# A Lagrangian Formulation of Zador's Entropy-Constrained Quantization Theorem[1]

Robert M. Gray Tamás Linder Jia Li

Information Systems Lab, Department of Electrical Engineering, Stanford University
Stanford, CA 94305. rmgray@stanford.edu
Department of Mathematics and Statistics
Queen's University
Kingston, Ontario, Canada K7L 3N6
linder@mast.queensu.ca
Department of Statistics, The Pennsylvania State University
University Park, PA 16802. jiali@stat.psu.edu

## Abstract

Zador's classic result for the asymptotic high-rate behavior of entropy-constrained vector quantization is recast in a Lagrangian form which better matches the Lloyd algorithm used to optimize such quantizers. The equivalence of the two formulations is shown and the result is proved for source distributions that are absolutely continuous with respect to the Lebesgue measure which satisfy an entropy condition, thereby generalizing the conditions stated by Zador under which the result holds.

*Keywords* asymptotic, entropy constrained, high rate, Langrangian, quantization

# 1 Introduction

In his classic Bell Labs Technical Memo of 1966, Paul Zador established the optimal tradeoff between average distortion and rate for $k$-dimensional quantization in the limit of large rate, where rate was measured either by the log of the number of quantization levels or by the Shannon entropy of the quantized vector [18]. The history and generality of the results may be found in [10]. Most notably, Bucklew and Wise [2] demonstrated Zador's fixed-rate result for $r$th power distortion measures of the form $||x-y||^r$, assuming only that $E(||X||^{r+\delta}) < \infty$ for some $\delta > 0$. Their result was subsequently simplified and elaborated by Graf and Luschgy [8]. Zador's entropy-constrained results, however, have not received similar attention in the literature.

Zador formulated the entropy-constrained problem as a minimization of average distortion over all quantizers with a constrained output entropy. Optimality properties and generalized Lloyd algorithms for quantizer design, however, require a Lagrangian formulation [4]. Specifically, Lagrangian optimization can be used to find the lower convex hull of the distortion-rate function, where rate is measured by output entropy. The Lagrangian form also turns out to be more convenient for problems involving multiple codebooks such as coding for mixtures since it obviates the need for the optimization of rate allocation among multiple codes such as occurs in

[1]This work was supported in part 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.

Zador's proof. We here recast Zador's theorem in a Lagrangian form and prove the result under the assumption that the distribution of the random vector is absolutely continuous with respect to Lebesgue measure, that the differential entropy exists and is finite, and that the entropy of a uniformly quantized version of the source is finite. These conditions generalize those stated by Zador in his entropy constrained quantization theorem. Our goal has been to extend Bucklew and Wise's results to entropy-constrained quantization while taking advantage of simplifications introduced by Graf and Luschgy.

## 2 Preliminaries

Consider the measurable space $(\Omega, \mathcal{B})$ consisting of the $k$-dimensional Euclidean space $\Omega = \Re^k$ and its Borel subsets. Assume that $X$ is a 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 $\alpha : \Omega \to \mathcal{I}$, where $\mathcal{I}$ is a countable index set, an associated measurable partition $\mathcal{S} = \{S_i;\ i \in \mathcal{I}\}$ such that $\alpha(x) = i$ if $x \in S_i$, a decoder $\beta : \mathcal{I} \to \Omega$, an associated reproduction codebook $\mathcal{C} = \{\beta(i);\ i \in \mathcal{I}\}$, an index coder $\psi : \mathcal{I} \to \{0, \ldots, D-1\}^*$, the space of all finite-length $D$-ary strings, and the associated length $L : \mathcal{I} \to \{1, 2, \ldots\}$ defined by $L(i) = \text{length}(\psi(i))$. $\psi$ is assumed to be uniquely decodable (a lossless or noiseless code). The overall quantizer is

$$q(x) = \beta(\alpha(x)). \tag{1}$$

Without loss of generality we assume that the codevectors $\beta(i);\ i \in \mathcal{I}$ are all distinct. From the Kraft inequality (e.g., [5]) the codelengths $L(i)$ must satisfy

$$\sum_i D^{-L(i)} \leq 1. \tag{2}$$

It is convenient to measure lengths in a normalized fashion and hence we define the length function of the code in nats as $\ell(i) = L(i) \ln D$ so that Kraft's inequality becomes

$$\sum_i e^{-\ell(i)} \leq 1. \tag{3}$$

A set of codelengths $\ell(i)$ is said to be *admissible* if (3) holds.

As do Cover and Thomas [5], it will also be convenient to remove the restriction of integer $D$-ary codelengths and hence we define any collection of nonnegative real numbers $\ell(i); i \in \mathcal{I}$ to be an *admissible length function* if it satisfies (3). The primary

reason for dropping the constraint is to provide a useful tool for proving results, but the general definition can be interpreted as an approximation since if $\ell(i)$ is an admissible length function, then for a code alphabet of size $D$ the actual integer codelengths $L(i) = \lceil \frac{\ell(i)}{\ln D} \rceil$ will satisfy the Kraft inequality. (Throughout this paper $\lceil t \rceil$ denotes the smallest integer not less than $t$, and $\lfloor t \rfloor$ denotes the largest integer not greater than $t$.) Furthermore, abbreviating $P_f(S_i)$ to $p_i$ the average length (in nats) will satisfy

$$\sum_i p_i \ell(i) \leq (\ln D) \sum_i p_i L(i) < \sum_i p_i \ell(i) + \ln D.$$

If this is normalized by $1/k$, then the actual average length can be made arbitrarily close to the average length function by choosing a sufficiently large dimension. We do not here require large dimension, the dropping of the integer constraint is simply a convenience and the above discussion is intended only to observe a coding interpretation of the unconstrained lengths.

Let $\mathcal{A}$ denote the collection of all admissible length functions $\ell$.

