

# High Rate Mismatch in Entropy Constrained Quantization<sup>1</sup>

Robert M. Gray

Tamás Linder

Information Systems Lab, Department of Electrical Engineering, Stanford University  
Stanford, CA 94305. [rmgray@stanford.edu](mailto:rmgray@stanford.edu)

Department of Mathematics and Statistics, Queen's University  
Kingston, Ontario, Canada K7L 3N6 [linder@mast.queensu.ca](mailto:linder@mast.queensu.ca)

## Abstract

It is shown that if an asymptotically (high rate) optimal sequence of variable rate codes is designed for a  $k$ -dimensional probability density function (pdf)  $g$  and then applied to another pdf  $f$  for which  $f/g$  is bounded, then the resulting mismatch or loss of performance from the optimal possible is given by the relative entropy or Kullback-Leibler divergence  $I(f||g)$ . It is also shown that under the same assumptions an asymptotically optimal code sequence for  $g$  can be converted to an asymptotically optimal code sequence for a mismatched source  $f$  by modifying only the lossless component of the code. The development does not require Gersho's conjecture.

## 1 Introduction

The optimal performance of high rate vector quantization using fixed-rate codes was established in Zador's classic Bell Labs Technical Memo [19] and generalized and simplified by Bucklew and Wise [1] and Graf and Luschgy [7]. The history and generality of the results may be found, e.g., in [12]. Bucklew [2] developed further asymptotic properties of high rate quantization, most notably providing a mismatch result that quantified the performance resulting when a sequence of quantizers that are asymptotically optimal for one source are applied instead to another "mismatched" source. Such mismatch results are important for theory and potentially important for practice as code designs are often based on source models which are estimated based on data and hence which are often inaccurate. Mismatch results provide a means of quantifying such performance variations. Another potential application of mismatch performance results is in the design of "robust" source codes. Sakrison [17] and Lapidoth [16] showed that for large dimensions Gaussian sources provide a "worst case" or "robust" approach to code design in that the Gaussian source has the largest (worst) Shannon rate-distortion function and, more importantly, that a code designed for a Gaussian model will yield approximately the same performance when applied to any source with the same mean and covariance. The Gaussian code will of course be suboptimal for the nonGaussian source, but it will provide "robust" or reliable performance in the sense that the resulting rate and distortion will be the same whichever source the code is applied to. The results of Sakrison and Lapidoth were asymptotic in the typical Shannon fashion, large dimensions were required in order

---

<sup>1</sup>This work was supported by the National Science Foundation under NSF Grants MIP-9706284-001 and CCR-0073050 and by Natural Sciences and Engineering Research Council (NSERC) of Canada.

to apply Shannon source coding arguments, in contrast to the fixed dimension and asymptotically large rate of Bucklew’s approach.

Zador also developed the rate-distortion tradeoffs for entropy-constrained vector quantizers [19], but these results have only recently been generalized [13] to conditions of comparable generality to the fixed rate results of [1, 7]. The goal of this paper is to describe a general variable rate mismatch result following the Lagrangian approach of [13]. We here sketch the key ideas of the result and its proof, the detailed proof may be found in [14]. Earlier versions of these results based on Gersho’s heuristic methods were reported in [10, 11].

## 2 Preliminaries

The required preliminaries follow [13].  $(\Omega, \mathcal{B})$  is the measurable space consisting of the  $k$ -dimensional Euclidean space  $\Omega = \mathfrak{R}^k$  and its Borel subsets. Assume that  $X$  is random vector with a distribution  $P_f$ , which is absolutely continuous with respect to the Lebesgue measure  $V$  and hence possesses a probability density function (pdf)  $f = dP_f/dV$  so that  $P_f(F) = \int_F f(x)dV(x) = \int_F f(x) dx$  for any  $F \in \mathcal{B}$ . The volume of a set  $F \in \mathcal{B}$  is given by its Lebesgue measure  $V(F) = \int_F dx$ . We assume that the differential entropy  $h(f) \triangleq - \int dx f(x) \ln f(x)$  exists and is finite. The unit of entropy is nats or bits according to whether the base of the logarithm is  $e$  or 2. Usually nats will be assumed, but bits will be used when entropies appear in an exponent of 2.

