

# A Vector Quantization Approach to Universal Noiseless Coding and Quantization

Philip A. Chou, *Member, IEEE*, Michelle Effros, *Member, IEEE*, and Robert M. Gray, *Fellow IEEE*

**Abstract**—A two-stage code is a block code in which each block of data is coded in two stages: the first stage codes the identity of a block code among a collection of codes, and the second stage codes the data using the identified code. The collection of codes may be noiseless codes, fixed-rate quantizers, or variable-rate quantizers. We take a vector quantization approach to two-stage coding, in which the first stage code can be regarded as a vector quantizer that “quantizes” the input data of length  $n$  to one of a fixed collection of block codes. We apply the generalized Lloyd algorithm to the first-stage quantizer, using induced measures of rate and distortion, to design locally optimal two-stage codes. On a source of medical images, two-stage variable-rate vector quantizers designed in this way outperform standard (one-stage) fixed-rate vector quantizers by over 9 dB. The tail of the operational distortion-rate function of the first-stage quantizer determines the optimal rate of convergence of the redundancy of a universal sequence of two-stage codes. We show that there exist two-stage universal noiseless codes, fixed-rate quantizers, and variable-rate quantizers whose per-letter rate and distortion redundancies converge to zero as  $(k/2)^{n^{-1}} \log n$ , when the universe of sources has finite dimension  $k$ . This extends the achievability part of Rissanen’s theorem on universal noiseless codes to universal quantizers. Further, we show that the redundancies converge as  $O(n^{-1})$  when the universe of sources is countable, and as  $O(n^{-1+\epsilon})$  when the universe of sources is infinite-dimensional, under appropriate conditions.

**Index Terms**—Two-stage, adaptive, compression, minimum description length, clustering.

## I. INTRODUCTION

**D**URING its “Grand Tour” of our solar system between 1978 and 1990, the spacecraft Voyager transmitted back to Earth millions of stunning images of encounters with the outer planets and their moons. Appropriately enough, responsible for Voyager’s image compression was a universal noiseless code. This early universal code, called the Rice machine after its inventor, independently encoded each block of 16 image pixels using the best of four memoryless entropy codes; the selected code was specified to the decoder using a two-bit prefix for each coded block [45], [46].

Manuscript received July 25, 1994; revised February 27, 1996. The material in this paper was presented in part at the 1991, 1993, and 1995 International Symposia on Information Theory. This paper is based on work partially supported by the NSF under Grant MIP-9501977, by an AT&T Bell Laboratories Ph.D. Scholarship, by a grant from the Center for Telecommunications at Stanford, and by an NSF Graduate Fellowship.

P. A. Chou is with the Xerox Palo Alto Research Center, Palo Alto, CA 94304 USA.

M. Effros is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA.

R. M. Gray is with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305-4055 USA. Publisher Item Identifier S 0018-9448(96)04589-0.

The Rice machine is an example of a two-stage noiseless code. The first stage describes a memoryless code and the second stage describes the data using the code. In essence, the first stage quantizes to two bits the space of all possible memoryless codes for the second stage. Thus there is a tradeoff between the average lengths of the first- and second-stage descriptions. If the first-stage description is long, then the space of second-stage codes is finely quantized, and there is a second-stage code fairly well matched to the data in the sense that it has low redundancy.<sup>1</sup> Therefore, the second-stage description is relatively short, on average. Conversely, if the first-stage description is short, then the second-stage description is relatively long, on average. By identifying redundancy as distortion, this tradeoff is formally equivalent to the distortion-rate tradeoff in ordinary quantization.

Two-stage quantizers can be similarly constructed. In a two-stage quantizer, the first stage describes a block memoryless quantizer and the second stage describes (quantizes) the data using the quantizer. The finer the description in the first stage, the lower the excess distortion (or “distortion redundancy”) in the second. Here again, this tradeoff is formally equivalent to the distortion-rate tradeoff in ordinary quantization. This paper examines both universal noiseless coding and universal quantization from this quantization point of view.

The quantization point of view enables us to solve two important problems in two-stage universal coding: 1) determining the overall redundancy of an optimal two-stage code as a function of data length, and 2) designing the optimal two-stage code for any particular data length. Our solution to the first problem is roughly the following, for noiseless coding. (The solution for quantization is similar.) Let  $\tilde{R}$  be the length of the first stage in bits, and let  $\tilde{D}(\tilde{R})$  be the expected redundancy of the second stage in bits per letter. Naturally,  $\tilde{D}(\tilde{R})$  is a decreasing function of  $\tilde{R}$ . When  $\tilde{R}$  is amortized by the length of the data  $n$ , the overhead of the first stage is  $\tilde{R}/n$  bits per letter. Thus

$$\rho(\tilde{R}) = \frac{1}{n}\tilde{R} + \tilde{D}(\tilde{R})$$

is the overall redundancy per letter of the universal noiseless code when the length of the first stage is  $\tilde{R}$  bits. The minimum of  $\rho(\tilde{R})$  occurs when the first derivative  $\rho'(\tilde{R}) = (1/n) + \tilde{D}'(\tilde{R})$  equals 0, or when the length of the first stage is  $\tilde{R}^* = (\tilde{D}')^{-1}(-1/n)$ . When  $n$  is large,  $\tilde{R}^*$  is also large, so we can apply asymptotic quantization theory. If the space of

<sup>1</sup>The redundancy of a code relative to a source is the excess number of bits to which it can encode the source relative to the optimal code for that source.

sources  $\{P_\theta\}$  is finite-dimensional, i.e.,  $\theta \in \mathbb{R}^k$ , then the space of codes  $\{-\log P_\theta\}$  is also finite-dimensional, and  $\tilde{D}(\tilde{R})$  has the familiar asymptotic form  $\tilde{D}(\tilde{R}) = A2^{-2\tilde{R}/k}$ , for large  $\tilde{R}$ . This implies that

$$\rho(\tilde{R}^*) = \frac{k \log n}{2n} + O(n^{-1}).$$

This is the familiar achievable redundancy result of Rissanen and others for universal noiseless coding of finitely parameterized sources [15], [16], [33], [47], [48]. We show the same result for universal quantization. Furthermore, we show that in infinite-dimensional cases (e.g., noiseless coding of countably infinite alphabets and variable-rate vector quantization) the distortion-rate tradeoff has the asymptotic power-law form  $\tilde{D}(\tilde{R}) = A\tilde{R}^{-b}$  for large  $\tilde{R}$ , whence

$$\rho(\tilde{R}^*) = O(n^{-b/(b+1)}).$$

That is, the overall redundancy per letter follows a power law. Rates of convergence for redundancies of universal quantizers were first reported in [11], and have also recently been examined in [20], [36], [37], and [53].

The second problem in two-stage universal coding that is solved, by regarding it as a quantization problem, is the problem of optimal design. That is, given a fixed data length  $n$  and first-stage length  $\tilde{R}$ , what collection of  $2^{\tilde{R}}$  codes minimizes the expected redundancy in the second stage? By identifying codes as codewords, this problem is formally equivalent to the problem of optimal quantizer design. For large  $n$ , the optimal codebook can be “designed” by appealing to asymptotic quantization theory. For small  $n$ , it is generally necessary to use a codebook optimization algorithm. We use an iterative descent algorithm formally equivalent to the generalized Lloyd algorithm to design locally optimal weighted universal codes. This procedure, for the case of weighted universal quantizers, was first reported in [9] and has more recently been examined in [1], [18], [19], and [21].

It is interesting to note that Block Truncation Coding [17], [43] can be regarded as a special case of two-stage fixed-rate quantization. Hence the methods described here are applicable to the design of optimal block truncation codes.

We now outline the remainder of the paper. Section II provides the necessary notational and theoretical preliminaries. Section III details our quantization approach and shows how it is used to determine the redundancies of two-stage weighted universal noiseless codes and quantizers. Section IV details our design algorithm. Section V presents our experimental results. In one case we report a 9-dB improvement of our weighted universal entropy-constrained vector quantizer over a standard vector quantizer matched to the mixture source. Section VI is a summary and conclusion. Readers who wish to forego the theoretical details of our approach can skip straight to the algorithm section.

## II. PRELIMINARIES

Let  $\{X_i\}, i = 1, 2, \dots$ , be a stationary random process with alphabet  $\mathcal{X}$  and let  $\Theta$  be a jointly distributed random

variable with alphabet  $\Lambda$ .<sup>2</sup> Denote the marginal distributions of  $\{X_i\}$  and  $\Theta$  by  $P$  and  $W$ , respectively, and denote the regular conditional distribution of  $\{X_i\}$  given  $\Theta = \theta$  by  $P_\theta$ . Assume  $P_\theta$  is stationary but not necessarily ergodic for each  $\theta \in \Lambda$ . Denote the distributions on the  $n$ -blocks  $X^n = (X_1, \dots, X_n)$  by  $P^n$  and  $P_\theta^n$ , as appropriate, with densities  $p^n$  and  $p_\theta^n$  (when they exist). The superscript  $n$  can be dropped when it is clear from the argument, e.g.,  $p(x^n)$ , and the subscript  $\theta$  can be written as the more usual conditional argument, e.g.

$$\begin{aligned} p(x^n|\theta) &= p_\theta(x^n) \\ E[X|\theta] &= E_\theta X \\ H(X^n|\theta) &= H_\theta(X^n) \end{aligned}$$

etc.

We wish to code  $\{X_i\}$  using a block code of length  $n$ . Let  $C^n = \beta \circ \alpha$  be such a code, with encoder  $\alpha: \mathcal{X}^n \rightarrow \mathcal{S}$  and decoder  $\beta: \mathcal{S} \rightarrow \hat{\mathcal{X}}^n$ , where  $\mathcal{S} \subseteq \{0, 1\}^*$  is a binary prefix code and  $\hat{\mathcal{X}}$  is the reproduction alphabet. If  $C^n$  is a noiseless code, then  $\hat{\mathcal{X}} = \mathcal{X}$ , and  $\beta(\alpha(x^n)) = x^n$  for all  $x^n \in \mathcal{X}$ . If  $C^n$  is a lossy block code, or equivalently a vector quantizer, then (in general)  $\hat{\mathcal{X}} \neq \mathcal{X}$ , and  $\hat{\mathcal{X}}$  are possibly uncountably infinite, and  $\beta(\alpha(x^n)) = \hat{x}^n$  with (instantaneous) distortion given by

$$d(x^n, \hat{x}^n) = \sum_i d(x_i, \hat{x}_i)$$

where  $d: \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}^+$  is a single-letter distortion measure or fidelity criterion [4]. If  $C^n$  is a *fixed-rate* vector quantizer, then  $\mathcal{S}$  is a finite collection of fixed-length binary strings; if  $C^n$  is a *variable-rate* (entropy-constrained) vector quantizer, then  $\mathcal{S}$  is a finite or countable collection of variable-length binary strings. In either case, the collection of reproduction vectors  $\Gamma = \{\beta(s): s \in \mathcal{S}\}$  is called the reproduction codebook. If  $C^n$  is a noiseless code, then  $\mathcal{S}$  is of course also a finite or countable collection of variable-length binary strings. In all cases, the (instantaneous) rate of  $C^n$  is denoted

$$r(x^n) = |\alpha(x^n)|$$

where  $|s|$  denotes the length in bits of  $s = \alpha(x^n) \in \mathcal{S}$ . The expected rate of  $C^n$ , with respect to  $P_\theta$ , is

$$R_\theta(C^n) = E_\theta r(X^n) = E_\theta |\alpha(X^n)|$$

while the expected distortion of  $C^n$ , when applicable, is

$$D_\theta(C^n) = E_\theta d(X^n, C(X^n)) = E_\theta d(X^n, \beta(\alpha(X^n))).$$

The expected rate and distortion of  $C^n$  with respect to the mixture  $P$  are simply

$$R(C^n) = ER(C^n|\Theta) = ER(X^n)$$

and

$$D(C^n) = ED(C^n|\Theta) = E d(X^n, C(X^n))$$

respectively.

<sup>2</sup>More precisely, assume that  $\{X_i\}$  and  $\Theta$  are defined on a standard measurable space  $(\Omega, \mathcal{B})$ , so that regular conditional probabilities exist and the ergodic decomposition holds [25]. A sufficient condition for  $(\Omega, \mathcal{B})$  to be standard is that  $\mathcal{X}$  be a complete separable metric (Polish) space, and that either  $\Lambda$  be Polish or  $\Theta$  be a function of  $X^\infty$ , e.g., the ergodic mode of  $\{X_i\}$ . This covers almost any situation encountered in practice.

A universal code for  $(\{X_i\}, \Theta)$  is a sequence of block codes  $\{C^n\}$ ,  $n = 1, 2, \dots$ , such that for each  $\theta \in \Lambda$ ,  $R_\theta(C^n)$  and (when applicable)  $D_\theta(C^n)$  converge to the optimal performance theoretically achievable (OPTA) for  $P_\theta$ , as  $n \rightarrow \infty$ . Thus a universal code is a single sequence of codes that eventually achieves optimal performance regardless of the source  $P_\theta$  in effect. The optimal performance theoretically achievable for  $P_\theta$  is of course measured differently for noiseless coding, fixed-rate coding, and variable-rate coding. Let us take each case in turn.

For noiseless coding, the optimal performance achievable by a blocklength- $n$  code  $C^n$  on a source  $P_\theta$  (the  $n$ th-order OPTA) is by definition

$$\hat{H}_\theta^n = \inf_{C^n} \frac{1}{n} R_\theta(C^n)$$

where the infimum is over all noiseless codes with blocklength  $n$ . It is well known [23, Theorem 3.3.2] that

$$\frac{1}{n} H_\theta(X^n) \leq \hat{H}_\theta^n \leq \frac{1}{n} H_\theta(X^n) + \frac{1}{n}$$

so that the optimal performance theoretically achievable (by any  $n$ ) is

$$\hat{H}_\theta = \inf_n \hat{H}_\theta^n = \inf_n \frac{1}{n} H_\theta(X^n) = \lim_{n \rightarrow \infty} \frac{1}{n} H_\theta(X^n) = \bar{H}_\theta$$

the entropy rate of  $P_\theta$ . We say that  $\{C^n\}$  is a *weakly minimax* universal sequence of noiseless codes if for each  $\theta \in \Lambda$  the per-letter redundancy<sup>3</sup>

$$\rho_\theta(C^n) = \frac{1}{n} R_\theta(C^n) - \bar{H}_\theta \quad (1)$$

(which is nonnegative) goes to zero as  $n \rightarrow \infty$ ;  $\{C^n\}$  is *strongly minimax* universal if the convergence is uniform in  $\theta$ ; and  $\{C^n\}$  is *weighted* universal if the convergence occurs in  $L^1(W)$ , i.e., if

$$E|\rho(C^n|\Theta)| = E\rho(C^n|\Theta) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

For fixed-rate quantization, the optimal performance achievable by a blocklength- $n$  code  $C^n$  on a source  $P_\theta$  (the  $n$ th-order OPTA) is the “operational  $n$ th-order distortion-rate function” (for fixed-rate quantizers)

$$\hat{D}_\theta^n(R) = \inf_{C^n} \left\{ \frac{1}{n} D_\theta(C^n) : \frac{1}{n} R(C^n) \leq R \right\}$$

where the infimum is over all fixed-rate quantizers with blocklength  $n$  and rate at most  $R$ . Thus the optimal performance theoretically achievable (for any  $n$ ) is the “operational distortion-rate function” (for fixed-rate quantizers)

$$\hat{D}_\theta(R) = \inf_n \hat{D}_\theta^n(R) = \lim_{n \rightarrow \infty} \hat{D}_\theta^n(R)$$

[25], [30]. It is well known (by the source-coding theorem and its converse [4, Theorems 7.2.4 and 7.2.5] that  $\hat{D}_\theta(R)$  equals the distortion-rate function  $D_\theta(R)$ , provided  $P_\theta$  is ergodic. If  $P_\theta$  is not ergodic, then one must consider its ergodic

<sup>3</sup>Davission [15] defines the redundancy of a universal noiseless code as  $n^{-1}[R_\theta(C^n) - H_\theta(X^n)]$ . We call this the  $n$ th-order redundancy  $\rho_\theta^n(C^n)$ , as in (8).

components  $\{P_x\}$  and their associated distortion-rate functions  $\{D_x(R)\}$ .<sup>4</sup> Gray, Davission, and Kieffer [26], [30], [25] have shown that the operational distortion-rate function  $\hat{D}_\theta(R)$  has the ergodic decomposition

$$\begin{aligned} \hat{D}_\theta(R) &= \int \hat{D}_x(R) dP_\theta(x) \\ &= \int D_x(R) dP_\theta(x) \\ &\triangleq \bar{D}_\theta(R) \end{aligned} \quad (2)$$

provided  $\mathcal{X}$  is a complete separable metric (Polish) space,  $d(x, \hat{x})$  is continuous in  $x$  for each  $\hat{x}$ , and there exists a “reference letter”  $a^* \in \mathcal{X}$  such that  $E_\theta d(X_1, a^*) < \infty$ . Now if  $\{C^n\}$  is a sequence of fixed-rate block quantizers with  $n^{-1}R(C^n) \leq R$ , we say that  $\{C^n\}$  is a *weakly minimax* universal sequence of fixed-rate quantizers at rate  $R$  if for each  $\theta \in \Lambda$  the per-letter “distortion redundancy”

$$\delta_\theta(C^n) = \frac{1}{n} D_\theta(C^n) - \bar{D}_\theta(R) \quad (3)$$

(which is nonnegative) goes to zero as  $n \rightarrow \infty$ ;  $\{C^n\}$  is *strongly minimax* universal if the convergence is uniform in  $\theta$ ; and  $\{C^n\}$  is *weighted* universal if the convergence occurs in  $L^1(W)$ , i.e., if<sup>5</sup>

$$E|\delta(C^n|\Theta)| = E\delta(C^n|\Theta) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

For variable-rate quantization, performance is more easily expressed in terms of Lagrangians than in terms of distortion subject to a rate constraint [12]. For any variable-rate quantizer  $C^n$ , define the (instantaneous) “Lagrangian performance” of  $C^n$  (at Lagrange multiplier  $\lambda > 0$ ) as

$$l(x^n, \lambda) = d(x^n, C(x^n)) + \lambda r(x^n) \quad (4)$$

and denote its expected performance (with respect to  $P_\theta$ ) as

$$L_\theta(C^n, \lambda) = E_\theta l(X^n, \lambda) = D_\theta(C^n) + \lambda R_\theta(C^n). \quad (5)$$

$L_\theta(C^n, \lambda)$  can be interpreted as the  $y$ -intercept of the line with slope  $-\lambda$  passing through the point  $(R_\theta(C^n), D_\theta(C^n))$  in the distortion-rate plane, as in Fig. 1. Thus the optimal performance achievable by a blocklength- $n$  code  $C^n$  on a

<sup>4</sup>For each infinite sequence  $x = (x_1, x_2, \dots)$ , the ergodic component  $P_x$  is the stationary measure induced by the limiting relative frequencies

$$P_x(F) = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=0}^{n-1} 1_F(x_{i+1}, x_{i+2}, \dots)$$

of events  $F$  in a countable generating field for the  $\sigma$ -field on which  $P$  and each  $P_\theta$  are defined. Note that  $P_x$  depends on the sequence  $x$  and not on  $P$  or  $P_\theta$ . Hence the ergodic decomposition is the same for each  $P_\theta$  as it is for  $P$ .

<sup>5</sup>Following Kieffer [31], we here require  $n^{-1}R(C^n) \leq R$ . It is possible to relax this restriction, and require only  $n^{-1}R(C^n) \rightarrow R$ , as in Neuhoff, Gray, and Davission [39]. However, in that case the distortion redundancy may no longer be positive, and in general  $E|\delta(C^n|\Theta)| \neq E\delta(C^n|\Theta)$ . Nevertheless, it is possible to show in that case that  $E|\delta(C^n|\Theta)| \rightarrow 0$  if and only if  $E\delta(C^n|\Theta) \rightarrow 0$  (see the Appendix, Lemma 1). Later in this paper we shall consider fixed-rate universal codes with  $n^{-1}R(C^n)$  converging to  $R$  from above.

source  $P_\theta$  (the  $n$ th-order OPTA) is the “operational  $n$ th-order distortion-rate Lagrangian”

$$\hat{L}_\theta^n(\lambda) = \inf_{C^n} \frac{1}{n} L_\theta(C^n, \lambda)$$

where the infimum is over all variable-rate quantizers with blocklength  $n$ , and the optimal performance theoretically achievable (for any  $n$ ) is the “operational distortion-rate Lagrangian”

$$\hat{L}_\theta(\lambda) = \inf_n \hat{L}_\theta^n(\lambda) = \lim_{n \rightarrow \infty} \hat{L}_\theta^n(\lambda)$$

[22]. Similarly, let the “distortion-rate Lagrangian”

$$L_\theta(\lambda) = \min_R [D_\theta(R) + \lambda R]$$

be the minimum possible  $y$ -intercept of a line with slope  $-\lambda$  passing through some point  $(R, D_\theta(R))$  on the graph of  $D_\theta(R)$ , as in Fig. 1. If  $P_\theta$  is ergodic, then clearly  $\hat{L}_\theta(\lambda) = L_\theta(\lambda)$ , since  $L_\theta(\lambda)$  can be approached arbitrarily closely by the normalized Lagrangian performance  $n^{-1}L_\theta(C^n, \lambda)$  of some blocklength- $n$  variable-rate quantizer  $C^n$ . (This is proved formally in the Appendix, Lemma 2.) If  $P_\theta$  is nonergodic, then consider its ergodic components  $\{P_x\}$  and their associated distortion-rate Lagrangians  $\{L_x(\lambda)\}$ . Effros, Chou, and Gray [22] have shown that the operational distortion-rate Lagrangian  $\hat{L}_\theta(\lambda)$  has the ergodic decomposition

$$\begin{aligned} \hat{L}_\theta(\lambda) &= \int \hat{L}_x(\lambda) dP_\theta(x) \\ &= \int L_x(\lambda) dP_\theta(x) \\ &= L_\theta(\lambda) \end{aligned} \quad (6)$$

which in turn equals the distortion-rate Lagrangian  $L_\theta(\lambda)$ , under the same conditions as the ergodic decomposition (2). Thus we say that  $\{C^n\}$  is a *weakly minimax* universal sequence of variable-rate quantizers if for each  $\theta \in \Lambda$  the per-letter “Lagrangian redundancy”

$$\ell_\theta(C^n, \lambda) = \frac{1}{n} L_\theta(C^n, \lambda) - L_\theta(\lambda) \quad (7)$$

(which is nonnegative) goes to zero as  $n \rightarrow \infty$ ;  $\{C^n\}$  is *strongly minimax* universal if the convergence is uniform in  $\theta$ ; and  $\{C^n\}$  is *weighted* universal if the convergence occurs in  $L^1(W)$ , i.e., if

$$E|\ell(C^n, \lambda|\Theta)| = E\ell(C^n, \lambda|\Theta) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Consider now the relationships between strong, weak, and weighted universal code sequences. These relationships do not depend in any essential way on whether we are discussing noiseless codes, fixed-rate quantizers, or variable-rate quantizers. If  $\{C^n\}$  is strongly minimax universal, then of course  $\{C^n\}$  is also weakly minimax universal. Furthermore, if  $\{C^n\}$  is weakly minimax universal, then there exists a code sequence  $\{C_w^n\}$  that is weighted universal, assuming a mild regularity condition. The regularity condition for noiseless coding is  $H(X_1) < \infty$ ; for quantization it is  $Ed(X_1, a^*) < \infty$  for some  $a^* \in \mathcal{X}$  (see Appendix, Lemma 3). This result was shown by Davisson for finite-alphabet noiseless codes (for which the

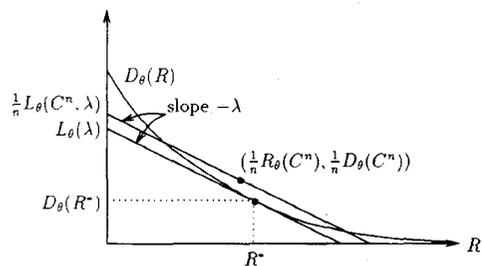


Fig. 1. Geometric interpretation of Lagrangians.

regularity condition always holds) [15, Theorem 1] and by Neuhoff, Gray, and Davisson for fixed-rate quantizers [39, Theorem 4.2]. Our Lemma 3 extends the approach implicit in [39] to variable-rate quantization and infinite-alphabet noiseless coding. Thus in some sense weighted universal coding is the weakest type of universal coding.

On the other hand, if  $\{C_w^n\}$  is weighted universal, so that its performance converges in  $L^1(W)$  to the optimal performance, then it also converges in probability to the optimal performance, and hence there is a subsequence  $\{C_w^{n_j}\}$  whose performance converges ( $W$ ) almost surely to the optimal performance [50, Proposition 5.17]. However, following [26, Lemma C.1], one can always construct a full sequence  $\{C^n\}$  from  $\{C_w^{n_j}\}$  such that  $C^n = C_w^{n_j}$  whenever  $n = n_j$  for some  $j$ , and such that the performance of  $\{C^n\}$  converges ( $W$ ) almost surely to the optimal performance. Hence if  $\{C_w^n\}$  is weighted universal, then there exists a code sequence  $\{C^n\}$  that is (almost) weakly minimax universal. Furthermore, since the performance of  $\{C^n\}$  converges ( $W$ ) almost everywhere on  $\Lambda$ , then for any  $\epsilon > 0$  there exists a subset  $\Lambda_0 \subseteq \Lambda$  such that  $W(\Lambda_0) > 1 - \epsilon$  and the convergence is uniform on  $\Lambda_0$ . Thus in a practical sense, there is not that much difference between the three types of universality. In this paper, we concentrate on the “weakest” type, weighted universal coding, although we also present results for weakly minimax universal coding.

Weighted universal codes exist under quite general conditions. The following is proved in the Appendix.

*Theorem 1:* Suppose  $\{X_i\}$  is stationary and ergodic for each  $\theta$ . Weighted universal codes exist for  $(\{X_i\}, \Theta)$  if and only if the OPTA has an ergodic decomposition.

The entropy rate has an ergodic decomposition if  $H(X_1) < \infty$  [29], while the distortion-rate function and its Lagrangian have ergodic decompositions if  $\mathcal{X}$  is Polish,  $d(x, \hat{x})$  is continuous in  $x$  for each  $\hat{x}$ , and there exists a reference letter  $a^*$  such that  $Ed(X_1, a^*) < \infty$  [22], [25]. These are the same or similar to conditions under which it has been possible to show directly that weighted universal codes exist [15, Theorem 8], [39, Theorem 4.1]. In principle, weighted universal codes are easy to construct:  $\{C^n\}$  is weighted universal if for each  $n$ ,  $C^n$  minimizes the expectation of the redundancy (1), (3), or (7). The  $C^n$  that minimizes the expectation of (1), for example, is the Huffman code matched to the mixture distribution  $P^n$ .

