can be chosen as a function of the number of training samples n so that the average distortion approaches $D(R)$ as $\sqrt{\log \log n/\log n}$.

We limit our discussion to squared-error distortion, though it seems possible to extend our results to distortion functions with certain regularity properties, such as "nice functions" of norms.

II. DESIGNING QUANTIZERS FROM TRAINING DATA

Let $\mathscr{R}^k$ denote k-dimensional Euclidean space. A k-dimensional, N-level vector quantizer $Q_{N,k}$ is a measurable function of the form $Q_{N,k} : \mathscr{R}^k \to \{y_1, \cdots, y_N\} \subset \mathscr{R}^k$, such that

$$Q_{N,k}(x) = y_i, \quad \text{if } \|x - y_i\| < \|x - y_j\| \quad \text{for all } j,$$

where $\|\cdot\|$ denotes the Euclidean norm. In this definition, the quantizer is determined by its code vectors $\{y_1, \cdots, y_N\}$ via the nearest-neighbor quantization rule, and ties are broken arbitrarily. The rate of this quantizer is $(\log N)/k$ bits per source letter. Throughout this paper, the notation "log" will always mean logarithm in base 2. Let $Z, Z_1, \cdots, Z_m \in \mathscr{R}^k$ be independent and identically distributed random vectors. Here Z is the source we want to quantize, and $Z_1, \cdots, Z_m$, which are i.i.d. copies of Z, represent the training data. Given a quantizer $Q_{N,k}$, de-

fine its *average distortion* (mean square) as

$$\Delta(Q_{N,k}) = E\|Z - Q_{N,k}(Z)\|^2,$$

and its *empirical distortion* as

$$\Delta_m(Q_{N,k}) = \frac{1}{m} \sum_{i=1}^{m} \|Z_i - Q_{N,k}(Z_i)\|^2.$$

Note that $\Delta_m(Q_{N,k})$ is a random variable depending on $Z^{(m)} = (Z_1, \cdots, Z_m)$. Throughout the paper, we will often refer to these distortions simply as the "MSE'" (mean-square error), when no confusion results.

Let $Q_{m,N,k}^*$ be a quantizer with minimal empirical distortion, which again depends on the training set $Z^{(m)}$, and let $Q_{N,k}^*$ be a quantizer of minimal average distortion. That is,

$$Q_{N,k}^* = \arg\min_{Q_{N,k}} \Delta(Q_{N,k}) \tag{1}$$

and

$$Q_{m,N,k}^* = \arg\min_{Q_{N,k}} \Delta_m(Q_{N,k}). \tag{2}$$

Note that $Q_{N,k}^*$ is a deterministic mapping, while $Q_{m,N,k}^*$ is a function of $Z_1, \cdots, Z_m$. We will refer to $Q_{N,k}^*$ as the *truly optimal quantizer* and to $Q_{m,N,k}^*$ as the *empirically optimal quantizer*. By the nearest-neighbor encoding rule, finding an optimal quantizer is equivalent to finding its codeword set $\{y_1, \cdots, y_N\}$.

The existence of a quantizer in (2) minimizing the empirical distortion is clear, although it may not be unique; in the case of (1) the existence of a minimizing quantizer is proved in [3] for $E\|Z\|^2 < \infty$. Since we deal only with the associated distortions, the issue of uniqueness is immaterial in what follows.

We are interested in two random quantities, each depending on $Z^{(m)}$. The first is

$$\Delta_m(Q_{m,N,k}^*) - \Delta(Q_{N,k}^*), \tag{3}$$

the difference between the empirical distortion of the empirically optimal quantizer $Q_{m,N,k}^*$ and the average distortion of the truly optimal quantizer $Q_{N,k}^*$. This is the difference between the distortion of a quantizer designed from training data when used to encode the very same training data, and the distortion of a "globally best" quantizer. This quantity is very important for analyzing the performance of certain *universal quantization* schemes, where the encoder determines $Q_{m,N,k}^*$ from the source input $Z_1, \cdots, Z_m$, and then transmits its description to the decoder along with the quantized sequence $Q_{m,N,k}^*(Z_i)$, $i = 1, \cdots, m$. Note that in this case $Q_{m,N,k}^*$ is used for the same data it was designed from.

The second quantity of interest is

$$\Delta(Q_{m,N,k}^*) - \Delta(Q_{N,k}^*), \tag{4}$$

the difference between the *average distortion* of the empirically optimal quantizer and the average distortion of the optimal quantizer, where

$$\Delta(Q_{m,N,k}^*) = E\big(\|Z - Q_{m,N,k}^*(Z)\|^2 | Z_1, \cdots, Z_m\big).$$

This quantity specifies the difference in performances (when encoding the true source) between a quantizer designed to minimize the distortion on a training set and a quantizer designed optimally for the true source. This quantity is important in the context of vector quantizer design (see, e.g., [16]), where the quantizer that was designed from the training data is used later for encoding data from a source which has the same distribution as the training data, but is independent of the training data. The difference is nonnegative and can be thought of as the penalty for designing a quantizer based on a training set rather than the actual source distribution.

Note that the generalized Lloyd algorithm (if it does not get trapped in a local optimum) produces quantizers whose empirical distortions converge to $\Delta_m(Q^*_{m,N,k})$ as the number of iterations in the algorithm increases. The random variable $\Delta(Q^*_{m,N,k})$, depending only on $Z^{(m)}$, represents the average distortion when this quantizer is used on a source with the same distribution but independent of $Z_1, Z_2, \cdots$. In what follows, we obtain exponential large deviation inequalities for (3) and (4).

The next lemma shows that if the Z_i's are independent and have bounded support, then the probability that $\Delta_m(Q^*_{m,N,k})$ is larger than $\Delta(Q^*_{N,k})$ plus any positive constant is exponentially small in m.

Lemma 1: Let $Z_1, Z_2, \cdots \in \mathscr{R}^k$ be an i.i.d. source such that $\Pr\{\|Z_1\|^2 \le B\} = 1$ for some $B < \infty$. Then for every $t > 0$, the difference between the empirical distortion of the empirically optimal quantizer and the average distortion of the truly optimal quantizer satisfies

$$\Pr\{\Delta_m(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) > t\} \le e^{-mt^2/(8B^2)}.$$

Proof: From (1) and (2), it follows that $\Delta_m(Q^*_{m,N,k}) \le \Delta_m(Q^*_{N,k})$. Thus we have

$$\Pr\{\Delta_m(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) > t\}$$
$$\le \Pr\{\Delta_m(Q^*_{N,k}) - \Delta(Q^*_{N,k}) > t\}$$
$$= \Pr\left\{\frac{1}{m}\sum_{i=1}^{m}\|Z_i - Q^*_{N,k}(Z_i)\|^2\right.$$
$$\left. - E\|Z - Q^*_{N,k}(Z)\|^2 > t\right\}. \quad (5)$$

The random variables $\|Z_i - Q^*_{N,k}(Z_i)\|^2$ are independent and bounded almost surely by $4B$ since the optimal quantizer for Z clearly has all its levels inside a sphere of radius $\sqrt{B}$. Thus, we can apply Hoeffding's inequality [17], which asserts that if $X_1, \cdots, X_m$ are independent random variables taking values in a bounded interval $[a, b]$, then for all $t > 0$, we have

$$\Pr\left\{\frac{1}{m}\sum_{j=1}^{m}(X_j - EX_j) \ge t\right\} \le e^{-2mt^2/(b-a)^2}. \quad (6)$$

Taking $X_i = \|Z_i - Q^*_{N,k}(Z_i)\|^2$ and $b - a = 4B$ in

Hoeffding's inequality, (5) gives

$$\Pr\{\Delta_m(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) > t\} \le e^{-mt^2/(8B^2)},$$

which was to be shown. □

The relevance of Lemma 1 will be made explicit in Section IV, where it is used to evaluate the empirical distortion performance of Ziv's universal lossy source coding scheme.