A vector quantizer  $Q$  can be described by the following mappings and sets:

- An *encoder*  $\alpha : \Omega \rightarrow \mathcal{I}$ , where  $\mathcal{I}$  is a countable index set, say  $\{0, 1, \dots, N - 1\}$  or  $\{0, 1, 2, \dots\}$ , and an associated measurable partition  $\mathcal{S} = \{S_i; i \in \mathcal{I}\}$  such that  $\alpha(x) = i$  if  $x \in S_i$ .
- A *decoder*  $\beta : \mathcal{I} \rightarrow \Omega$  and an associated reproduction codebook  $\mathcal{C} = \{\beta(i); i \in \mathcal{I}\}$ . Without loss of generality we assume that the codevectors  $\beta(i); i \in \mathcal{I}$  are all distinct.
- A *length function*  $\ell$  satisfying

$$\sum_i e^{-\ell(i)} \leq 1 \tag{1}$$

which describes (in nats) the lengths of the codewords of a uniquely decodable lossless index coder. A set of codelengths  $\ell(i)$  is said to be *admissible* if (1) holds.

We abbreviate the overall action of producing a reproduction from an input, the cascade of decoder and encoder, using a lower case  $q$ :  $q(x) = \beta(\alpha(x))$ .

The instantaneous rate of a quantizer is defined by  $r(\alpha(x)) = \ell(\alpha(x))$ . The average rate is

$$R_f(Q) = R_f(\alpha, \ell) = E_f r(\alpha(X)) = \sum_i p_i \ell(i)$$

where  $p_i = P_f(S_i)$ . Given a quantizer  $Q$ , the entropy of the quantizer is defined in the usual fashion by

$$H_f(Q) = H_f(\alpha) = - \sum_i p_i \ln p_i$$

and we assume that  $p_i > 0$  for all  $i$ .

For any admissible length function  $\ell$  the divergence inequality implies that  $R_f(Q) \geq H_f(Q)$  with equality if and only if  $\ell(i) = -\ln p_i$ . Thus in particular

$$H_f(Q) = \inf_{\ell \in \mathcal{A}} R_f(\alpha, \ell). \quad (2)$$

We assume a squared-error distortion measure  $d(x, \hat{x}) = \|x - \hat{x}\|^2 = \sum_{i=1}^k |x_i - \hat{x}_i|^2$ , where  $x = (x_1, \dots, x_k)$  and  $\hat{x} = (\hat{x}_1, \dots, \hat{x}_k)$ , and measure performance by average distortion

$$D_f(Q) = D_f(\alpha, \beta) = Ed(X, \beta(\alpha(X))).$$

A quantizer of particular interest is the uniform quantizer with side length  $\Delta$ : For  $\Delta > 0$  let  $Q_\Delta$  denote a quantizer of  $\Omega$  into contiguous cubes of side  $\Delta$ . In other words,  $Q_\Delta$  can be viewed as a uniform scalar quantizer with bin size  $\Delta$  applied  $k$  successive times. We assume the axes of the cubes align with the coordinate axes (and that point 0 is touched by corners of cubes). In particular,  $Q_1$  is a cubic lattice quantizer with unit volume cells.

The traditional distortion-rate approach defines the optimal performance as the minimum distortion achievable for a given rate:

$$\delta_f(R) = \inf_{q: R_f(Q) \leq R} D_f(Q). \quad (3)$$

The traditional form of Zador's theorem states that under suitable assumptions on  $f$ ,

$$\lim_{R \rightarrow \infty} 2^{\frac{2}{k}R} \delta_f(R) = b(2, k) 2^{\frac{2}{k}h(f)} \quad (4)$$

where  $b(2, k)$  is Zador's constant, which depends only on  $k$  and not  $f$ . Zador did not evaluate the constant  $b(2, k)$  but he did provide upper and lower bounds that become tight for large  $k$ . Zador's basic results contained technical errors and restrictive conditions on the allowed densities. These problems were discussed and fixed and the results generalized [13] using a Lagrangian approach, which we turn to next.