Conditions for the existence of weakly minimax universal codes are only slightly more stringent. For example, weakly minimax universal noiseless codes exist if there exists a

probability mass function  $p(x_1)$  such that for all  $\theta \in \Lambda$ ,  $-E_\theta \log p(X_1) \leq \infty$  [15, Theorem 8]. Nevertheless, weakly minimax universal codes do not exist for important classes of sources, for instance, the class of all finite-entropy memoryless sources over a countably infinite alphabet [27], [31].<sup>6</sup> Kieffer gives conditions under which weakly minimax universal fixed- and variable-rate quantizers exist [31]. For example, weakly minimax universal fixed-rate quantizers exist if  $d(x, \hat{x})$  is a metric on  $\mathcal{X} \cup \hat{\mathcal{X}}$ , and  $\mathcal{X}$  is separable.

In the present paper, the existence of weakly minimax codes (or weighted universal codes, as appropriate) is assumed throughout. In particular, throughout the paper we assume in the noiseless case that there exists a probability mass function  $p(x_1)$  such that  $-E_\theta \log p(X_1)$  is finite for all  $\theta \in \Lambda$  (or integrable with respect to  $W$  as appropriate), and we assume in the quantization case that  $\mathcal{X}$  is Polish,  $d(x, \hat{x})$  is continuous in  $x$  for each  $\hat{x}$ , and there exists a reference letter  $a^*$  such that  $E_\theta d(X_1, a^*)$  is finite for all  $\theta \in \Lambda$  (or integrable with respect to  $W$ , as appropriate). At the very least, these conditions ensure that the necessary OPTA's are finite.

Once one establishes that universal codes exist and that the redundancy of a code (per letter) converges to zero in an appropriate sense, it is natural to ask next about the optimal rate of convergence. The redundancy consists of two terms: the difference between the performance of  $C^n$  and the  $n$ th-order OPTA, and the difference between the  $n$ th-order OPTA and the OPTA. Both terms are nonnegative. We name the first term the  *$n$ th-order redundancy*

$$\rho_\theta^n(C^n) = \frac{1}{n} R_\theta(C^n) - \hat{H}_\theta^n \quad (8)$$

$$\delta_\theta^n(C^n) = \frac{1}{n} D_\theta(C^n) - \hat{D}_\theta^n(R) \quad (9)$$

or

$$\ell_\theta^n(C^n, \lambda) = \frac{1}{n} L_\theta(C^n, \lambda) - \hat{L}_\theta^n(\lambda). \quad (10)$$

The  $n$ th-order redundancy converges to zero if and only if the “infinite order” redundancy does, since the  $n$ th-order OPTA converges to the OPTA for each  $\theta \in \Lambda$ . Hence, when determining the existence of universal codes, we can consider either the  $n$ th-order redundancy or the infinite-order redundancy.<sup>7</sup> However, when determining the rate of convergence, we prefer to consider the  $n$ th-order redundancy over the infinite-order redundancy. The rate of convergence of the  $n$ th-order OPTA to the OPTA is best left as a separate issue of interest in its own right [36], [40], [41].

In the case of noiseless coding, Rissanen and others have provided the optimal rate of convergence of the  $n$ th-order rate redundancy to zero, when  $\Lambda$  is a subset of  $\mathbb{R}^k$  [15], [16], [33],

<sup>6</sup>On the other hand, we show in the Appendix, Lemma 10 that weighted universal noiseless codes exist and moreover have expected redundancy converging to zero as  $O(n^{-b/(b+1)})$ , for the class of all memoryless sources over a countably infinite alphabet for which the probability mass function is  $O(n^{-(1+b)})$ , for some  $b > 0$ .

<sup>7</sup>When determining the existence of *strongly* minimax universal codes, however, the  $n$ th-order and infinite-order redundancy may not be equivalent. In the noiseless case, for example, while  $\rho_\theta(C^n) \rightarrow 0$  uniformly implies  $\rho_\theta^n(C^n) \rightarrow 0$  uniformly, the converse is not necessarily true, unless  $n^{-1}H_\theta(X^n) \rightarrow \bar{H}_\theta$  uniformly. Kieffer [32] gives some conditions under which this uniformity holds.

[47], [48]. The following theorem is a slight generalization of [47, Theorem 1b], which we shall call Rissanen’s achievability theorem.

*Theorem 2 (Rissanen):* Let  $\Lambda$  be an open subset of  $\mathbb{R}^k$ . Suppose that for each  $\theta \in \Lambda$  the  $n$ th-order relative entropies  $D_n(\theta||\hat{\theta})$  as functions of  $\hat{\theta} \in \Lambda$  are twice continuously differentiable, and their normalized second derivatives  $D_n''(\theta||\hat{\theta})/n$  as functions of  $\hat{\theta} \in \Lambda$  are uniformly bounded in  $n$  on some open neighborhood of  $\theta$ . Then there exists a weakly minimax universal noiseless code  $\{C^n\}$  with

$$\rho_\theta^n(C^n) \leq \frac{k \log n + c_\theta}{2n}.$$

In [47, Theorem 1a], Rissanen also shows the converse is true (subject to a technical condition on the  $P_\theta$ ’s): if  $\Lambda$  is compact in  $\mathbb{R}^k$ , then for any sequence of noiseless codes  $\{C^n\}$  and for all  $\theta$  (except in a “bad” set  $B \subseteq \Lambda$  having arbitrarily small Lebesgue measure)

$$\rho_\theta^n(C^n) \geq \frac{k \log n + O(\log n)}{2n}.$$

Rissanen later strengthens this result to hold for almost all  $\theta$  (Lebesgue) [48]. Thus  $(k/2)(\log n/n)$  is essentially the minimum possible per-letter redundancy for finitely parametrized sources.

In the next section we extend Theorem 2 to include fixed-rate and variable-rate quantizers as well as noiseless codes. We also obtain achievable rate-of-convergence results for noiseless codes, fixed-rate quantizers, and variable-rate quantizers when  $\Lambda$  is not necessarily finitely parametrized. We do not present converse rate-of-convergence results, but conjecture that our achievable rates of convergence are optimal.

To facilitate the development in the next section, we collapse noiseless coding, fixed-rate quantization, and variable-rate quantization into a single case by considering only the  $n$ th-order Lagrangian redundancy

$$\begin{aligned} \ell_\theta^n(C^n, \lambda) &= \frac{1}{n} L_\theta(C^n, \lambda) - \hat{L}_\theta^n(\lambda) \\ &= \frac{1}{n} [D_\theta(C^n) + \lambda R_\theta(C^n)] - \hat{L}_\theta^n(\lambda) \end{aligned} \quad (11)$$

where

$$\hat{L}_\theta^n(\lambda) = \inf_{C^n \in \mathcal{C}^n} \frac{1}{n} E_\theta [d(X^n, \beta(\alpha(X^n))) + \lambda |\alpha(X^n)|] \quad (12)$$

that is,  $\hat{L}_\theta^n(\lambda)$  is the optimal Lagrangian performance achievable by a code  $C^n = \beta \circ \alpha \in \mathcal{C}^n$ , where  $\mathcal{C}^n$  is the set of all blocklength- $n$  noiseless codes, fixed-rate quantizers at rate  $R$  bits per letter, or variable-rate quantizers, as appropriate. Clearly, (11) is identical to (10) in the case of variable-rate quantization. In the case of noiseless coding, for which  $d \equiv 0$ , (11) reduces to (8) when  $\lambda = 1$ . Finally, in the case of fixed-rate quantization, (11) reduces to

$$\begin{aligned} \ell_\theta^n(C^n, \lambda) &= \left( \frac{1}{n} D_\theta(C^n) - \hat{D}_\theta^n(R) \right) + \lambda \left( \frac{1}{n} R(C^n) - R \right) \\ &= \delta_\theta^n(C^n) + \lambda \rho(C^n) \end{aligned}$$

where  $\rho(C^n) = n^{-1}R(C^n) - R$  can be considered the rate redundancy for a fixed-rate code. This reduces to the

distortion redundancy in the following sense (see the Appendix, Lemma 4). (Recall that the distortion redundancy of a fixed-rate code  $C^n$  can be measured fairly only when  $n^{-1}R(C^n) \leq R$ , or  $\rho(C^n) \leq 0$ .) Suppose there exists a sequence of fixed rate codes  $C^n$  with  $\rho(C^n) \geq 0$  such that  $\ell_\theta(C^n, \lambda) = \delta_\theta(C^n) + \lambda\rho(C^n)$  is upper-bounded by a sequence, say  $f_\theta^n(\lambda)$ , converging to zero as  $n \rightarrow \infty$ . Then there exists a sequence of fixed-rate codes  $C^n$  with  $\rho(C^n) \leq 0$  such that  $\delta_\theta(C^n)$  by itself is upper-bounded by  $f_\theta^n(\lambda)$ , for any  $\lambda > -(d/dR)\bar{D}_\theta(R)$ . Thus an upper bound on the achievable infinite-order Lagrangian redundancy is also an upper bound on the achievable infinite-order distortion redundancy (although this is not necessarily true for  $n$ th-order redundancies). This justifies the use of the Lagrangian redundancy for fixed-rate as well as variable-rate codes.

### III. THE VECTOR QUANTIZATION APPROACH

Consider the Rice machine. The Rice machine is a two-stage noiseless code that encodes, in a first stage, the identity of a block noiseless code with blocklength  $l$ , and then encodes, in the second stage, the data  $x^n$  using the identified noiseless code  $n/l$  times. The Rice machine, with effective blocklength  $n$ , is little more complex than a code with blocklength  $l$ . This practical two-stage coding strategy can also be applied to quantization, using a block quantizer instead of a block noiseless code [9], [55], [56]. For example, a block truncation code [17], [43] can be regarded as a two-stage quantizer with  $l = 1$ .<sup>8</sup>

Mathematically, a two-stage code can be described as follows. Let  $\tilde{\alpha}: \mathcal{X}^n \rightarrow \tilde{\mathcal{S}}$  be a mapping from  $\mathcal{X}^n$  into a finite or countable prefix code  $\tilde{\mathcal{S}}$ , and let  $\{C_{\tilde{s}}^n\}$  be a collection of block codes  $C_{\tilde{s}}^n = \beta_{\tilde{s}} \circ \alpha_{\tilde{s}}$  indexed by  $\tilde{s} \in \tilde{\mathcal{S}}$ . Then a block code  $C^n = \beta \circ \alpha$  is a *two-stage code* if its encoder  $\alpha$  and decoder  $\beta$  have the following form:

$$\alpha(x^n) = \tilde{\alpha}(x^n)\alpha_{\tilde{\alpha}(x^n)}(x^n)$$

and

$$\beta(\tilde{\alpha}(x^n)\alpha_{\tilde{\alpha}(x^n)}(x^n)) = \beta_{\tilde{\alpha}(x^n)}(\alpha_{\tilde{\alpha}(x^n)}(x^n)).$$

That is,  $C_n$  is a two-stage code if it codes each  $x^n$  by first coding (in a first stage) the identity of a block code and then coding (in a second stage) the data  $x^n$  using the identified code.

Two-stage codes have been historically useful for proving theorems on universal coding. If  $\{C_{\tilde{s}}^n\}$  is a collection of codes, each one good for a different source, then under the right conditions, the two-stage code  $C^n$  (or rather a sequence of such two-stage codes) can be shown to be universal. Two-stage codes have also been useful in practice under various guises (such as block truncation codes). One of the goals of this paper is to bring the theory and practice of two-stage coding together.

In this paper we focus on the case where the codes  $\{C_{\tilde{s}}^n\}$  are product codes. For simplicity, we assume that each  $C_{\tilde{s}}^n$

<sup>8</sup>In its standard form, a block truncation code encodes in a first stage a "low value" and a "high value," and encodes in a second stage the data using one bit per symbol to specify whether each symbol should be reproduced as the low or high value.

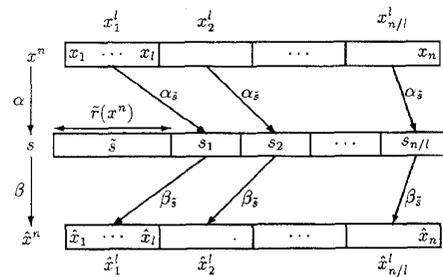


Fig. 2. Two-stage encoding and decoding.

is the product of  $n/l$   $l$ -dimensional block codes  $C_{\tilde{s}}^l$ , where  $l$  evenly divides  $n$ . This includes the theoretically important special case where (as  $n$  grows)  $l = n$  and also the practically important special case where  $l$  is fixed. We denote the resulting two-stage code  $C^n$  as  $C^{n,l}$ . We elect to suppress many of the structural details of the two-stage code in using this simple notation.

Fig. 2 illustrates the action of a two-stage code  $C^{n,l} = \beta \circ \alpha$  on a data sequence  $x^n$ . The encoder  $\alpha$  maps  $x^n$  into a binary description  $s$  with two stages. The first stage description, or header, is the string  $\tilde{s}$  describing a selected *second-stage code*  $C_{\tilde{s}}^l = \beta_{\tilde{s}} \circ \alpha_{\tilde{s}}$  having blocklength  $l$ . The second stage description, or body, is the concatenation of the strings  $s_i = \alpha_{\tilde{s}}(x_i^l)$ ,  $i = 1, \dots, n/l$ , that describe the sub-blocks  $\{x_i^l\}$  using the selected encoder  $\alpha_{\tilde{s}}$ . The decoder  $\beta$  maps the two-stage description  $s = \tilde{s}s_1s_2 \dots s_{n/l}$  into a reproduction sequence  $\hat{x}^n$  by concatenating the sub-blocks  $\hat{x}_i^l = \beta_{\tilde{s}}(s_i)$ ,  $i = 1, \dots, n/l$ , that reproduce the sub-blocks  $\{x_i^l\}$  using the selected decoder  $\beta_{\tilde{s}}$ .

Our approach to two-stage universal coding is based on the idea that the function that selects a second-stage code  $C_{\tilde{s}}^l$  for each input sequence  $x^n$  can be regarded as a kind of quantizer, which we shall call the *first-stage quantizer*  $\tilde{C}^{n,l}$ . Like any quantizer, the first-stage quantizer maps an input space to an output or reproduction space via a fixed-length or variable-length noiseless code, say  $\tilde{\mathcal{S}}$ , which we assume is binary. For a first-stage quantizer, the input space is the set  $\mathcal{X}^n$  of all data vectors of length  $n$ , and the output space is the set  $\mathcal{C}^l$  of all block codes of length  $l$ . (As appropriate,  $\mathcal{C}^l$  is a space of noiseless codes, fixed-rate quantizers at rate  $R$  bits per letter, or variable-rate quantizers.) The first-stage quantizer  $\tilde{C}^{n,l}: \mathcal{X}^n \rightarrow \mathcal{C}^l$  is thus composed of an encoder  $\tilde{\alpha}: \mathcal{X}^n \rightarrow \tilde{\mathcal{S}}$  and a decoder  $\tilde{\beta}: \tilde{\mathcal{S}} \rightarrow \mathcal{C}^l$ . The string  $\tilde{s} = \tilde{\alpha}(x^n) \in \tilde{\mathcal{S}}$  is precisely the first-stage description of  $x^n$ , and its length  $|\tilde{s}|$  is the first-stage description length. The reproduction  $C_{\tilde{s}}^l = \tilde{\beta}(\tilde{s}) \in \mathcal{C}^l$  is precisely the code  $C_{\tilde{s}}^l = \beta_{\tilde{s}} \circ \alpha_{\tilde{s}}$  used to encode and decode the length- $l$  sub-blocks  $x_1^l, \dots, x_{n/l}^l$  of  $x^n$ , in the second stage. As with any quantizer, the collection of reproductions  $\Gamma^l = \{\tilde{\beta}(\tilde{s}): \tilde{s} \in \tilde{\mathcal{S}}\}$  is known as the reproduction codebook. The codebook of a first-stage quantizer is simply a collection of codes.

Fig. 3 illustrates the input and output spaces of the first-stage quantizer. The input space  $\mathcal{X}^n$  is partitioned into encoding regions. The output space  $\mathcal{C}^l$  contains a finite or countable collection  $\Gamma^l$  of codewords  $C_{\tilde{s}}^l$  (actually, codes), one codeword

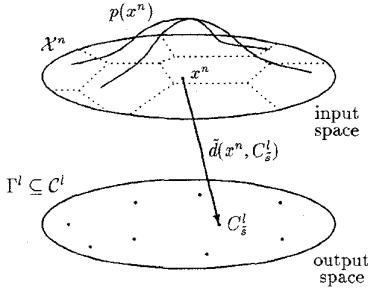


Fig. 3. First-stage quantizer.

per encoding region. Within an encoding region, every input vector  $x^n$  maps to the code  $C_s^l$  for that encoding region, via the string  $\tilde{s} = \tilde{\alpha}(x^n)$ .

Like any quantizer, the first-stage quantizer  $\tilde{C}^{n,l}$  has associated notions of rate and distortion. The instantaneous rate  $\tilde{r}(x^n)$  of  $\tilde{C}^{n,l}$  is defined as the length of the first-stage description,

$$\tilde{r}(x^n) = |\tilde{\alpha}(x^n)| \quad (13)$$

while the instantaneous distortion  $\tilde{d}(x^n, \tilde{\beta}(\tilde{s}))$  incurred by reproducing  $x^n$  as a code  $\tilde{\beta}(\tilde{s}) = C_s^l = \beta_s \circ \alpha_s$  is defined

$$\tilde{d}(x^n, \tilde{\beta}(\tilde{s})) = \frac{1}{n} \sum_{i=1}^{n/l} [d(x_i^l, C_s^l(x_i^l)) + \lambda |\alpha_s(x_i^l)|] \quad (14)$$

the average Lagrangian performance of  $C_s^l$  on the sub-blocks of  $x^n$ , per letter. If  $C_s^l$  is a noiseless code, then  $d(x_i^l, C_s^l(x_i^l)) \equiv 0$ , so that (14) equals  $\lambda$  times the length of the second-stage description of  $x^n$ , per letter. If  $C_s^l$  is a fixed-rate quantizer, then  $|\alpha_s(x_i^l)| \equiv lR$ , so that (14) equals a constant plus the distortion of  $C_s^l$  on the sub-blocks of  $x^n$ , per letter. Thus  $\tilde{r}$  reflects the first-stage encoding of  $x^n$ , and  $\tilde{d}$  reflects the second stage.

Now the instantaneous Lagrangian performance of  $\tilde{C}^{n,l}$  at Lagrange multiplier  $\lambda/n$  can be defined using (13) and (14)

$$\tilde{l}(x^n, \lambda/n) = \tilde{d}(x^n, \tilde{\beta}(\tilde{\alpha}(x^n))) + \frac{\lambda}{n} |\tilde{\alpha}(x^n)| \quad (15)$$

$$= \frac{1}{n} \left[ \left( \sum_{i=1}^{n/l} d(x_i^l, C_{\tilde{\alpha}(x^n)}^l(x_i^l)) \right) + \lambda \left( |\tilde{\alpha}(x^n)| + \sum_{i=1}^{n/l} |\alpha_{\tilde{\alpha}(x^n)}(x_i^l)| \right) \right]. \quad (16)$$

The terms in parentheses

$$d(x^n, C^{n,l}(x^n)) = \left( \sum_{i=1}^{n/l} d(x_i^l, C_{\tilde{\alpha}(x^n)}^l(x_i^l)) \right) \quad (17)$$

and

$$r(x^n) = \left( |\tilde{\alpha}(x^n)| + \sum_{i=1}^{n/l} |\alpha_{\tilde{\alpha}(x^n)}(x_i^l)| \right) \quad (18)$$

are, respectively, the total distortion on  $x^n$  and the total description length of  $x^n$  by the two-stage code  $C^{n,l}$ , for which,

as defined in (4), the instantaneous Lagrangian is

$$l(x^n, \lambda) = d(x^n, C^{n,l}(x^n)) + \lambda r(x^n). \quad (19)$$

Combining (16)–(19), we obtain

$$\tilde{l}(x^n, \lambda/n) = \frac{1}{n} l(x^n, \lambda) \quad (20)$$

the instantaneous Lagrangian performance of  $C^{n,l}$  per letter. If  $C^{n,l}$  is a noiseless code, then  $d(x^n, C^{n,l}(x^n)) \equiv 0$ , so that (20) equals  $\lambda$  times the two-stage description length for  $x^n$ , per letter. If  $C^{n,l}$  is a fixed-rate quantizer, then  $r(x^n)$  is a constant, so that (20) equals a constant plus the distortion of  $C^{n,l}$  on  $x^n$ , per letter. Thus  $\tilde{l}$  reflects both the first- and second-stage coding of  $x^n$ . Now using (20), the expected Lagrangian performance of  $\tilde{C}^{n,l}$  at Lagrange multiplier  $\lambda/n$  (with respect to  $P_\theta$ ) is

$$\begin{aligned} L_\theta(\tilde{C}^{n,l}, \lambda/n) &= E_\theta \tilde{l}(X^n, \lambda/n) = E_\theta \frac{1}{n} l(X^n, \lambda) \\ &= \frac{1}{n} L_\theta(C^{n,l}, \lambda). \end{aligned} \quad (21)$$

We can now express the  $n$ th-order Lagrangian redundancy (11) of a two-stage code  $C^{n,l}$  in terms of the expected Lagrangian performance of its first-stage quantizer  $\tilde{C}^{n,l}$ , using (21) and (15)

$$\begin{aligned} \ell_\theta^n(C^{n,l}, \lambda) &\triangleq \frac{1}{n} L_\theta(C^{n,l}, \lambda) - \hat{L}_\theta^n(\lambda) \\ &= L_\theta(\tilde{C}^{n,l}, \lambda/n) - \hat{L}_\theta^n(\lambda) \\ &= \int \left[ \tilde{d}(x^n, \tilde{\beta}(\tilde{\alpha}(x^n))) + \frac{\lambda}{n} |\tilde{\alpha}(x^n)| \right] dP_\theta(x^n) \\ &\quad - \hat{L}_\theta^n(\lambda). \end{aligned} \quad (22)$$

Because the first-stage encoder  $\tilde{\alpha}$  that minimizes (22) minimizes the integrand in (22) pointwise, we have the following:

*Proposition 1:* Given any two-stage code  $C^{n,l}$  with first-stage quantizer  $\tilde{C}^{n,l} = \tilde{\beta} \circ \tilde{\alpha}$ , there is a two-stage code  $C_*^{n,l}$  with first-stage quantizer  $\tilde{C}_*^{n,l} = \tilde{\beta} \circ \tilde{\alpha}_*$  such that for all  $\theta \in \Lambda$ ,  $\ell_\theta^n(C_*^{n,l}, \lambda) \leq \ell_\theta^n(C^{n,l}, \lambda)$ , where

$$\tilde{\alpha}_*(x^n) = \arg \min_{\tilde{s} \in \tilde{\mathcal{S}}} \left[ \tilde{d}(x^n, \tilde{\beta}(\tilde{s})) + \frac{\lambda}{n} |\tilde{s}| \right] \quad (23)$$

$$= \arg \min_{\tilde{s} \in \tilde{\mathcal{S}}} \left[ \left( \sum_{i=1}^{n/l} d(x_i^l, C_{\tilde{s}}^l(x_i^l)) \right) + \lambda \left( |\tilde{s}| + \sum_{i=1}^{n/l} |\alpha_{\tilde{s}}(x_i^l)| \right) \right] \quad (24)$$

for each  $x^n$ .

We shall call such a first-stage encoder  $\tilde{\alpha}_*$  *optimal* with respect to the first-stage decoder  $\tilde{\beta}$ .

Equation (24) generalizes the minimum description length (MDL) rule for noiseless codes [47], since in the noiseless case  $d \equiv 0$  and  $\tilde{\alpha}(x^n)$  selects the first-stage description  $\tilde{s}$  that minimizes the two-stage description length of  $x^n$ . A detail that must be checked is that the minimum in (24) can always be achieved even when  $\tilde{\mathcal{S}}$  is countably infinite. Barron and Cover

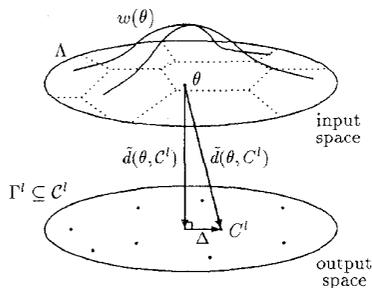


Fig. 4. Omniscient first-stage quantizer.

prove this for noiseless coding [3, Proposition 1]; we prove it for lossy coding as well (see the Appendix, Lemma 5).

Condition (23) corresponds to the biased nearest neighbor condition in entropy-constrained vector quantization [12, eq. (3)]. This analogy is carried further in Section IV, where a Lloyd-like algorithm is used to design a locally optimal weighted universal two-stage code  $C_*^{n,l}$  by iteratively optimizing  $\tilde{\alpha}$  for a fixed  $\tilde{\beta}$  and  $\tilde{\beta}$  for a fixed  $\tilde{\alpha}$ , given a mixture distribution  $P$ . Our aim in this section is simply to analyze the rate of convergence to zero of the redundancy of such optimal universal codes (both minimax and weighted). Toward that end we next find an upper bound for the redundancy (22) using a suboptimal first-stage encoder  $\tilde{\alpha}_o$ , which we call an omniscient first-stage encoder.