The random variable $\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k})$ gives the difference between the distortion of a quantizer that best fits the training sequence and the distortion of a best theoretically achievable quantizer. To obtain the counterpart of Lemma 1 for this quantity, a stronger technique is needed. We begin with the following simple observation.

Lemma 2: If $\Pr\{\|Z\|^2 \le B\} = 1$ for some $B < \infty$, then

$$\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{k,N}) \le 2 \sup_{Q_{N,k}} |\Delta_m(Q_{N,k}) - \Delta(Q_{N,k})|$$

a.s.,

where the supremum is taken over all quantizers $Q_{N,k}$ with $\sup_{z \in \mathscr{R}^k} \|Q_{N,k}(z)\|^2 \le B$.

Proof: The inequality follows from

$$\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k})$$
$$= \Delta(Q^*_{m,N,k}) - \Delta_m(Q^*_{m,N,k}) + \Delta_m(Q^*_{m,N,k})$$
$$\quad - \Delta(Q^*_{N,k})$$
$$\le \Delta(Q^*_{m,N,k}) - \Delta_m(Q^*_{m,N,k}) + \Delta_m(Q^*_{N,k})$$
$$\quad - \Delta(Q^*_{N,k})$$
$$\le 2 \sup_{Q_{N,k}} |\Delta_m(Q_{N,k}) - \Delta(Q_{N,k})|.$$

It is enough to take the supremum over the $Q_{N,k}$'s with $\|Q_{N,k}(z)\|^2 \le B$, since we clearly have $\|Q^*_{m,N,k}(z)\|^2 \le B$ a.s. and $\|Q^*_{N,k}(z)\|^2 \le B$ a.s. for all $z \in \mathscr{R}^k$. □

For the sake of notational simplicity, define the function $f_Q: \mathscr{R}^k \to \mathscr{R}$ by

$$f_Q(z) = \|z - Q_{N,k}(z)\|^2.$$

By the bounded support assumption on Z and the restriction on the quantizers over which the supremum is taken, the random variables $f_Q(Z_i)$ are uniformly bounded (a.s.) by $4B$. Therefore,

$$\sup_{Q_{N,k}} |\Delta_m(Q_{N,k}) - \Delta(Q_{N,k})|$$
$$= \sup_Q \left|\frac{1}{m}\sum_{i=1}^{m} f_Q(Z_i) - E(f_Q(Z))\right|$$
$$= \sup_Q \left|\int_0^\infty \left(\frac{1}{m}\sum_{i=1}^{m} I_{\{f_Q(Z_i) > u\}} - \Pr\{f_Q(Z) > u\}\right) du\right|$$
$$\le 4B \sup_{Q,u} \left|\frac{1}{m}\sum_{i=1}^{m} I_{\{f_Q(Z_i) > u\}} - \Pr\{f_Q(Z) > u\}\right| \text{ a.s.,}$$
$$\quad (7)$$

where I_A is the indicator of the set A. We see, for example, that $\sup_{Q_{N,k}} |\Delta_m(Q_{N,k}) - \Delta(Q_{N,k})|$ converges to zero with probability 1, if the averages of the Bernoulli

random variables $I_{\{f_Q(Z_i)>u\}}$ converge to the probabilities $\Pr\{f_Q(Z)>u\}$ uniformly (over all Q's and u's), almost surely. At this point we recall a celebrated inequality of Vapnik and Chervonenkis.

Lemma 3 (Vapnik and Chervonenkis [18]): Let $\{S_\theta\colon \theta \in \Theta\}$ be an arbitrary collection of measurable subsets of $\mathscr{R}^k$, and let $Z_1,\cdots,Z_m$ be i.i.d. random variables in $\mathscr{R}^k$. Then for all m and $t>0$ such that $mt^2 \ge 2$,

$$\Pr\left\{ \sup_\theta \left| \frac{1}{m}\sum_{i=1}^m I_{\{Z_i \in S_\theta\}} - \Pr\{Z_1 \in S_\theta\}\right| > t\right\}$$
$$\le 4V(2m)e^{-mt^2/8},$$

where the *shatter coefficient* $V(m)$ is defined as follows. Fix m points, $z_1,\cdots,z_m \in \mathscr{R}^k$, and count how many distinct binary vectors

$$(I_{\{z_1 \in S_\theta\}},\cdots, I_{\{z_m \in S_\theta\}})$$

can be obtained by varying θ over Θ. Taking the maximum over $z_1,\cdots,z_m \in \mathscr{R}^k$ of the number of distinct binary vectors gives the value of $V(m)$.

Note that $V(m)$ is the maximum number of different subsets of m points $\{z_1,\cdots,z_m\}$ that are of the form $S_\theta \cap \{z_1,\cdots,z_m\}$ for some $\theta \in \Theta$. Clearly, $V(m) \le 2^m$ and $V(m)$ is monotone increasing in m. If $V(m)$ is upper bounded by a polynomial in m, then we have an exponential upper bound in Lemma 3, and, for example,

$$\sup_\theta \left| \frac{1}{m}\sum_{i=1}^m I_{\{Z_i \in S_\theta\}} - \Pr\{Z_1 \in S_\theta\}\right| \to 0, \quad \text{a.s.}$$

by the Borel–Cantelli lemma.

Now, define the index set $\Theta = \{(u,Q)\}$, where u runs through the positive reals and Q runs through all k-dimensional N-level quantizers, and for each $\theta = (u,Q) \in \Theta$, let $S_\theta = \{z\colon f_Q(z) > u\}$. In order to apply Lemma 3 to (7), we need an upper bound on the shatter coefficient of the class of sets of the form $\{z\colon f_Q(z) > u\}$.

Lemma 4: The shatter coefficient $V(m)$ of the class of sets $\{z\colon f_Q(z) > u\}$, where u runs through all positive reals and Q through all k-dimensional quantizers with N levels, is bounded as

$$V(m) \le m^{N(k+1)}.$$

Proof: It is easy to see that for each Q and u, the set $\{z\colon f_Q(z) > u\}$ is just the intersection of the complements of N spheres of radius $\sqrt{u}$. Thus $\{S_\theta\colon \theta \in \Theta\}$ is just the collection of all sets that are the intersections of the complements of N k-dimensional spheres of equal radii.

Consider now two collections $\mathscr{A}_1$ and $\mathscr{A}_2$ of subsets of $\mathscr{R}^k$, and denote their shatter coefficients by $V(\mathscr{A}_1,m)$ and $V(\mathscr{A}_2,m)$, respectively. Let $\mathscr{A} = \{A_1 \cap A_2\colon A_1 \in \mathscr{A}_1, A_2 \in \mathscr{A}_2\}$. We will prove that

$$V(\mathscr{A},m) \le V(\mathscr{A}_1,m)V(\mathscr{A}_2,m). \tag{8}$$

Fix the points $z_1,\cdots,z_m$, and let $B_1,\cdots,B_l$ be an enumeration of all the distinct sets of the form $\{z_1,\cdots,z_m\} \cap A$, where $A \in \mathscr{A}_1$. Then $l \le V(\mathscr{A}_1,m)$, and since $\mathscr{A}_2$ can pick

at most $V(\mathscr{A}_2,|B_i|)$ subsets of B_i, and since $|B_i| \le m$, we have that the number of sets of the form $\{z_1,\cdots,z_m\} \cap A_1 \cap A_2$, where $A_1 \in \mathscr{A}_1$ and $A_2 \in \mathscr{A}_2$, is upper bounded as

$$|\{\{z_1,\cdots,z_m\} \cap A_1 \cap A_2\colon A_1 \in \mathscr{A}_1, A_2 \in \mathscr{A}_2\}|$$
$$\le \sum_{i=1}^l V(\mathscr{A}_2,|B_i|)$$
$$\le V(\mathscr{A}_1,m)V(\mathscr{A}_2,m),$$

which proves (8).