With a slight abuse of notation we will use the symbol $q$ to denote both the composite of encoder $\alpha$ and decoder $\beta$ as in (1) and the complete quantizer comprising the triple $(\alpha, \beta, \ell)$. The meaning should be clear from context.

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

$$R_f(q) = R_f(\alpha, \ell) = Er(\alpha(X)) = \sum_i p_i \ell(i).$$

Given a quantizer $q$, the entropy of the quantizer is defined in the usual fashion by

$$H_f(q) = -\sum_i p_i \ln p_i$$

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

For any admissible length function $\ell$ the divergence inequality [5] implies that

$$\begin{aligned} R_f(q) - H_f(q) &= E\ell(\alpha(X)) - H_f(q) \\ &= \sum_i p_i(\ell(i) + \ln p_i) \\ &= \sum_i p_i \ln \frac{p_i}{e^{-\ell(i)}} \\ &\geq \sum_i p_i \ln \frac{p_i}{e^{-\ell(i)} / \sum_j e^{-\ell(j)}} \\ &\geq 0 \end{aligned}$$

with equality if and only if $\ell(i) = -\ln p_i$. Thus in particular

$$H_f(q) = \inf_{\ell \in \mathcal{A}} R_f(\alpha, \ell). \tag{4}$$

We assume a distortion measure $d(x, \hat{x}) \geq 0$ and measure performance by average distortion

$$D_f(q) = D_f(\alpha, \beta)$$

$$
\begin{aligned}
&= Ed(X, q(X)) \\
&= Ed(X, \beta(\alpha(X)).
\end{aligned}
$$

For simplicity we assume squared error distortion with average

$$d(x, \hat{x}) = ||x - \hat{x}||^2 = \sum_{i=1}^{k} |x_i - \hat{x}_i|^2$$

for $x = (x_1, \ldots, x_k)$ and $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_k)$.

The optimal performance is the minimum distortion achievable for a given rate:

$$\delta_f(R) = \inf_{q: R_f(q) \leq R} D_f(q) \tag{5}$$

$$= \inf_{\alpha, \beta, \ell: R_f(\alpha, \ell) \leq R} D_f(\alpha, \beta). \tag{6}$$

Zador [18] defines rate as $R_f(q) = H_f(q)$ and uses this rate to define the optimal performance by (5). These two definitions of optimal performance are easily seen to be equivalent in view of (4).

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

$$\lim_{R \to \infty} 2^{\frac{2}{k}R} \delta_f(R) = b(2, k) 2^{\frac{2}{k}h(f)} \tag{7}$$

where $b(2, k)$ is Zador's constant, which depends only on $k$ and not $f$. The "2" in $b(2, k)$ reflects the use of squared error distortion, Zador also considered powers other than two. This is often stated loosely as

$$\delta_f(R) \approx b(2, k) 2^{-\frac{2}{k}(R - h(f))}.$$

Zador's argument explicitly requires that his asymptotic result for fixed-rate coding holds and that $h(f)$ is finite. Zador's fixed rate conditions have been generalized through the years (see, e.g., [2], [8]), but his variable results have not been similarly extended. Furthermore, there are problems with Zador's proofs. In particular, as described in the proof of Lemma 2, Zador incorrectly assumes that a conditional entropy term is zero in his proof of his Corollary 3.3, an error which invalidates the remainder of the proof. Another serious problem occurs in the proof of the main entropy constrained quantization theorem, Theorem 3.1, where he assumes that all events in the sigma-field of $\Omega$ have finite volume, an assumption which is invalid for pdfs with infinite support. The first problem is corrected in our consideration of disjoint mixtures. The second is avoided by using a method closer to that of Bucklew and Wise than that of Zador.

As a final preliminary, 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.

# 3 The Lagrangian Formulation

The Lagrangian formulation of variable rate vector quantization [4] defines for each value of a Lagrangian multiplier $\lambda > 0$ a Lagrangian distortion

$$\begin{aligned} \rho_\lambda(x,i) &= d(x,\beta(i)) + \lambda r(\alpha(i)) \\ &= d(x,\beta(i)) + \lambda \ell(i) \end{aligned}$$

and corresponding performance

$$\begin{aligned} \rho(f,\lambda,q) &= Ed(X,q(X)) + \lambda E\ell(\alpha(X)) \\ &= D_f(q) + \lambda R_f(q) \end{aligned}$$

and an optimal performance

$$\rho(f,\lambda) = \inf_q \rho(f,\lambda,q)$$

where the infimum is over all quantizers $q = (\alpha,\beta,\ell)$ where $\ell$ is assumed admissible. Unlike the traditional formulation, 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. In particular, for a given decoder $\beta$ and length function $\ell$, the optimal encoder is

$$\alpha(x) = \operatorname*{argmin}_i \left(d(x,\beta(i)) + \lambda\ell(i)\right)$$

(ties are broken arbitrarily). The optimal decoder for a given encoder and length function is the usual Lloyd centroid

$$\beta(i) = \operatorname*{argmin}_y E(d(X,y)|\alpha(X) = i),$$

and the optimal length function for the given encoder and decoder is, as we have seen, the negative log probabilities of the encoder output. Unlike Zador's proof, our proof will take advantage of these properties.

Our main result is the following.