An *omniscient first-stage quantizer* is a first-stage quantizer  $\tilde{C}_o^l = \tilde{\beta} \circ \tilde{\alpha}_o$  whose encoder  $\tilde{\alpha}_o$  is omniscient, i.e., has direct access to  $\theta$  (rather than to  $x^n$ ). Thus its encoder is a map  $\tilde{\alpha}_o: \Lambda \rightarrow \tilde{\mathcal{S}}$  rather than a map  $\tilde{\alpha}: \mathcal{X}^n \rightarrow \tilde{\mathcal{S}}$ . Fig. 4 illustrates the input and output spaces of an omniscient first-stage quantizer. Regardless of how the map  $\tilde{\alpha}_o$  is defined, by (22) and the definition of  $\tilde{\alpha}_*$  (23)

$$\begin{aligned} \ell_\theta^n(C_*^{n,l}, \lambda) &= \int \left[ \tilde{d}(x^n, \tilde{\beta}(\tilde{\alpha}_*(x^n))) + \frac{\lambda}{n} |\tilde{\alpha}_*(x^n)| \right] \\ &\quad \cdot dP_\theta(x^n) - \hat{L}_\theta^n(\lambda) \\ &\leq \int \left[ \tilde{d}(x^n, \tilde{\beta}(\tilde{\alpha}_o(\theta))) + \frac{\lambda}{n} |\tilde{\alpha}_o(\theta)| \right] \\ &\quad \cdot dP_\theta(x^n) - \hat{L}_\theta^n(\lambda) \end{aligned} \quad (25)$$

so we have the following:

**Proposition 2:** Given any two-stage code  $C_o^{n,l}$  with omniscient first-stage quantizer  $\tilde{C}_o^l = \tilde{\beta} \circ \tilde{\alpha}_o$ , there is a two-stage code  $C_*^{n,l}$  with first-stage quantizer  $\tilde{C}_*^{n,l} = \tilde{\beta} \circ \tilde{\alpha}_*$  such that for all  $\theta \in \Lambda$

$$\ell_\theta^n(C_*^{n,l}, \lambda) \leq \ell_\theta^n(C_o^{n,l}, \lambda).$$

This proposition corresponds to [3, Proposition 4] of Barron and Cover, who call the redundancy  $\ell_\theta^n(C_o^{n,l}, \lambda)$  the *index of resolvability* in the noiseless case when  $\lambda = 1, l = 1$ , and  $P_\theta$  is independent and identically distributed (iid).

Armed with Proposition 2, we are free to focus on the redundancy of two-stage codes that have omniscient first-stage quantizers. The rate of convergence of the redundancy of  $C_o^{n,l}$  is an upper bound to the rate of convergence of the redundancy of  $C_*^{n,l}$ .

Every omniscient first-stage quantizer  $\tilde{C}_o^l$  has associated with it a rate and a distortion. The instantaneous rate  $\tilde{r}(\theta)$  of  $\tilde{C}_o^l$  is defined as usual as the length of the first-stage description,  $\tilde{r}(\theta) = |\tilde{\alpha}_o(\theta)|$ , while the instantaneous distortion  $\tilde{d}(\theta, C_s^l)$  incurred by reproducing  $\theta$  as a code  $C_s^l \in C^l$  is defined as the *expected value with respect to  $P_\theta$*  of the instantaneous distortion  $\tilde{d}(X^n, C_s^l)$

$$\begin{aligned} \tilde{d}(\theta, C_s^l) &= E_\theta \tilde{d}(X^n, C_s^l) \\ &= E_\theta \frac{1}{n} \sum_{i=1}^{n/l} [d(X_i^l, C_s^l(X_i^l)) + \lambda |\alpha_s(X_i^l)|] \\ &= \frac{1}{l} E_\theta [d(X^l, C_s(X^l)) + \lambda |\alpha_s(X^l)|]. \end{aligned} \quad (26)$$

The key quantity associated with omniscient first-stage quantizers, however, is the instantaneous *divergence*  $\Delta(\theta, C_s^l)$ , which is defined as the difference between the instantaneous distortion  $\tilde{d}(\theta, C_s^l)$  and the minimum possible instantaneous distortion

$$\tilde{d}(\theta, C^l) = \inf_{C^l \in C^l} \tilde{d}(\theta, C^l) \quad (27)$$

which upon inspection is the  $l$ th-order OPTA (12). That is

$$\begin{aligned} \Delta(\theta, C_s^l) &= \tilde{d}(\theta, C_s^l) - \tilde{d}(\theta, C^l) \\ &= \tilde{d}(\theta, C_s^l) - \hat{L}_\theta^l(\lambda) \end{aligned} \quad (28)$$

as depicted in Fig. 4. This divergence is equivalent to Kullback's information divergence [34] in the noiseless case. To see that, let  $C^l$  be the space of noiseless codes with blocklength  $l$ , and associate each noiseless code  $C_s^l = \beta_s \circ \alpha_s \in C^l$  with the probability density  $q_s(x^l) = 2^{-|\alpha_s(x^l)|}$ . (Here and in the sequel, we assume the codelengths  $|\alpha_s(x^l)|$  satisfy the Kraft inequality with equality, and in fact may be ideal, noninteger lengths.) Then

$$\tilde{d}(\theta, C_s^l) = (\lambda/l) E_\theta |\alpha_s(X^l)| = -(\lambda/l) E_\theta \log q_s(X^l)$$

and

$$\tilde{d}(\theta, C^l) = \lambda \hat{H}_\theta^l = (\lambda/l) H_\theta(X^l) = -(\lambda/l) E_\theta \log p_\theta(X^l)$$

so that

$$\Delta(\theta, C_s^l) = (\lambda/l) E_\theta \log (p_\theta(X^l)/q_s(X^l)) = (\lambda/l) D_l(p_\theta || q_s)$$

which is the information divergence or relative entropy. In the decision theory and machine learning literature,  $C^l$  is more generally the hypothesis space,  $\tilde{d}(X^n, C^l)$  is the empirical or average loss,  $\tilde{d}(\theta, C^l)$  is the expected loss, or risk, and  $\Delta(\theta, C^l)$  is the regret or divergence [10], [28].

Now the  $n$ th-order redundancy of a two-stage code  $C_o^{n,l}$  with omniscient first-stage quantizer  $\tilde{C}_o^l$  can be expressed using (25) and (26)

$$\begin{aligned} \ell_\theta^n(C_o^{n,l}, \lambda) &= E_\theta \left[ \tilde{d}(X^n, \tilde{\beta}(\tilde{\alpha}_o(\theta))) + \frac{\lambda}{n} |\tilde{\alpha}_o(\theta)| \right] - \hat{L}_\theta^n(\lambda) \\ &= \tilde{d}(\theta, \tilde{C}_o^l(\theta)) + \frac{\lambda}{n} |\tilde{\alpha}_o(\theta)| - \hat{L}_\theta^n(\lambda). \end{aligned} \quad (29)$$

To simplify (29) further, define the  $(n, l)$ th-order redundancy

$$\ell_\theta^{n,l}(C^m, \lambda) \triangleq \frac{1}{n} L_\theta(C^m, \lambda) - \hat{L}_\theta^n(\lambda) \quad (30)$$

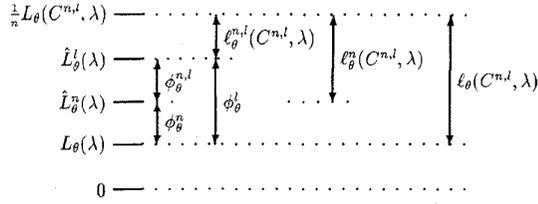


Fig. 5. Relationships between various redundancies.

which is the difference between the performance of a code with blocklength  $n$  and the  $l$ th-order OPTA (rather than the  $n$ th-order OPTA) (see Fig. 5). Then (29) becomes

$$\ell_\theta^n(C_o^{n,l}, \lambda) = \ell_\theta^{n,l}(C_o^{n,l}, \lambda) + \phi_\theta^{n,l}$$

where (using (28))

$$\begin{aligned} \ell_\theta^{n,l}(C_o^{n,l}, \lambda) &= \tilde{d}(\theta, \tilde{C}_o^l(\theta)) + \frac{\lambda}{n} |\tilde{\alpha}_o(\theta)| - \hat{L}_\theta^l(\lambda) \\ &= \Delta(\theta, \tilde{C}_o^l(\theta)) + \frac{\lambda}{n} |\tilde{\alpha}_o(\theta)| \end{aligned} \quad (31)$$

and

$$\phi_\theta^{n,l} = \hat{L}_\theta^l(\lambda) - \hat{L}_\theta^n(\lambda).$$

Clearly, if  $l = n$  then  $\phi_\theta^{n,l} = 0$  so that  $\ell_\theta^{n,l}(C_o^{n,l}, \lambda) = \ell_\theta^n(C_o^n, \lambda)$ , and if  $l < n$  then  $\phi_\theta^{n,l}$  is bounded above by

$$\phi_\theta^l = \hat{L}_\theta^l(\lambda) - \hat{L}_\theta^n(\lambda)$$

a quantity that can be independently evaluated.<sup>9</sup>

We now extend Rissanen's achievable redundancy result (Theorem 2) from noiseless codes to the general case. The trick is to identify each  $\theta \in \Lambda$  with its optimal code in  $C^l$ . We therefore assume that for each  $\theta \in \Lambda$  there exists a code  $C_\theta^l \in C^l$  with  $\tilde{d}(\theta, C_\theta^l) = \hat{L}_\theta^l(\lambda)$ ; that is, the  $l$ th-order OPTA (27) is achieved by  $C_\theta^l$ . Then we can naturally define the  $l$ th-order divergence between  $\theta, \hat{\theta} \in \Lambda$  as

$$\Delta_l(\theta || \hat{\theta}) = \Delta(\theta, C_{\hat{\theta}}^l) \quad (32)$$

which is the excess distortion (on average with respect to  $P_\theta$ ) if we code  $X^n$  with a code matched to  $\hat{\theta}$  instead of  $\theta$ . Note that in the case of noiseless coding with  $\lambda$  set to 1,  $\Delta_l(\theta || \hat{\theta}) = (1/l)D_l(p_\theta || p_{\hat{\theta}})$ . In that case, Rissanen's theorem requires the second derivative of  $\Delta_l(\theta || \hat{\theta})$  with respect to  $\hat{\theta}$  be bounded on some neighborhood of  $\theta$ . We replace this condition by the more general condition that  $\Delta_l(\theta || \hat{\theta})$  be locally quadratic.

<sup>9</sup>For iid finite alphabet sources, Pilc has shown that  $\phi_\theta^n = O(n^{-1} \log n)$  [40], [41], [58]. For iid real alphabet sources and the squared-error distortion measure, Linder, Lugosi, and Zeger have shown that  $\phi_\theta^n = O(\sqrt{n^{-1} \log n})$  [36]. In the latter case we conjecture that  $\phi_\theta^n = O(n^{-1} \log n)$  since the coefficient of quantization  $G_n$  [24] converges to its limit as  $O(n^{-1} \log n)$  [54].

**Theorem 3:** Let  $\Lambda$  be a subset of  $\mathbb{R}^k$  (bounded if we are considering fixed-rate coding but possibly unbounded otherwise). Suppose that for each  $\theta, l$  there exists a code  $C_\theta^l$  achieving the  $l$ th-order OPTA at  $\theta$ , and that the resulting  $l$ th-order divergence  $\Delta_l(\theta || \hat{\theta})$  is locally quadratic. More precisely, suppose that for each  $\theta, l$  there exists a neighborhood  $S_\theta^l$  of  $\theta$  and a constant  $m_\theta^l$  such that  $\Delta_l(\theta || \hat{\theta}) \leq m_\theta^l \|\theta - \hat{\theta}\|^2$  for all  $\hat{\theta} \in S_\theta^l$ . Then for all  $n$  and  $l$  dividing  $n$ , there exists a two-stage code  $C^{n,l}$  such that for all  $\theta \in \Lambda$

$$\ell_\theta^{n,l}(C^{n,l}, \lambda) \leq \lambda \frac{k \log n + c_\theta^l}{2n}. \quad (33)$$

In particular, if  $S_\theta^l$  and  $m_\theta^l$  do not depend on  $l$ , then neither does  $c_\theta^l = c_\theta$ , and (setting  $l = n$ ) there exists a weakly minimax universal sequence of codes  $\{C^n\}$  such that

$$\ell_\theta^n(C^n, \lambda) \leq \lambda(k/2)(\log n + c_\theta)/n.$$

*Proof:* The proof is by construction of a two-stage code  $C_o^{n,l}$  with an omniscient first-stage quantizer  $\tilde{C}_o^l = \tilde{\beta} \circ \tilde{\alpha}_o$ , such that  $C_o^{n,l}$  satisfies (33). Then by Proposition 2, there exists some realizable, nonomniscient code  $C^{n,l}$  also satisfying (33).

To construct  $C_o^{n,l}$ , partition  $\mathbb{R}^k$  into a grid of hypercubes  $A_{n1}, A_{n2}, \dots$  each of side  $1/\lceil n^{1/2} \rceil$ , such that for each  $n \geq 1$ , the partition  $\{A_{n1}, A_{n2}, \dots\}$  refines the partition  $\{A_{11}, A_{12}, \dots\}$ . For each hypercube  $A_n \in \{A_{n1}, A_{n2}, \dots\}$  that intersects  $\Lambda$ , choose a representative  $\hat{\theta}_n \in A_n \cap \Lambda$ . Further, represent each *unit* hypercube  $A_1 \in \{A_{11}, A_{12}, \dots\}$  that intersects  $\Lambda$  by a unique string  $\tilde{s}_1$  from a binary prefix code  $\tilde{S}_1$ , which can be fixed-length if  $\Lambda$  is bounded. Then if  $\theta \in A_n \subset A_1$ , define  $\tilde{\alpha}_o(\theta)$  to be the string  $\tilde{s}_1$  that specifies  $A_1$  followed by some string  $\tilde{s}'_n$  of fixed length that specifies  $A_n$  within  $A_1$ . The length of  $\tilde{s}'_n$  is  $\log \lceil n^{1/2} \rceil^k$  bits, so that the first-stage description length  $|\tilde{s}_1 \tilde{s}'_n|$  is

$$|\tilde{\alpha}_o(\theta)| = b_\theta + \frac{k}{2} \log n$$

bits, for some constant  $b_\theta$  possibly depending on  $\theta$ . Also, define  $\tilde{\beta}(\tilde{s}) = \hat{\theta}_n$  whenever  $\tilde{s} = \tilde{s}_1 \tilde{s}'_n$  specifies  $A_n \subset A_1$ . Now the two-stage code  $C_o^{n,l}$  with omniscient first-stage quantizer  $\tilde{C}_o^l = \tilde{\beta} \circ \tilde{\alpha}_o$  has  $(n, l)$ th-order redundancy (by (31) and (32))

$$\ell_\theta^{n,l}(C_o^{n,l}, \lambda) = \Delta_l(\theta || \hat{\theta}_n) + \frac{\lambda}{n} \left( b_\theta + \frac{k}{2} \log n \right)$$

and by assumption,  $\Delta_l(\theta || \hat{\theta}) \leq m_\theta^l \|\theta - \hat{\theta}\|^2$  for all  $\hat{\theta}$  in a neighborhood  $S_\theta^l$  of  $\theta$ . Since  $\hat{\theta}_n \rightarrow \theta$  with  $\|\theta - \hat{\theta}_n\|^2 \leq k/n$ , there exists a constant  $a_\theta^l$  such that  $\Delta_l(\theta || \hat{\theta}_n) \leq a_\theta^l k/n$  for all  $n$ . Thus the theorem is proved with  $c_\theta^l = 2a_\theta^l/\lambda + 2b_\theta/k$ .  $\square$

*Remark 1:* Theorem 3 reduces to Theorem 2 in the noiseless case with  $\lambda = 1$  and  $l = n$ . To see this, suppose the conditions of Theorem 2 are satisfied; that is, suppose  $(1/l)D_l(\theta || \hat{\theta}) = \Delta_l(\theta || \hat{\theta})$  is twice continuously differentiable as a function of  $\hat{\theta}$ , and  $\|\Delta_l''(\theta || \hat{\theta})\| \leq 2m_\theta$  for all  $\hat{\theta}$  in some open neighborhood  $S_\theta$  of  $\theta$ ; then by Taylor's theorem, there exists  $\bar{\theta}$  between  $\theta$  and  $\hat{\theta}$  such that

$$\begin{aligned} \Delta_l(\theta || \hat{\theta}) &= \frac{1}{2} (\theta - \hat{\theta})^t \Delta_l''(\theta || \bar{\theta}) (\theta - \hat{\theta}) \\ &\leq m_\theta \|\theta - \hat{\theta}\|^2 \end{aligned}$$

for all  $\hat{\theta} \in S_\theta$ , and hence the conditions of Theorem 3 are satisfied with  $S_\theta$  and  $m_\theta$  independent of  $l$ .

*Remark 2:* Our proof of Theorem 3 closely follows Rissanen's proof of Theorem 2, the essential difference being that we explicitly quantize the parameter space using a distortion measure induced by the problem at hand, while Rissanen implicitly quantizes the parameter space by truncating the maximum-likelihood estimate.

*Remark 3:* The reason that we wish to treat the case  $l < n$  is that in practice, for complexity reasons,  $l$  will usually be small, or constant, while  $n$  is large, or growing. If  $l$  remains constant (as in the "universal" Gold Washing scheme of [57], for example) we cannot hope to beat the  $l$ th-order OPTA, even if  $n$  goes to infinity. However, Theorem 3 ensures that at least the  $l$ th-order OPTA can be approached as  $\lambda(k/2)n^{-1} \log n$ .

*Remark 4:* The supremum of the constant  $c_\theta^l$  in the bound (33) can be tightened by quantizing the parameter space  $\Lambda$  not in cubes, but according to a minimax criterion. Specifically, the omniscient two-stage code  $C_\theta^{n,l}$  can partition  $\Lambda$  into encoding regions  $\{A_{n,i}\}$  with corresponding reproductions  $\{\hat{\theta}_{n,i}\}$  such that the maximum divergence  $\max_{x_i, \theta \in A_{n,i}} \Delta_l(\theta | \hat{\theta}_{n,i})$  is minimized. This is similar to the approaches taken in [2], [3], [13], [49] in the noiseless case, and leads to value for the constant that is related to the Fisher information or its equivalent in the lossy case.

*Remark 5:* Theorems 3 and 2 also hold for parameter spaces  $\Lambda$  that are a countable union  $\cup_k \Lambda_k$  of  $k$ -dimensional parameter spaces  $\Lambda_k \subseteq \mathbb{R}^k$  (or a finite union if we are considering fixed-rate coding). To be specific, if  $\Lambda = \cup_k \Lambda_k$ , then (33) holds for all  $k$  and all  $\theta \in \Lambda_k$ . This implies that it is not necessary to know the dimensionality of the parameter space *a priori* in order to get the right constant in front of the bound, since there exists a two-stage code  $C_\theta^{n,l}$  that achieves (33) regardless of what dimensionality  $\theta$  turns out to have. The proof of this extension changes the above proof by having the omniscient encoder prepend a code for  $k$  in front of its usual encoding.

*Remark 6:* Rissanen remarks in the noiseless case that there exist parametric sources satisfying the conditions of his theorem [47]. This is also true in the quantization case. As a simple example, consider any real-valued stationary random process  $X_1, X_2, \dots$  with unknown mean  $\mu$  and unknown standard deviation  $0 < \sigma < \infty$ , and let  $P_\theta, \theta = (\mu, \sigma) \in \Lambda$  be its process measure. Then as shown in the Appendix, Lemma 6, under the squared-error distortion measure, for all  $l, \theta$ , and  $\hat{\theta}$ ,  $\Delta_l(\theta | \hat{\theta}) \leq \|\theta - \hat{\theta}\|^2$ . Hence for any stationary source with unknown mean and variance, there exists a weakly minimax universal sequence of variable-rate quantizers (and also fixed-length quantizers if  $\Lambda$  is bounded) for which the  $n$ th-order Lagrangian redundancy is at most  $\lambda(k/2)(\log n + c_\theta)/n$ , where  $k = 2$ . This result can be generalized to any class of sources for which the mapping  $T_{\theta, \hat{\theta}}^l: C_\theta^l \mapsto C_{\hat{\theta}}^l$  satisfies  $\|I - T_{\theta, \hat{\theta}}^l\|^2 \leq m_\theta^l \|\theta - \hat{\theta}\|^2$ . The difficulty in the generalization is explicitly identifying  $T_{\theta, \hat{\theta}}^l$  for anything other than scale and translation families.

*Remark 7:* In the noiseless case, Rissanen's theorem applies immediately to iid finite alphabet sources with alphabet

size  $k + 1$ , for then the parameter space has dimension  $k$ . One might hope that our theorem might apply similarly, but in the lossy case, to iid finite alphabet sources with alphabet size  $k + 1$ . Unfortunately, in the lossy case, the divergence may not in general be bounded locally by a quadratic that is uniform in  $l$ . Hence our theorem cannot be applied directly to obtain a  $(k/2)n^{-1} \log n$  redundancy. However, using other techniques, Linder *et al.* [37] have shown that in the iid finite alphabet case, the distortion redundancy is indeed  $O(n^{-1} \log n)$ .

Although Rissanen has proved a converse theorem in the noiseless case, we have so far been unable to prove a converse theorem in the quantization case. Indicative of the problem is the following theorem, which shows that if  $\Lambda$  is countable, then the  $n^{-1} \log n$  rate can be easily beaten.

*Theorem 4:* Let  $\Lambda$  be finite if we are considering fixed-rate coding but possibly countably infinite otherwise. Suppose that for each  $\theta, l$  there exists a code  $C_\theta^l$  achieving the  $l$ th-order OPTA at  $\theta$ . Then for all  $n$  and  $l$  dividing  $n$ , there exists a two-stage code  $C_\theta^{n,l}$  such that for all  $\theta \in \Lambda$

$$\ell_\theta^{n,l}(C_\theta^{n,l}, \lambda) \leq \lambda \frac{b_\theta}{n}.$$

In particular, setting  $l = n$ , there exists a weakly minimax universal sequence of codes  $\{C_\theta^n\}$  such that  $\ell_\theta^n(C_\theta^n, \lambda) \leq \lambda b_\theta/n$ .

*Proof:* Since  $\Lambda$  is already finite or countable, there is no need to construct a finite or countable subset. Let  $\Gamma^l$  be the codebook of blocklength- $l$  codes matched to each  $\theta \in \Lambda$ , namely,  $\Gamma^l = \{C_\theta^l \in \mathcal{C}^l: \theta \in \Lambda\}$ , and let  $\tilde{\alpha}_\theta$  be the omniscient first-stage encoder that maps each  $\theta \in \Lambda$  to a string  $\tilde{s} \in \tilde{\mathcal{S}}$  describing  $C_\theta^l = C_\theta^l$ , using an arbitrary prefix code  $\tilde{\mathcal{S}}$ . The first-stage description length  $|\tilde{s}|$  is thus  $|\tilde{\alpha}_\theta(\theta)| = b_\theta$  bits, for some constant  $b_\theta$  not depending on  $l$ , and  $\Delta(\theta, \tilde{C}_\theta^l(\theta)) \equiv 0$ , so that by (31),  $\ell_\theta^{n,l}(C_\theta^{n,l}, \lambda) = \lambda b_\theta/n$ .  $\square$

Notice that Theorem 4 does not contradict Rissanen's converse for lossless coding, since the converse holds only for almost any  $\theta$  (Lebesgue) and therefore does not cover the case of  $\Lambda$  countable. Theorem 4 applies in the important case that the  $P_\theta$ 's are computable, i.e., describable by finite-length computer programs.<sup>10</sup> In this case,  $\Lambda$  is countable, and any computer program for  $P_\theta$  can be used as the first-stage description  $\tilde{\alpha}_\theta(\theta)$ . If we use the shortest computer program for each  $P_\theta$ , then  $|\tilde{\alpha}_\theta(\theta)|$  is the Kolmogorov complexity of  $P_\theta$ , although in that case computing  $\tilde{\alpha}_\theta(\theta)$  will be undecidable [14].

We now shift our focus from weakly minimax universal coding to weighted universal coding, for which we can obtain more detailed rate-of-convergence results based on the

<sup>10</sup>According to Barron and Cover [3], a probability distribution  $P$  is *computable* (relative to a countable collection of sets  $A_1, A_2, \dots$  generating the measurable space on which  $P$  is defined) if the set  $\{(r_1, r_2, k): r_1 < P(A_k) < r_2 \text{ for rational } r_1, r_2 \text{ and } k = 1, 2, \dots\}$  is recursively enumerable. The Kolmogorov complexity of  $P$  (relative to a universal Turing machine  $U$ ) is the length of the shortest program (for  $U$ ) that recursively enumerates  $P$ . Such a program can compute each  $P(A_k)$  to any pre-assigned degree of precision.

distortion-rate function of the omniscient first-stage quantizer.<sup>11</sup>

For any omniscient first-stage quantizer  $\tilde{C}_o^l$ , let

$$R(\tilde{C}_o^l) = E\tilde{r}(\Theta) = E|\tilde{\alpha}_o(\Theta)|$$

be the expected rate, and let

$$\Delta(\tilde{C}_o^l) = E\Delta(\Theta, \tilde{C}_o^l(\Theta)) = E\Delta(\Theta, \tilde{\beta}(\tilde{\alpha}_o(\Theta)))$$

be the expected distortion (actually, the expected divergence) of  $\tilde{C}_o^l$ . Then define

$$\tilde{D}_l(\tilde{R}) = \inf_{\tilde{C}_o^l} \{\Delta(\tilde{C}_o^l) : R(\tilde{C}_o^l) \leq \tilde{R}\}$$

to be the operational ‘‘distortion-rate function’’ for omniscient first-stage quantizers  $\tilde{C}_o^l : \Lambda \rightarrow \mathcal{C}^l$ . Thus for any  $\epsilon > 0$  and  $\tilde{R} \geq 0$ , there exists an omniscient first-stage quantizer  $\tilde{C}_o^l$  such that  $R(\tilde{C}_o^l) \leq \tilde{R}$  and  $\Delta(\tilde{C}_o^l) < \tilde{D}_l(\tilde{R}) + \epsilon$ .

Suppose  $C_o^{n,l}$  is a two-stage code with omniscient first-stage quantizer  $\tilde{C}_o^l$ . Then its expected  $(n, l)$ th-order redundancy (31) can be expressed

$$\begin{aligned} E\ell^{n,l}(C_o^{n,l}, \lambda|\Theta) &= \int \left[ \Delta(\theta, \tilde{C}_o^l(\theta)) + \frac{\lambda}{n} |\tilde{\alpha}_o(\theta)| \right] dW(\theta) \\ &= \Delta(\tilde{C}_o^l) + \frac{\lambda}{n} R(\tilde{C}_o^l). \end{aligned}$$

Therefore, for any  $\epsilon > 0$ ,  $\tilde{R}$ ,  $n$ , and  $l$  dividing  $n$ , there exists a two-stage code  $C_o^{n,l}$  with omniscient first-stage quantizer  $\tilde{C}_o^l$  such that

$$E\ell^{n,l}(C_o^{n,l}, \lambda|\Theta) \leq \tilde{D}_l(\tilde{R}) + \frac{\lambda}{n} \tilde{R} + \epsilon. \quad (34)$$

Thus if for each  $l$ ,  $\tilde{D}_l(\tilde{R}) \rightarrow 0$  as  $\tilde{R} \rightarrow \infty$ , we can construct a weighted universal sequence of codes  $\{C_o^{n,l}\}$ , since we can make the expected  $n$ th-order redundancy  $E\ell^n(C_o^{n,l}, \lambda|\Theta)$  arbitrarily small by first choosing  $l$  large enough so that the difference between  $E\ell^{n,l}(C_o^{n,l}, \lambda|\Theta)$  and  $E\ell^n(C_o^{n,l}, \lambda|\Theta)$  must be small, then by choosing  $\tilde{R}$  large enough so that  $\tilde{D}_l(\tilde{R})$  is small, then by choosing  $n$  large enough so that  $\lambda\tilde{R}/n$  is small, and finally by choosing  $\epsilon$  small. Thus weighted universal codes exist if the distortion-rate function of the omniscient first-stage quantizer goes to zero for each  $l$ .