Let $\mathscr{S}$ be the collection of complements of all spheres. For $N = 2$, since $\{S_\theta\colon \theta \in \Theta\} \subset \{S_1 \cap S_2\colon S_1, S_2 \in \mathscr{S}\}$, (8) shows that $V(m) \le V(\mathscr{S},m)^2$, and by induction we get, for all $N \ge 1$,

$$V(m) \le V(\mathscr{S},m)^N. \tag{9}$$

The shatter coefficient of the collection of all k-dimensional spheres is at most m^{k+1} (see, e.g., Devroye [19] or Dudley [20]). Since the shatter coefficient does not change if each set in the collection is replaced by its complement, we have

$$V(\mathscr{S},m) \le m^{k+1},$$

which, together with (9), proves the lemma. $\square$

The following theorem, the main result of this section, gives a nonasymptotic large deviation inequality concerning the performance of empirically designed quantizers.

Theorem 1: Let $Z_1, Z_2,\cdots \in \mathscr{R}^k$ be i.i.d. random vectors such that $\Pr\{\|Z_1\|^2 \le B\} = 1$ and $m(t/8B)^2 \ge 2$. Suppose an N-level, k-dimensional quantizer $Q^*_{m,N,k}$ is designed to minimize the empirical MSE over a training set of m vectors $Z_1,\cdots,Z_m$. Then the difference between the MSE of this quantizer for the true source and the MSE of the best quantizer, $Q^*_{N,k}$, for the true source, satisfies

$$\Pr\{\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) > t\}$$
$$\le 4(2m)^{N(k+1)}e^{-mt^2/(512B^2)}. \tag{10}$$

Proof: From Lemma 2, (7), Lemma 3, and Lemma 4, we have

$$\Pr\{\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) > t\}$$
$$\le \Pr\left\{2 \sup_{Q_{N,k}} |\Delta_m(Q_{N,k}) - \Delta(Q_{N,k})| > t\right\}$$
$$\le \Pr\left\{8B \sup_{Q,u} \left| \frac{1}{m}\sum_{i=1}^m I_{\{f_Q(Z_i)>u\}} - \Pr\{f_Q(Z) > u\}\right| > t\right\}$$
$$\le 4(2m)^{N(k+1)}e^{-mt^2/(512B^2)}. \qquad \square$$

Note that the upper bound in the theorem depends on the dimension k. This reflects the fact that quantizers of larger dimensions (and thus of smaller distortion) require a larger training sequence size m. This matter will be pursued in Section V.

For N and k fixed, a simple corollary of Theorem 1 follows by the Borel–Cantelli lemma, namely,

$$\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) \to 0 \quad \text{a.s.} \quad \text{as } m \to \infty, \quad (11)$$

This is Pollard's theorem [7], [3], who showed (11) under the more general condition $E\|Z\|^2 < \infty$. However, unlike Pollard's result, Theorem 1 also yields an upper bound on the rate of convergence, which will be important in the sequel. To obtain this almost sure rate of convergence for $\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k})$ from Theorem 1 via the Borel–Cantelli lemma, the parameter t should be chosen (as a function of m) so that the right-hand side of (10) becomes summable in m. It is easy to see that the choice $t = c\sqrt{\log m/m}$ satisfies this if $c > 16B\sqrt{N(k+1)}$, but the upper bound fails to be summable in m when $c \leq 16B\sqrt{N(k+1)}$. In this sense, $c = 16B\sqrt{N(k+1)} + \epsilon$ is an asymptotically optimal choice for a small $\epsilon > 0$. The almost sure convergence implies convergence in expectation, since the random variables here are bounded. However, the rate of convergence in expectation is not immediately implied by the almost sure rate of convergence.

Given a sequence of random variables $V_1, V_2, \cdots$ and a sequence of nonnegative numbers $a_1, a_2, \cdots$, the notation $V_n = O(a_n)$ a.s. denotes that there exist a positive constant c and a nonnegative integer-valued random variable M such that $V_n \leq ca_n$ if $n \geq M$. Using this notation, we can state the following corollary.

Corollary 1: Let $Z_1, Z_2, \cdots \in \mathscr{R}^k$ be an i.i.d. source that is bounded with probability 1 and suppose an N-level, k-dimensional quantizer $Q^*_{m,N,k}$ is designed to minimize the empirical MSE over a training set of m vectors $Z_1, \cdots, Z_m$. Then its MSE for the true source converges almost surely as $m \to \infty$ to the minimum MSE of the best quantizer, $Q^*_{N,k}$, for the true source, at a rate bounded as

$$\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) = O\left(\sqrt{\frac{\log m}{m}}\right) \quad \text{a.s.}$$

Remarks (Convergence in Expectation): (a) Theorem 1 also implies that the rate of convergence given in Corollary 1 holds for convergence in expectation. To see this, note that

$$E|\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k})|$$

$$\leq a + \int_a^\infty \Pr\{\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) > t\} \, dt \quad (12)$$

$$\leq a + 4(2m)^{N(k+1)}\left(\frac{16B}{\sqrt{m}}\right)\int_{\frac{a\sqrt{m}}{16B}}^\infty e^{-u^2/2} \, du \quad (13)$$

$$< c\sqrt{\frac{\log m}{m}} + \frac{B^2}{c} 2^{Nk+N+10}$$

$$\cdot \left[\frac{m^{N(k+1)-(1/2)-(c^2/(512B^2))}}{\sqrt{\log m}}\right] \quad (14)$$

$$= 16B\sqrt{N(k+1)}\sqrt{\frac{\log m}{m}} + o\left(\sqrt{\frac{\log m}{m}}\right). \quad (15)$$

Inequality (12) holds for all $a \geq 0$ with equality when $a = 0$. Inequality (13) follows from inequality (10) in Theorem 1 along with the substitution $u = t\sqrt{m}/(16B)$. Inequality (14) follows from the inequality $\int_x^\infty e^{-u^2/2} \, du < (1/x)e^{-x^2/2}, \forall x > 0$, and choosing $a = c\sqrt{\log m/m}$, and (15) follows from choosing $c = 16B\sqrt{N(k+1)}$.

(b) Following our proof of Theorem 1, it is also possible to obtain the result

$$E|\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k})| \leq c(N,k,B)\frac{1}{\sqrt{m}}, \quad (16)$$

where the constant $c(N,k,B)$ does not depend on m. To derive this upper bound, we have to replace the basic Vapnik–Chervonenkis inequality (Lemma 3) by a result of Alexander [21] stating that, for $mt^2 \geq 64$,

$$\Pr\left\{\sup_\theta \left|\frac{1}{m}\sum_{i=1}^m I_{\{Z_i \in S_\theta\}} - \Pr\{Z_1 \in S_\theta\}\right| > t\right\}$$

$$\leq 16(t\sqrt{m})^{4096V_d}e^{-2mt^2}, \quad (17)$$

where V_d is the so-called *Vapnik–Chervonenkis dimension* of the class $\{S_\theta: \theta \in \Theta\}$, i.e., the smallest integer m such that the shatter coefficient satisfies $V(m) < 2^m$. The fact that in this case the expected value of the supremum in (17) decreases as $1/\sqrt{m}$ was pointed out in [22]. However, the constant $c(N,k,B)$ is so enormous, even for small values of N and k, that the bound (16) is worse than (15) for any practical values of m.

(Convergence Almost Surely): (c) It was pointed out [23] that one could slightly improve our almost sure rate $O(\sqrt{\log n/n})$ in Corollary 1 to $O(\sqrt{\log\log n/n})$ by using another result of Alexander [21]. However, in this case the constant is not identified, nor is any exact upper bound in the form of (10) obtained.