**Theorem 1** *Assume that the distribution of $X$ is absolutely continuous with respect to Lebesgue measure with pdf $f$, that $h(f)$ exists and is finite, and that $H_f(Q_1) < \infty$. Then*

$$\lim_{\lambda\to 0}\left(\frac{\rho(f,\lambda)}{\lambda} + \frac{k}{2}\ln\lambda\right) = \theta_k + h(f) \tag{8}$$

*where $\theta_k$ is the finite constant defined by*

$$\theta_k \triangleq \inf_{\lambda>0}\left(\frac{\rho(u_1,\lambda)}{\lambda} + \frac{k}{2}\ln\lambda\right) \tag{9}$$

*and $u_1$ is the uniform pdf on the $k$-dimensional unit cube $C_1 = [0,1)^k$.*

Analogous to the approximate interpretation of the traditional Zador result, the interpretation here is that for small $\lambda$,

$$\rho(f,\lambda) = D_f(q_n) + \lambda H_f(q_n) \approx \lambda\theta_k + \lambda h(f) - \frac{k}{2}\lambda \ln \lambda. \tag{10}$$

Note in particular that the asymptotic performance depends on the input pdf only through its differential entropy.

As an example of the conditions of the theorem, the divergence inequality can be used to show that if the random vector $X$ has a finite second moment, then $H_f(Q_1) < \infty$ and $h(f) < \infty$. Thus the theorem holds for pdfs with finite second moment if $h(f) > -\infty$. This implies, for example, that the theorem holds for nonsingular Gaussian pdfs. Thus for example if the vector is a Gaussian vector with mean $m = E(X)$ and nonsingular covariance $K = E((X-m)(X-m)^t)$, then the approximation for the optimal codes becomes

$$D_f(q_n) + \lambda H_f(q_n) \approx \lambda\theta_k + \lambda\frac{k}{2}\ln(2\pi e|K|^{1/k}) - \frac{k}{2}\lambda \ln \lambda \tag{11}$$

for small $\lambda$. (See, e.g., [5], p. 230.) Thus the asymptotic performance depends on the size of the determinant of the covariance. This has an interesting interpretation: Since the pdf having the largest differential entropy of all those having a given covariance matrix is the Gaussian (see, e.g., Theorem 9.6.5 of [5]), this says that for small $\lambda$ the Gaussian source is the "worst case" in the sense of having the largest optimum average Lagrangian distortion over all such pdfs. This provides a quantization analog to Sakrison's result for Shannon rate-distortion functions [13]. More generally, suppose that a class of pdfs contains all pdfs having a specified partial covariance, that is, only a subset of the entries of the covariance is known, but it is known to be consistent with a complete covariance. The MAXDET algorithm [15] can then be used to find the covariance which agrees with the constrained partial covariance and has the largest possible determinant of all such matrices. The Gaussian pdf with this maximum determinant (provided it exists) then provides the worst case for this class of sources for quantization (as $\lambda$ goes to zero).

As another example under which the conditions hold, consider a uniform pdf on a bounded measurable set with positive volume $V$. Once again the conditions of the theorem hold and the asymptotic approximation for the optimal codes becomes

$$D_f(q_n) + \lambda H_f(q_n) \approx \lambda\theta_k + \lambda \ln V - \frac{k}{2}\lambda \ln \lambda \tag{12}$$

for small $\lambda$. Again this provides an example of a worst case source since the divergence inequality shows that the uniform pdf yields the largest differential entropy of all pdfs constrained to have the same finite volume support region.

The following result relates the traditional and Lagrangian form of Zador's results for variable rate vector quantization so that the two forms will hold under equally general conditions. The result is proved in Appendix A using tools developed in the proof of the theorem.

**Lemma 1** *Under the conditions of Theorem 1, the limit of (7) exists if and only if the limit of (8) exists, in which case*

$$\theta_k = \frac{k}{2} \ln \frac{2e}{k} b(2,k). \tag{13}$$

Thus in particular Zador's formula holds under the conditions given in the theorem. This provides a generalization of the results claimed in Zador [18] since Zador requires tail conditions on the marginal densities induced by the pdf. In particular, he requires that the marginal pdfs $f_i;\ i = 1, \ldots, k$ each have the property that $f_i(t) \leq |t|^{-l}$ for $|t| \geq c_i$, where the $c_i$ are nonzero, finite constants, and where $l > 3$. As noted in [10], the variable rate results reported by Zador in his PhD thesis [17] and in [19] are incorrect as they are for the fixed-rate case and do not include the needed differential entropy term, so it is the results in his Bell Labs Technical Report [18] which are considered here. The conditions of the theorem are more general than those of the current most general fixed rate results (see [8]) in the sense that no moment condition is required, but they are less general in the sense that it is assumed that the probability measure is absolute continuous with respect to Lebesgue measure, the density has finite differential entropy, and that a uniformly quantized version of the random vector has finite Shannon entropy.

The result shows that the Lloyd algorithm can be used to estimate Zador's constant, as suggested in [14]. By choosing a decreasing sequence of $\lambda$ and using the Lloyd properties to design an entropy constrained vector quantizer for any pdf, the limiting performance should approach $\theta_k$ and hence yield an estimate of $b_k$. Since the pdf is not important, a uniform pdf on the uniform cube can be used. This was done for various dimensions in [14] and the results compared with known bounds on the zador function.

The theorem will be proved in a series of steps. We begin in the next section with a study of the performance of quantizers for mixture sources, which play an important role in the development by permitting us to "divide and conquer" a complicated source by decomposing it into simpler sources. The subsequent section develops several fundamental properties and bounds on the measures of quantizer performance. These properties are used to quantify relevant asymptotics in the subsequent section. In the final section the theorem is proved by showing that successively more general densities yield the conclusions of the theorem. We consider first uniform densities on cubes, then disjoint mixtures of such densities, then general densities defined on a unit cube, and finally general densities. Anticipating these steps we say that $f \in \mathcal{Z}$ (or $f$ has the Zador property) if the conclusions of the theorem hold for a density $f$.