The tails of  $\tilde{D}_l(\tilde{R})$  actually tell us much more than whether there exist codes  $C_o^{n,l}$  such that  $E\ell^n(C_o^{n,l}, \lambda|\Theta) \rightarrow 0$ . They also tell us about the optimal rate of convergence. Indeed, the expected redundancy (34) can be minimized for each  $n, l$  by optimizing the first-stage rate  $\tilde{R}$

$$E\ell^{n,l}(C_o^{n,l}, \lambda|\Theta) \leq \min_{\tilde{R}} \left[ \tilde{D}_l(\tilde{R}) + \frac{\lambda}{n} \tilde{R} \right] + \epsilon. \quad (35)$$

The minimum of the bracketed term can be regarded as the Lagrangian (at Lagrange multiplier  $\lambda/n$ ) of the function  $\tilde{D}_l(\tilde{R})$ , as illustrated in Fig. 6. That is,  $\min [\tilde{D}_l(\tilde{R}) + (\lambda/n)\tilde{R}]$  is the  $y$ -intercept of the line supporting the graph of  $\tilde{D}_l(\tilde{R})$  at slope  $-\lambda/n$ . It is obvious geometrically that as  $n \rightarrow \infty$

<sup>11</sup>It may also be possible to obtain similarly general results for weakly minimax universal coding using the inverse of the  $\epsilon$ -entropy  $H(\epsilon)$ , but we have not tried this approach. The  $\epsilon$ -entropy is the minimum possible entropy of a quantizer with maximum distortion  $\epsilon$  [5], [42].

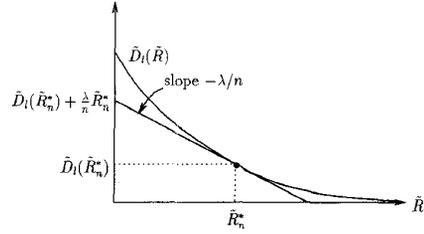


Fig. 6. Geometric interpretation of the Lagrangian of  $\tilde{D}_l(\tilde{R})$ .

for any fixed  $\lambda$ , the slope of the supporting line goes to zero, and hence the minimizing  $\tilde{R}_n^*$  shoots off to infinity while the Lagrangian  $\tilde{D}_l(\tilde{R}_n^*) + (\lambda/n)\tilde{R}_n^*$  goes to zero, and with it, both  $\tilde{D}_l(\tilde{R}_n^*)$  and  $\tilde{R}_n^*/n$ . For each  $n$ , the optimal first-stage rate  $\tilde{R}_n^*$  can be explicitly found by taking derivatives. To ensure that the derivative exists, let  $\check{D}_l(\tilde{R})$  be any decreasing, differentiable, convex  $\cup$  upper bound to the lower convex hull of  $\tilde{D}_l(\tilde{R})$ , such that  $\check{D}_l(\tilde{R}) \rightarrow 0$  as  $\tilde{R} \rightarrow \infty$ . Then the bound (35) becomes

$$E\ell^{n,l}(C_o^{n,l}, \lambda|\Theta) \leq \min_{\tilde{R}} \left[ \check{D}_l(\tilde{R}) + \frac{\lambda}{n} \tilde{R} \right] + \epsilon \quad (36)$$

which achieves its unique minimum when  $\check{D}_l'(\tilde{R}) + \lambda/n = 0$ , or when  $\tilde{R}$  satisfies

$$\check{D}_l'(\tilde{R}_n^*) = -\frac{\lambda}{n}. \quad (37)$$

Thus the rate at which (36) can be made to go to zero as  $n \rightarrow \infty$  depends on the tail of  $\check{D}_l(\tilde{R})$  as  $\tilde{R} \rightarrow \infty$ . We have proved the following:

*Theorem 5* Let  $\check{D}_l(\tilde{R})$  be a decreasing, differentiable, convex  $\cup$  upper bound to the lower convex hull of the operational distortion-rate function  $\tilde{D}_l(\tilde{R})$  for omniscient first-stage quantizers  $\tilde{C}_o^l : \Lambda \rightarrow \mathcal{C}^l$  such that  $\tilde{D}_l(\tilde{R}) \rightarrow 0$  as  $\tilde{R} \rightarrow \infty$ . Then for each  $n$  and  $l$  dividing  $n$ , for any  $\epsilon > 0$  there exists a two-stage code  $C_o^{n,l}$  such that

$$E\ell^{n,l}(C_o^{n,l}, \lambda|\Theta) \leq \min_{\tilde{R}} \left[ \check{D}_l(\tilde{R}) + \frac{\lambda}{n} (\tilde{R} + \epsilon) \right] \quad (38)$$

$$= \check{D}_l(\tilde{R}_n^*) + \frac{\lambda}{n} (\tilde{R}_n^* + \epsilon) \quad (39)$$

where  $\tilde{R}_n^*$  satisfies  $\check{D}_l'(\tilde{R}_n^*) = -\lambda/n$ . In particular, if  $\check{D}_l(\tilde{R})$  does not depend on  $l$ , then (setting  $l = n$ ) there exists a weighted universal sequence of codes  $\{C_o^n\}$  such that

$$E\ell^n(C_o^n, \lambda|\Theta) \leq \check{D}(\tilde{R}_n^*) + (\lambda/n)(\tilde{R}_n^* + \epsilon).$$

We apply the theorem in three cases.

*Case I:*  $\check{D}_l(\tilde{R}) = A2^{-b\tilde{R}}$ ,  $b > 0$ . This is the case if either  $\Lambda$  or  $\mathcal{C}^l$  has finite dimension  $k$ , whence  $b = 2/k$ . (This is shown for  $\Lambda$  finite-dimensional in the Appendix, Lemma 7, and, for  $\mathcal{C}^l$  finite-dimensional, at the end of this section.) Then

$$\check{D}_l'(\tilde{R}) = -\lambda c 2^{-b\tilde{R}}$$

where  $c = Ab/(\lambda \log e)$ , so that  $\check{D}_l'(\tilde{R}_n^*) = -\lambda/n$  (as in (37)) implies

$$\tilde{R}_n^* = \frac{1}{b} \log cn \quad (40)$$

and hence

$$\check{D}_l(\tilde{R}_n^*) = \frac{A}{cn}. \quad (41)$$

Thus (39) becomes

$$\begin{aligned} E\ell^{n,l}(C^{n,l}, \lambda|\Theta) &\leq \frac{A}{cn} + \frac{\lambda}{n} \left( \frac{1}{b} \log cn + \epsilon \right) \\ &= \frac{\lambda}{bn} (\log n + (\log ce + \epsilon')). \end{aligned} \quad (42)$$

Thus as a corollary to Theorem 5, we obtain a weighted version of Theorem 3, with the constant identified in terms of  $A$  and  $b$ .  $\square$

*Case 2:*  $\check{D}_l(\tilde{R}) = 0$  for all sufficiently large  $\tilde{R}$  ( $\tilde{R} \geq \tilde{R}_0$ ). This is the case if  $\Lambda$  is countable and  $H(\Theta)$  is finite (since then there exists an omniscient first-stage quantizer that maps each  $\theta \in \Lambda$  to its optimal  $C_\theta^l \in \mathcal{C}^l$  with zero first-stage distortion at rate  $\tilde{R}_0 = H(\Theta)$ ). Then by (38)

$$\begin{aligned} E\ell^{n,l}(C_o^{n,l}, \lambda|\Theta) &\leq \min_{\tilde{R}} \left[ \check{D}_l(\tilde{R}) + \frac{\lambda}{n} (\tilde{R} + \epsilon) \right] \\ &\leq \check{D}_l(\tilde{R}_0) + \frac{\lambda}{n} (\tilde{R}_0 + \epsilon) \\ &\leq \frac{\lambda}{n} (\tilde{R}_0 + \epsilon). \end{aligned} \quad (43)$$

Thus as a corollary to Theorem 5, we obtain the weighted version of Theorem 4, with the constant  $b_\theta$  identified as the optimal codelength  $-\log w(\theta)$ , whose expectation is the entropy  $\tilde{R}_0 = H(\Theta)$ .  $\square$

*Case 3:*  $\check{D}_l(\tilde{R}) = A\tilde{R}^{-b}$ ,  $b > 0$ . This is the case if  $\Lambda$  is infinite-dimensional, under some conditions (as shown in the Appendix, Lemma 10). Then

$$\check{D}'_l(\tilde{R}) = -Ab\tilde{R}^{-b-1}$$

so that  $\check{D}'_l(\tilde{R}_n^*) = -\lambda/n$  implies

$$\tilde{R}_n^* = \left( \frac{nAb}{\lambda} \right)^{1/(b+1)}$$

and hence

$$\check{D}_l(\tilde{R}_n^*) = A \left( \frac{nAb}{\lambda} \right)^{-b/(b+1)}$$

Thus (39) becomes

$$\begin{aligned} E\ell^{n,l}(C^{n,l}, \lambda|\Theta) &\leq A \left( \frac{nAb}{\lambda} \right)^{-b/(b+1)} \\ &\quad + \frac{\lambda}{n} \left[ \left( \frac{nAb}{\lambda} \right)^{1/(b+1)} + \epsilon \right] \\ &\leq n^{-b/(b+1)} \left( \frac{Ab}{\lambda} \right)^{1/(b+1)} \left( \frac{\lambda}{b} + \lambda + \epsilon' \right). \end{aligned} \quad (44)$$

Thus as a corollary to Theorem 5, we obtain a rate of convergence of the expected redundancy in an infinite dimensional case.  $\square$

In the remainder of this section, we show in the case of fixed-rate quantization under the squared-error distortion

measure that if  $\mathcal{C}^l$  is finite-dimensional with dimension  $k = l2^{lR}$  (where  $R$  is the rate of each code  $C^l \in \mathcal{C}^l$ , so that  $k$  is the number of components in each  $C^l$ ) then  $\check{D}_l(\tilde{R}) \leq A2^{-2\tilde{R}/k}$ . Here, both  $A$  and  $k$  depend on  $l$ .

Assume first that  $X^l$  has bounded support, i.e.,  $\|X^l\| \leq B$  almost surely ( $P^l$ ). Then for almost all  $\theta$ , if  $C_\theta^l$  is an optimal quantizer for  $P_\theta^l$ , each component of each codeword of  $C_\theta^l$  lies in  $[-B, B]$ . Partition the range  $[-B, B]$  uniformly into  $2^{\tilde{R}/k}$  bins, so that  $[-B, B]^k$  is partitioned uniformly into  $2^{\tilde{R}}$  bins. Assign a representative to each  $k$ -dimensional bin, and let  $\Gamma^l$  be the collection of  $2^{\tilde{R}}$  codes corresponding to the  $2^{\tilde{R}}$  representatives. Then for any code  $C_\theta^l = \beta_\theta \circ \alpha_\theta$ , there exists a code  $C_{\tilde{s}}^l = \beta_{\tilde{s}} \circ \alpha_{\tilde{s}} \in \Gamma^l$  such that  $\|\beta_\theta(s) - \beta_{\tilde{s}}(s)\|^2 \leq l[(2B)2^{-\tilde{R}/k}]^2$  for all  $s \in \mathcal{S}$ ,  $|\mathcal{S}| = 2^{lR}$ . Moreover

$$\begin{aligned} \Delta(\theta, C_{\tilde{s}}^l) &= l^{-1} E_\theta [\|X^l - \beta_{\tilde{s}}(\alpha_{\tilde{s}}(X^l))\|^2 \\ &\quad - \|X^l - \beta_\theta(\alpha_\theta(X^l))\|^2] \\ &\leq l^{-1} E_\theta [\|X^l - \beta_{\tilde{s}}(\alpha_\theta(X^l))\|^2 \\ &\quad - \|X^l - \beta_\theta(\alpha_\theta(X^l))\|^2] \\ &= l^{-1} E_\theta [\|\beta_{\tilde{s}}(\alpha_\theta(X^l)) - \beta_\theta(\alpha_\theta(X^l))\|^2] \\ &\leq (4B^2)2^{-2\tilde{R}/k}. \end{aligned}$$

Thus there exists an omniscient two-stage code  $\tilde{C}_o^l$  with fixed rate  $\tilde{R}$  such that

$$\begin{aligned} \check{D}_l(\tilde{R}) &\leq E\Delta(\Theta, \tilde{C}_o^l(\Theta)) \leq \sup_{\theta \in \Lambda} \Delta(\theta, \tilde{C}_o^l(\theta)) \\ &\leq (4B^2)2^{-2\tilde{R}/k}. \end{aligned}$$

Note we get a minimax result for free when  $X^l$  has bounded support.

When  $X^l$  does not have bounded support, but still satisfies a moment condition, we can appeal to asymptotic quantization theory. Suppose  $E\|X^l\|^{2+\epsilon} < \infty$  for some  $\epsilon > 0$ . Then  $\check{D}_l(\tilde{R}) \leq A2^{-2\tilde{R}/k}$  for some  $A < \infty$ , as proved in the Appendix, Lemma 8. Hence using (40) and (41) in *Case 1* we have the following corollary to Theorem 5, first reported in [11]:

*Theorem 6:* Fix any  $l$  and  $R$  such that  $2^{lR}$  is an integer, and fix any constant  $c$ . Suppose  $E\|X^l\|^{2+\epsilon} < \infty$  for some  $\epsilon > 0$ . For each  $n$  that is a multiple of  $l$ , there exists a two-stage fixed-rate quantizer  $C^{n,l}$  with  $n^{-1}R(C^{n,l}) \rightarrow R$  as  $n \rightarrow \infty$  such that under the squared-error distortion measure

$$\begin{aligned} \rho(C^{n,l}) &\triangleq \frac{1}{n} R(C^{n,l}) - R \\ &= \frac{1}{n} \left[ \frac{k}{2} \log cn \right] \end{aligned}$$

and

$$\begin{aligned} E\delta^{n,l}(C^{n,l}|\Theta) &\triangleq \int [D_\theta(C^{n,l}) - \hat{D}_\theta^l(R)] dW(\theta) \\ &\leq \frac{A}{cn} \end{aligned}$$

where  $k = l2^{lR}$  is the number of components (parameters) in each blocklength- $l$ , rate- $R$ , second-stage quantizer, and  $A$  is a

constant possibly depending on  $l$  and  $R$ . If the source lies in  $[-B, B]$  almost surely ( $P$ ), then  $A \leq 4B^2$  for all  $l$  and  $R$ .<sup>12</sup>

As mentioned at the end of Section II, it is always possible to trade rate redundancy for distortion redundancy. Hence we may restrict  $C^{n,l}$  to have rate at most  $R$  (i.e., rate redundancy less than or equal to zero), in which case the distortion redundancy bound increases to about  $A/cn + (\lambda/n)(k/2) \log n$ , where  $\lambda$  is the (negative) slope of the distortion-rate function at  $R$ . Thus for fixed  $l$ , the performance of a two-stage fixed-rate quantizer converges to the  $l$ th-order OPTA at rate  $O(n^{-1} \log n)$ . This is perhaps our most important result for applications, since in practice the dimension  $l$  of the second-stage quantizer will remain fixed, rather than go to infinity. Of course, in theory it is also possible to let  $l$  go to infinity with  $n$  to obtain results on the convergence to zero of the expected  $n$ th-order redundancy

$$E\delta^n(C^{n,l}|\Theta) \leq E\delta^{n,l}(C^{n,l}|\Theta) + \phi^l$$

when bounds on

$$\phi^l = \int \phi_\theta^l dW(\theta)$$

are known, e.g., for bounded iid sources [36, Theorem 2]. For bounded iid sources, this approach yields a convergence result equal to  $O(\sqrt{(\log n)^{-1} \log \log n})$ , consistent with [36, Theorem 3].

We now turn to a method for actually designing two-stage noiseless codes, fixed-rate quantizers, and variable-rate quantizers with the minimum possible expected redundancy for each  $n$  and  $l$ .

#### IV. THE WEIGHTED UNIVERSAL CODE DESIGN ALGORITHM

A two-stage code with the minimum possible expected redundancy we call a *weighted universal code*, since a sequence of such codes is a weighted universal sequence of codes whose redundancies converge to zero at the fastest possible rate. In this section we describe an algorithm for designing weighted universal noiseless codes (WUNC's) and weighted universal vector quantizers (WUVQ's), both fixed-rate and variable-rate.

Recall the quantization interpretation of a two-stage coding system. The system comprises two parts: the first-stage quantizer maps the input space of possible data blocks of length  $n$  to the output space of possible block codes of length  $l$ , where  $l$  divides  $n$ ; the chosen second-stage block code then maps each sub-block of the data to the output space of possible reproductions. In universal noiseless coding, these block codes are block noiseless codes; in universal fixed-rate quantization, the block codes are block quantizers at rate  $R$ ; and in universal variable-rate quantization, the block codes are variable-rate block quantizers.

Specifically, the first-stage quantizer,  $\tilde{C}^{n,l} = \tilde{\beta} \circ \tilde{\alpha}$ , contains an encoder  $\tilde{\alpha}: \mathcal{X}^n \rightarrow \tilde{\mathcal{S}}$ , and a decoder  $\tilde{\beta}: \tilde{\mathcal{S}} \rightarrow \mathcal{C}^l$ , which

<sup>12</sup>We conjecture that  $A = \bar{D}^l(R)$  suffices, where

$$\bar{D}^l(R) = \int \bar{D}_\theta^l(R) dW(\theta)$$

is the operational distortion-rate function for the mixture. Note that  $\bar{D}^l(R) \leq EX_1^2$  regardless of  $l$ .

together map the input space of possible data vectors to the output space of length- $l$  block codes via a noiseless code  $\tilde{\mathcal{S}}$ . The first-stage quantizer's reproduction of a vector  $x^n$  is the length- $l$  block code,  $C_{\tilde{\alpha}(x^n)}^l = \tilde{\beta}(\tilde{\alpha}(x^n)) \in \mathcal{C}^l$ , used to code the length- $l$  sub-blocks  $x_1^l, \dots, x_{n/l}^l$  of  $x^n$ . We use  $|\tilde{\alpha}(x^n)|$  to denote the first-stage description length of  $x^n$ . For any  $\tilde{s} \in \tilde{\mathcal{S}}$ , the second-stage block code  $C_{\tilde{s}}^l = \beta_{\tilde{s}} \circ \alpha_{\tilde{s}}$  contains an encoder  $\alpha_{\tilde{s}}: \mathcal{X}^l \rightarrow \mathcal{S}_{\tilde{s}}$  and a decoder  $\beta_{\tilde{s}}: \mathcal{S}_{\tilde{s}} \rightarrow \hat{\mathcal{X}}^l$ , which together map the input space of possible  $l$ -vectors to the output space of possible reproductions via a noiseless code  $\mathcal{S}_{\tilde{s}}$ . Given a first-stage description  $\tilde{s}$  of  $x^n$ ,  $\alpha_{\tilde{s}}(x^l) \in \mathcal{S}_{\tilde{s}}$  is the second-stage description of a sub-block  $x^l$  of  $x^n$ ,

$$\sum_{i=1}^{n/l} |\alpha_{\tilde{s}}(x_i^l)|$$

is the second-stage description length of  $x^n$ , and

$$\sum_{i=1}^{n/l} d(x^l, \beta_{\tilde{s}}(\alpha_{\tilde{s}}(x_i^l)))$$

is the distortion incurred by using the block code  $C_{\tilde{s}}^l$  on the sub-blocks of  $x^n$ . The distortion is always zero in noiseless coding.

In the case of two-stage fixed-rate quantization,  $\tilde{\mathcal{S}}$  and  $\mathcal{S}_{\tilde{s}}$  are finite collections of fixed-length binary strings. The string lengths in  $\mathcal{S}_{\tilde{s}}$  are identical for all  $\tilde{s} \in \tilde{\mathcal{S}}$  but need not equal the string length in  $\tilde{\mathcal{S}}$ . For two-stage noiseless codes,  $\tilde{\mathcal{S}}$  and  $\mathcal{S}_{\tilde{s}}$  for each  $\tilde{s} \in \tilde{\mathcal{S}}$  are countable collections of variable-length binary strings. For two-stage variable-rate codes,  $\tilde{\mathcal{S}}$  and  $\mathcal{S}_{\tilde{s}}$  may be finite collections of fixed-length binary strings where the string lengths in  $\mathcal{S}_{\tilde{s}}$  vary with  $\tilde{\mathcal{S}}$ , or countable collections of variable-length binary strings.

As defined in Section III, the instantaneous rate  $\tilde{r}(x^n)$  of a first-stage quantizer is the length of the first-stage description,  $\tilde{r}(x^n) = |\tilde{\alpha}(x^n)|$ , while the instantaneous distortion  $\tilde{d}(x^n, \tilde{\beta}(\tilde{s}))$  is the Lagrangian performance (per letter) associated with coding the sub-blocks of  $x^n$  with the block code  $C_{\tilde{s}}^l = \tilde{\beta}(\tilde{s})$

$$\tilde{d}(x^n, \tilde{\beta}(\tilde{s})) = \frac{1}{n} \sum_{i=1}^{n/l} [d(x_i^l, C_{\tilde{s}}^l(x_i^l)) + \lambda |\alpha_{\tilde{s}}(x_i^l)|].$$

Thus the corresponding instantaneous Lagrangian  $\tilde{l}(x^n, \lambda/n)$  associated with the first-stage quantizer is

$$\tilde{l}(x^n, \lambda/n) = \tilde{d}(x^n, \tilde{\beta}(\tilde{\alpha}(x^n))) + \frac{\lambda}{n} |\tilde{\alpha}(x^n)|.$$

We showed in Section III, (15)–(22), that the two-stage code  $C^{n,l}$  minimizing the expected Lagrangian  $D(C^{n,l}) + \lambda R(C^{n,l})$  (and hence also minimizing the expected redundancy  $E\ell^n(C^{n,l}, \lambda|\Theta)$ ) has a first-stage quantizer  $\tilde{C}^{n,l}$  minimizing the expected Lagrangian

$$L(\tilde{C}^{n,l}, \lambda/n) = \int \left[ \tilde{d}(x^n, \tilde{\beta}(\tilde{\alpha}(x^n))) + \frac{\lambda}{n} |\tilde{\alpha}(x^n)| \right] dP(x^n). \quad (45)$$

To minimize (45), we employ an iterative descent technique formally equivalent to the generalized Lloyd algorithm [35],

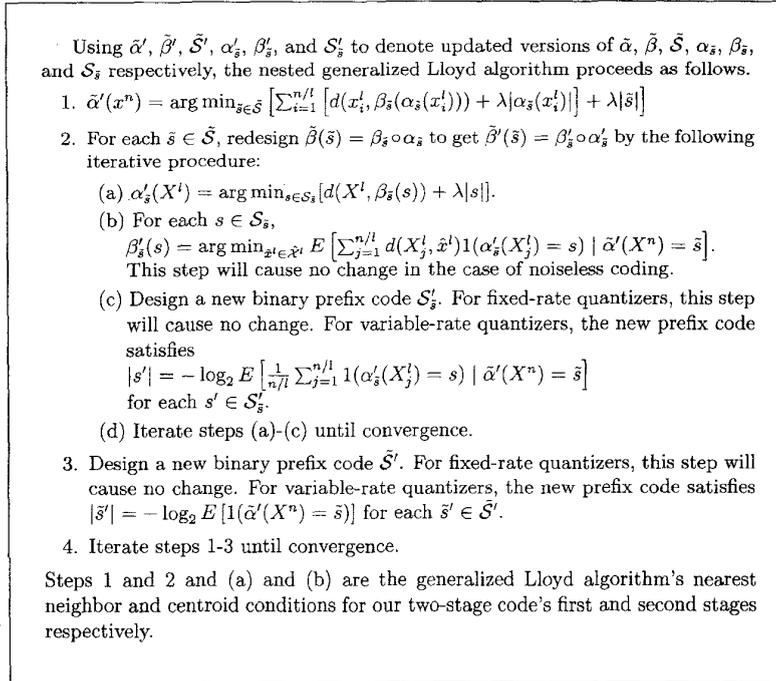


Fig. 7. The design algorithm for a two-stage code may be described as a nested version of the generalized Lloyd algorithm, as detailed above.

or rather its entropy-constrained variation [12]. The algorithm is initialized with an arbitrary prefix code  $\tilde{\mathcal{S}}$  and collection  $\{\tilde{\beta}(\tilde{s}) : \tilde{s} \in \tilde{\mathcal{S}}\}$  of block codes with blocklength  $l$ . The prefix code is either fixed-length or variable-length, and the block codes are either noiseless codes, fixed-rate quantizers, or variable-rate quantizers, as appropriate. Each iteration in the algorithm is accomplished in three steps, which are enumerated below.

#### 1) Nearest Neighbor Encoding

The optimal first-stage encoder  $\tilde{\alpha}^*$  for a given first-stage decoder  $\tilde{\beta}$  and prefix code  $\tilde{\mathcal{S}}$  follows a nearest neighbor law, as discussed in Proposition 1

$$\tilde{\alpha}^*(x^n) = \arg \min_{\tilde{s} \in \tilde{\mathcal{S}}} \left[ \tilde{d}(x^n, \tilde{\beta}(\tilde{s})) + \frac{\lambda}{n} |\tilde{s}| \right]$$

which maps each incoming block  $x^n$  to the index of the second stage code  $\tilde{\beta}(\tilde{s})$  that encodes  $x^n$  to the lowest value of  $\tilde{l}(x^n, \lambda/n)$ .