III. RATE OF CONVERGENCE IN THE SOURCE CODING THEOREM

In the previous section, we derived bounds on how much the average and sample distortions of the empirically designed quantizer of dimension k differ from the distortion of the truly best quantizer of the same dimension. However, if we are to compare the performance of the empirically designed quantizers with the distortion-rate bound $D(R)$, we have to estimate the excess distortion caused by the finiteness of the quantizer dimension. In this section, we establish an upper bound for memoryless sources on the rate at which the distortion $D_k(R)$ of the best k-dimensional quantizer of rate R converges to the distortion-rate function $D(R)$. For finite-alphabet i.i.d. sources, the problem was addressed by Pilc [9], who showed a $c \log k/k$-type upper bound on the difference $D_k(R) - D(R)$.

Unfortunately, there are no results as strong as Pilc's for wide classes of real sources. Sakrison [24] has demon-

strated that, for any i.i.d. Gaussian source, there exists a sequence of k-dimensional quantizers having MSE $D(R) + O(1/k)$ and normalized output entropy $R + O(\log k/k)$. For memoryless Gaussian sources, Wyner [25] proved that $D_k(R) - D(R) = O(\log k/k)$, and he later showed in [26] that the bound $D_k(R) - D(R) = O(\sqrt{\log k/k})$ holds for any stationary Gaussian source with a spectral density having a Lipschitz-continuous derivative.

Remark: Pilc also reported a $c \log k/k$ type lower bound on the same quantity, but it has been pointed out by several researchers that his proof of this lower bound is flawed. Thus, at present, finding a lower bound is an open question and there is no indication if the $O(\log k/k)$ upper bound on $D_k(R) - D(R)$ is sharp.

In what follows, we provide a rate of convergence result for i.i.d. sources with bounded support but otherwise arbitrary distribution. The method of proof is an extension of an earlier result of Gallager [27], who showed $O(\sqrt{\log k/k})$ rate of convergence for finite-alphabet sources. We consider only mean-square distortion; the result straightforwardly extends to distortion functions $\rho: \mathscr{R} \times \mathscr{R} \to \mathscr{R}^+$ such that $\rho(x, x) = 0$ and ρ is bounded over compact subsets of $\mathscr{R}^2$.

The *distortion-rate function* $D(R)$ for the mean-square distortion function of a real i.i.d. source $X_1, X_2, \cdots$ is defined for $R \geq 0$ as

$$D(R) = \inf_{Y: I(X; Y) \leq R} E|X - Y|^2,$$

where X has the common distribution of the X_i's and the infimum is taken over all real random variables Y such that the mutual information between X and Y is at most R. For the definition of the mutual information in this case, see [28].

Let $X^{(k)} = (X_1, \cdots, X_k)$. If $E|X_1|^2 < \infty$, the source coding theorem for memoryless sources [27, Thm. 9.6.2] asserts that

$$D(R)$$

$$= \lim_{k \to \infty} \left\{ \inf_{Q_{N,k}: \frac{1}{k} \log N \leq R} \left\{ \frac{1}{k} E \|Q_{N,k}(X^{(k)}) - X^{(k)}\|^2 \right\} \right\}$$

$$= \lim_{k \to \infty} D_k(R), \tag{18}$$

where the infimum is taken over all k-dimensional, N-level quantizers $Q_{N,k}$, with rate $(1/k) \log N$ not exceeding R. With the notation of the previous section,

$$D_k(R) = \frac{1}{k} \Delta \big(Q_{N,k}^*(X^{(k)}) \big),$$

where $N = \lfloor 2^{kR} \rfloor$.

For bounded memoryless sources, an upper bound on the rate of the convergence in (18) is given in the following theorem.

Theorem 2: Let $X_1, X_2, \cdots$ be a real-valued i.i.d. source that is bounded with probability 1 and has distortion-rate function $D(R)$. Then for every $R > 0$ with $D(R) > 0$, there is a constant c such that, for every k, the difference between the per-letter MSE of the best k-dimensional quantizer of rate R and $D(R)$ satisfies

$$D_k(R) - D(R) \leq c \sqrt{\frac{\log k}{k}}.$$

The essential condition $D(R) > 0$ holds for all $R > 0$ when P_X has a continuous component. The proof of Theorem 2 is given in the Appendix and is based on Csiszár's generalization [29] of the parametric representation of the distortion-rate function for abstract-alphabet sources.

IV. UNIVERSAL FIXED-RATE LOSSY CODING

In this section, we apply the previous results to fixed-rate universal lossy coding of i.i.d. sources. The notation of universality we consider is defined next.

Definition 1: For $R > 0$, a sequence of pairs of functions (f_n, ϕ_n) of the form

$$f_n: R^n \to \{0, 1\}^{\lfloor nR \rfloor} \quad \text{and} \quad \phi_n: \{0, 1\}^{\lfloor nR \rfloor} \to \mathscr{R}^n$$

is called a *fixed-rate universal source coding scheme* of rate R with respect to a family of real stationary and ergodic sources, if for each source $X_1, X_2, \cdots$ in the family, the n-blocks $(Y_{1,n}, \cdots, Y_{n,n}) = \phi_n(f_n(X_1, \cdots, X_n))$ satisfy

$$\lim_{n \to \infty} E \left[\frac{1}{n} \sum_{i=1}^{n} (X_i - Y_{i,n})^2 \right] = D(R), \tag{19}$$

where $D(R)$ is the distortion-rate function of the source $X_1, X_2, \cdots$.

In [13], various types of universality of fixed-rate lossy source codes are defined and analyzed in detail. Our Definition 1 is similar to their "weakly minimax" universality. While [13] deals with general alphabets and distortion functions, and stationary sources, we restrict our investigation to memoryless sources, which allows us to obtain a stronger result at the expense of lesser generality.

Ziv [10] considered stationary sources with metric space source and reproduction alphabets such that every bounded subset is compact. The distortion criterion was a "nice" function of the metric. Under these general conditions, he gave a construction of a sequence of source codes of rate less than or equal to R such that the expected distortion converges to $D(R)$ for all stationary sources. (Here $D(R) = \lim_n D_n(R)$, where $D_n(R)$ is the minimum distortion a quantizer of fixed dimension n and rate R can achieve.)

In what follows, we restrict the class of sources to memoryless real sources with bounded support, that is, for i.i.d. sources such that there exists a bounded set $K \subset \mathscr{R}$ such that $\Pr\{X_1 \in K\} = 1$. We establish an almost sure rate of convergence result for the sample distortion of Ziv's scheme when used on such a source. Given a se-

quence of random variables $V_1, V_2, \cdots$ and a sequence of nonnegative reals $a_1, a_2, \cdots$, we say that $V_n = o(a_n)$ a.s., if $\limsup_{n \to \infty}(V_n/a_n) \leq 0$ a.s.

Theorem 3: For every rate $R > 0$, there exists a fixed-rate universal source coding scheme for the family of stationary and ergodic real sources with bounded support that satisfies the following: For any i.i.d. source $X_1, X_2, \cdots$, in this family with distortion-rate function $D(R) > 0$, there is a constant $c > 0$ such that the difference between the per-letter expected MSE and $D(R)$ decays as

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_{i,n})^2\right] - D(R) \leq c\sqrt{\frac{\log\log n}{\log n}}. \quad (20)$$

Moreover, the per-letter sample squared error satisfies

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_{i,n})^2 - D(R)$$

$$\leq c\sqrt{\frac{\log\log n}{\log n}} + o\left(\left(\frac{\log n}{n}\right)^{1/2-\epsilon}\right) \quad \text{a.s.} \quad (21)$$

for any $\epsilon \in (0, \frac{1}{2})$.

Here c is a deterministic constant depending only on the source statistics and the second term on the right-hand side of (21) is a random variable.