The Lagrangian formulation of variable rate vector quantization [3] defines for each value of a Lagrangian multiplier  $\lambda > 0$  a Lagrangian distortion  $\rho_\lambda(x, i) = d(x, \beta(i)) + \lambda \ell(i)$  and corresponding performance

$$\rho(f, \lambda, Q) = Ed(X, q(X)) + \lambda E \ell(\alpha(X)) = D_f(Q) + \lambda R_f(Q)$$

and an optimal performance

$$\rho(f, \lambda) = \inf_Q \rho(f, \lambda, Q)$$

where the infimum is over all quantizers  $Q = (\alpha, \beta, \ell)$  with  $\ell$  admissible. The Lagrangian formulation yields Lloyd optimality conditions for vector quantizers, that is, a necessary condition for optimality is that each of the three components of the quantizer be optimal for the other two:

- For a given decoder  $\beta$  and length function  $\ell$ , the optimal encoder is  $\alpha(x) = \operatorname{argmin}_i (d(x, \beta(i)) + \lambda \ell(i))$  (ties are broken arbitrarily).

- The optimal decoder for a given encoder and length function is the usual Lloyd centroid  $\beta(i) = \operatorname{argmin}_y E(d(X, y) | \alpha(X) = i)$ .
- The optimal length function for the given encoder and decoder is  $\ell(i) = -\ln p_i$ .

Note that smaller values of  $\lambda$  correspond to higher rates. The next result characterizes the asymptotic performance of optimal entropy constrained vector quantizers as  $\lambda \rightarrow 0$ , i.e., for asymptotically high rates.

**Theorem 1** [13]. *Assume that the distribution  $P_f$  of  $X$  is absolutely continuous with respect to Lebesgue measure  $V$  with pdf  $f = dP_f/dV$ , that the differential entropy  $h(f)$  exists and is finite, and that  $H_f(Q_1) < \infty$ . Then*

$$\lim_{\lambda \rightarrow 0} \left( \frac{\rho(f, \lambda)}{\lambda} + \frac{k}{2} \ln \lambda \right) = h(f) + \theta_k \quad (5)$$

where the finite constant  $\theta_k$  is defined by

$$\theta_k \triangleq \inf_{\lambda > 0} \left( \frac{\rho(u_1, \lambda)}{\lambda} + \frac{k}{2} \ln \lambda \right) \quad (6)$$

and  $u_1$  is the uniform pdf on the  $k$ -dimensional unit cube  $[0, 1)^k$ .

In particular, the limiting constant  $\theta_k$  depends only on the dimension and not on the pdf. It is also shown in [13] that under the stated assumptions, Zador's original result (4) and the Lagrangian formulation are equivalent and

$$\theta_k = \frac{k}{2} \ln \frac{2e}{k} b(2, k). \quad (7)$$

In order to state two final preliminary results we introduce the following notation:

$$\theta(f, \lambda, \alpha, \beta, \ell) = \theta(f, \lambda, Q) \triangleq \frac{E_f d(X, q(X))}{\lambda} + E_f \ell(\alpha(X)) + \frac{k}{2} \ln \lambda - h(f) \quad (8)$$

so that the theorem states that under suitable conditions

$$\liminf_{\lambda \rightarrow 0} \theta(f, \lambda, Q) = \theta_k. \quad (9)$$

If one or more of the components is optimized, then it is dropped from the argument of  $\theta$ , e.g.,

$$\theta(f, \lambda, \alpha, \beta) = \inf_{\ell} \theta(f, \lambda, \alpha, \beta, \ell) = \frac{D_f(\alpha, \beta)}{\lambda} + H_f(\alpha) + \frac{k}{2} \ln \lambda - h(f) \quad (10)$$

$$\theta(f, \lambda) = \inf_{\alpha, \beta, \ell} \theta(f, \lambda, \alpha, \beta, \ell). \quad (11)$$