#### 2) Decoding to the Centroid

The optimal first-stage decoder  $\tilde{\beta}^*$  for a given first-stage encoder  $\tilde{\alpha}$  and prefix code  $\tilde{\mathcal{S}}$  satisfies

$$\tilde{\beta}^*(\tilde{\mathcal{S}}) = \arg \min_{C^l \in \mathcal{C}^l} E \left[ \tilde{d}(X^n, C^l) + \frac{\lambda}{n} |\tilde{s}| \mid \tilde{\alpha}(X^n) = \tilde{s} \right]$$

which assigns to each  $\tilde{s} \in \tilde{\mathcal{S}}$  the blocklength- $l$  code  $C^l \in \mathcal{C}^l$  that minimizes the expected value of  $\tilde{l}(X^n, \lambda/n)$  given that  $X^n$  mapped to  $\tilde{s}$  in the first step.

#### 3) Optimizing the Prefix Code

The optimal prefix code  $\tilde{\mathcal{S}}^*$  for a given first-stage encoder  $\tilde{\alpha}$  and decoder  $\tilde{\beta}$  is the entropy code matched to the

probabilities  $P\{\tilde{\alpha}(X^n) = \tilde{s}\}$ , whose ideal codelengths are

$$|\tilde{s}^*| = -\log P\{\tilde{\alpha}(X^n) = \tilde{s}\}.$$

It is simple to show that in each step, the objective (45) decreases or remains the same. Since it is bounded below by zero, it must decrease to a limit as the number of iterations grows.

For noiseless coding, the centroiding step is accomplished for each  $\tilde{s}$  by designing a Huffman code  $C^l_s$  for the data that it must noiselessly code. For quantization, the centroiding step is accomplished for each  $\tilde{s}$  by using the generalized Lloyd algorithm or its entropy-constrained variation to design a locally optimal fixed-rate or variable-rate vector quantizer  $C^l_s$  for the data that it must quantize. The resulting algorithm may be described as a nested generalized Lloyd algorithm, as detailed in Fig. 7. This nested process easily generalizes to design a large array of two-stage codes by replacing either or both of the uses of the Lloyd algorithm by other codebook design algorithms. Alternative codebook design algorithms that may be considered include tree-structured vector quantization, deterministic annealing, fast codebook search algorithms, and so on.

Our algorithm is not the first to use the generalized Lloyd algorithm (GLA) with another optimization algorithm nested inside. For example, Buzo *et al.* [7] used the GLA to cluster linear predictors for speech, using the prediction error as the distortion measure and the Levinson algorithm to optimize the predictors. Rabiner *et al.* [44] used the GLA to cluster hidden Markov models for speech, using the log likelihood

as the distortion measure and the Baum/Welch algorithm to optimize the models; Safranek and Johnston [51] used the GLA to cluster entropy codes for subband image coding, using bit rate as the distortion measure and the Huffman algorithm to optimize the entropy codes; Chou [10] used the GLA to cluster probability mass functions for classification tree design, using the Kullback–Leibler divergence as the distortion measure and conditional expectation to calculate the pmf’s; and Chan and Gersho [8] used the GLA to cluster vector quantizers for residual VQ, using the overall distortion as the distortion measure and a nested GLA to optimize the individual quantizers. Our algorithm may be the first, however, to incorporate bit rates from both the first and second stages into the design. This is necessitated by the fact that both the first and second stages contribute to the redundancy in two-stage universal coding.

As compared to traditional one-stage codes, two-stage codes exhibit a growth in computational complexity that is roughly proportional to  $2^{\tilde{R}}$ , the number of codebooks in the two-stage code. In particular, for fixed-rate quantization with respect to the squared-error distortion measure, encoding a sequence  $x^n$  with a two-stage code  $C^{n,l}$  requires approximately  $2^{\tilde{R}}(2n2^{lR} + n/l)$  additions and  $2^{\tilde{R}}n2^{lR}$  multiplies as compared to the  $2n2^{lR}$  additions and  $n2^{lR}$  multiplies required in a one-stage code of rate  $R$ . Of course, the complexity can be considerably reduced by the use of tree structures or other means.

## V. EXPERIMENTAL RESULTS

In this section we examine the empirical performance of weighted universal noiseless codes, weighted universal fixed-rate quantizers, and weighted universal variable-rate quantizers designed using the algorithm described in the previous section. We compare the performance of these codes on synthetic sources to the sources’ theoretical performance bounds and examine the convergence properties as we increase the data length  $n$ . On real image sources (for which the optimal performance cannot be calculated) we also compare the performance of the weighted universal fixed-rate and variable-rate quantizers to the performance of standard VQ and entropy-constrained VQ (ECVQ) with the same second-stage dimension.

### A. Weighted Universal Noiseless Coding

In order to test the performance of noiseless codes generated by the weighted universal code design algorithm, we compare the total rate redundancies achieved by the noiseless codes, as a function of data length, to the optimal performance and convergence results described in Section III. In particular, we expect the total (i.e., unnormalized) rate redundancy to grow as  $(k/2) \log n$ , where  $k$  is the dimension of the parameter space and  $n$  is the data length.

Experiments on two classes of sources are performed. The first class is the class of iid sources over the binary alphabet  $\{0, 1\}$ , for which the probabilities  $p_0$  and  $p_1$  of a 0 and 1, respectively, are distributed uniformly on the open 1-dimensional probability simplex

$$\Lambda = \{(p_0, p_1): p_0 + p_1 = 1, p_0 > 0, p_1 > 0\}$$

for which  $k = 1$ . The second class is the class of iid sources over the ternary alphabet  $\{0, 1, 2\}$ , for which the probability vector  $(p_0, p_1, p_2)$  associated with the three elements is distributed uniformly over the open 2-dimensional probability simplex

$$\Lambda = \{(p_0, p_1, p_2): p_0 + p_1 + p_2 = 1, p_0 > 0, p_1 > 0, p_2 > 0\}$$

for which  $k = 2$ .

For each class, a data set of training and test sequences is generated as follows: 1024 parameter vectors (i.e., probability vectors), say  $\theta_1, \dots, \theta_{1024}$ , are chosen at random from the parameter space  $\Lambda$ , and for each source  $P_{\theta_i}$ , 2048 independent samples are generated and split evenly between training and test sequences. The 1024 training sequences are concatenated together to form a training sequence for the mixture, and likewise for the test sequences.

Weighted universal noiseless codes  $C^{n,l}$  are designed on the training sequence for the mixture, and evaluated on the test sequence, starting from an initial codebook of 1024 randomly chosen noiseless codes with blocklength  $l = 1$ . For the first class of sources, the data length  $n$  varies from 1 to 1024 (in powers of 2), while for the second class, the data length is allowed to vary from 1 to 128, since the optimal number of codebooks predicted by the  $(k/2) \log n$  result becomes prohibitively large for larger  $n$  due to the higher value of  $k$ .

For each class, and for each value of  $n$ , we evaluate the total expected rate redundancy

$$nE\rho(C^{n,l}|\Theta) = R(C^{n,l}) - nH(X|\Theta).$$

(Note that the redundancy, the  $n$ th-order redundancy, and the  $(n, l)$ th-order redundancy, as defined in (1), (8), and (30), respectively, are all essentially the same, within  $1/n$  bits, in the iid case.) We estimate  $R(C^{n,l})$  by the performance of  $C^{n,l}$  on the test sequence for the mixture, and we estimate  $H(X|\Theta)$  by the average

$$\frac{1}{1024} \sum_{i=1}^{1024} H(X|\theta_i)$$

where  $H(X|\theta_i)$  is the analytically calculated entropy of the  $i$ th source. We then plot the total expected rate redundancy as a function of  $\log n$ , as shown in Fig. 8. The dashed lines indicate the theoretical asymptotic slopes of the graphs for  $k = 1$  and  $k = 2$  (predicted by the  $(k/2) \log n$  result), while solid lines represent the achieved results.

Final codebooks for the first class with  $n = 1024$  and for the second class with  $n = 128$  are shown in Figs. 9 and 10. Notice that the codebooks are more densely populated toward the edges of the parameter space  $\Lambda$ . This behavior is explained by the shape of the relative entropy function  $D(p||q)$  between two distributions. In both cases,  $D(p||q)$  is relatively flat in the center of the  $pq$  space but becomes very steep (and finally infinite) toward the edges.

### B. Weighted Universal Quantization of Synthetic Sources

Here we test the performance of quantizers generated by the weighted universal code design algorithm by comparing

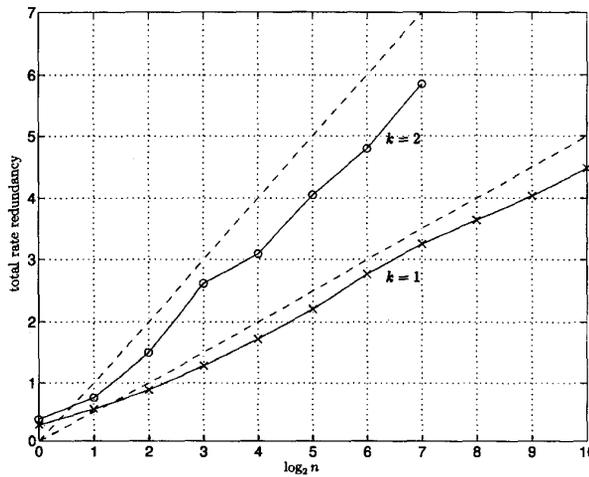


Fig. 8. Total rate redundancies on iid samples from two collections of sources:  $X \in \{0, 1\}$  with probability vector  $(p_0, p_1)$  distributed uniformly on  $\{(p_0, p_1): p_0 + p_1 = 1, p_0 > 0, p_1 > 0\}$  giving  $k = 1$  and  $X \in \{0, 1, 2\}$  with probability vector  $(p_0, p_1, p_2)$  distributed uniformly on  $\{(p_0, p_1, p_2): p_0 + p_1 + p_2 = 1, p_0 > 0, p_1 > 0, p_2 > 0\}$  giving  $k = 2$ . The dashed lines show the theoretical slopes for the two cases.

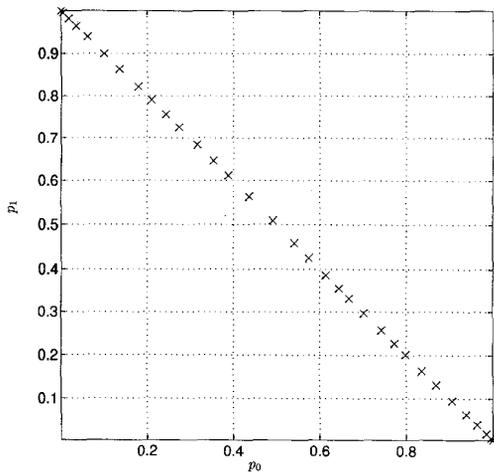


Fig. 9. Codebook of noiseless codes designed on iid samples from sources on the alphabet  $\{0, 1\}$  with probability vector  $(p_0, p_1)$  distributed uniformly on  $\{(p_0, p_1): p_0 + p_1 = 1, p_0 > 0, p_1 > 0\}$  and  $(n, l) = (1024, 1)$ .

the total rate and distortion redundancies achieved by the quantizers, as a function of data length, to the theory. We expect the total rate redundancy to grow as  $(k/2) \log n$ , and the total distortion redundancy to remain at a constant  $A$ .

We again include experiments on two classes of sources. The first class is the class of iid Gaussian sources  $N(\mu, \frac{1}{4})$  with mean  $\mu \sim N(0, 1)$  and constant variance equal to  $\frac{1}{4}$ , for which  $k = 1$ . The second class is the class of iid Gaussian sources  $N(\mu, \sigma^2)$  with mean  $\mu \sim N(0, 1)$  and scale factor  $\sigma \sim N(0, \pi/2)$  (where the distribution of the scale factor was chosen to give a symmetric two-dimensional distribution), for which  $k = 2$ .

As in the noiseless experiments, a data set of training and test sequences is generated for each class as follows: 1024 parameter vectors  $\theta_1, \dots, \theta_{1024}$  are chosen at random

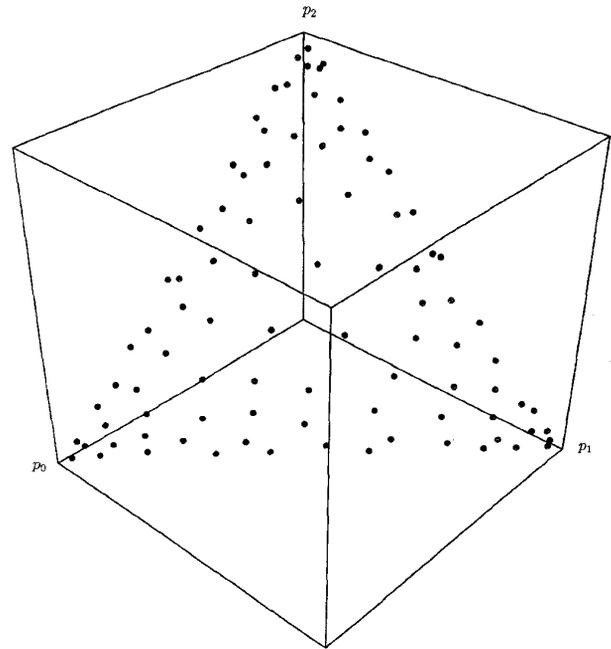


Fig. 10. Codebook of codes designed on iid samples from sources on the alphabet  $\{0, 1, 2\}$  with probability vector  $(p_0, p_1, p_2)$  distributed uniformly on  $\{(p_0, p_1, p_2): p_0 + p_1 + p_2 = 1, p_0 > 0, p_1 > 0, p_2 > 0\}$  and  $(n, l) = (128, 1)$ .

from the parameter space  $\Lambda$ , and from each source  $P_{\theta_i}$ , 2048 independent samples are generated and split evenly between training and test sequences. The 1024 training sequences are concatenated together to form a training sequence for the mixture, and likewise for the test sequences.

First we investigate fixed-rate quantizers. Weighted universal fixed-rate quantizers  $C^{n,l}$  are designed on the mixture training sequence, and evaluated on the mixture test sequence, starting from an initial codebook of  $2^{\tilde{R}}$  randomly chosen quantizers with blocklength  $l = 1$  (i.e., scalar quantizers), each at rate  $R$ . We again allow the data length  $n$  to vary from 1 to 1024 (in powers of 2) for the first class, and to vary from 1 to 128 for the second class. Further, we allow the first-stage rate  $\tilde{R}$  to take values in  $\{0, 1, \dots, 9\}$ , and we allow the second-stage rate  $R$  to take values in  $\{0, 1, 2, 3\}$ , for both classes of sources.

For each class, and for each value of  $n, \tilde{R}$  and  $R$ , we evaluate the total rate redundancy

$$n\rho(C^{n,l}) = R(C^{n,l}) - nR = \tilde{R}$$

and the total expected distortion redundancy

$$nE\delta^{n,l}(C^{n,l}|\Theta) = D(C^{n,l}) - n\hat{D}^l(R).$$

We estimate  $D(C^{n,l})$  by the performance of  $C^{n,l}$  on the test sequence for the mixture, and we estimate  $\hat{D}^l(R)$  at each rate  $R$  by the average

$$\frac{1}{1024} \sum_{i=1}^{1024} \hat{D}_{\theta_i}^l(R)$$

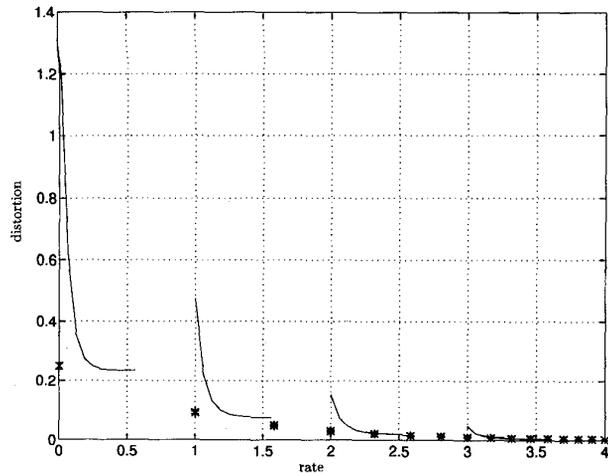


Fig. 11. Distortion-rate results for fixed-rate coding of  $N(\mu, \frac{1}{4})$  with  $\mu \sim N(0, 1)$  and  $n = 16$  compared to the optimal performance for the same source (indicated by \*s).

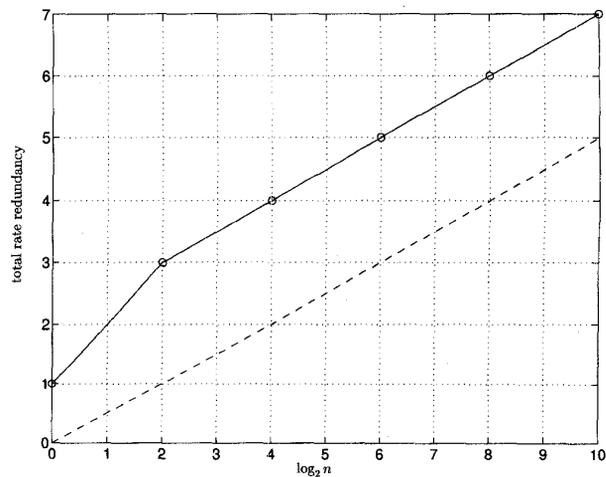


Fig. 13. Total rate redundancy for fixed-rate coding on  $N(\mu, \frac{1}{4})$  with  $\mu \sim N(0, 1)$ . The dashed line shows the asymptotic slope for  $k = 1$ .

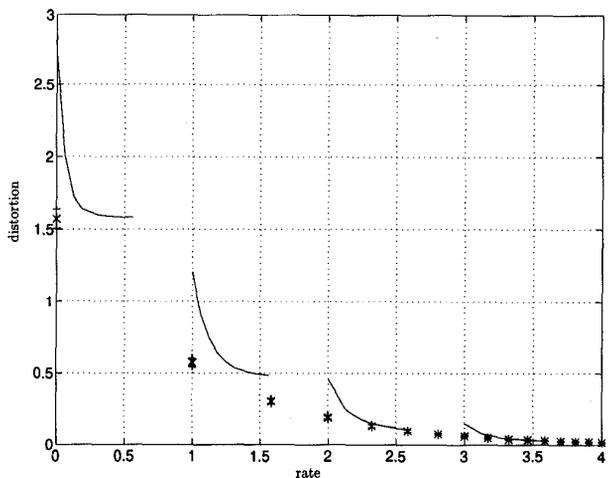


Fig. 12. Distortion-rate results for fixed-rate coding of  $N(\mu, \sigma^2)$  with  $\mu \sim N(0, 1)$ ,  $\sigma \sim N(0, \pi/2)$  and  $n = 16$  compared to the optimal performance for the same source (indicated by \*s).

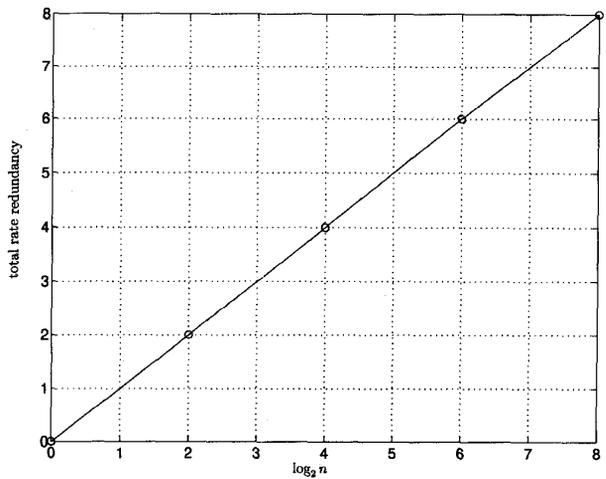


Fig. 14. Total rate redundancy for fixed-rate coding on  $N(\mu, \sigma^2)$  with  $\mu \sim N(0, 1)$  and  $\sigma \sim N(0, \pi/2)$ . The line has exactly the predicted slope for  $k = 2$ .

where  $\hat{D}_{\theta_i}^l(R)$  is in turn estimated by the distortion on the  $i$ th test sequence of the scalar quantizer optimized for the  $i$ th training sequence using the Lloyd algorithm run to convergence. (Lloyd–Max quantizers are optimal for Gaussian sources under the squared error distortion measure [52].) As this “optimal” distortion is based on a finite number of training samples and is therefore only an estimate of the true optimum, we include bars indicating one standard deviation above and below the calculated averages.

Figs. 11 and 12 show the distortion-rate curves for fixed values of  $R$  and  $n = 16$ . The lower convex hull of this set of curves approaches the lower convex hull of  $\hat{D}^l(R)$  as  $n \rightarrow \infty$ . The gap between these convex hulls can be measured by the gap between their supporting lines at slope  $-\lambda$ , for any given  $\lambda > 0$ . We choose  $\lambda$  to support  $\hat{D}^l(R)$  at  $R = 1$ , and then for each value of  $n$ , we choose the two-stage quantizer  $C^{n,l}$  having first-stage rate  $\tilde{R}$  and second-stage rate  $R$  such that

the expected Lagrangian  $EL(C^{n,l}, \lambda | \Theta)$  is minimized. The expected Lagrangian is estimated by an average performance over the test sequences, as usual. Using this sequence of quantizers, we then plot the achieved total rate redundancies for the two classes in Figs. 13 and 14. Notice that the slope of the rate redundancy line, as a function of  $\log n$ , changes with  $k$  in the predicted manner. In this example, the rate redundancies conform almost exactly to the specified slope. This exactness is more likely in fixed-rate coding, where the rate redundancy  $\tilde{R}$  is a function only of the number of codebooks rather than the entropy of those codebooks. The total distortion redundancy, as a function of  $n$ , is expected to remain constant. Figs. 15 and 16 show the achieved total distortion redundancies on the two sources with indications of the change in distortion redundancies when the optimal distortion estimates are varied by one standard deviation in either direction. The sum of distortion redundancy and rate redundancy, weighted by the  $\lambda$

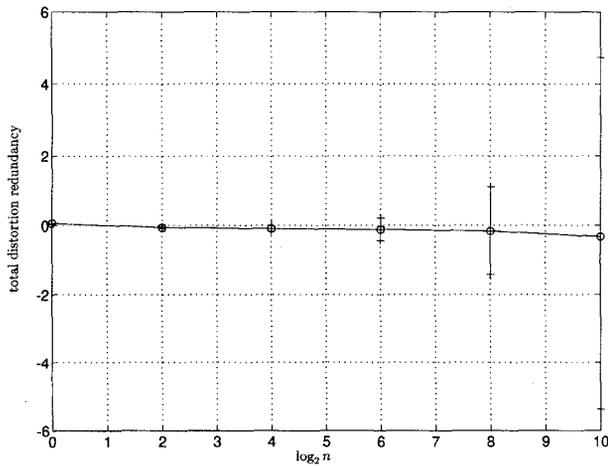


Fig. 15. Total distortion redundancy for fixed-rate coding on  $N(\mu, \frac{1}{4})$  with  $\mu \sim N(0, 1)$  and  $k = 1$ .

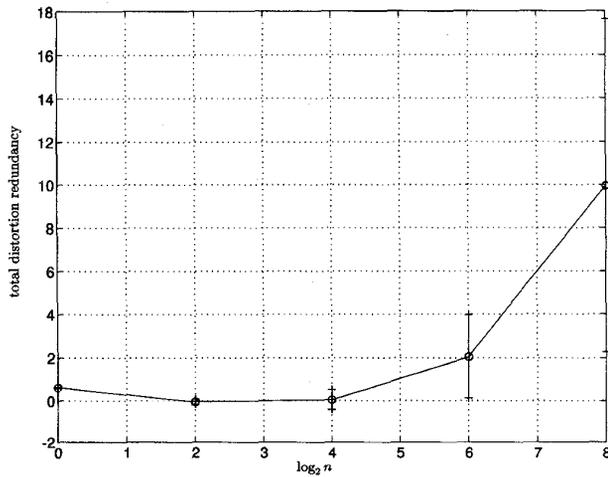


Fig. 16. Total distortion redundancy for fixed-rate coding on  $N(\mu, \sigma^2)$  with  $\mu \sim N(0, 1)$  and  $\sigma \sim N(0, \pi/2)$  and  $k = 2$ .

corresponding to  $R = 1$  for each class of sources, is included in Figs. 17 and 18.

We next investigate variable-rate quantizers. Weighted universal variable-rate quantizers  $C^{n,l}$  are trained and tested on the same two classes of sources as described above for the fixed-rate case. The initial first and second-stage codebook sizes are fixed at 512 and 8, respectively. We again code scalars ( $l = 1$ ) and allow the data length  $n$  to vary from 1 to 1024 for  $k = 1$  and from 1 to 256 for  $k = 2$ . Codebooks are designed for each value of  $n$  using a sequence of  $\lambda$ 's.

We evaluate for each  $C^{n,l}$  the total expected Lagrangian redundancy

$$nE\ell^{n,l}(C^{n,l}, \lambda|\Theta) = L(C^{n,l}, \lambda) - n\hat{L}^l(\lambda)$$

where

$$L(C^{n,l}, \lambda) = D(C^{n,l}) + \lambda R(C^{n,l})$$

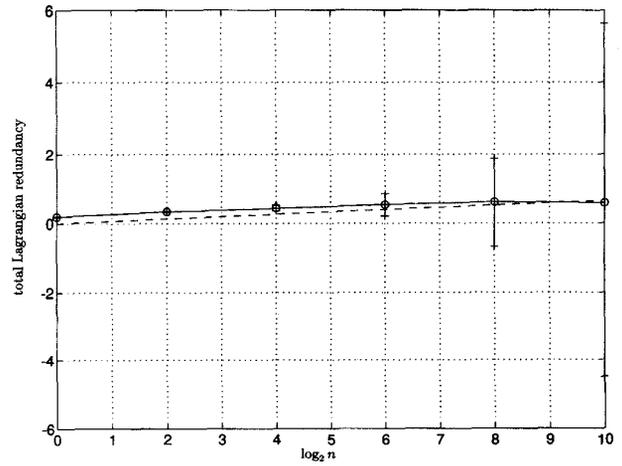


Fig. 17. Total Lagrangian redundancy for fixed-rate coding on  $N(\mu, \frac{1}{4})$  with  $\mu \sim N(0, 1)$  and  $\lambda = 0.1276$ . The dashed line shows the predicted slope for  $k = 1$ .