We prove the theorem by constructing a universal source coding scheme similar to Ziv's [10]. Ziv proved that his scheme is universal for the class of stationary and ergodic sources. Our scheme is also universal for the class of stationary and ergodic sources, which easily follows from Ziv's proof. In the sequel we will omit the subscripts from the positive integers b_n, k_n, N_n, and m_n, but they always will be understood as functions of n. Now, we are prepared to describe the universal schemes that provide a constructive proof for Theorem 3. Split the data $X^{(n)} = (X_1, \cdots, X_n)$ into k-long blocks $Z_1, \cdots, Z_m$ ($m = \lfloor n/k \rfloor$), where $Z_i = (X_{(i-1)k+1}, \cdots, X_{ki})$ is the k-dimensional vector formed by the ith block of $X^{(n)}$. Consider the k-dimensional quantizer with N code vectors that minimizes the empirical distortion over $Z_1, \cdots, Z_m$, namely,

$$Q^*_{m,N,k} = \arg\min_{Q_{N,k}} \frac{1}{m}\sum_{i=1}^{m}\|Z_i - Q_{N,k}(Z_i)\|^2,$$

and let c_0 be a positive number such that $K \subset [-c_0, c_0]$. Quantize uniformly each of the coordinates of the code vectors of $Q^*_{m,N,k}$ with b bits, and transmit them. The uniform quantizer partitions the interval $[-c_0, c_0]$ into 2^b subintervals of equal length. Denote the quantized quantizer by $\hat{Q}^*_{m,N,k}$, that is, $\hat{Q}^*_{m,N,k}$ is the quantizer whose code vectors are the code-vectors of $Q^*_{m,N,k}$ quantized uniformly using b bits per coordinate. Finally, quantize the vectors Z_i with $\hat{Q}^*_{m,N,k}$ and transmit them. The remaining letters of $X^{(n)}$ are not transmitted, and will be decoded as zeros. Observe that to transmit the quantizer $\hat{Q}^*_{m,N,k}$ (the approximation to the empirically optimal quantizer $Q^*_{m,N,k}$), we use bkN bits, since each of the k coordinates of the N code vectors is quantized with b bits.

To transmit the Z_i's, we use $m \log N$ bits. Therefore, the overall per-letter rate is

$$r_n = \frac{m\log N}{n} + \frac{bkN}{n}. \quad (22)$$

The reproduction value for a block

$$Z_i = (X_{(i-1)k+1}, \cdots, X_{ki})$$

will be the code vector of the quantized version of the $Q^*_{m,N,k}$ which is closet to Z_i, and zero for $X_{mk+1}, \cdots, X_n$. Formally,

$$(Y_{(i-1)k+1,n}, \cdots, Y_{ki,n}) = \hat{Q}_{m,N,k}(Z_i), \quad \text{for } i = 1, \cdots, m,$$

and

$$Y_j = 0, \quad \text{for } j = mk + 1, \cdots, n.$$

Our aim is to choose the parameters k, N, and b (as functions of n) such that $r_n \leq R$ and the overall per-letter distortion converges to $D(R)$ at the fastest possible rate. Clearly, a necessary condition for Theorem 3 is that $k \to \infty$ and $(\log N)/k \to R$ as $n \to \infty$, i.e., the overhead bkN/n should become asymptotically negligible.

Proof of Theorem 3: Assume, for the sake of simplicity, that a positive constant c_0 is known to the receiver for which $\Pr\{|X_1| \leq c_0\} = 1$. Otherwise a bound for $\max(|X_1|, \cdots, |X_n|)$ can be transmitted using a constant number of bits, and therefore, will not affect the rate asymptotically. The overall per-letter sample distortion is

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_{i,n})^2$$

$$= \frac{1}{mk}\sum_{j=1}^{m}\|Z_j - \hat{Q}^*_{m,N,k}(Z_j)\|^2 + \frac{1}{n}\sum_{i=km+1}^{n}|X_i|^2$$

$$= \frac{1}{k}\Delta_m(\hat{Q}^*_{m,N,k}) + \frac{1}{n}\sum_{i=km+1}^{n}|X_i|^2.$$

We want to prove that there is a choice of the parameters k, N, and b such that (20) and (21) are satisfied while $r_n \leq R$. In order to show this, we use the following decomposition:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_{i,n})^2 - D(R)$$

$$= \frac{1}{n}\sum_{i=km+1}^{n}|X_i|^2$$

$$+ \left(\frac{1}{k}\Delta_m(\hat{Q}^*_{m,N,k}) - \frac{1}{k}\Delta_m(Q^*_{m,N,k})\right)$$

$$+ \left(\frac{1}{k}\Delta_m(Q^*_{m,N,k}) - \frac{1}{k}\Delta(Q^*_{N,k})\right)$$

$$+ \left(\frac{1}{k}\Delta(Q^*_{N,k}) - D(R)\right). \quad (23)$$

In what follows, we give upper bounds for the four terms on the right-hand side of (23). For the first term, the

boundedness assumption obviously implies

$$\frac{1}{n} \sum_{i=km+1}^{n} \|X_i\|^2 \le \frac{kc_0^2}{n} \quad \text{a.s.} \tag{24}$$

Therefore, $(1/n)\sum_{i=km+1}^{n}\|X_i\|^2 \to 0$ a.s., if

$$k/n \to 0. \tag{25}$$

For the second term, observe first that by the triangle inequality,

$$
\begin{aligned}
\|Z &- \hat{Q}_{m,N,k}^*(Z)\|^2 \\
&= \|Z - Q_{m,N,k}^*(Z) + Q_{m,N,k}^*(Z) - \hat{Q}_{m,N,k}^*(Z)\|^2 \\
&\le \|Z - Q_{m,N,k}^*(Z)\|^2 + 2\|Z - Q_{m,N,k}^*(Z) \\
&\quad \cdot \|Q_{m,N,k}^*(Z) - \hat{Q}_{m,N,k}^*(Z)\| \\
&\quad + \|Q_{m,N,k}^*(Z) - \hat{Q}_{m,N,k}^*(Z)\|^2 \\
&\le \|Z - Q_{m,N,k}^*(Z)\|^2 \\
&\quad + 6c_0\sqrt{k}\,\|Q_{m,N,k}^*(Z) - \hat{Q}_{m,N,k}^*(Z)\|, \quad \text{a.s.,}
\end{aligned}
$$

where $c_0\sqrt{k}$ is a bound on $\|Z\|$ (a.s.). On the other hand, since each coordinate of the code vector of $Q_{m,N,k}^*$ is uniformly quantized with b bits,

$$\|Q_{m,N,k}^*(Z) - \hat{Q}_{m,N,k}^*(Z)\| \le \frac{c_0\sqrt{k}}{2^b} \quad \text{a.s.}$$

Therefore, we can bound the second term of (23) as

$$
\begin{aligned}
\frac{1}{k}\Delta_m\big(\hat{Q}_{m,N,k}^*\big) &- \frac{1}{k}\Delta_m(Q_{m,N,k}^*) \\
&= \frac{1}{k}\left(\frac{1}{m}\sum_{j=1}^{m}\Big(\|Z_j - \hat{Q}_{m,N,k}^*(Z_j)\|^2 \right. \\
&\qquad\quad \left. -\|Z_j - Q_{m,N,k}^*(Z_j)\|^2\Big)\right) \\
&\le \frac{1}{km}\sum_{j=1}^{m} 6c_0\sqrt{k}\,\frac{c_0\sqrt{k}}{2^b} = \frac{6c_0^2}{2b} \quad \text{a.s.,} \tag{26}
\end{aligned}
$$

which goes to zero if $b \to \infty$ as $n \to \infty$. For the third term, first we use definitions (1) and (2) to conclude that

$$
\begin{aligned}
\frac{1}{k}E[\Delta_m(Q_{m,N,k}^*)] &- \frac{1}{k}\Delta(Q_{N,k}^*) \\
&= E\left[\frac{1}{k}\Delta_m(Q_{m,N,k}^*) - \frac{1}{k}\Delta_m(Q_{N,k}^*)\right] \le 0. \tag{27}
\end{aligned}
$$

As for the almost sure statement for the same term, we have by Lemma 1 that for every $t > 0$,

$$
\begin{aligned}
\Pr\left\{\frac{1}{k}(\Delta_m(Q_{m,N,k}^*) - \Delta(Q_{N,k}^*)) > t\right\} \\
\le e^{-mk^2t^2/(8k^2c_0^4)} = e^{-mt^2/(8c_0^4)}. \tag{28}
\end{aligned}
$$

Therefore, taking $t = 1/\sqrt{m^{1-\epsilon}}$, we get $(1/k)\sqrt{m^{1-\epsilon}}$ $(\Delta_m(Q_{m,N,k}^*) - \Delta(Q_{N,k}^*)) \to 0$ a.s. for any $\epsilon > 0$ by the Borel–Cantelli lemma.