The following notation will be used throughout the paper:

$$\begin{aligned}
\theta(f,\lambda,q,\ell) &= \frac{D_f(q)}{\lambda} + R_f(\alpha,\ell) - h(f) + \frac{k}{2}\ln\lambda \\
\theta(f,\lambda,q) &= \frac{D_f(q)}{\lambda} + H_f(q) - h(f) + \frac{k}{2}\ln\lambda \\
\theta(f,\lambda) &= \inf_q \theta(f,\lambda,q) \\
\overline{\theta}(f) &= \limsup_{\lambda\to 0} \theta(f,\lambda)
\end{aligned}$$

$$\underline{\theta}(f) = \liminf_{\lambda \to 0} \theta(f, \lambda).$$

Thus the theorem is a statement of conditions under which

$$\underline{\theta}(f) = \overline{\theta}(f) = \theta_k.$$

# 4 Disjoint Mixtures

A mixture source is a random pair $\{X, Z\}$, where $Z$ is a discrete random variable with pmf $w_m = P(Z = m)$, $m = 1, 2, \ldots$ and conditional pdfs $f_{X|Z}(x|m) = f_m(x)$ such that $P_{f_m}(\Omega_m) = 1$ for some $\Omega_m \in \mathcal{B}$, $m = 1, 2, \ldots$ The pdf for $X$ is given by

$$f(x) = f_X(x) = \sum_m w_m f_m(x).$$

In the special case where the $\Omega_m$ are disjoint, the mixture is said to be *orthogonal* or *disjoint*.

Suppose that $f$ is a disjoint mixture and that for each $f_m$ we have a quantizer $q_m$ defined on $\Omega_m$, i.e., an encoder $\alpha_m : \Omega_m \to \mathcal{I}$ (recall that $\mathcal{I}$ is a countable index set usually taken to be the positive integers unless indicated otherwise), a partition of $\Omega_m$, $\{S_{m,i};\ i = 1, 2, \ldots\}$, and a decoder $\beta_m : \mathcal{I} \to \mathcal{C}_m$. The component quantizers $\{q_m\}$ together imply an overall composite quantizer $q$ with an encoder $\alpha$ that maps $x$ into a pair $(m, i)$ if $x \in \Omega_m$ and $\alpha_m(x) = i$, a partition of $\Omega$, $\{S_{m,i};\ i = 1, 2, \ldots, m = 1, 2, \ldots\}$, and a decoder $\beta$ that maps $(m, i)$ into $\beta_m(i)$, $q(x) = \sum_m q_m(x) 1_{\Omega_m}(x)$, where $1_A(x)$ denotes the indicator function of $A \subset \Omega$. Conversely, an overall quantizer $q : \Omega \to \mathcal{I}$ can be applied to every component in the mixture, effectively implying component quantizers $q_m(x) = q(x) 1_{\Omega_m}(x)$ for all $m$. In this case the structure is not so simple as quantization cells can straddle boundaries of $\Omega_m$. Here the partition of $\Omega_m$ is $\{S_i \cap \Omega_m;\ i = 1, 2, \ldots\}$ and many of the cells may be empty.

Begin by observing that since $f_m(x) = f(x)/w_m$ for $x \in \Omega_m$,

$$\begin{aligned} h(f) &= -\int f(x) \ln f(x)\, dx \\ &= -\sum_m w_m \int_{\Omega_m} f_m(x) \ln w_m f_m(x)\, dx \\ &= -\sum_m w_m \int_{\Omega_m} f_m(x) \ln f_m(x)\, dx - \sum_m w_m \ln w_m \\ &= \sum_m w_m h(f_m) + H(Z). \end{aligned} \tag{14}$$

Note that the equality holds whenever at least two of the three quantities $h(f)$, $H(Z)$, and $\sum_m w_m h(f_m)$ are finite. Similarly, if $q$ is a quantizer defined for the entire space $\Omega = \bigcup_m \Omega_m$ with partition $\{S_l\}$ and codebook $\{y_l\}$, then

$$H_f(q) = \sum_l P_f(S_l) \ln \frac{1}{P_f(S_l)}$$

$$
\begin{aligned}
&= \sum_m w_m \sum_l P_{f_m}(S_l) \ln \frac{1}{P_f(S_l)} \\
&= \sum_m w_m \sum_l P_{f_m}(S_l) \ln \frac{1}{w_m P_{f_m}(S_l)} \\
&\quad + \sum_m w_m \sum_l P_{f_m}(S_l) \ln \frac{w_m P_{f_m}(S_l)}{P_f(S_l)} \\
&= \sum_m w_m \sum_l P_{f_m}(S_l) \ln \frac{1}{P_{f_m}(S_l)} - \sum_m w_m \ln w_m \\
&\quad + \sum_m \sum_l P(Z = m, q(X) = y_l) \ln \frac{P(Z = m, q(X) = y_l)}{P(q(X) = y_l)} \\
&= \sum_m w_m H_{f_m}(q) + H(Z) - H(Z|q(X)). \qquad (15)
\end{aligned}
$$

The above equality holds whenever $H_f(q)$ and $H(Z)$ are finite.

**Lemma 2** *Suppose $f$ is a disjoint mixture $\{f_m, \Omega_m, w_m\}$ such that $h(f)$ is finite and $H(Z) < \infty$ ($Z = m$ if $x \in \Omega_m$). If $q$ is an overall quantizer (not necessarily a composite quantizer) such that $H_f(q) < \infty$, then*

$$
\begin{aligned}
H_f(q) - h(f) &= \sum_m w_m [H_{f_m}(q) - h(f_m)] - H(Z|q(X)) \qquad (16) \\
\theta(f, \lambda, q) &= \sum_m w_m \theta(f_m, \lambda, q) - H(Z|q(X)). \qquad (17)
\end{aligned}
$$

*It follows that*

$$
\theta(f, \lambda) \leq \sum_m w_m \theta(f_m, \lambda). \qquad (18)
$$

*Proof:* Subtracting (14) from (15) gives (16), which in turn implies (17). Zador is missing the $H(Z|q(X))$ term in his analogous formula on p. 29 in the proof of his Lemma 3.3(b) [18]; he tacitly assumes it is 0. If $q$ is a composite quantizer, then $\theta(f_m, \lambda, q) = \theta(f_m, \lambda, q_m)$. Thus (18) follows from (17) since $H(Z|q(X)) \geq 0$ and for a given $\lambda$ and $\epsilon > 0$, $q_m$ can be chosen so that $\theta(f_m, \lambda, q_m) \leq \theta(f_m, \lambda) + \epsilon$ for all $m$ and hence

$$
\begin{aligned}
\sum_m w_m \theta(f_m, \lambda) + \epsilon &\geq \sum_m w_m \theta(f_m, \lambda, q_m) \\
&\geq \theta(f, \lambda, q) \\
&\geq \theta(f, \lambda).
\end{aligned}
$$

□

# 5 Bounds

The following lemmas provide useful lower and upper bounds.