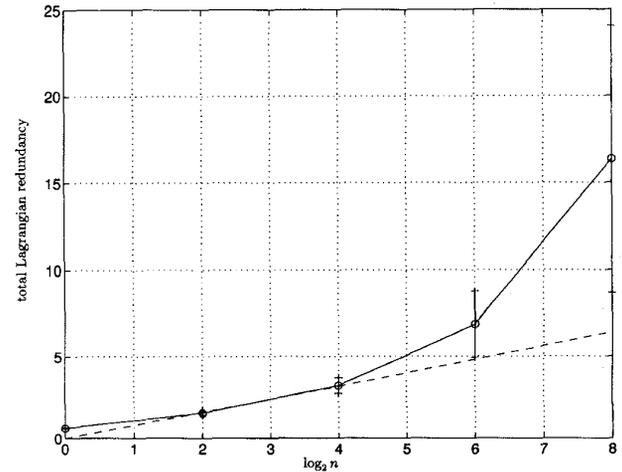


Fig. 18. Total Lagrangian redundancy for fixed-rate coding on  $N(\mu, \sigma^2)$  with  $\mu \sim N(0, 1)$ ,  $\sigma \sim N(0, \pi/2)$ , and  $\lambda = 0.8016$ . The dashed line shows the predicted slope for  $k = 2$ .

and

$$\hat{L}^l(\lambda) = \min_{C^l} L(C^l, \lambda).$$

We estimate  $D(C^{n,l})$  and  $R(C^{n,l})$  as usual on the test sequence for the mixture, and we estimate  $\hat{L}^l(\lambda)$  by the average

$$\frac{1}{1024} \sum_{i=1}^{1024} \hat{L}_{\theta_i}^l(\lambda)$$

where  $\hat{L}_{\theta_i}^l(\lambda)$  is in turn estimated by the Lagrangian on the  $i$ th test sequence of the entropy-constrained scalar quantizer optimized for the  $i$ th training sequence using the algorithm of [12]. As indicated, all averages are taken over points of constant  $\lambda$  (rather than, say, constant rate), since the operational distortion-rate function of a stationary nonergodic source is equal to a mixture of the operational distortion-rate

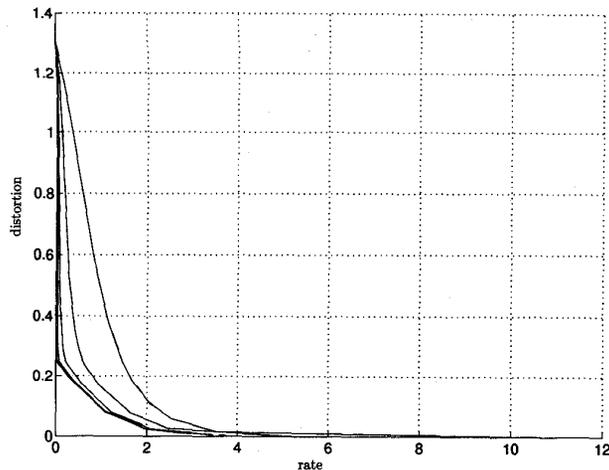


Fig. 19. Distortion-rate results for variable-rate coding of  $N(\mu, \frac{1}{4})$  with  $\mu \sim N(0, 1)$ . Progressing from top to bottom, the curves show the performance with increasing values of  $n$ . The lowest curve shows the optimal performance.

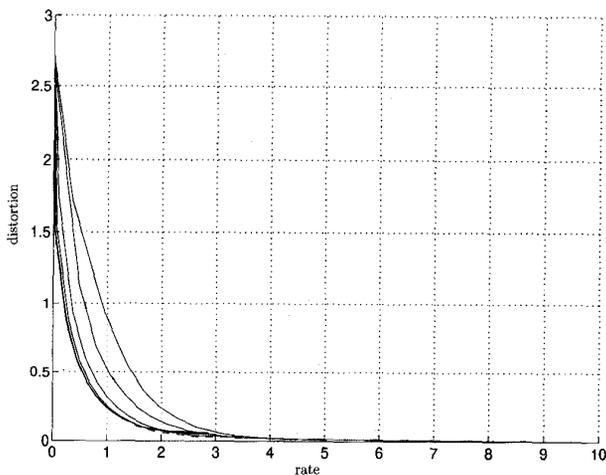


Fig. 20. Distortion-rate results for variable-rate coding of  $N(\mu, \sigma^2)$  with  $\mu \sim N(0, 1)$  and  $\sigma \sim N(0, \pi/2)$ . Progressing from top to bottom, the curves show the performance with increasing values of  $n$ . The lowest curve shows the optimal performance.

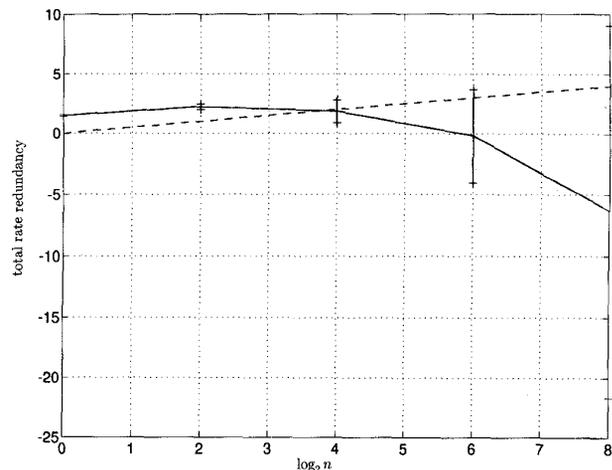


Fig. 21. Total rate redundancy for variable-rate coding on  $N(\mu, \frac{1}{4})$  with  $\mu \sim N(0, 1)$  and  $\lambda = 0.1105$ . The dashed line shows the predicted slope for  $k = 1$ .

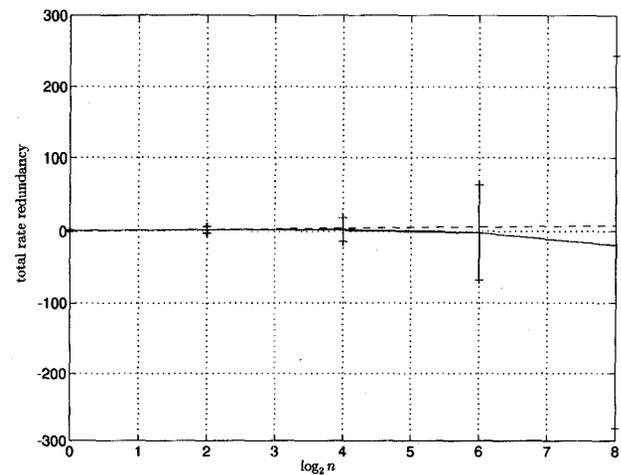


Fig. 22. Total rate redundancy for fixed-rate coding on  $N(\mu, \sigma^2)$  with  $\mu \sim N(0, 1)$ ,  $\sigma \sim N(0, \pi/2)$ , and  $\lambda = 0.393$ . The dashed line shows the predicted slope for  $k = 2$ .

functions of the source's stationary ergodic components where the averages are taken over points of constant  $\lambda$  [22]. We again include bars indicating one standard deviation above and below the calculated rate and distortion redundancy points to indicate a reasonable margin of error.

Distortion-rate curves for constant values of  $n$  and varying  $\lambda$  are shown in Figs. 19 and 20. Figs. 21–26 show total rate, distortion, and Lagrangian redundancies for the two classes. In each case, we choose the  $\lambda$  from our sequence of  $\lambda$ 's which supports the operational distortion-rate function at  $R = 1$ .

*C. Weighted Universal Vector Quantization of Real Sources*

In studying WUVQ performance on real sources, we first consider fixed-rate quantization of a mixture of sources which

includes magnetic resonance (MR) brain scans, airport images, human portraits, and text documents. For each image source, we use a single image for training and set aside a separate image for testing. The image sizes are  $512 \times 512$  for all but the airport images, where the image sizes are  $480 \times 512$ . Each  $512 \times 512$  brain image is made up of four distinct  $256 \times 256$  brain slices. For these examples, we use  $4 \times 4$  vectors.

Fig. 27 shows the results of training a separate full search standard VQ of dimension 16 on each image source and then testing the performance of that quantizer on the test image from the same source. These performances are identical to those that would be achieved on the test set of images in a hypothetical multi-codebook system in which both the encoder and decoder are omniscient, i.e., both know the type of image being coded at any given time and hence know which of the multiple codebooks to use for that image. (Notice that

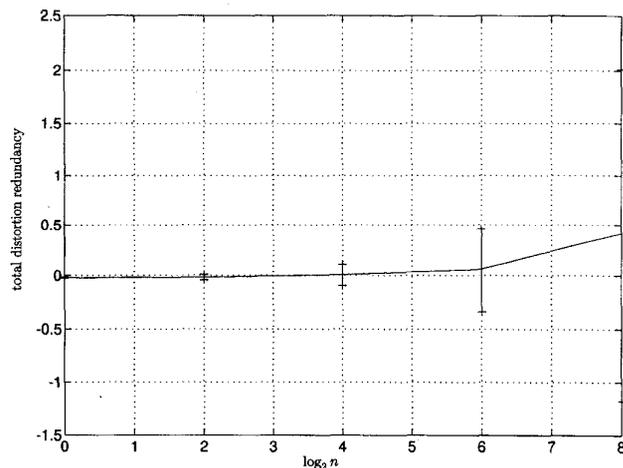


Fig. 23. Total distortion redundancy for variable-rate coding on  $N(\mu, \frac{1}{4})$  with  $\mu \sim N(0, 1)$  and  $\lambda = 0.1105$ .

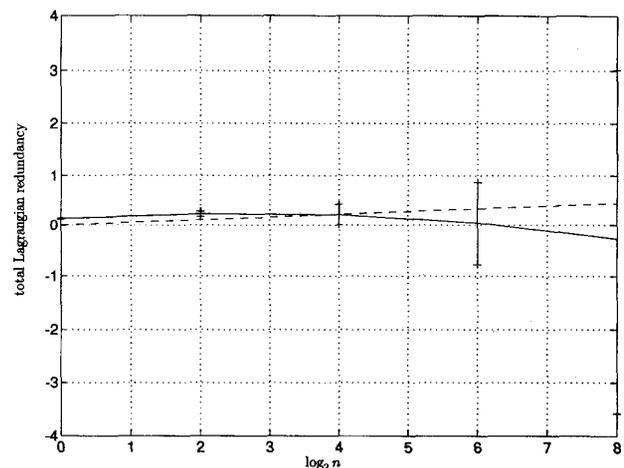


Fig. 25. Total Lagrangian redundancy for variable-rate coding on  $N(\mu, \frac{1}{4})$  with  $\mu \sim N(0, 1)$  and  $\lambda = 0.1105$ . The dashed line shows the predicted slope for  $k = 1$ .

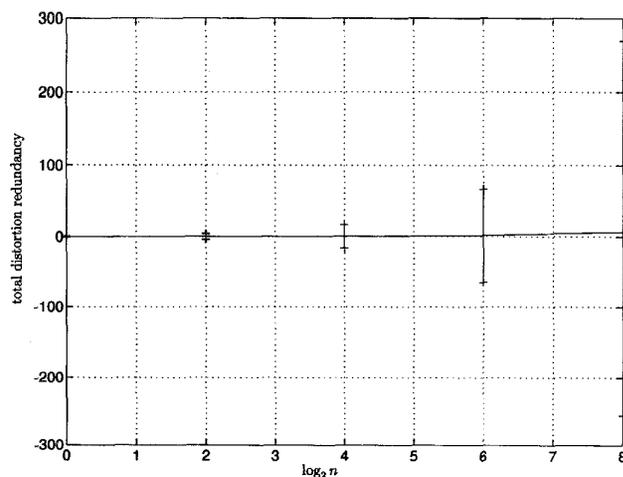


Fig. 24. Total distortion redundancy for variable-rate coding on  $N(\mu, \sigma^2)$  with  $\mu \sim N(0, 1)$ ,  $\sigma \sim N(0, \pi/2)$ , and  $\lambda = 0.393$ .

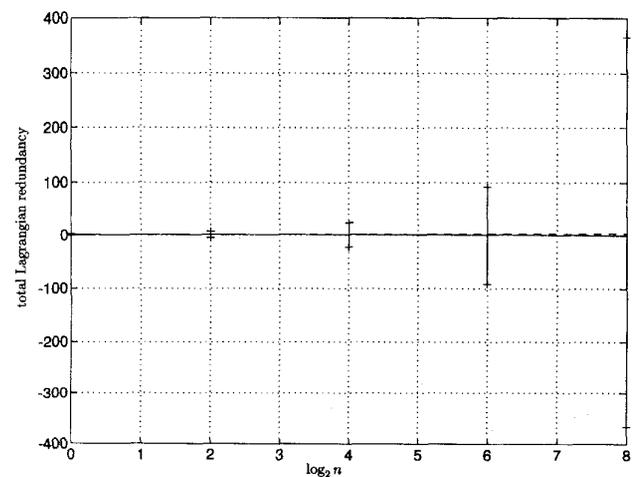


Fig. 26. Total Lagrangian redundancy for variable-rate coding on  $N(\mu, \sigma^2)$  with  $\mu \sim N(0, 1)$ ,  $\sigma \sim N(0, \pi/2)$ , and  $\lambda = 0.393$ . The dashed line shows the predicted slope for  $k = 2$ .

while the distortion achieved by such a system is equal to the distortion that would be achieved by a system with only an omniscient encoder, the rate is slightly lower, since the cost of describing the codebook to the decoder is not included.) If all of the images are concatenated together and sent to this hypothetical system, the performance on the mixture source would be the average of the performances on the individual graphs (weighted by the images' sizes); this average performance is shown in Fig. 28.

The earlier assumption that the encoder and decoder are both omniscient is unrealistic. We now remove this assumption and again consider vector quantizer performance on the mixture of sources. If a standard full search VQ is trained on one image source and tested on another, mismatch will occur, resulting in suboptimal performance, as shown in Fig. 29, where the performance of a VQ trained on MR brain scans and tested on a portrait is compared with the performance of a VQ trained and tested on portraits. Fig. 30 compares the performance of a

standard full-search VQ to a WUVQ with four codebooks and  $n = 64$ . Both quantizers are trained and tested on the mixture of sources. While the standard VQ is trained to do well on average across the images, it will not truly match any of the sources. The WUVQ, on the other hand, can better match the sources given, and therefore more closely approximates the "optimal" results of Fig. 28, here reproduced as individual points.

Notice that in this example, as in all examples where we compare the performance of one- and two-stage codes, the performance of an  $(n, l)$ -dimensional two-stage code is always compared to the performance of its  $l$ -, rather than  $n$ -, dimensional counterpart. That is, the performance of an  $(n, l)$ -dimensional code is compared either to the performance of the best possible  $l$ -dimensional code or to the performance of a suboptimal  $l$ -dimensional code. This choice is made despite the

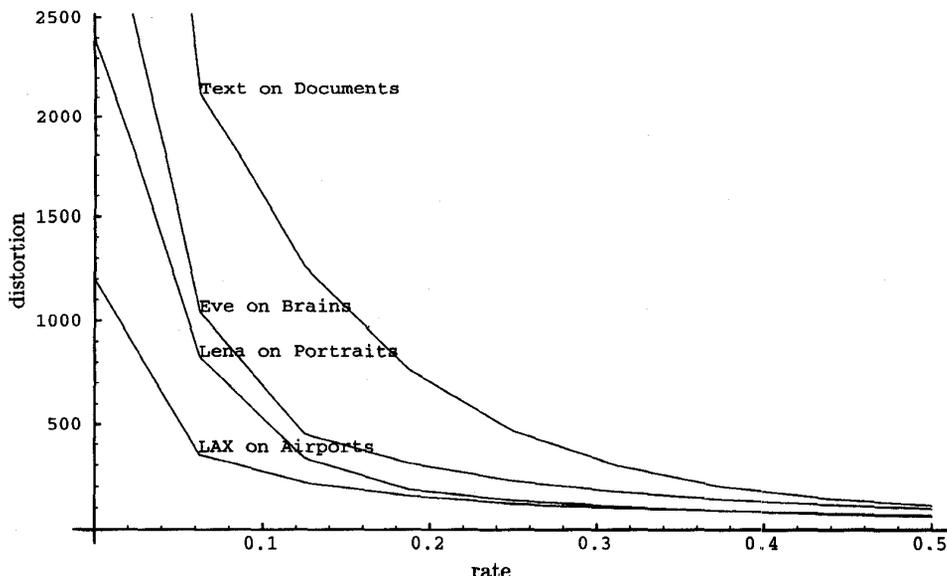


Fig. 27. Distortion-rate results for fixed-rate coding of each member of the mixture test set with a standard full search VQ trained for that data type.

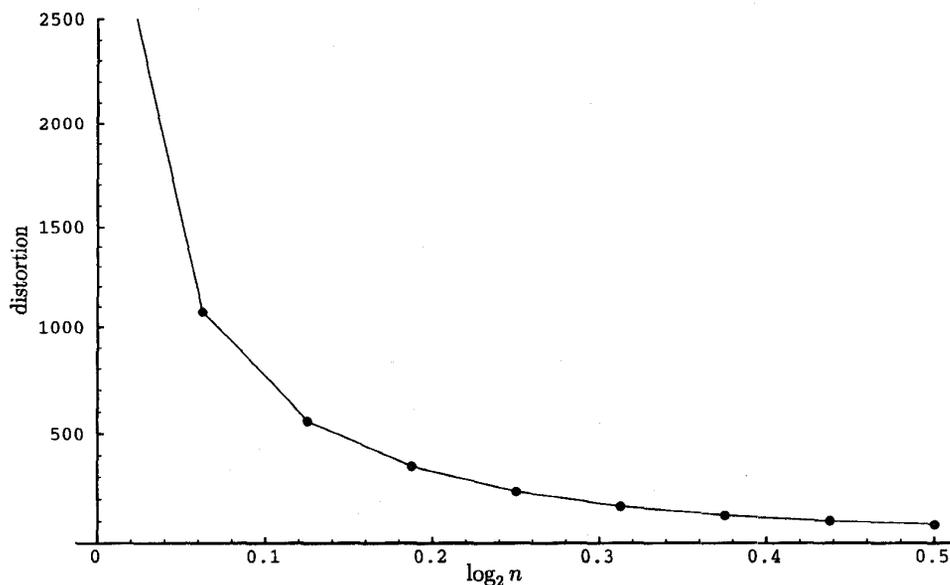


Fig. 28. Average distortion-rate results for fixed-rate coding of each member of the mixture test set with a standard full search VQ trained for that data type.

fact that the  $(n, l)$ -dimensional code is, technically speaking, an  $n$ -dimensional block code, suffering a coding delay of length  $n$ . The motivation for this choice is two-fold. First, the performance of any  $(n, l)$ -dimensional two-stage code is bounded below by the optimal  $l$ th-order performance, as described in Section III. Second, since  $(n, l)$ -dimensional codes use  $l$ -dimensional codewords, the computational complexity of an  $(n, l)$ -dimensional WUVQ is proportional to the computational complexity of an  $l$ -dimensional VQ. (The computational complexity of a vector quantizer grows exponentially with the dimension of its codewords.)

The previous example illustrates the performance improvements to be gained by using WUVQ on mixed sources, but is itself based on a questionable assumption: the assumption that each of its component subsources is a single source. In actuality, each type of image is produced by a variety of subsources. A brain image, for example, is the mixture of signals produced by bone, muscle, fat, etc. Thus the quantization of brain images by a single codebook designed on a mixture of brain images is subject to the same shortcomings as the quantization of the previous mixture of sources by any single codebook system. While it is difficult to break up

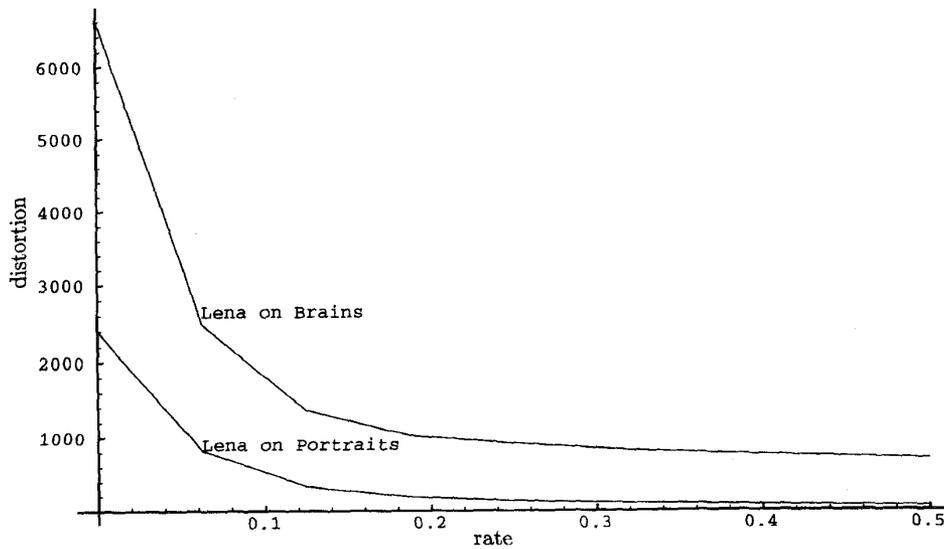


Fig. 29. Distortion-rate results for fixed-rate coding of a portrait (Lena) with a standard full search VQ trained on brain images compared to fixed-rate coding of the same image with a standard full search VQ trained on portraits.

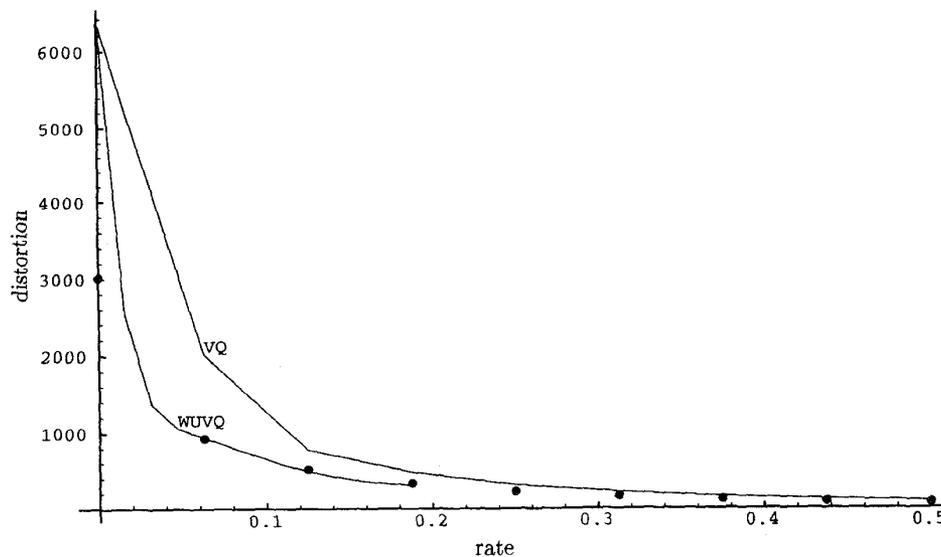


Fig. 30. Comparison of average distortion-rate results on the mixture of sources for three fixed-rate VQ systems. From top to bottom: a single full search VQ trained on all training images, WUVQ trained on all training images, and a collection of VQ's, one trained for each source type, in which each test image is coded with the VQ designed for its type.

the brain images into their component subsources, we can, nonetheless illustrate the performance gains to be achieved by moving from single to multi-codebook VQ systems.

We tested fixed-rate and variable-rate WUVQ on a sequence of medical images and compared them to standard full-search VQ and ECVQ, respectively. The training sequence in each case consisted of twenty  $256 \times 256$  MR images and the test sequence consisted of five MR images outside the training sequence. Performance results are reported in terms of signal-to-quantization-noise ratio (SQNR). In this case, we choose a vector dimension of  $l = 4$  and consider two by two collections of vectors at the first-stage level, giving  $n = 16$ .

Fixed rate systems consist of up to 512 codebooks with up to 16 codewords per codebook. Codebook sizes for first- and second-stage codebooks in the case of variable-rate coding were initialized to 256 and 4, respectively.

Fig. 31 shows the distortion-rate results of fixed-rate coding on the test sequence from the medical data set. Each curve represents a constant value of  $R$ , and is labeled with the appropriate  $R$  value. The lower dashed curve corresponds to the performance of a standard full-search VQ, which is equivalent to a fixed-rate WUVQ at  $\tilde{R} = 0$ . Fig. 32 shows the corresponding entropy-constrained results. In this case, we consider both fixed-rate first-stage codes with fixed-

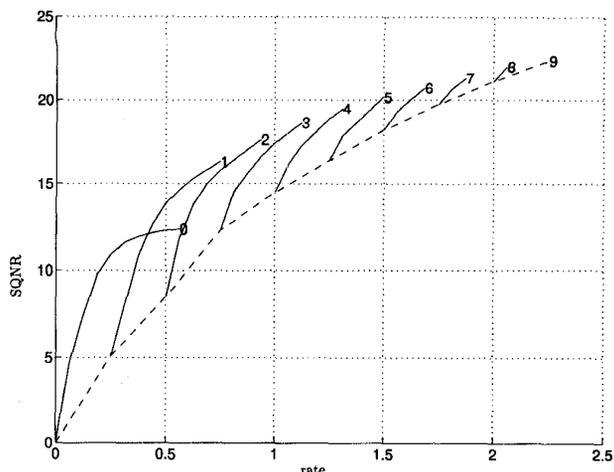


Fig. 31. Distortion-rate results for fixed-rate coding of medical test sequence. WUVQ performance (solid lines) is compared to standard full search VQ performance (dashed line).

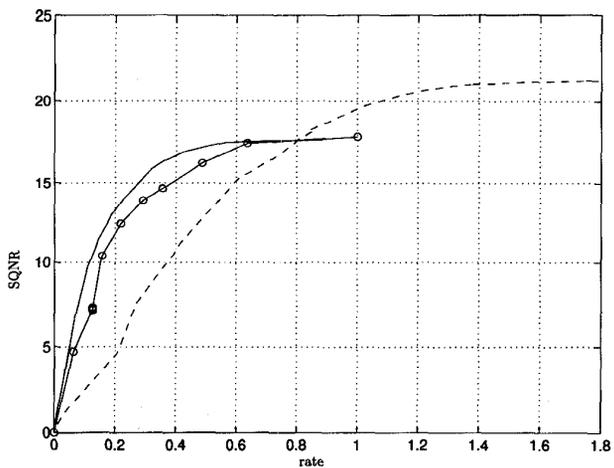


Fig. 32. Distortion-rate results for variable-rate coding of medical test sequence. WUVQ performance (solid line) and a collection of fixed rate codes where the codes are not constrained to be of the same rate (circles) are compared to ECVQ performance (dashed line).

rate second-stage codes of varying rates and variable-rate first-stage codes with variable-rate second-stage codes and compare the resulting performance to that of an ECVQ. In both fixed-rate and variable-rate coding, significant gains are demonstrated. Indeed, at bit rates around 0.25–0.50 bits per pixel, there is a 4–5-dB gain of fixed-rate WUVQ over standard VQ, and about 7–8-dB gain of variable-rate WUVQ over ECVQ. Since ECVQ already represents a substantial improvement over entropy-coded standard VQ, variable-rate WUVQ achieves more than a 9-dB gain over entropy-coded standard VQ.