With this notation the theorem statement can be simplified to

$$\lim_{\lambda \rightarrow 0} \theta(f, \lambda) = \theta_k. \quad (12)$$

The theorem guarantees that if a pdf  $f$  satisfies the conditions of the theorem, then there is an *asymptotically optimal* sequence of quantizers  $q_n$  for  $f$  in the sense that for any decreasing sequence  $\lambda_n$  converging to 0 there exists a sequence of quantizers  $q_n$  such that

$$\lim_{n \rightarrow \infty} \theta(f, \lambda_n, q_n) = \theta_k. \quad (13)$$

Given two probability measures  $P$  and  $G$  on  $(\Omega, \mathcal{B})$  for which  $P \ll G$  (i.e.,  $P$  is absolutely continuous with respect to  $G$ ) and a finite measurable partition  $\mathcal{S} = \{S_i\}$ , define the *relative entropy* of  $P$  with respect to  $G$  of the partition  $\mathcal{S}$  as  $H_{P||G}(\mathcal{S}) = \sum_i P(S_i) \ln \frac{P(S_i)}{G(S_i)}$  and the *relative entropy* of  $P$  with respect to  $G$  as  $I(P||G) = \sup_{\mathcal{S}} H_{P||G}(\mathcal{S})$ , where the supremum is over all finite measurable partitions. The relative entropy is also known as the Kullback-Leibler number or Kullback-Leibler  $I$ -divergence or directed divergence or discrimination. The reader is referred to [15, 4, 5, 6, 9] for thorough treatments of relative entropy and its properties. If the two measures are induced by pdf's  $f$  and  $g$  it can be shown that

$$I(P_f||P_g) = I(f||g) = \int dx f(x) \ln \frac{f(x)}{g(x)}$$

where we have abbreviated the notation to emphasize the dependence on the densities.

### 3 Quantizer Mismatch

Our principal result is the following high rate variable-rate quantizer mismatch theorem.

**Theorem 2 (*The mismatch theorem*)** *Suppose that a probability measure  $P_g$  on  $\mathbb{R}^k$  satisfies the conditions of Theorem 1 and has pdf  $g$ . Suppose that  $Q_n = (q_n, \ell_n)$  is an asymptotically optimal sequence of variable-rate quantizers for  $P_g$ , where  $\ell_n$  is the optimal length function for  $P_g$  and  $q_n$ . Suppose also that  $P_f \ll P_g$  and that  $dP_f/dP_g = f/g$  is bounded. Then*

$$\lim_{n \rightarrow \infty} \frac{E_f d(X, q_n(X))}{\lambda_n} + E_f \ell_n(\alpha_n(X)) + \frac{k}{2} \ln \lambda_n = \theta_k - \int dx f(x) \ln g(x) \quad (14)$$

or, equivalently,

$$\lim_{n \rightarrow \infty} \theta(f, \lambda_n, Q_n) = \theta_k + I(f||g). \quad (15)$$

The second form of the theorem provides a characterization of the *mismatch* resulting from applying an asymptotically optimal quantizer sequence for one pdf to another: the mismatch is exactly the relative entropy of the mismatched pdf to the design pdf, a continuous analog to the mismatch formula arising in noiseless coding. The result provides a new interpretation of relative entropy as a measure of mismatch for high rate fixed dimension lossy data compression.

The development of the mismatch result for entropy-constrained vector quantization parallels that of Bucklew's fixed-rate result in that the beginning and core of the proof of the lemma of the next section provides what can be interpreted as a local form of Theorem 1, an entropy-constrained variation on Bucklew's Lemma 2. Our proof is simpler, however, and the result is used in a different manner.