**Lemma 3** *For any $f, \lambda, q$*

$$\theta(f, \lambda, q) \geq -\frac{k}{2} \ln \pi.$$

*Therefore also $\underline{\theta}(f) \geq -\frac{k}{2} \ln \pi$.*

*Proof:* Let the partition associated with $q$ be $\{S_i;\ i \in \mathcal{I}\}$, let $\{y_i;\ i \in \mathcal{I}\}$ be the associated codebook, and assume $p_i = P_f(S_i) > 0$ all $i$ (otherwise merge cells into cells with positive probability). Define $f_i(x) = f(x)1_{S_i}(x)/p_i$ and

$$g_i(x) = \frac{e^{-\frac{1}{\lambda}||x-y_i||^2}}{(\pi\lambda)^{\frac{k}{2}}}$$

the pdf for a $k$-dimensional Gaussian random vector with mean $y_i$ and covariance $\sigma^2 I_k = (\lambda/2) I_k$, where $I_k$ is the $k \times k$ identity matrix.

Then

$$\begin{aligned}
\theta(f, \lambda, q) &= \sum_i \int_{S_i} dx\, f(x) \left( \frac{1}{\lambda} ||x - y_i||^2 + \ln \frac{f(x)\lambda^{\frac{k}{2}}}{p_i} \right) \\
&= \sum_i p_i \int_{S_i} dx\, f_i(x) \ln \frac{f_i(x)}{g_i(x)\pi^{\frac{k}{2}}} \\
&= \sum_i p_i D(f_i || g_i) - \frac{k}{2} \ln \pi
\end{aligned}$$

where $D(f_i||g_i)$ is the relative entropy between the densities $f_i$ and $g_i$ defined by

$$D(f_i||g_i) = \int f_i(x) \ln \frac{f_i(x)}{g_i(x)}\, dx.$$

The lemma follows from the nonnegativity of relative entropy (see, e.g., [5]). This result also follows from the classic Shannon lower bound, but the above proof accomplishes the goal without recourse to Shannon rate-distortion theory. □

**Lemma 4** *For any $f$ satisfying the conditions of the theorem, and any $\lambda$*

$$\theta(f, \lambda) \leq k(\frac{1}{4} + \sqrt{\lambda}) + H_f(Q_1) - h(f).$$

*Therefore, if $\lambda \leq 1/16$ ,*

$$\theta(f, \lambda) \leq \frac{k}{2} + H_f(Q_1) - h(f) \tag{19}$$

*and hence*

$$\overline{\theta}(f) \leq \frac{k}{2} + H_f(Q_1) - h(f).$$

*Proof:* Fix $\lambda > 0$ and overbound $\theta(f, \lambda)$ by the performance using the uniform quantizer $Q_\Delta$ where

$$\Delta = \frac{1}{N} \tag{20}$$

$$N = \lceil \lambda^{-1/2} \rceil. \tag{21}$$

The quantizer $Q_\Delta$ divides the unit cube into $N^k$ cubes of side $\Delta$ and volume $\Delta^k$. Consider the density $f$ as a disjoint mixture of densities $f_m$ on disjoint unit cubes $C_m$ with $w_m = P_f(C_m)$. Let $q_m$ be the restriction of $Q_\Delta$ to $C_m$. By construction, $Q_\Delta$ is a composite quantizer with component quantizers $q_m$, all of which are uniform quantizers with $N^k$ small cubes of volume $\Delta^k$. If $f$ satisfies the conditions of the theorem, then $f$, $\{f_m, C_m, w_m\}$, and $Q_\Delta$ satisfy the conditions of Lemma 2. Since $H(Z|Q_\Delta(X)) = 0$ in this case, from (17) we obtain

$$\theta(f, \lambda, q) = \sum_m w_m \theta(f_m, \lambda, q_m).$$

The maximum squared error within a cube using the uniform quantizer is

$$k\left(\frac{\Delta}{2}\right)^2 \leq k\left(\frac{\sqrt{\lambda}}{2}\right)^2 = \frac{k}{4}\lambda.$$

Also, $q_m$ has $N^k$ codevectors, and so $H_{f_m}(q_m) \leq \ln N^k$. Since $N \leq \lambda^{-1/2} + 1$, we obtain

$$\begin{aligned}
\theta(f_m, \lambda, q_m) &= \frac{D_{f_m}(q_m)}{\lambda} + H_{f_m}(q_m) + \frac{k}{2}\ln\lambda - h(f_m) \\
&\leq \frac{k}{4} + k\ln(\lambda^{-1/2} + 1) + k\ln\sqrt{\lambda} - h(f_m) \\
&= \frac{k}{4} + k\ln(1 + \sqrt{\lambda}) - h(f_m) \\
&\leq \frac{k}{4} + k\sqrt{\lambda} - h(f_m)
\end{aligned}$$

where the final inequality uses the $\ln r \leq r - 1$ inequality. Thus from (14) applied to the partition into unit cubes

$$\begin{aligned}
\theta(f, \lambda, q) &= \sum_m w_m \theta(f_m, \lambda, q_m) \\
&\leq \frac{k}{4} + k\sqrt{\lambda} - \sum_m w_m h(f_m) \\
&= \frac{k}{4} + k\sqrt{\lambda} - h(f) + H_f(Q_1).
\end{aligned}$$