To handle the last term of (23), write

$$
\begin{aligned}
\frac{1}{k}\Delta(Q_{N,k}^*) - D(R) &= \frac{1}{k}\Delta(Q_{N,k}^*) - D\left(\frac{1}{k}\log N\right) \\
&\quad + D\left(\frac{1}{k}\log N\right) - D(R). \tag{29}
\end{aligned}
$$

By a first-order Taylor expansion of $D(R)$, we have

$$D\left(\frac{1}{k}\log N\right) - D(R) = (D'(R) + o(1))\left(R - \frac{1}{k}\log N\right), \tag{30}$$

as $(\log N)/k \to R$, where the existence of the derivative $D'(R)$ follows from Lemma 5 in the Appendix. Furthermore, from Theorem 2, we have

$$\frac{1}{k}\Delta(Q_{N,k}^*) - D\left(\frac{1}{k}\log N\right) \le (c(N,k) + o(1))\sqrt{\frac{\log k}{k}}, \tag{31}$$

where $c(N,k)$ depends on N and k. It follows from the proof of Theorem 2 that $c(N,k) \to c_2$, for some constant c_2, as $(\log N)/k \to R$. Thus we have from (29), (30), and (31) that for some constant c_3,

$$
\begin{aligned}
\frac{1}{k}\Delta(Q_{N,k}^*) &- D(R) \\
&\le (c_3 + o(1))\left(\sqrt{\frac{\log k}{k}} + R - \frac{\log N}{k}\right), \tag{32}
\end{aligned}
$$

as $N, k \to \infty$ such that $(\log N)/k \uparrow R$. Combining these bounds, we have

$$
\begin{aligned}
\frac{1}{n}E\left[\sum_{i=1}^{n}(X_i - Y_{i,n})^2\right] &- D(R) \\
&\le (c_4 + o(1))\left(\sqrt{\frac{\log k}{k}} + 2^{-b} + R - \frac{\log N}{k}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_{i,n})^2 &- D(R) \\
&\le (c_4 + o(1))\left(\sqrt{\frac{\log k}{k}} + 2^{-b} + R - \frac{\log N}{k}\right) \\
&\quad + o\left(\frac{1}{\sqrt{m^{1-\epsilon}}}\right) \quad \text{a.s.}
\end{aligned}
$$

for some constant c_4. Now, the parameters are to be chosen to minimize this upper bound, subject to the constraint

$$r_n = \frac{\log N}{k} + \frac{bkN}{n} \le R.$$

The choices

$$k = \frac{1}{R}(1 - \epsilon) \log n,$$

$$N = \lfloor 2^{k(R - (1/\log n))} \rfloor, \quad b = \log\log n$$

give the desired results, completing the proof. □

Remarks: (a) By the proofs of Theorem 2 and Theorem 3, the constant c in (20) depends on the derivative $D'(R)$ of the distortion-rate function and the diameter of the support of the source.

(b) It follows from [10] that a scheme almost identical to what we used in the preceding proof is universal (in the sense that the expected squared distortion converges with increasing block length to the infimum of the distortions of all rate-R quantizers of arbitrary dimension) over the class of stationary sources with marginals of finite second moment. It is easy to conclude that the scheme we used is also universal in the above sense over the class of stationary sources with marginals of bounded support. In fact, if we only require universality over the class of all i.i.d. sources of bounded support, then a faster convergence rate can be achieved [30].

V. AVERAGE DISTORTION OF EMPIRICALLY DESIGNED QUANTIZERS

The evaluation of the performance of quantizers designed from a set of training data is a fundamental problem of vector quantizer design. Here we give an exact upper bound for the case when the source is memoryless and of bounded support.

As defined in (2), let $Q^*_{m,N,k}$ be the N-level, k-dimensional quantizer minimizing the sample distortion over the k-vectors $Z_1, \cdots, Z_m$. The usual analysis (see, e.g., [1], [7], [3], and [8]), examines the consistency of $Q^*_{m,N,k}$, i.e., the convergence of its average distortion to the distortion of the optimal quantizer $Q^*_{N,k}$.

However, a different question of equal importance can be addressed. Suppose we are given the n scalar samples $X_1, \cdots, X_n$ from a source. What is the minimal distortion (for encoding the true source) over all possible vector quantizers, of rate of most R, designed from this training set? This problem differs from the previous one in that the dimension of the quantizer (and therefore the number of quantization levels) is a free parameter that needs to be determined. That is, the n samples are to be partitioned into $m = \lfloor n/k \rfloor$ vectors, each of dimension k. Clearly, a quantizer of large dimension can produce smaller distortion but needs much more training data to fit the source statistics properly. As the dimension k increases, the number m of training vectors decreases. This in turn decreases the ability to design a quantizer from the m training vectors that will perform well on the true source, since the resulting quantizer will tend to overfit the training data. Hence, for a fixed rate R, there is a trade-off in determining the optimal vector dimension k as a function of the training set size n, between using longer blocks for better quantization and using more training vectors for a better fit of the quantizer to the true source. The next theorem examines this trade-off and gives an upper bound on the rate of convergence of the minimal distortion to the distortion-rate function.

Theorem 4: Let $X_1, \cdots, X_n$ be n samples from a real-valued i.i.d. source that are bounded with probability 1, and suppose these samples are blocked into k-dimensional "training" vectors $Z_1, \cdots, Z_m$, where $Z_i = (X_{(i-1)k+1}, \cdots, X_{ki})$ and $m = \lfloor n/k \rfloor$. Let $Q^*_{m,N,k}$ be a k-dimensional vector quantizer designed to minimize the empirical MSE for the m training vectors. Then by choosing $k = \lfloor (1/R)(1 - \epsilon) \log n \rfloor$, for any $\epsilon \in (0, 1)$, the per-letter MSE of $Q^*_{m,N,k}$ for the true source, converges to the distortion-rate function at the rate

$$\frac{1}{k}\Delta(Q^*_{m,N,k}) - D(R) = O\left(\sqrt{\frac{\log\log n}{\log n}}\right) \quad \text{a.s.} \quad (33)$$

Remark: We can also obtain the same rate of convergence for convergence in expectation by following the same technique as in the remark after Corollary 1.

Proof of Theorem 4: First we note that if $n > mk$, then the samples $X_{mk+1}, \cdots, X_n$ are not used. Setting the quantizer dimension as the logarithm of the sample size is not necessarily optimal; it is a choice required by our proof technique.

The left-hand side of (33) can be written as

$$\frac{1}{k}\Delta(Q^*_{m,N,k}) - D(R) = \frac{1}{k}\Delta(Q^*_{m,N,k}) - \frac{1}{k}\Delta(Q^*_{N,k})$$

$$+ \frac{1}{k}\Delta(Q^*_{N,k}) - D(R), \quad (34)$$

where, by the rate constraint, we have $N \leq 2^{kR}$. We will use Theorem 1 to upper bound $(1/k)\Delta(Q^*_{m,N,k}) - (1/k)\Delta(Q^*_{N,k})$. Since $Z_i = (X_{(i-1)k+1}, \cdots, X_{ki})$, the conditions of Theorem 1 are satisfied with $B = kc_0^2$, and we have

$$\Pr\left\{\frac{1}{k}\Delta(Q^*_{m,N,k}) - \frac{1}{k}\Delta(Q^*_{N,k}) > t\right\}$$

$$\leq 4(2m)^{N(k+1)}e^{-mt^2/(512c_0^4)}. \quad (35)$$

For the sake of simplicity, we assume that $n = mk$ in the following argument.