## 4 Asymptotically Optimal Quantization

We return to the setting of Theorem 1. We assume that  $f$  is the source pdf and that an asymptotically optimal quantizer sequence is designed for  $f$  and we investigate properties of the sequence. More formally, recall that for any decreasing sequence  $\lambda_n$  converging to 0, Theorem 1 implies the existence of an asymptotically optimal sequence of encoders and decoders  $\alpha_n, \beta_n$ , and the corresponding optimized length function  $\ell_n^*(i) = -\ln P_f(\alpha_n(X) = i)$  for which

$$\lim_{n \rightarrow \infty} \theta(f, \lambda_n, \alpha_n, \beta_n) = \theta_k. \quad (16)$$

**Lemma 1** *Suppose that  $Q_n = (\alpha_n, \beta_n, \ell_n^*)$  is an asymptotically optimal sequence of variable-rate quantizers for  $P_f$  in the sense of (16), where  $\ell_n^*$  is the optimal length function for  $P_f$ . Then for every measurable set  $F$*

$$\lim_{n \rightarrow \infty} \int_F dx f(x) \left( \frac{d(x, \beta_n(\alpha_n(x)))}{\lambda_n} + \ell_n^*(\alpha_n(x)) + \frac{k}{2} \ln \lambda_n + \ln f(x) \right) = P_f(F) \theta_k. \quad (17)$$

*Proof sketch:* If  $P_f(F) = 0$  or  $1$  the claim is immediate, so assume that  $0 < P_f(F) < 1$ . The lemma can be stated simply adopting Bucklew's notation. Define

$$M_f^n(F) \triangleq \int_F dx f(x) \left( \frac{d(x, \beta_n(\alpha_n(x)))}{\lambda_n} + \ell_n^*(\alpha_n(x)) + \frac{k}{2} \ln \lambda_n + \ln f(x) \right) \quad (18)$$

where  $\ell_n^*$  is the optimal length function for  $\alpha_n$  and  $\beta_n$   $\ell_n^*(i) = -\ln P_f(\alpha_n(X) = i)$  and  $M_f(F) = P_f(F) \theta_k$  so that the claim of the lemma becomes

$$\lim_{n \rightarrow \infty} M_f^n(F) = M_f(F). \quad (19)$$

By construction and Theorem 1 as  $n \rightarrow \infty$

$$M_f^n(F) + M_f^n(F^c) = \theta(f, \lambda_n, \alpha_n, \beta_n) \rightarrow \theta_k. \quad (20)$$

Let  $w_1 = P_f(F)$ ,  $w_2 = P_f(F^c)$ ,  $f_1(x) = f(x)/w_1$  for  $x \in F$  and 0 otherwise, and  $f_2(x) = f(x)/w_2$  for  $x \in F^c$  and 0 otherwise. It can be shown that

$$M_f^n(F) = w_1 \int dx f_1(x) \left( \frac{d(x, \beta_n(\alpha_n(x)))}{\lambda_n} + \ell_n^*(\alpha_n(x)) + \frac{k}{2} \ln \lambda_n + \ln f_1(x) w_1 \right)$$

and

$$\begin{aligned} & M_f^n(F) + M_f^n(F^c) \\ &= \sum_{m=1}^2 w_m \int dx f_m(x) \left( \frac{d(x, \beta_n(\alpha_n(x)))}{\lambda_n} + \ell_n^*(\alpha_n(x)) + \frac{k}{2} \ln \lambda_n + \ln f_m(x) w_m \right). \end{aligned}$$

Since  $\lim_{n \rightarrow \infty} \theta(f, \lambda_n, \alpha_n, \beta_n) = \theta_k$ , it is easy to see that

$$\lim_{n \rightarrow \infty} D_f(q_n) = 0 \quad (21)$$

which implies (with some work) that

$$\lim_{n \rightarrow \infty} \theta(f, \lambda_n, \alpha_n, \beta_n) = \lim_{n \rightarrow \infty} \sum_{m=1}^2 w_m \theta(f_m, \lambda_n, \alpha_n, \beta_n) = \theta_k.$$