## VI. SUMMARY AND CONCLUSION

In Section II (Preliminaries) we defined redundancy and universality for noiseless coding, fixed-rate quantization, and variable-rate quantization. We showed the relationships between various definitions of redundancy and universality, and

gave conditions for the existence of weighted universal and weakly minimax universal codes. In particular, we proved that weighted universal codes exist for a mixture of ergodic sources (i.e., for a nonergodic source) if and only if the OPTA for the mixture source has an ergodic decomposition. Finally, we argued that the Lagrangian redundancy can be used to characterize universal noiseless codes and universal fixed-rate quantizers as well as universal variable-rate quantizers.

In Section III (The Vector Quantization Approach) we described two-stage universal coding and introduced the notion of the first-stage quantizer. The first-stage quantizer is a vector quantizer whose codebook is a collection of second-stage codes. We showed that the two-stage code with minimum redundancy has a first-stage quantizer satisfying a nearest neighbor condition: the input vector is mapped to the second-stage code that best codes the data. This condition reduces to the minimum description length principle in the noiseless case, and it parallels the optimal encoder condition for ordinary and entropy-constrained vector quantization. We introduced the notion of an omniscient first-stage quantizer, and showed that an omniscient first-stage quantizer can perform no better than a nonomniscient quantizer. We defined the divergence of an omniscient first-stage quantizer. The divergence is equivalent to the information divergence, or relative entropy, in the noiseless case. Using the divergence, we generalized the achievability part of Rissanen's theorem from noiseless coding to quantization. Our theorem states that under appropriate smoothness conditions on the divergence, if the collection of sources  $\{P_\theta: \theta \in \Lambda\}$  is parametric with  $\Lambda \subseteq \mathbb{R}^k$ , then the optimal  $n$ th-order redundancy is at most  $(k/2)n^{-1}(\log n + c)$  for some constant  $c$  depending on  $\theta$ . We provided an example of a collection of stationary ergodic sources satisfying the smoothness conditions of the theorem. We also showed that if the collection of sources is finite or countable, then the optimal  $n$ th-order redundancy is at most  $\lambda b/n$ , for some constant  $b$  depending on  $\theta$ . We then shifted from minimax to weighted universal results. We showed that, in general, the tail of the divergence-rate function of the omniscient first-stage quantizer determines the rate at which the expected redundancy of the overall two-stage quantizer can converge to zero. We examined three cases. First, when  $\Lambda$  is finite-dimensional, and the divergence is locally quadratic (i.e., smooth),  $\Theta$  can be quantized to an average divergence decaying exponentially with rate  $A2^{-2\bar{R}/k}$ . This leads to a  $(k/2)n^{-1}(\log n + c)$  rate of convergence of the expected  $n$ th-order redundancy to zero. Second, when  $\Lambda$  is countable,  $\Theta$  can be coded losslessly at a sufficiently high rate (equal to the entropy of  $\Theta$ ). This leads to a rate of convergence equal to  $H(\Theta)/n$ . And third, when  $\Lambda$  is infinite-dimensional, under appropriate conditions  $\Theta$  can be quantized to an average divergence decaying with rate as a power law  $A\bar{R}^{-b}$ . This leads to a  $O(1/n^{1-c})$  rate of convergence. The convergence of the expected  $n$ th-order redundancy to zero at these rates requires the dimension  $l$  of the second-stage code to equal the overall two-stage blocklength  $n$ . If  $l$  is less than  $n$ , then these same rates of convergence hold for the difference between the expected performance of the two-stage code with blocklength  $n$  and the expected  $l$ th-order OPTA (rather than the  $n$ th-order OPTA). If  $l$  is fixed, and  $k$  is the number of parameters in the second-

stage code (i.e.,  $k$  is the dimension of the output space of the omniscient first-stage quantizer rather than the dimension of the input space), then the difference between the expected performance of the two-stage code with blocklength  $n$  and the expected  $l$ th-order OPTA is again at most  $(k/2)n^{-1}(\log n + c)$ . This is the most practical result for applications.

In Section IV (The Design Algorithm) we described an algorithm for designing locally optimal two-stage codes for any fixed overall blocklength  $n$  and second-stage dimension  $l$ . The algorithm is formally equivalent to the generalized Lloyd algorithm, but is applied to the first-stage quantizer of the two-stage code. Since the codebook of the first-stage quantizer is a collection of second-stage codes with dimension  $l$ , the nearest neighbor encoding step of the algorithm assigns each data block of length  $n$  to the second-stage code that best codes the data according to the appropriate performance measure. The centroiding step of the algorithm redesigns the set of second-stage codes so that each second-stage code optimally codes those vectors of dimension  $l$  that were assigned to it in the encoding step. The centroiding step is accomplished by the Huffman algorithm if the second-stage codes are noiseless codes, and by the generalized Lloyd algorithm if the second-stage codes are quantizers. If the second-stage codes are variable-rate quantizers, then the appropriate entropy-constrained version of the generalized Lloyd algorithm is used.

In Section V (Experimental Results) we experimentally confirmed our theoretical rate-of-convergence results on sources with one and two parameters chosen randomly according to a set prior. We observed for two-stage noiseless codes, fixed-rate quantizers, and variable-rate quantizers designed by our algorithm that the total average redundancy grows as  $(k/2)\log n$  for  $k = 1, 2$ , within reasonable confidence limits. We applied the algorithm to real image data, and observed that two-stage vector quantizers designed by our algorithm with  $n = 8 \times 8$  and  $l = 4 \times 4$  perform as well on each type of image modality considered (portraits, aerial views, medical images, and text) as a  $4 \times 4$  vector quantizer designed explicitly for the source, at a variety of rates. Trained on medical images only, a two-stage fixed-rate vector quantizer with  $n = 4 \times 4$  and  $l = 2 \times 2$  outperforms a standard  $2 \times 2$  vector quantizer on medical images outside the training set by 4–5 dB. If the two-stage vector quantizer is entropy-constrained, it gains another 4–5 dB for a total performance gain more than 9 dB over entropy-coded standard vector quantization.

One of the most important contributions of this paper may be a rule of thumb for how finely to quantize a codebook when transmitting it as side information and how often to transmit it. Suppose that for each block of data containing  $n = 2^{16}$  samples (such as a  $256 \times 256$  image), we wish to design a fixed-rate vector quantizer with dimension  $l = 4$  at rate  $R = 1$  bit per pixel, and transmit the codebook along with the encoded data. How finely should we quantize the codebook? Equivalently, how many bits should we spend on the codebook? To answer that, we invoke Corollary 6 of Section III. The number of parameters in the codebook is  $k = l2^{lR} = 64$ , so the optimal number of bits to spend on the codebook is approximately  $\tilde{R}_n^* = (k/2)\log cn$  in order to

achieve excess distortion (due to quantization of the codebook) approximately  $\tilde{D}_l(\tilde{R}_n^*) = A/cn$ . Here,  $c$  is an arbitrarily chosen constant and  $A$  is at most the variance of the mixture source. Choosing  $c = 1$ , we use  $\tilde{R}_n^* = (64/2)16 = 512$  bits, or 8 bits per parameter. In principle we could vector-quantize the codebook to 512 bits using a codebook of codebooks designed using the algorithm of Section IV. However, that would be impractical for such a large  $\tilde{R}_n^*$ . More practical would be to use a product quantizer, at the expense of a slight increase in the constant  $A$  due to loose sphere packing (see [38]). Zeger, Bist, and Linder [55], [56] independently vector-quantize each  $l$ -dimensional codevector. However, even this might be impractical at a rate of 32 bits per 4-dimensional codevector. The simplest scheme is to independently scalar-quantize each component of each codevector to 8 bits of precision. Indeed, this scheme is the basis of the following heuristic derivation of the rule of thumb. Of the  $n/l$  vectors in the block of data, approximately  $(n/l)/2^{lR} = n/k$  will be encoded to each of the  $2^{lR}$  codewords. Each codeword is the average of the vectors falling into its cell. The variance of the components of the vectors falling into a cell is somewhere between  $A/(2^R)^2$  and  $A/(2^{lR})^2$ , approximately, depending on whether the width of a cell is about  $2^{-R}$  times the range of data or only  $2^{-lR}$  times the range. (The latter occurs when the data are highly correlated and all  $2^{lR}$  codewords lie along a line.) Thus the variance of each component of a codeword is  $A/cn$ , where  $c$  is somewhere between  $2^{2R}/k$  and  $2^{2lR}/k$ . Therefore, each component is known to a precision approximately  $\sqrt{A/cn}$  (one standard deviation), and there are about  $\sqrt{A}/\sqrt{A/cn} = \sqrt{cn}$  bins of that precision. Hence about  $(1/2)\log cn$  bits are required to specify a bin for each component, or  $(k/2)\log cn$  bits to specify the whole codebook. Conversely, if we know *a priori* that we want to spend 8 bits per component, then it is likewise possible to determine how often the codebook should be transmitted, by solving for  $n$ .

In sum, this paper presents theory, algorithms, and practical guidelines for soundly developing systems for two-stage universal coding.

#### APPENDIX

*Lemma 1:* Assume the conditions for the ergodic decomposition (2) hold, and assume  $n^{-1}R(C^n) \rightarrow R$ . Then  $E|\delta(C^n|\Theta)| \rightarrow 0$  if and only if  $E\delta(C^n|\Theta) \rightarrow 0$ .

*Proof:* The “only if” part is trivial. For the “if” part, suppose  $E\delta(C^n|\Theta) \rightarrow 0$ . Now (see the top of the following page) where  $\epsilon_n = 0$  if  $R_n \triangleq n^{-1}R(C^n) \leq R$ , and if  $R_n > R$

$$\begin{aligned} 0 &\leq \epsilon_n \\ &= \int_{\theta: n^{-1}D_\theta(C^n) < \bar{D}_\theta(R)} \left[ \bar{D}_\theta(R) - \frac{1}{n}D_\theta(C^n) \right] dW(\theta) \\ &\leq \int_{\theta: n^{-1}D_\theta(C^n) < \bar{D}_\theta(R)} [\bar{D}_\theta(R) - \bar{D}_\theta(R_n)] dW(\theta) \\ &\leq \int [\bar{D}_\theta(R) - \bar{D}_\theta(R_n)] dW(\theta) \\ &= \int \bar{D}_\theta(R) dW(\theta) - \int \bar{D}_\theta(R_n) dW(\theta) \\ &= \bar{D}(R) - \bar{D}(R_n) \end{aligned}$$

$$\begin{aligned}
 E|\delta(C^n|\Theta)| &= \int \left| \frac{1}{n} D_\theta(C^n) - \bar{D}_\theta(R) \right| dW(\theta) \\
 &= \int_{\theta: n^{-1} D_\theta(C^n) \geq \bar{D}_\theta(R)} \left[ \frac{1}{n} D_\theta(C^n) - \bar{D}_\theta(R) \right] dW(\theta) \\
 &\quad + \int_{\theta: n^{-1} D_\theta(C^n) < \bar{D}_\theta(R)} \left[ \bar{D}_\theta(R) - \frac{1}{n} D_\theta(C^n) \right] dW(\theta) \\
 &= \int \left[ \frac{1}{n} D_\theta(C^n) - \bar{D}_\theta(R) \right] dW(\theta) \\
 &\quad + 2 \int_{\theta: n^{-1} D_\theta(C^n) < \bar{D}_\theta(R)} \left[ \bar{D}_\theta(R) - \frac{1}{n} D_\theta(C^n) \right] dW(\theta) \\
 &= E\delta(C^n|\Theta) + 2\epsilon_n
 \end{aligned}$$

where the last equality follows from the ergodic decomposition (2). Since  $\bar{D}(R)$  is continuous in  $R$  [25], [30],  $\bar{D}(R_n) \rightarrow \bar{D}(R)$  as  $R_n \rightarrow R$ , and so  $\epsilon_n \rightarrow 0$ . Hence  $E|\delta(C^n|\Theta)| \rightarrow 0$ .  $\square$

*Lemma 2:*

$$\min_R [D(R) + \lambda R] = \inf_{n, C^n} \left[ \frac{1}{n} D(C^n) + \frac{\lambda}{n} R(C^n) \right].$$

*Proof:*  $\geq$ : Choose  $R^*$  such that

$$\hat{D}(R^*) + \lambda R^* = \min_R [\hat{D}(R) + \lambda R]$$

and choose  $n^*, C_{n^*}^*$  such that

$$(1/n^*) R(C_{n^*}^*) \leq R^*$$

and

$$(1/n^*) D(C_{n^*}^*) \leq \hat{D}(R^*) + \epsilon.$$

Then

$$\begin{aligned}
 \min_R [\hat{D}(R) + \lambda R] &= \hat{D}(R^*) + \lambda R^* \geq (1/n^*) D(C_{n^*}^*) \\
 &\quad + (\lambda/n^*) R(C_{n^*}^*) - \epsilon \\
 &\geq \inf_{n, C^n} [n^{-1} D(C^n) + \lambda n^{-1} R(C^n)] - \epsilon
 \end{aligned}$$

$\epsilon$  arbitrary.

$\leq$ : For any  $n, C^n$

$$\begin{aligned}
 n^{-1} D(C^n) + \lambda n^{-1} R(C^n) \\
 &\geq D(n^{-1} R(C^n)) + \lambda n^{-1} R(C^n) \\
 &\geq \min_R [D(R) + \lambda R].
 \end{aligned}$$

Take the infimum over  $n, C^n$ .  $\square$

*Lemma 3:* Let  $\{C^n\}$  be a weakly minimax universal code, with  $H(X_1) < \infty$  if  $\{C^n\}$  is a noiseless code, and  $E d(X_1, a^*) < \infty$  for some  $a^* \in \hat{\mathcal{X}}$  if  $\{C^n\}$  is a quantizer. Then there exists a weighted universal code  $\{C_w^n\}$  with performance within one bit of  $\{C^n\}$  for each  $n$ .

*Proof:* Let  $C_e^n$  be a code whose expected performance given  $\theta$  is dominated by an integrable function  $f$  of  $\theta$ . In particular, for noiseless coding, let  $C_e^n$  be a code matched to an iid source of  $X$ 's each with distribution  $P^1$ . Then the instantaneous rate of  $C_e^n$  is

$$r_e(x^n) \leq - \sum_{i=1}^n \log p^1(x_i) + 1$$

so that

$$n^{-1} R_\theta(C_e^n) = n^{-1} E_\theta r_e(X^n) \leq -E_\theta \log p(X_1) + 1 \triangleq f(\theta)$$

where

$$\int f(\theta) dW(\theta) = H(X_1) + 1 < \infty.$$

For fixed-rate quantization, let  $C_e^n$  be a quantizer mapping each  $x^n$  to  $(a^*, a^*, \dots, a^*)$ , with rate 0 if  $C^n$  is a variable-rate quantizer and with rate  $R(C^n)$  if  $C^n$  is a fixed-rate quantizer. Then

$$d(x^n, C_e(x^n)) = \sum_{i=1}^n d(x_i, a^*)$$

so that

$$\begin{aligned}
 n^{-1} D_\theta(C_e^n) &= n^{-1} E_\theta d(X^n, C_e^n(X^n)) \\
 &= E_\theta d(X_1, a^*) \triangleq f(\theta)
 \end{aligned}$$

where again

$$\int f(\theta) dW(\theta) = E d(X_1, a^*) < \infty.$$

Thus the existence of a dominating integrable function  $f(\theta)$  is guaranteed by the applicable regularity condition. Then let  $C_w^n$  be  $C^n$  with a 1-bit escape

$$\alpha_w(x^n) = \begin{cases} 0\alpha(x^n), & \text{if } C^n \text{ codes } x^n \text{ better than } C_e^n \\ 1\alpha_e(x^n), & \text{otherwise} \end{cases}$$

where  $\alpha_w, \alpha_e$ , and  $\alpha$  are the encoders for  $C_w^n, C_e^n$ , and  $C^n$ , respectively,  $0\alpha(x^n)$  denotes the concatenation of the binary strings "0" and  $\alpha(x^n)$ , and  $1\alpha_e(x^n)$  denotes the concatenation of the binary strings "1" and  $\alpha_e(x^n)$ . Here, " $C^n$  codes  $x^n$ "

better than  $C_e^n$  means  $r(x^n) < r_e(x^n)$  for noiseless coding,  $d(x^n, C(x^n)) < d(x^n, C_e(x^n))$  for fixed-rate quantization, and  $d(x^n, C(x^n)) + \lambda r(x^n) < d(x^n, C_e(x^n)) + \lambda r_e(x^n)$  for variable-rate quantization. Then, clearly, the performance of  $C_w^n$  is at least as good as the performance of  $C^n$ , and no worse than the performance of  $C_e^n$  (except for the small penalty in rate for the 1-bit escape). Hence, like  $C^n$ ,  $C_w^n$  is weakly minimax, and like  $C_e^n$ , its expected performance given  $\theta$  is dominated by an integrable function of  $\theta$ . Thus by the dominated convergence theorem, the expected performance of  $C_w^n$  given  $\theta$  converges in  $L^1(W)$ , and hence  $C_w^n$  is weighted universal.  $\square$

*Theorem 1:* Suppose  $\{X_i\}$  is stationary and ergodic for each  $\theta$ . Weighted universal codes exist for  $(\{X_i\}, \Theta)$  if and only if the OPTA has an ergodic decomposition.

*Proof:* For noiseless coding,  $\{C^n\}$  is weighted universal if and only if

$$\begin{aligned} E\rho(C^n|\Theta) &= \int \left[ \frac{1}{n} R_\theta(C^n) - \bar{H}_\theta \right] dW(\theta) \\ &= \frac{1}{n} R(C^n) - \int \bar{H}_\theta dW(\theta) \end{aligned} \quad (46)$$

goes to zero as  $n \rightarrow \infty$ . Thus weighted universal sequences of noiseless codes exist if and only if the entropy rate of  $P$  satisfies

$$\bar{H} = \int \bar{H}_\theta dW(\theta) \quad (47)$$

since if (47) holds, then because there exists a sequence  $\{C^n\}$  with  $n^{-1}R(C^n) \rightarrow \bar{H}$ , (46) goes to zero. On the other hand,  $n^{-1}R(C^n) \geq \bar{H}$  (by the converse to the source coding theorem), and

$$\bar{H} \geq \int \bar{H}_\theta dW(\theta)$$

(by the concavity of  $H(X^n|\theta)$  and Fatou's lemma), so that if (46) goes to zero, then the ergodic decomposition (47) must hold.

For fixed-rate quantization, the situation is similar.  $\{C^n\}$  is weighted universal if and only if

$$\begin{aligned} E\delta(C^n|\Theta) &= \int \left[ \frac{1}{n} D_\theta(C^n) - \bar{D}_\theta(R) \right] dW(\theta) \\ &= \frac{1}{n} D(C^n) - \int \bar{D}_\theta(R) dW(\theta) \end{aligned} \quad (48)$$

goes to zero as  $n \rightarrow \infty$ . But (48) goes to zero if and only if the operational distortion-rate function of  $P$  satisfies

$$\hat{D}(R) = \int \bar{D}_\theta(R) dW(\theta) \quad (49)$$

since if (49) holds, then because there exists a sequence  $\{C^n\}$  with  $n^{-1}R(C^n) \leq R$  and  $n^{-1}D(C^n) \rightarrow \hat{D}(R)$ , (48) goes to

zero. On the other hand,  $n^{-1}D(C^n) \geq \hat{D}(R)$  and

$$\begin{aligned} \hat{D}(R) &= \lim_n \inf_{\{C^n\}} \int D_x(C^n) dP(x) \\ &\geq \lim_n \int \inf_{\{C^n\}} D_x(C^n) dP(x) \\ &\geq \int \lim_n \inf_{\{C^n\}} D_x(C^n) dP(x) \\ &= \int D_x(R) dP(x) \\ &= \int \int D_x(R) dP_\theta(x) dW(\theta) \\ &= \int \bar{D}_\theta dW(\theta) \end{aligned}$$

so that if (48) goes to zero, then the ergodic decomposition (49) must hold.

For variable-rate quantization,  $\{C^n\}$  is weighted universal if and only if

$$\begin{aligned} E\ell(C^n, \lambda|\Theta) &= \int \left[ \frac{1}{n} L_\theta(C^n, \lambda) - L_\theta(\lambda) \right] dW(\theta) \\ &= \frac{1}{n} L(C^n, \lambda) - \int L_\theta(\lambda) dW(\theta) \end{aligned} \quad (50)$$

goes to zero as  $n \rightarrow \infty$ . But (50) goes to zero if and only if the distortion-rate Lagrangian of  $P$  satisfies

$$L(\lambda) = \int L_\theta(\lambda) dW(\theta) \quad (51)$$

since if (51) holds, then because there exists a sequence  $\{C^n\}$  with  $n^{-1}L(C^n, \lambda) \rightarrow L(\lambda)$ , (50) goes to zero. On the other hand,  $n^{-1}L(C^n, \lambda) \geq L(\lambda)$  (by the converse to the source coding theorem), and  $L(\lambda) \geq \int L_\theta(\lambda) dW(\theta)$  (by the concavity of the mutual information  $I(X^n; Y^n)$  in  $p_\theta(x^n)$  and Fatou's lemma [22, proof of Theorem 2]), so that if (50) goes to zero, then the ergodic decomposition (51) must hold.  $\square$

*Lemma 4:* Suppose that  $\delta_\theta^n$  and  $\rho^n$  are two sequences of positive numbers converging to zero such that for each rate  $R$  there exists a sequence of fixed-rate block quantizers  $C^n$  for which  $n^{-1}R(C^n) \leq R + \rho^n$  and  $n^{-1}D_\theta(C^n) \leq \bar{D}_\theta(R) + \delta_\theta^n$ . Then for each rate  $R_0 > 0$  there exists a sequence of fixed-rate block quantizers  $C^n$  for which  $n^{-1}R(C^n) \leq R_0$  and

$$n^{-1}D_\theta(C^n) \leq \bar{D}_\theta(R_0) + \delta_\theta^n + \lambda\rho^n$$

for sufficiently large  $n$ , for any  $\lambda > -(d/dR)\bar{D}_\theta(R_0)$ .

*Proof:* Since  $D(R)$  is convex  $\cup$ , there exists a rate  $R_1 < R_0$  such that for all  $R \in [R_1, R_0]$

$$\bar{D}_\theta(R) \leq \bar{D}_\theta(R_0) + \lambda(R_0 - R).$$

Let  $n$  be sufficiently large so that  $R = R_0 - \rho^n$  lies in  $[R_1, R_0]$ . Then there exists a fixed-rate block quantizer  $C^n$  for which  $n^{-1}R(C^n) \leq R + \rho^n = R_0$  and

$$n^{-1}D_\theta(C^n) \leq \bar{D}_\theta(R) + \delta_\theta^n \leq \bar{D}_\theta(R_0) + \lambda\rho^n + \delta_\theta^n. \quad \square$$

*Lemma 5:* For each  $x^n$ , there exists some  $\tilde{s} \in \tilde{S}$  for which

$$\inf_{\tilde{s} \in \tilde{S}} [d(x^n, \tilde{\beta}(\tilde{s})) + (\lambda/n)|\tilde{s}|]$$

is achieved.

*Proof:* Arbitrarily pick  $\tilde{s}_0 \in \tilde{\mathcal{S}}$ . There can be only a finite number of distinct strings  $\tilde{s} \in \tilde{\mathcal{S}}$  with

$$d(x^n, \tilde{\beta}(\tilde{s})) + (\lambda/n)|\tilde{s}| \leq d(x^n, \tilde{\beta}(\tilde{s}_0)) + (\lambda/n)|\tilde{s}_0|.$$

Hence the infimum is achieved.  $\square$

*Lemma 6:* Let  $X_1, X_2, \dots$  be any real-valued stationary random process with unknown mean  $\mu$  and unknown standard deviation  $0 < \sigma < \infty$ , and let  $P_\theta, \theta = (\mu, \sigma) \in \Lambda$  be its process measure. Then under the squared-error distortion measure, for all  $l, \theta$ , and  $\hat{\theta}$ ,  $\Delta_l(\theta || \hat{\theta}) \leq \|\theta - \hat{\theta}\|^2$ .