Since both differences in (34) are nonnegative, the left-hand side can converge to zero only if $(1/k)\Delta(Q^*_{N,k}) - D(R) \to 0$ as $n \to \infty$. This means that the number of quantization levels is of the form $N = 2^{k(R-\epsilon_n)}$, where $\epsilon_n \downarrow 0$ as $n \to \infty$. Thus, the right-hand side of (35) equals

$$4\left(\frac{2n}{k}\right)^{2^{k(R-\epsilon_n)}(k+1)} e^{-t^2 n/(512kc_0^4)}. \quad (36)$$

It is easy to check that, for any $\epsilon \in (0, 1)$ and $\delta \in (0, \epsilon/4)$, if $k = (1/R)(1 - \epsilon)\log n$ and $t = n^{-\delta}$, then (36) is summable over n, and we have, by the Borel–Cantelli lemma, that

$$\frac{1}{k}\Delta(Q^*_{m,N,k}) - \frac{1}{k}\Delta(Q^*_{N,k}) = O(n^{-\delta}) \quad \text{a.s.} \quad (37)$$

Note that for any $\epsilon > 0$, the choice $k = (1/R)(1 + \epsilon)\log n$ makes the sum infinite for fixed $t > 0$, so that our choice of k is the best possible for the given proof technique. Setting $N = \lfloor 2^{kR} \rfloor$ and using Theorem 2 gives

$$\frac{1}{k}\Delta(Q^*_{N,k}) - D(R) = O\left(\sqrt{\frac{\log k}{k}}\right), \qquad (38)$$

and we have

$$\frac{1}{k}\Delta(Q^*_{m,N,k}) - D(R) = O\left(\sqrt{\frac{\log \log n}{\log n}}\right) + O(n^{-\delta}),$$

a.s.

Also note that the order of the term $O(n^{-\delta})$ does not change for the more general choice $N = 2^{k(R-\epsilon_n)}$, but the convergence in (38) can only be made slower. Thus for $N = \lfloor 2^{kR} \rfloor$ and $k = (1/R)(1 - \epsilon)\log n$, we have

$$\frac{1}{k}\Delta(Q^*_{m,N,k}) - D(R) = O\left(\sqrt{\frac{\log \log n}{\log n}}\right) \quad \text{a.s.,}$$

which proves the theorem. $\qquad\qquad\square$

Remarks: (a) There is a certain duality between the proofs of Theorems 3 and 4. When proving Theorem 3, we were forced by the overall rate constraint to limit the quantizer dimension to be proportional to the logarithm of the block length. The same choice is made in Theorem 4, but for a different reason, namely that, for larger k, our large deviations inequality does not guarantee that the quantizer is well trained.

(b) It may be possible to improve our convergence rate result in a couple of ways. First, if Pilc's upper bound could be proven for the nondiscrete case, the square root could be eliminated in (33). Second, one might attempt to improve the factor $4(2m)^{N(k+1)}$ in Theorem 1 by exploiting the fact that not only are the training vectors $Z_1, \cdots, Z_m$ independent, but so are the coordinates themselves.

(c) Theorem 4 (as well as Theorem 3) could be slightly generalized by considering i.i.d. vectors sources, i.e., a sequence of independent and identically distributed random vectors (the coordinates need not be i.i.d. random variables). It is clear that the methods of proof carry over to this setup. Lemma 1 and Theorem 1 apply without any change; only Theorem 2 needs to be generalized for i.i.d. vector sources.

APPENDIX

In the course of the proof of Theorem 2, we will make use of Csiszár's generalization of the parametric representation of the distortion-rate curve. The following lemma summarizes some general results from [29] adopted here for real-valued sources.

Lemma 5 (Csiszár [29]): Let X be a real-valued random variable such that $E|X|^2 < \infty$ and consider the squared error distortion measure. Then for each point $(D(R), R)$ of the distortion-rate curve, there exists a random variable Y such that

$$I(X;Y) = R \quad \text{and} \quad E|X - Y|^2 = D(R). \qquad (A1)$$

Furthermore, the Radon–Nikodym derivative $a(x, y)$ of P_{XY} with respect to $P_{X \times Y}$ is given as

$$a(x, y) = \alpha(x)2^{-s|x-y|^2}, \qquad (A2)$$

where $s \geq 0$ and $\alpha(x) \geq 1$ satisfy

$$\int \alpha(x)2^{-s|x-y|^2}P_X(dx) \leq 1, \quad \text{for all } y \in \mathscr{R}^1. \qquad (A3)$$

Here $-1/s$ is the derivative of the distortion-rate function at R.

Remark: When $|X|$ is bounded by c_0, the same is true for the Y achieving the distortion-rate bound in (A1). This is easily seen by defining $g(y) = yI_{\{|y| \leq c_0\}} + c_0 I_{\{y > c_0\}} - c_0 I_{\{y < -c_0\}}$, where I_A is the indicator of the set A. Then we have, for any random variable Y,

$$E|X - g(Y)|^2 \leq E|X - Y|^2$$

and

$$I(X; g(Y)) \leq I(X, Y).$$

The claim now follows because the first inequality is strict when $\Pr\{|Y| > c_0\} > 0$. Also, recall that the definition of $I(X; Y)$ in this case is given as follows. If P_{XY} is absolutely continuous with respect to $P_{X \times Y}$ with Radon–Nikodym derivative $dP_{XY}/dP_{X \times Y} = a(x, y)$, then

$$I(X;Y) = \int \log a(x, y)P_{XY}(dx, dy);$$

otherwise $I(X;Y) = \infty$.

Proof of Theorem 2: We follow the classical random coding proof of the source coding theorem [27]. Let $N = \lfloor 2^{k(R+u)} \rfloor$, where $u > 0$ will be specified later. Let $A_1, \cdots, A_N$ be k-dimensional random vectors, $A_i = (A_{i1}, \cdots, A_{ik})$, such that the A_{ij}'s are i.i.d. and have the same distribution as the Y which achieves the distortion-rate bound in (A1).

Furthermore, let the collection $A_1, \cdots, A_N$ be independent of $X^{(k)} = (X_1, \cdots, X_k)$. The vectors $A_1, \cdots, A_N$ are the codevectors of a k-dimensional random quantizer Q using the nearest-neighbor rule. Then by independence,

$$D_k(R + u) \leq E(D(Q))$$

$$= \int_{\mathscr{R}^k} E\left(\min_{1 \leq i \leq N} \frac{1}{k}\|x - A_i\|^2\right) P_X^k(dx), \quad (A4)$$

where the measure P_X^k is the k-fold product of P_X. Now, as in the finite-alphabet case, we have an a.s. upper bound on the distortion of Q, which is $4c_0^2$ by the above remark. It is easily checked that, using the same steps for bounding the right-hand side of (A4) as in the proof of the finite-alphabet case (converting sums to integrals with respect to the corresponding probability measures), we get

$$D_k(R + u) \leq D(R) + u + 4c_0^2 \Pr\{Y^{(k)} \notin S_{X^{(k)}}\}$$
$$+ 4c_0^2 e^{-2^{ku/2}}, \quad (A5)$$

where $Y^{(k)} = (Y_1, \cdots, Y_N)$, and the pairs (X_i, Y_i) are i.i.d. with the same joint distribution as (X, Y). The sets $S_{x^{(k)}}$, for $x^{(k)} = (x_1, \cdots, x_n) \in \mathscr{R}^k$, are defined as

$$S_{x^{(k)}} = \left\{ y^{(k)} = (y_1, \cdots, y_k): \frac{1}{k}\sum_{j=1}^{k} \log a(x_i, y_i) \leq R + u/2, \right.$$