It can be shown by contradiction that each of the two component compression functions must individually converge to  $\theta_k$ , not just the overall weighted average.  $\square$

We now change notation slightly: Replace the single pdf  $f$  above by a “design” pdf  $g$  for which the optimal code sequence is designed and now let  $f$  denote the pdf to which the code will be applied. Define the set function  $\mu_n(F) = M_g^n(F)$  with  $M_g^n$  defined as  $M_f^n$  in (18) with  $g$  in place of  $f$ . The set function  $\mu_n$  is a signed measure (see e.g., Doob [8]). From Lemma 1 we know that  $\mu_n(F)$  converges to  $\mu(F) = M_g(F) = \theta_k P_g(F)$  for all measurable sets  $F$ . We now explore the consequences of this convergence.

Given a signed measure  $\mu$ , for any measurable set  $F$  define the positive variation  $\mu^+(F) = \sup_{G \subset F} \mu(G)$ , the negative variation  $\mu^-(F) = -\inf_{G \subset F} \mu(G)$ , and the total variation  $|\mu|(F) = \mu^+(F) + \mu^-(F)$ . The total variation of  $\mu$  on  $\mathfrak{R}^k$  is  $\|\mu\| = |\mu|(\mathfrak{R}^k)$ .

For all measurable sets  $F$ ,  $\mu_n(F) < \infty$  and  $\lim_{n \rightarrow \infty} \mu_n(F) = P_g(F)\theta_k < \infty$ , and hence also  $\lim_{n \rightarrow \infty} |\mu_n(F)| = P_g(F)\theta_k < \infty$ . Thus from the discussion following Theorem IX.9 of Doob [8] it follows that

$$\sup_n \|\mu_n\| < \infty. \quad (22)$$

From Lemma 1 it follows that for any simple function  $\phi$

$$\lim_{n \rightarrow \infty} \int \phi d\mu_n = \int \phi d\mu. \quad (23)$$

It is straightforward to show by standard methods that the limit will also hold for any bounded nonnegative function  $\phi$ .

**Proof of the mismatch theorem:** Since the Radon-Nikodym derivative  $\phi = f/g$  is assumed to be bounded,

$$\lim_{n \rightarrow \infty} \int \frac{f}{g} d\mu_n = \int \frac{f}{g} d\mu \quad (24)$$

and the two sides of the equation evaluate as

$$\begin{aligned} \lim_{n \rightarrow \infty} \int \frac{f}{g} d\mu_n &= \int dx f(x) \left( \frac{d(x, \beta_n(\alpha_n(x)))}{\lambda_n} + \ell_n^*(\alpha_n(x)) + \frac{k}{2} \ln \lambda_n + \ln g(x) \right) \\ \int \frac{f}{g} d\mu &= \theta_k \int dx g(x) \frac{f(x)}{g(x)} = \theta_k. \end{aligned}$$

$\square$

## 5 High Rate Universal Codes

The mismatch theorem shows the asymptotic performance that is lost when an asymptotically optimal sequence of quantizers  $Q_n$  designed for a pdf  $g$  is applied to a pdf  $f$  and that this loss is just the relative entropy  $I(f||g)$ . This performance loss can be eliminated by modifying only the length function to match the pdf  $f$ . This implies that the asymptotically optimal sequence of reproduction codebooks for the design pdf  $g$  remains asymptotically optimal for any  $f$  meeting the conditions of the theorem. Thus, for example, if  $f$  has bounded support, one could design an asymptotically optimal sequence of codes for a uniform pdf on the support set and it would also be optimal for  $f$ . If  $f$  has unbounded support, one could design an asymptotically optimal quantizer sequence for a Gaussian pdf and its reproduction codebook would be optimal for  $f$ .