*Proof:* If  $C_{(0,1)}^l$  is the optimal  $l$ -dimensional quantizer for  $P_{(0,1)}$ , then the quantizer obtained by scaling each component of  $C_{(0,1)}^l$  by  $\sigma$  and then translating by  $\mu$  is the optimal  $l$ -dimensional quantizer  $C_{(\mu,\sigma)}^l$  for  $P_{(\mu,\sigma)}$ . Thus

$$\begin{aligned} \Delta_l(\theta || \hat{\theta}) &= \tilde{d}(\theta, C_{\hat{\theta}}^l) - \tilde{d}(\theta, C_\theta^l) \\ &= l^{-1} E_\theta [\|X^l - \beta_{\hat{\theta}}(\alpha_{\hat{\theta}}(X^l))\|^2 \\ &\quad - \|X^l - \beta_\theta(\alpha_\theta(X^l))\|^2] \\ &\leq l^{-1} E_\theta [\|X^l - \beta_{\hat{\theta}}(\alpha_\theta(X^l))\|^2 \\ &\quad - \|X^l - \beta_\theta(\alpha_\theta(X^l))\|^2] \\ &= l^{-1} E_\theta [\|\beta_\theta(\alpha_\theta(X^l)) - \beta_{\hat{\theta}}(\alpha_\theta(X^l))\|^2] \\ &= l^{-1} E_\theta \sum_{i=1}^l [\beta_{\theta,i}(\alpha_\theta(X^l)) - \beta_{\hat{\theta},i}(\alpha_\theta(X^l))]^2 \\ &= l^{-1} E_\theta \sum_{i=1}^l [\beta_{\theta,i}(\alpha_\theta(X^l)) \\ &\quad - \left( \frac{\hat{\sigma}}{\sigma} (\beta_{\theta,i}(\alpha_\theta(X^l)) - \mu) + \hat{\mu} \right)]^2 \\ &= l^{-1} E_\theta \sum_{i=1}^l \left[ \left( 1 - \frac{\hat{\sigma}}{\sigma} \right) \beta_{\theta,i}(\alpha_\theta(X^l)) \right. \\ &\quad \left. + \left( \frac{\mu \hat{\sigma}}{\sigma} - \hat{\mu} \right) \right]^2 \\ &= l^{-1} E_\theta \sum_{i=1}^l \left[ \left( 1 - \frac{\hat{\sigma}}{\sigma} \right)^2 \beta_{\theta,i}^2(\alpha_\theta(X^l)) \right. \\ &\quad \left. + 2 \left( \frac{\mu \hat{\sigma}}{\sigma} - \hat{\mu} \right) \left( 1 - \frac{\hat{\sigma}}{\sigma} \right) \beta_{\theta,i}(\alpha_\theta(X^l)) \right. \\ &\quad \left. + \left( \frac{\mu \hat{\sigma}}{\sigma} - \hat{\mu} \right)^2 \right] \\ &\leq \left( 1 - \frac{\hat{\sigma}}{\sigma} \right)^2 E_\theta X_1^2 + \left( \frac{\mu \hat{\sigma}}{\sigma} - \hat{\mu} \right)^2 \\ &\quad + 2 \left( \frac{\mu \hat{\sigma}}{\sigma} - \hat{\mu} \right) \left( 1 - \frac{\hat{\sigma}}{\sigma} \right) E_\theta X_1 \\ &= \left( 1 - \frac{\hat{\sigma}}{\sigma} \right)^2 (\sigma^2 + \mu^2) + \left( \frac{\mu \hat{\sigma}}{\sigma} - \hat{\mu} \right)^2 \\ &\quad + 2 \left( \frac{\mu \hat{\sigma}}{\sigma} - \hat{\mu} \right) \left( 1 - \frac{\hat{\sigma}}{\sigma} \right) \mu \\ &= (\sigma - \hat{\sigma})^2 + (\mu - \hat{\mu})^2. \end{aligned}$$

*Lemma 7:* Let  $\Lambda$  be a subset of  $\mathbb{R}^k$  (bounded if we are considering fixed-rate coding but possibly unbounded otherwise). If  $\Lambda$  is unbounded suppose the differential entropy

$$h(\Theta) = - \int w(\theta) \log w(\theta) d\theta$$

is finite. Further suppose that for each  $l$  there is a constant  $m_l$  such that for all  $\theta$  and  $\hat{\theta}$ ,  $\Delta_l(\theta || \hat{\theta}) \leq m_l \|\theta - \hat{\theta}\|^2$ . Then there is a constant  $A_l$  depending on  $l$  such that

$$\tilde{D}_l(\tilde{R}) \leq \tilde{D}_l(\tilde{R}) = A_l 2^{-2\tilde{R}/k}.$$

If the constants  $\{m_l\}$  are bounded, then  $A_l = A$  does not depend on  $l$ .

*Proof:* The proof is similar to the proof of Theorem 3. Partition  $\mathbb{R}^k$  into a grid of hypercubes  $A_{n1}, A_{n2}, \dots$  each of side  $2^{-n}$ , such that for each  $n \geq 1$ , the partition  $\{A_{n1}, A_{n2}, \dots\}$  refines the partition  $\{A_{n-1,1}, A_{n-1,2}, \dots\}$ , where  $\{A_{01}, A_{02}, \dots\}$  is a partition of unit hypercubes. For each hypercube  $A_n \in \{A_{n1}, A_{n2}, \dots\}$  that intersects  $\Lambda$ , choose a representative  $\hat{\theta}_n \in A_n \cap \Lambda$ . Further represent the *unit* hypercubes that intersect  $\Lambda$  by a fixed-length code (if  $\Lambda$  is bounded) or by a variable-length code with finite entropy (if  $h(\Theta) < \infty$ ). Then  $\tilde{C}_{\theta}^{n,l}$  can quantize  $\theta$  using the code to specify  $A_0 \in \{A_{01}, A_{02}, \dots\}$  and then a fixed-length code to specify  $A_n$  within  $A_0$  (and hence to specify  $\hat{\theta}_n$  and its corresponding optimal length- $l$  code  $C_{\hat{\theta}_n}^l$ ). The expected rate of  $\tilde{C}_{\theta}^{n,l}$  is thus

$$\tilde{R}_n = R(\tilde{C}_{\theta}^{n,l}) = \tilde{R}_0 + kn$$

bits, where  $\tilde{R}_0$  is the expected length of the code for  $A_0$ , and the expected distortion is

$$\begin{aligned} D(\tilde{C}_{\theta}^{n,l}) &= E \Delta_l(\Theta || \hat{\Theta}_n) \\ &\leq E m_l \|\Theta - \hat{\Theta}_n\|^2 \\ &\leq m_l k (2^{-n})^2 \\ &= m_l k (2^{-(\tilde{R}_n - \tilde{R}_0)/k})^2. \end{aligned}$$

Hence for each rate  $\tilde{R} \in [\tilde{R}_n, \tilde{R}_{n+1}] = [\tilde{R}_n, \tilde{R}_n + k]$

$$\begin{aligned} \tilde{D}_l(\tilde{R}) &\leq D(\tilde{C}_{\theta}^{n,l}) \\ &\leq m_l k (2^{-(\tilde{R}_n - \tilde{R}_0)/k})^2 \\ &\leq m_l k (2^{-(\tilde{R} - k - \tilde{R}_0)/k})^2 \\ &= m_l k 2^{2(\tilde{R}_0 + k)/k} 2^{-2\tilde{R}/k}. \end{aligned}$$

The theorem is proved with  $A_l = m_l k 2^{2(\tilde{R}_0 + k)/k}$ . It is possible to obtain the optimal  $A_l$  with a more refined analysis involving asymptotic quantization theory.  $\square$

*Lemma 8:* Suppose  $E\|X^l\|^{2+\epsilon} < \infty$  for some  $\epsilon > 0$ . Then  $\tilde{D}_l(\tilde{R}) \leq A 2^{-2\tilde{R}/k}$  for some  $A < \infty$ ,

*Proof:* By Lemma 9 in this Appendix, for each  $s$  the random  $l$ -vector  $\beta_\Theta(s)$  also has a finite  $(2 + \epsilon)$  moment

$$E[\|\beta_\Theta(s)\|^{2+\epsilon} | \alpha_\Theta(X^l) = s] < \infty.$$

Then by [6, Theorem 2], for each  $s$  there exists a constant  $A_s < \infty$  such that for each rate, say  $\tilde{R}/k$  bits per sample,  $\square$

there exists a collection of  $\lfloor 2^{l\tilde{R}/k} \rfloor$   $l$ -vectors, say  $\Gamma_s^l$ , such that

$$l^{-1} E \left[ \min_{\beta \in \Gamma_s^l} \|\beta - \beta_{\Theta}(s)\|^2 \mid \alpha_{\Theta}(X^l) = s \right] \leq A_s 2^{-2\tilde{R}/k}.$$

Note there are  $2^{lR}$  such collections, one for each fixed length- $lR$  string  $s$ . Now let  $\Gamma^l$  be the collection of

$$\lfloor 2^{l\tilde{R}/k} \rfloor^{2^{lR}} \leq 2^{l2^{lR}\tilde{R}/k} = 2^{\tilde{R}}$$

codes corresponding to the Cartesian product of the  $\Gamma_s^l$ 's. Then for any code  $C_{\theta}^l = \beta_{\theta} \circ \alpha_{\theta}$ , there exists a code  $C_{\tilde{s}}^l = \beta_{\tilde{s}} \circ \alpha_{\tilde{s}} \in \Gamma^l$  such that

$$\beta_{\tilde{s}}(s) = \arg \min_{\beta \in \Gamma_s^l} \|\beta - \beta_{\theta}(s)\|^2$$

for each  $s$ . Thus for each  $\tilde{R}$  there exists an omniscient first-stage quantizer  $\tilde{C}_o^l$  with encoder  $\tilde{\alpha}_o(\theta) = \tilde{s}$  at fixed rate at most  $\tilde{R}$  such that

$$\begin{aligned} \tilde{D}_l(\tilde{R}) &\leq E \Delta(\Theta, \tilde{C}_o^l(\Theta)) \\ &\leq l^{-1} E \|\beta_{\tilde{s}}(\alpha_{\Theta}(X^l)) - \beta_{\Theta}(\alpha_{\Theta}(X^l))\|^2 \\ &= \sum_s P\{\alpha_{\Theta}(X^l) = s\} l^{-1} \\ &\quad \cdot E \|\beta_{\tilde{s}}(s) - \beta_{\Theta}(s)\|^2 \mid \alpha_{\Theta}(X^l) = s \\ &\leq \left( \sum_s P\{\alpha_{\Theta}(X^l) = s\} A_s \right) 2^{-2\tilde{R}/k}. \quad \square \end{aligned}$$

*Lemma 9:* For any  $\epsilon > 0$

$$E \|X^l\|^{2+\epsilon} \geq E \|\beta_{\Theta}(\alpha_{\Theta}(X^l))\|^{2+\epsilon}.$$

Hence if  $E \|X^l\|^{2+\epsilon} < \infty$ , then

$$\sum_s P\{\alpha_{\Theta}(X^l) = s\} E \|\beta_{\Theta}(s)\|^{2+\epsilon} \mid \alpha_{\Theta}(X^l) = s < \infty.$$

*Proof:* By Jensen's inequality,

$$E_{\theta} \|\beta_{\Theta}(s)\|^{2+\epsilon} \mid \alpha_{\Theta}(X^l) = s \geq (E_{\theta} \|X^l\|^2 \mid \alpha_{\Theta}(X^l) = s)^{1+\epsilon/2}$$

for each  $\theta$  and  $s$ . Hence

$$\begin{aligned} E_{\theta} \|X^l\|^{2+\epsilon} &= \sum_s P_{\theta}\{\alpha_{\Theta}(X^l) = s\} \\ &\quad \cdot E_{\theta} \|\beta_{\Theta}(s)\|^{2+\epsilon} \mid \alpha_{\Theta}(X^l) = s \\ &\geq \sum_s P_{\theta}\{\alpha_{\Theta}(X^l) = s\} \\ &\quad \cdot (E_{\theta} \|X^l\|^2 \mid \alpha_{\Theta}(X^l) = s)^{1+\epsilon/2} \\ &= \sum_s P_{\theta}\{\alpha_{\Theta}(X^l) = s\} \\ &\quad \cdot (E_{\theta} \|\beta_{\Theta}(s)\|^2 \mid \alpha_{\Theta}(X^l) = s \\ &\quad + \|\beta_{\Theta}(s)\|^2)^{1+\epsilon/2} \\ &\geq \sum_s P_{\theta}\{\alpha_{\Theta}(X^l) = s\} \|\beta_{\Theta}(s)\|^{2+\epsilon} \\ &= E_{\theta} \|\beta_{\Theta}(\alpha_{\Theta}(X^l))\|^{2+\epsilon}. \end{aligned}$$

Taking expectations over  $\theta$  yields the desired result.  $\square$

*Lemma 10:* Fix  $b > 0$  and  $K > 0$ . Let  $\Lambda$  be the set of all infinite-dimensional probability mass functions  $p = (p(1), p(2), \dots)$  such that for all  $n > K$ ,  $p(n) \leq u(n)$ , where  $u$  is the probability mass function  $u(n) = cn^{-(1+b)}$ . Then using relative entropy as the divergence measure, for any  $b' < b$ , there is a constant  $A_1$  such that

$$\tilde{D}_1(\tilde{R}) \leq A_1 \tilde{R}^{-b'}.$$

*Proof:* First we sketch the idea of the proof. Each probability mass function (pmf)  $p$  is associated with an optimal noiseless code with ideal codelengths  $-\log p(n)$ ,  $n = 1, 2, \dots$ . We shall quantize this optimal noiseless code, denoted  $-\log p$ , to another noiseless code, denoted  $-\log q$ , from a collection of  $2^{\tilde{R}}$  noiseless codes, and we shall show that the divergence

$$D(p \parallel q) = \sum_n p(n) \log(p(n)/q(n))$$

is at most  $A_1 \tilde{R}^{-b}$ . This will prove the theorem.

We quantize  $-\log p$  to  $-\log q$  using  $2^{\tilde{R}}$  bits as follows. The first  $k$  codelengths of  $-\log p$ , namely  $-\log p(1), \dots, -\log p(k)$ , are independently scalar quantized to  $\tilde{R}/k$  bits each and the remaining codelengths are set to the default codelength  $-\log u(n)$ , such that the resulting codelengths, say  $-\log q(1), -\log q(2), \dots$  satisfy the Kraft inequality, and a corresponding noiseless code  $-\log q$  can be constructed from them. For simplicity we allow ideal noninteger codelengths in this construction.

More precisely, let

$$P_k = \sum_{n=1}^k p(n)$$

and let

$$U_k = \sum_{n=1}^k u(n)$$

be the probabilities, according to pmfs  $p$  and  $u$ , respectively, that one of the first  $k$  symbols occurs, and let  $\bar{P}_k = 1 - P_k$  and  $\bar{U}_k = 1 - U_k$  be their complements. Let  $p_k(n) = p(n)/P_k$  for  $n = 1, \dots, k$  be the conditional probability (under  $p$ ) of the symbol given that it is at most  $k$ , and let  $\bar{u}_k(n) = u(n)/\bar{U}_k$  for  $n = k+1, k+2, \dots$  be the conditional probability (under  $u$ ) of the symbol given that it is at least  $k+1$ . Let  $p'_k(n) = \alpha p_k(n) + (1-\alpha)/k$  be a mixture of the conditional pmf  $p_k$  and the uniform pmf  $1/k$  such that  $p'_k(n) \geq \epsilon$ , where  $\epsilon < 1/k$ . Then  $-\log p'_k(n) \leq -\log \epsilon$ . To achieve  $\alpha p_k(n) + (1-\alpha)/k \geq \epsilon$ , set  $(1-\alpha)/k = \epsilon$ , or  $\alpha = 1 - k\epsilon$ , so that  $p'_k(n) = (1 - k\epsilon)p_k(n) + \epsilon$ , and thus

$$\begin{aligned} -\log p'_k(n) &= -\log [(1 - k\epsilon)p_k(n) + \epsilon] \\ &\leq -\log [(1 - k\epsilon)p_k(n)] \\ &= -\log p_k(n) + \log \frac{1}{1 - k\epsilon} \\ &\leq -\log p_k(n) + \left( \frac{1}{1 - k\epsilon} - 1 \right) \log e \\ &= -\log p_k(n) + \left( \frac{k\epsilon}{1 - k\epsilon} \right) \log e. \end{aligned}$$

Quantize  $-\log p'_k(n)$  to one of  $2^{\tilde{R}/k}$  levels uniformly in  $(0, -\log \epsilon]$ , but always quantize upwards to the next higher level, say  $-\log \hat{p}'_k(n)$ , so that

$$0 \leq -\log \hat{p}'_k(n) + \log p'_k(n) \leq (-\log \epsilon)2^{-\tilde{R}/k}.$$

Note that  $\hat{p}'_k(n)$  may sum to less than 1. Then let

$$-\log q(n) = \begin{cases} -\log U_k - \log \hat{p}'_k(n), & \text{for } n \leq k \\ -\log u(n), & \text{for } n > k. \end{cases}$$

Check to see that  $-\log q$  satisfies the Kraft inequality

$$\begin{aligned} \sum q(n) &= \sum_1^k U_k \hat{p}'_k(n) + \sum_{k+1}^{\infty} u(n) \\ &\leq U_k \sum_1^k p'_k(n) + \bar{U}_k = U_k + \bar{U}_k = 1. \end{aligned}$$

Note that  $q(n)$  may sum to less than 1, too. Nevertheless, we may compute the divergence between  $p$  and  $q$

$$D(p||q) = \sum_1^{\infty} p(n) \log(p(n)/q(n))$$

which is the expected redundancy of the noiseless code  $-\log q$  on source  $p$ . For  $n > k \geq K$ ,  $p(n) \leq u(n) = q(n)$ , so that for  $k > K$

$$\begin{aligned} D(p||q) &\leq \sum_1^k p(n) \log \frac{p(n)}{q(n)} \\ &= \sum_1^k p(n) \left( \log \frac{p(n)}{U_k} - \log \hat{p}'_k(n) \right) \\ &\leq \sum_1^k p(n) \left( \log \frac{p(n)}{U_k} - \log p'_k(n) \right. \\ &\quad \left. + (-\log \epsilon)2^{-\tilde{R}/k} \right) \\ &\leq \sum_1^k p(n) \left( \log \frac{p(n)}{U_k} - \log p_k(n) + \left( \frac{k\epsilon}{1-k\epsilon} \right) \right. \\ &\quad \left. \cdot \log e + (-\log \epsilon)2^{-\tilde{R}/k} \right) \\ &\leq \sum_1^k p(n) \log \frac{P_k}{U_k} + \left( \frac{k\epsilon}{1-k\epsilon} \right) \log e \\ &\quad + (-\log \epsilon)2^{-\tilde{R}/k}. \end{aligned}$$

But the first term is

$$P_k \log(P_k/U_k) \leq \log(1/U_k) \leq [(1/U_k) - 1] \log e = (\bar{U}_k/U_k) \log e$$

which for sufficiently large  $k$  is less than  $(1^+) \bar{U}_k \log e$ , where  $1^+ > 1$  is a constant arbitrarily close to 1. Furthermore, since  $u(n) = cn^{-(1+b)}$ , we have

$$\begin{aligned} \bar{U}_k &= \sum_{k+1}^{\infty} u(n) \leq \int_k^{\infty} cx^{-(1+b)} dx \\ &= -(c/b)x^{-b}|_k^{\infty} = (c/b)k^{-b}. \end{aligned}$$

Similarly

$$\int_1^{\infty} cx^{-(1+b)} dx \leq \sum_1^{\infty} cn^{-(1+b)} = 1$$

so that  $c/b \leq 1$ . Hence  $\bar{U}_k \leq k^{-b}$ . Now for sufficiently large  $k$  (uniformly in  $p$ )

$$D(p||q) \leq (1^+ \log e)k^{-b} + \left( \frac{k\epsilon}{1-k\epsilon} \right) \log e + (-\log \epsilon)2^{-\tilde{R}/k}.$$

We need to choose  $k$  and  $\epsilon$  properly as a function of  $\tilde{R}$  so as to minimize the bound on  $D(p||q)$ . We choose  $k = \tilde{R}/(b \log \tilde{R})$  and  $\epsilon = \tilde{R}^{-(1+b)}$ . Thus for sufficiently large  $\tilde{R}$  (uniformly in  $p$ )

$$\begin{aligned} D(p||q) &\leq (1^+ \log e) \left[ \frac{\tilde{R}}{b \log \tilde{R}} \right]^{-b} + (1^+ \log e) \frac{\tilde{R}^{-b}}{b \log \tilde{R}} \\ &\quad + (1+b)(\log \tilde{R})2^{-b \log \tilde{R}} \end{aligned}$$

which goes to zero as  $O(\tilde{R}^{-b'})$  for any  $b' < b$ .  $\square$

#### ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for suggestions that improved the quality of this paper.

#### REFERENCES

- [1] B. D. Andrews, M. Effros, P. A. Chou, and R. M. Gray, "A mean-removed variation of weighted universal vector quantization for image coding," in *Proc. Data Compression Conf.*, (IEEE Computer Soc., Snowbird, UT, Mar. 1993), pp. 302-309.
- [2] V. Balasubramanian, "A geometric formulation of Occam's razor for inference of parametric distributions," *IEEE Trans. Inform. Theory*, 1995, submitted for publication.
- [3] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 3, no. 4, pp. 1034-1054, July 1991.
- [4] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
- [5] J. A. Bucklew, "A note on the absolute epsilon entropy," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 142-144, Jan. 1991.
- [6] J. A. Bucklew and G. L. Wise, "Multidimensional asymptotic quantization theory with  $r$ th power distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 239-247, Mar. 1982.
- [7] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-28, pp. 562-574, Oct. 1980.
- [8] W.-Y. Chan and A. Gersho, "Constrained-storage quantization of multiple vector sources by codebook sharing," *IEEE Trans. Commun.*, vol. 39, no. 1, pp. 11-13, Jan. 1991.
- [9] P. A. Chou, "Code clustering for weighted universal VQ and other applications," in *Proc. IEEE Int. Symp. on Information Theory*, (Budapest, Hungary, June 1991), p. 253.
- [10] ———, "Optimal partitioning for classification and regression trees," *IEEE Trans. Pat. Anal. Machine Intell.*, vol. 13, no. 4, pp. 340-354, Apr. 1991.
- [11] P. A. Chou and M. Effros, "Rate and distortion redundancies for source coding with respect to a fidelity criterion," in *Proc. IEEE Int. Symp. on Information Theory* (San Antonio, TX, Jan. 1993).
- [12] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 37, no. 1, pp. 31-42, Jan. 1989.
- [13] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 453-471, May 1990.
- [14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [15] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 6, pp. 783-795, Nov. 1973.

- [16] ———, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 2, pp. 211–215, Mar. 1983.
- [17] E. J. Delp and O. R. Mitchell, "Image compression using block truncation coding," *IEEE Trans. Commun.*, vol. COM-27, no. 9, pp. 1335–1342, 1979.
- [18] M. Effros and P. A. Chou, "Weighted universal bit allocation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Detroit, MI, May 1995), vol. 4, pp. 2343–2346.
- [19] ———, "Weighted universal transform coding: universal image compression with the Karhounen Loève transform," in *Proc. IEEE Int. Conf. on Image Processing* (Washington, DC, Oct. 1995), to appear.
- [20] M. Effros, P. A. Chou, and R. M. Gray, "Rates of convergence in adaptive universal vector quantization," in *Proc. IEEE Int. Symp. on Information Theory* (Trondheim, Norway, June 1994).
- [21] ———, "Variable dimension weighted universal vector quantization and noiseless coding," in *Proc. IEEE Data Compression Conf.*, (Snowbird, UT, Mar. 1994), pp. 2–11.
- [22] ———, "Variable-rate source coding theorems for stationary nonergodic sources," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1920–1925, Nov. 1994.
- [23] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [24] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 4, pp. 373–380, July 1979.
- [25] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.
- [26] R. M. Gray and L. D. Davisson, "Source coding theorems without the ergodic assumption," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 502–526, July 1974.
- [27] L. Györfi, I. Páli, and E. C. van der Meulen, "There is no universal source code for an infinite source alphabet," *IEEE Trans. Inform. Theory*, vol. 40, no. 1, pp. 267–271, Jan. 1994.
- [28] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. Comput.*, vol. 100, no. 1, pp. 78–150, Sept. 1992.
- [29] K. Jacobs, "The ergodic decomposition of the Kolmogorov–Sinai invariant," in F. B. Wright and F. B. Wright, Eds., *Ergodic Theory*. New York: Academic Press, 1963.
- [30] J. C. Kieffer, "On the optimum average distortion attainable by a fixed-rate coding of a nonergodic source," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 2, pp. 190–193, Mar. 1975.
- [31] ———, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 6, pp. 674–682, Nov. 1978.
- [32] ———, "On the minimum rate for strong universal block coding of a class of ergodic sources," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 6, pp. 693–702, Nov. 1980.
- [33] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 2, pp. 199–207, Mar. 1981.
- [34] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959; republished: New York: Dover, 1968.
- [35] 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.
- [36] T. Linder, G. Lugosi, and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1728–1740, Nov. 1994.
- [37] ———, "Fixed rate universal lossy source coding and rates of convergence for memoryless sources," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 665–676, May 1995.
- [38] T. D. Lookabaugh and R. M. Gray, "High resolution quantization theory and the vector quantization advantage," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1020–1033, Sept. 1989.
- [39] 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, no. 5, pp. 511–523, Sept. 1975.
- [40] R. J. Pile, "Coding theorems for discrete source–channel pairs," Ph.D. dissertation, MIT, Cambridge, MA, 1967.
- [41] ———, "The transmission distortion of a source as a function of the encoding block length," *Bell Syst. Tech. J.*, vol. 47, pp. 827–885, 1968.
- [42] E. Posner, E. Rodemich, and H. Rumsey, "Epsilon entropy of stochastic processes," *Ann. Math. Stat.*, vol. 38, pp. 1000–1020, 1967.
- [43] M. Rabbani and P. W. Jones, *Digital Image Compression Techniques*. Bellingham, WA: SPIE, 1991.
- [44] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A segmental  $K$ -means training procedure for connected word recognition," *AT&T Tech. J.*, vol. 64, no. 3, pp. 21–40, May 1986.
- [45] R. F. Rice and J. R. Plaunt, "The Rice Machine: Television data compression," Tech. Rep. 900-408, Jet Propulsion Lab., Pasadena, CA, Sept. 1970.
- [46] ———, "Adaptive variable-length coding for efficient compression of spacecraft television data," *IEEE Trans. Commun.*, vol. COM-19, no. 12, pp. 889–897, Dec. 1971.
- [47] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 4, pp. 629–636, July 1984.
- [48] ———, "Stochastic complexity and modeling," *Ann. Stat.*, vol. 14, pp. 1080–1100, Sept. 1986.
- [49] ———, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 40–47, Jan. 1996.
- [50] H. L. Royden, *Real Analysis*, 2nd ed. New York: Macmillan, 1968.
- [51] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Glasgow, Scotland, May 1989), pp. 1945–1948.
- [52] A. V. Trushkin, "Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 187–198, Mar. 1982.
- [53] B. Yu and T. P. Speed, "A rate of convergence result for a universal  $d$ -semifaithful code," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 813–820, May 1993.
- [54] R. Zamir and M. Feder, "On lattice quantization noise," in *Proc. IEEE Int. Symp. on Information Theory* (Trondheim, Norway, June 1994).
- [55] K. Zeger and A. Bist, "Universal adaptive vector quantization using codebook quantization," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (San Francisco, CA, Mar. 1992), pp. III.381–384.
- [56] K. Zeger, A. Bist, and T. Linder, "Universal source coding with codebook transmission," *IEEE Trans. Commun.*, vol. 42, pp. 336–346, Feb. 1994.
- [57] Z. Zhang and V. K. Wei, "An on-line universal lossy data compression algorithm via continuous codebook refinement—Pt. I: Basic results," *IEEE Trans. Inform. Theory*, vol. 42, pp. 803–821, May 1996.
- [58] Z. Zhang, E.-H. Yang, and V.K. Wei, "The redundancy of source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, 1994, submitted.