□

# 6 Asymptotics

**Lemma 5** *Suppose that $f$ is a disjoint mixture $\{f_m, \Omega_m, w_m\}$ which satisfies the conditions of the theorem and that $H(Z) < \infty$ ($Z = m$ if $x \in \Omega_m$). Then*

$$\overline{\theta}(f) \leq \sum_m w_m \overline{\theta}(f_m). \tag{22}$$

*Proof:*

$$\begin{aligned} \overline{\theta}(f) &= \limsup_{\lambda \to 0} \theta(f, \lambda) \\ &\leq \limsup_{\lambda \to 0} \sum_m w_m \theta(f_m, \lambda) \\ &\leq \sum_m w_m \limsup_{\lambda \to 0} \theta(f_m, \lambda) \\ &= \sum_m w_m \overline{\theta}(f_m). \end{aligned} \tag{23}$$

The interchange of the limit superior and sum follows by the upper bound to $\theta(f, \lambda)$ for small $\lambda$ of Lemma 4. Specifically, choosing $\lambda \leq 1/16$ as in the lemma, then invoking (19)

$$w_m \theta(f_m, \lambda) \leq w_m \left( \frac{k}{2} + H_{f_m}(Q_1) - h(f_m) \right).$$

(Note that $H_{f_m}(Q_1)$ and $h(f_m)$ are finite for all $m$ by (14) and (15) since $H(Q_1)$, $h(f)$, and $H(Z)$ are finite by assumption.) We have by (16)

$$\sum_m w_m \left( \frac{k}{2} + H_{f_m}(Q_1) - h(f_m) \right) = \frac{k}{2} + H_f(Q_1) - h(f) + H(Z|Q_1(X)).$$

Since the right side is finite by assumption, (23) follows from the corresponding inequality for finite sums. □

The next lemma proves that if a sequence of quantizers is approximately optimal for a disjoint mixture, then in the limit the conditional entropy of $Z$, the random variable indicating which component of the mixture is in effect, given the quantizer output tends to 0. The result plays a key role in quantifying the asymptotics of the quantities considered in Lemma 2.

**Lemma 6** *Suppose $\lambda_n, q_n$, $n = 1, 2, \ldots$ satisfy $\lim_{n\to\infty} \lambda_n = 0$, where the $\lambda_n$ are decreasing, and $\lim_{n\to\infty} \theta(f, \lambda_n, q_n) = \underline{\theta}(f) < \infty$. Suppose also that $f$ is a disjoint mixture $\{f_m, \Omega_m, w_m\}$ such that $H(Z) < \infty$ ($Z = m$ if $x \in \Omega_m$). Then $\lim_{n\to\infty} H(Z|q_n(X)) = 0$.*

*Proof:* Since the mixture is disjoint, $Z$ is a function of $X$ and hence $H(Z|q_n(X)) = I(X; Z|q_n(X))$. Thus

$$\begin{aligned} I(X; Z|q_n(X)) &= I(X, q_n(X); Z) - I(q_n(X); Z) \\ &= I(X; Z) + I(q_n(X); Z|X) - I(q_n(X); Z) \\ &= I(X; Z) - I(q_n(X); Z) \end{aligned}$$

using the fact that $I(q_n(X); Z|X) = 0$ since both $q_n(X)$ and $Z$ are functions of $X$. The assumptions of the lemma imply that

$$D_f(q_n) \leq \lambda_n \underline{\theta}(f) - \frac{k}{2}\lambda_n \ln \lambda_n + \lambda_n h(f) + \lambda_n o(1) \tag{24}$$

where $o(1) \to 0$ as $n \to \infty$, and hence

$$\lim_{n\to\infty} D_f(q_n) = 0$$

which means that $(q_n(X), Z) \to (X, Z)$ in distribution as $n \to \infty$. Since mutual information is lower semicontinuous [7],

$$\liminf_{n\to\infty} I(q_n(X); Z) \geq I(X; Z)$$

and hence

$$\begin{aligned} \limsup_{n\to\infty} H(Z|q_n(X)) &= \limsup_{n\to\infty} I(X; Z|q_n(X)) \\ &= I(X; Z) - \liminf_{n\to\infty} I(q_n(X); Z) \\ &\leq 0. \end{aligned}$$

Since entropy is nonnegative, the lemma is proved. □

Combining the lemmas yields the following corollary.

**Corollary 1** *Suppose that $f$ is a disjoint mixture $\{f_m, \Omega_m, w_m\}$ which satisfies the conditions of the theorem and that $H(Z) < \infty$ ($Z = m$ if $x \in \Omega_m$). Then*

$$\sum_m w_m \underline{\theta}(f_m) \leq \underline{\theta}(f) \leq \overline{\theta}(f) \leq \sum_m w_m \overline{\theta}(f_m). \tag{25}$$

Thus if $f_m \in \mathcal{Z}$ for all $m$, then also $f \in \mathcal{Z}$.