**Corollary 1** *Suppose that  $Q_n = (q_n, \ell_n)$  is a sequence of variable-rate quantizers that is asymptotically optimal for a pdf  $g$  in the sense that*

$$\lim_{n \rightarrow \infty} \theta(g, \lambda_n, Q_n) = \theta_k$$

for some decreasing sequence  $\lambda_n$  converging to 0. Assume also that  $f$  is a pdf that meets the condition of the mismatch theorem and that  $h(f) > -\infty$ . Define  $\ell'_n$  to be the optimal length function for  $q_n$  and  $P_f$ . Then  $Q'_n = (q_n, \ell'_n)$  is asymptotically optimal for  $P_f$ , i.e.,

$$\lim_{n \rightarrow \infty} \theta(f, \lambda_n, Q'_n) = \theta_k. \quad (25)$$

*Proof:* Since  $Q_n$  and  $Q'_n$  share the same encoder  $\alpha_n$  and decoder  $\beta_n$  and differ only in their length functions, we have from (8) that

$$\begin{aligned} \theta(f, \lambda_n, Q'_n) &= \theta(f, \lambda_n, Q_n) - (E_f \ell(\alpha(X)) - E_f \ell'(\alpha(X))) \\ &= \theta(f, \lambda_n, Q_n) - \left( \sum_i P_f(S_{n,i}) \ln \frac{P_f(S_{n,i})}{P_g(S_{n,i})} \right) \end{aligned}$$

where we have plugged in the definitions for  $\ell$  and  $\ell'$  as the optimal length function for  $g$  and  $f$ , respectively, and where  $\{S_{n,i}\}$  is the partition corresponding to  $\alpha_n$ :  $S_{n,i} = \{x : \alpha_n(x) = i\}$ . We have immediately from Theorem 1

$$\liminf_{n \rightarrow \infty} \theta(f, \lambda_n, Q'_n) \geq \liminf_{n \rightarrow \infty} \inf_Q \theta(f, \lambda_n, Q) = \theta_k. \quad (26)$$

For the other direction we have, using the mismatch theorem, that

$$\limsup_{n \rightarrow \infty} \theta(f, \lambda_n, Q'_n) = \theta_k + I(f||g) - \liminf_{n \rightarrow \infty} \sum_i P_f(S_{n,i}) \ln \frac{P_f(S_{n,i})}{P_g(S_{n,i})}. \quad (27)$$

Define the discrete distribution  $P_f^n$  by  $P_f^n(y_{n,i}) = P_f(S_{n,i})$ , where  $y_{n,i} = \beta_n(i)$ , i.e., for any measurable set  $F$

$$P_f^n(F) = \sum_{i: y_{n,i} \in F} P_f^n(y_{n,i})$$

and  $P_f^n$  is the distribution for the random vector  $q_n(X)$  when  $X$  is described by the pdf  $f$ . Similarly define the discrete distribution  $P_g^n$  by  $P_g^n(y_{n,i}) = P_g(S_{n,i})$ . With this notation (27) becomes

$$\limsup_{n \rightarrow \infty} \theta(f, \lambda_n, Q'_n) = \theta_k + I(f||g) - \liminf_{n \rightarrow \infty} I(P_f^n || P_g^n). \quad (28)$$

From (21) applied to  $g$ ,

$$\lim_{n \rightarrow \infty} D_g(q_n) = \lim_{n \rightarrow \infty} \int dx g(x) \|x - q_n(x)\|^2 = 0 \quad (29)$$

i.e.,  $q_n(X)$  converges to  $X$  in mean square (here  $X$  has pdf  $g$ ). This implies that  $P_g^n \rightarrow P_g$  in the sense of weak convergence (see, e.g, Theorem 4.2 of [18]).