$$\left. \frac{1}{k}\|x^{(k)} - y^{(k)}\|^2 \leq D(R) + u \right\}.$$

Here $a(x, y)$ is the Radon–Nikodym derivative $dP_{XY}/dP_{X \times Y}$, and we have $E[\log a(X, Y)] = I(X; Y)$. We then get the bound

$$\Pr\{Y^{(k)} \notin S_{X^{(k)}}\} \le \Pr\left\{\frac{1}{k}\sum_{j=1}^{k}\log a(X_i, Y_i) \ge R + u/2\right\}$$

$$+ \Pr\left\{\frac{1}{k}\|X^{(k)} - Y^{(k)}\|^2 \ge D(R) + u\right\}.$$

$$\text{(A6)}$$

At this point, we depart from Gallager's proof and instead use Hoeffding's inequality to give exponential upper bounds on the probabilities on the right-hand side of (A6). The second term in (A6) is easily handled using the $4c_0^2$ upper bound on the distortion and Hoeffding's inequality (6) to get

$$\Pr\left\{\frac{1}{k}\|X^{(k)} - Y^{(k)}\|^2 \ge D(R) + u\right\} \le e^{-ku^2/(8c_0^2)}. \quad \text{(A7)}$$

To apply Hoeffding's inequality to the first term, we must show that $|\log a(X, Y)|$ is a bounded random variable. Here we use Lemma 5. By (A2), we have $\log a(x, y) = \log \alpha(x) - s|x - y|^2$, where $\alpha(x) \ge 1$. Integrating both sides of (A2) with respect to $P_Y(dy)$ yields

$$\alpha(x) = \left(\int 2^{-s|x-y|^2} P_Y(dy)\right)^{-1}, \quad \text{for } P_X \text{ a.e.} \quad \text{(A8)}$$

But $|x - y|^2 \le 4c_0^2 P_{XY}$ a.e., thus

$$1 \ge 2^{-s|x-y|^2} \ge 2^{-s4c_0^2} \quad \text{and} \quad 1 \le \alpha(x) \le 2^{s4c_0^2},$$

$$\text{for } P_X \text{ a.e.,} \quad \text{(A9)}$$

which gives us

$$2^{-s4c_0^2} \le a(X, Y) \le 2^{s4c_0^2} \quad \text{a.s.} \quad \text{(A10)}$$

and hence

$$|\log a(x, y)| \le s4c_0^2 \quad \text{a.s.} \quad \text{(A11)}$$

Now, Hoeffding's inequality implies

$$\Pr\left\{\frac{1}{k}\sum_{j=1}^{k}\log a(X_i, Y_i) \ge R + u/2\right\} \le e^{-ku^2/(s^2 256 c_0^4)}. \quad \text{(A12)}$$

In view of (A5), (A6), (A7), and (A12), we can upper bound $D_k(R + u)$ as

$$D_k(R + u) \le D(R) + u + 4c_0^2(e^{-ku^2/(8c_0^2)}$$

$$+ e^{-ku^2/(s^2 256 c_0^4)}) + e^{-2ku/2}. \quad \text{(A13)}$$

Putting $c_s = \max\{2c_0, s\sqrt{128}\,c_0^2\}$ and $u = c_s\sqrt{\ln k/k}$, we get

$$D_k\left(R + c_s\sqrt{\ln k/k}\right) - D(R) = (c_s + o(1))\sqrt{\frac{\ln k}{k}}. \quad \text{(A14)}$$

Since $-1/s$ is the slope of the distortion-rate curve at the point $(R, D(R))$, for $c_{s'}$ corresponding to a rate $R' < R$ we have $c_{s'} \le c_s$. This means that (A14) also holds when R is replaced by any $R' \le R$. This implies

$$D_k(R) - D\left(R - c_s\sqrt{\ln k/k}\right) \le (c_s + o(1))\sqrt{\frac{\ln k}{k}}, \quad \text{(A15)}$$

and it follows from the first-order Taylor expansion of $D(R)$ at R, and the fact that $D'(R) = -1/s$, that

$$D_k(R) - D(R) \le c_s\left(1 + \frac{1}{s} + o(1)\right)\sqrt{\frac{\ln k}{k}},$$

or, equivalently, using a base-2 logarithm,

$$D_k(R) - D(R) \le c(1 + o(1))\sqrt{\frac{\log k}{k}},$$

where $c = c_s(1 + 1/s)\ln 2$, which completes the proof.

ACKNOWLEDGMENT

REFERENCES

[1] E. A. Abaya and G. L. Wise, "Convergence of vector quantizers with applications to optimal quantization," *SIAM J. Appl. Math.*, vol. 44, pp. 183–189, 1984.
[2] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Boston, Kluwer, 1992.
[3] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 199–205, Mar. 1982.
[4] M. J. Sabin and R. M. Gray, "Global convergence and empirical consistency of the generalized Lloyd algorithm," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 148–155, March 1986.
[5] P. C. Cosman, K. O. Perlmutter, S. M. Perlmutter, R. A. Olshen, and R. M. Gray, "Training sequence size and vector quantizer performance," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, pp. 434–438, 1991.
[6] D. Cohn, E. Riskin, and R. Ladner, "Theory and practice of vector quantizers trained on small training sets," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, pp. 54–65, Jan. 1994.
[7] D. Pollard, "Strong consistency of k-means clustering," *Ann. Stat.*, vol. 9, no. 1, pp. 135–140, 1981.
[8] M. J. Sabin, "Global convergence and empirical consistency of the generalized Lloyd algorithm," Ph.D. thesis, Stanford Unv., 1984.
[9] 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.
[10] 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.
[11] K. Zeger, A. Bist, and T. Linder, "Universal source coding with codebook transmission," *IEEE Trans. Commun.*, vol. 42, pp. 336–346, Feb. 1994
[12] 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. Inform. Theory, San Antonio, TX, 1992.
[13] 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.
[14] M. B. Pursley and L. D. Davisson, "Variable rate coding for nonergodic sources and classes of ergodic sources subject to a fidelity constraint," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 324–337, May 1976.
[15] 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.
[16] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.
[17] W. Hoeffding, Probability inequalities for sums of bounded random variables," *J. Am. Stat. Assoc.*, vol. 58, pp. 13–30, 1963.
[18] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, pp. 264–280, 1971.
[19] L. Devroye, "The uniform convergence of the Narayada–Watson regression function estimate," *Can. J. Stat.*, vol. 6, pp. 179–191, 1978.
[20] R. M. Dudley, "Balls in $\mathscr{R}^k$ do not cut all subsets of $k + 2$ points," *Adv. Math.*, vol. 31, no. 3, pp. 306–308, 1979.
[21] K. Alexander, "Probability inequalities for empirical processes and a law of the iterated logarithm," *Ann. Probab.*, vol. 4, pp. 1041–1067, 1984.
[22] L. Devroye, "Automatic pattern recognition: A study of the probability of error," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, pp. 530–543, 1988.

[23] A. Nobel and R. Olshen, personal communication.
[24] D. J. Sakrison, "A geometric treatment of the source encoding of a Gaussian random variable," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 481–486, May 1968.
[25] 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.
[26] A. D. Wyner, "On the transmission of correlated Gaussian data over a noisy channel with finite encoding block length," *Inf. Contr.*, vol. 20, pp. 193–215, 1972.

[27] R. G. Gallager, *Information Theory and Reliable Communication.* New York, Wiley, 1968.
[28] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ. Prentice-Hall, 1971.
[29] I. Csiszár, "On an extremum problem of information theory," *Stud. Sci. Math. Hung.*, pp. 57–70, 1974.
[30] T. Linder, G. Lugosi, and K. Zeger, "Fixed rate universal lossy source coding and rate of convergence for memoryless sources," *IEEE Trans. Inform. Theory*, to appear.