*Proof:* The limit superior result was proved in Lemma 5. To prove the limit inferior result, suppose that $q_n, \lambda_n$ are chosen so that $\lambda_n$ is decreasing to 0 and

$$\lim_{n\to\infty} \theta(f, \lambda_n, q_n) = \underline{\theta}(f).$$

From Lemma 2

$$\begin{aligned} \theta(f, \lambda_n, q_n) &= \sum_m w_m \theta(f_m, \lambda_n, q_n) - H(Z|q_n(X)) \\ &\geq \sum_m w_m \theta(f_m, \lambda_n) - H(Z|q_n(X)). \end{aligned}$$

The rightmost term goes to zero as $n$ grows from Lemma 6. Hence

$$\begin{aligned} \underline{\theta}(f) &\geq \liminf_{n\to\infty} \sum_m w_m \theta(f_m, \lambda_n) \\ &\geq \sum_m w_m \underline{\theta}(f_m). \end{aligned}$$

The interchange of limit inferior and sum is justified because of the finite uniform lower bound to $\theta(f, \lambda)$ of Lemma 3. □

# 7 Proof of the Theorem

The conclusions of the theorem were originally stated with incomplete conditions and a sketch of a proof in [11]. The general approach of the first, second, and fourth steps is followed with corrections and details here. The proof of the third step in [11] was incorrect and a new approach is adopted here.

## First Step: Uniform pdfs on cubes

We begin by showing that if $f$ is a uniform pdf on a cube of any size, then $f \in \mathcal{Z}$. The approach is a natural variation on Zador's original proof.

Define a cube in $\Omega$ with side $a$ and location $r$ as $C_{a,r} = \{x : r \leq x_i < r + a;\ i = 1, \ldots, k\} = [r, r+a)^k$. Abbreviate $C_{a,0}$ to $C_a$, the cube of side $a$ in the positive quadrant with one corner at the origin. In particular, any translation $C_{1,r}$ of $C_1 = [0,1)^k$ is called a unit cube. Define the corresponding uniform pdf $u_{a,r}(x) = V(C_{a,r})^{-1} 1_{C_{a,r}}(x)$. Then $V(C_{a,r}) = a^k$, $h(u_{a,r}) = \ln V(C_{a,r}) = k \ln a$, and $u_{a,r}(x) = a^{-k} 1_{C_a}(x - r) = a^{-k} u_1(\frac{x-r}{a})$. As with cubes, we simplify the notation $u_{a,0}$ to $u_a$.

We first show that shifting does not effect performance so we can confine interest to cubes located at the origin.

**Lemma 7** *Suppose that a random vector $X$ has a pdf $f$ and that $g$ is the pdf of the random vector $X - r$ for a fixed constant $r$. Then $\theta(f, \lambda) = \theta(g, \lambda)$.*

*Proof:* Given a quantizer $q$ for $X$, define a shifted quantizer $Q$ for $X - r$ by $Q(x) = q(x + r) - r$. Then a simple change of variables immediately gives

$$\theta(f, \lambda, q) = \theta(g, \lambda, Q). \tag{26}$$

Conversely, given any quantizer $Q$ for $X - r$ the shifted quantizer $q(x) = Q(x - r) + r$ also satisfies (26). Taking infima over quantizers proves the lemma. □

**Lemma 8** *Suppose we have a quantizer $q_1$ with encoder $\alpha_1 : \mathcal{C}_1 \to \mathcal{I}$ and decoder $\beta_1 : \mathcal{I} \to \mathcal{C}$ defined for the unit cube $C_1$. Define a quantizer $q_a$ with encoder $\alpha_a$ and decoder $\beta_a$ for $C_a$ by straightforward variable changes $\alpha_a(x) = \alpha_1(\frac{x}{a})$, $\beta_a(l) = a\beta_1(l)$, $q_a(x) = a q_1(\frac{x}{a})$. Then $\theta(u_a, \lambda, q_a) = \theta(u_1, a^{-2}\lambda, q_1)$, $\theta(u_a, \lambda) = \theta(u_1, a^{-2}\lambda)$.*

*Proof:* Since Shannon entropy is not changed by scaling, $H_{u_a}(q_a) = H_{u_1}(q_1)$. Changing variables yields $h(u_a) = \ln a^k + h(u_1) = \ln a^k$, $\int ||x - q_a(x)||^2 u_a(x)\, dx = a^2 \int ||x - q_1(x)||^2 u_1(x)\, dx$ and hence $\theta(u_a, \lambda) = \theta(u_1, \lambda/a^2)$. □

The lemma allows us to concentrate on $u_1(x) = 1_{C_1}(x)$, the uniform pdf on the unit cube $C_1$.

**Lemma 9** $\lim_{\lambda \to 0} \theta(u_1, \lambda) = \theta_k$, *i.e.,* $u_1 \in \mathcal{Z}$.

*Proof:* Partition the unit cube $C_1$ into $m^k$ disjoint cubes $C_{1/m}$. For each of the small cubes have a uniform pdf $f_{1/m}(x) = m^k$ on the cube. All of the small cubes have the same $\rho(f_{1/m}, \lambda)$. From Lemma 8, $\theta(f_{1/m}, \lambda) = \theta(u_1, m^2\lambda)$. From Lemma 2, $\theta(u_1, \lambda) \leq \sum_{i=1}^{m^k} \frac{1}{m^k}\theta(f_{1/m}, \lambda) = \theta(f_{1/m}, \lambda)$, which with the previous equation implies $\theta(u_1, \lambda) \leq \theta(u_1, m^2\lambda)$. Replacing $m^2\lambda$ by $\lambda$, $\theta(u_1, \lambda) \geq \theta(u_1, m^{-2}\lambda)$. Fix $\lambda$ and note that $(0, \lambda] = \bigcup_{m=1}^{\infty}(\frac{\lambda}{(m+1)^2}, \frac{\lambda}{m^2}]$ so for any $\lambda' \in (0, \lambda]$ there is an integer $m$ such that $\lambda/(m+1)^2 < \lambda' \leq \lambda/m^2$. $\rho(f, \lambda)$ is nonincreasing with decreasing $\lambda$, hence

$$\theta(u_1, \lambda) \geq \frac{\rho(u_1, \lambda')}{(\frac{m+1}{m})^2\lambda'} + \frac{k}{2}\ln\lambda' = \left(\frac{m}{m+1}\right)^2 \theta(u_1, \lambda') + \left(\frac{2m+1}{m^2+2m+1}\right)\frac{k}{2}\ln\lambda'.$$