Furthermore, since by assumption there is a finite  $M$  such that  $f(x)/g(x) \leq M$ ,

$$\begin{aligned} \lim_{n \rightarrow \infty} D_f(q_n) &= \lim_{n \rightarrow \infty} \int dx g(x) \frac{f(x)}{g(x)} \|x - q_n(x)\|^2 \\ &\leq M \lim_{n \rightarrow \infty} D_g(q_n) = 0 \end{aligned}$$

and hence by the same argument  $P_f^n \rightarrow P_f$  (weak convergence). From [6], relative entropy is lower semicontinuous with respect to weak convergence of distributions so that

$$\liminf_{n \rightarrow \infty} I(P_f^n || P_g^n) \geq I(f||g) \quad (30)$$

which with (27) yields

$$\limsup_{n \rightarrow \infty} \theta(f, \lambda_n, Q'_n) \leq \theta_k$$

which completes the proof.  $\square$

Although the length function (and hence the lossless component) of the quantizer has been matched to the true source, the encoder has not been optimized for the new length function. Thus there remains a mismatch in the code sequence, which nonetheless is asymptotically optimal.

#### *Acknowledgements*

The authors gratefully acknowledge the comments and corrections of the Stanford Compression and Classification group and Ken Zeger.

## References

- [1] J. A. Bucklew and G. L. Wise, "Multidimensional asymptotic quantization theory with  $r$ th power distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 239–247, Mar. 1982.
- [2] J. A. Bucklew, "Two results on the asymptotic performance of quantizers," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 341–348, Mar. 1984.
- [3] P.A. Chou, T. Lookabaugh, and R.M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust. Speech and Signal Proc.*, Vol. 37, pp. 31–42, Jan. 1989.

- [4] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observations,” *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.
- [5] I. Csiszár, “Generalized entropy and quantization problems,” *Proc. Sixth Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes*, pp. 159–174, 1973.
- [6] I. Csiszár. “On an extremum problem of information theory,” *Studia Scientiarum Mathematicarum Hungarica*, pp. 57–70, 1974.
- [7] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Springer, Lecture Notes in Mathematics, 1730, Berlin, 2000.
- [8] J.L. Doob, *Measure Theory*, Springer, 1994.
- [9] R. M. Gray, *Entropy and Information Theory*, Springer–Verlag, 1990.
- [10] R. M. Gray, “Gauss mixture vector quantization,” *Proceedings 2001 IEEE ICASSP*, Salt Lake City, May 2001.
- [11] R.M. Gray, “Gauss mixture quantization: clustering Gauss mixtures,” to appear in a collection of papers based on a Math Sciences Research Institute Workshop on “Non-linear Estimation and Classification,” March 17th–March 29th 2002.
- [12] R.M. Gray and D.L. Neuhoff, “Quantization,” *IEEE Transactions on Information Theory*, vol. 44, pp. 2325–2384, Oct. 1998.
- [13] R.M. Gray, T. Linder, and J. Li, “A Lagrangian formulation of Zador’s entropy-constrained quantization theorem,” *IEEE Trans. Inform. Theory*, pp. 695–707, vol. 48, Mar. 2002.
- [14] R.M. Gray and T. Linder, “Mismatch in high rate entropy constrained vector quantization,” submitted for possible publication. Preprint available at <http://ee.stanford.edu/~gray/mismatch.pdf>
- [15] S. Kullback. *Information Theory and Statistics*, Dover, New York, 1968. (Reprint of 1959 edition published by Wiley.)
- [16] A. Lapidoth, “On the role of mismatch in rate distortion theory,” *IEEE Trans. Inform. Theory*, vol. 43, pp. 38–47, Jan. 1997.
- [17] D. J. Sakrison, “Worst sources and robust codes for difference distortion measures,” *IEEE Trans. Inform. Theory*, vol. 21, pp. 301–309, May 1975.
- [18] A. N. Shiryaev, *Probability*. New York: Springer-Verlag, 2nd ed., 1996.
- [19] P. L. Zador, “Topics in the asymptotic quantization of continuous random variables,” Bell Laboratories Technical Memorandum, 1966.