Choose any subsequence of $\lambda'$ tending to zero. The largest possible value of the limit superior of the right side is $\overline{\theta}(u_1)$ and hence $\theta(u_1, \lambda) \geq \overline{\theta}(u_1)$ which means that $\theta_k \triangleq \inf_\lambda \theta(u_1, \lambda) \geq \overline{\theta}(u_1)$. Hence $\underline{\theta}(u_1) \geq \theta_k \geq \overline{\theta}(u_1)$ and hence the limit $\lim_{\lambda\to 0}\theta(u_1, \lambda)$ must exist and equal $\theta_k$. Note that $\theta_k$ is finite by Lemmas 3 and 4. □

## Second step: Piecewise constant pdfs on cubes

Suppose that $\{C_1(m)\}$ is a collection of disjoint unit cubes, $\{w_m\}$ is a pmf with finite entropy, and

$$f(x) = \sum_m w_m \frac{1}{V(C_1(m))} 1_{C_1(m)}(x).$$

Combining Lemmas 7-9 and Corollary 1 implies that $f \in \mathcal{Z}$.

## Third Step: Distributions on a Unit Cube

In this step it is shown that if $f$ is supported on a unit cube, then $\underline{\theta}(f) = \overline{\theta}(f) = \theta_k$ and hence $f \in \mathcal{Z}$. Suppose that $P_f(C) = 1$ for a unit cube $C$ and that $h(f) > -\infty$.

From Lemma 7 it is enough to prove the statement for $C = C_1$. For any positive integer $M$ we can partition $C_1$ into $M^k$ cubes of side length $1/M$, say $\mathcal{S}_M = \{C(m);\ 1, 2, \ldots, M^k\}$. Given a pdf $f$, form a piecewise constant approximation

$$\begin{aligned}\hat{f}^{(M)}(x) &= \sum_{m=1}^{M^k} \frac{P_f(C(m))}{V(C(m))} 1_{C(m)}(x) \\ &= \sum_{m=1}^{M^k} w_m M^k 1_{C(m)}(x).\end{aligned}$$

The use of the piecewise constant approximation to the original pdf follows that of [2, 8]. This is a disjoint mixture source with $w_m = P_f(C(m))$ and component pdfs $\hat{f}_m(x) = M^k 1_{C(m)}(x)$. If $\hat{P}_M$ denotes the distribution induced by $\hat{f}^{(M)}$, i.e., $\hat{P}_M(F) = \int_F \hat{f}^{(M)}(x)\,dx$, then $\hat{f}^{(M)} = d\hat{P}_M/dV(x)$.

**Lemma 10** $\lim_{M\to\infty}\hat{f}^{(M)}(x) = f(x)$, $V - a.e.$, $\lim_{M\to\infty}||\hat{f}^{(M)} - f||_1 = 0$, $\lim_{M\to\infty} h(\hat{f}^{(M)}) = h(f)$.

*Proof:* The first result follows by differentiation of measures (see, e.g., [16] p. 108), the second from Scheffé's lemma (see, e.g., [3]). The third result follows from the convergence of entropy for uniform scalar quantizers, e.g., [6]. For completeness we also provide a direct proof: Let $P_f$ and $P_g$ be two distributions corresponding to pdfs $f$ and $g$. The relative entropy of a measurable partition $\mathcal{S} = \{S_l;\ l = 1, 2, \ldots\}$ with distribution $P_f$ with respect to a distribution $P_g$ is

$$H_{f||g}(\mathcal{S}) = \sum_i P_f(S_i) \ln \frac{P_f(S_i)}{P_g(S_i)} \geq 0$$

where the inequality follows from the divergence inequality. If $\mathcal{S}_M,\ M = 1, 2, \ldots$ asymptotically generates the Borel field of $\Omega$, then

$$\lim_{M\to\infty} H_{f||g}(\mathcal{S}_M) = \int f(x) \ln \frac{f(x)}{g(x)}\, dx. \tag{27}$$

(See, e.g., [9].) Now $f$ and $\hat{f}^{(M)}$ are pdfs on the unit cube. We have that

$$\begin{aligned} h(\hat{f}^{(M)}) - h(f) &= \int f(x) \ln f(x)\, dx - \sum_m \int_{C(m)} \hat{f}^{(M)}(x) \ln \hat{f}^{(M)}(x)\, dx \\ &= \int f(x) \ln f(x)\, dx - \sum_m w_m \ln(w_m M^k)\, dx \\ &= \int f(x) \ln \frac{f(x)}{u_1(x)}\, dx - H_{f||u_1}(\mathcal{S}_M). \end{aligned}$$

This goes to zero from (27) and the fact that the sequence of partitions $\mathcal{S}_M$ of the unit cube into $M^k$ cubes of side length $1/M$ generates the Borel field of the unit cube. □

Suppose $q_1$ is a quantizer on $C_1$ with corresponding encoder $\alpha_1$, index set $\mathcal{I}_1$, partition $\{S_i^1 : i \in \mathcal{I}_1\}$, and decoder $\beta_1$. Let $g$ be a pdf on $C_1$ (which will be either $f$ or $\hat{f}^{(M)}$). Fix $0 < \epsilon < 1$. The next lemma (proved in Appendix B) shows that if $\lambda$ is small enough and $q_1$ is (approximately) optimal for the pdf $g$, then $q_1$ has a collection of cells with total probability between $\epsilon/2$ and $\epsilon$.

**Lemma 11** *Let $g$ be any pdf such that $\overline{\theta}(g) < \infty$ and $h(g)$ is finite, and fix $0 < \epsilon < 1$. Then there is a threshold $\lambda_0 = \lambda_0(g, \epsilon) > 0$ such that if $\lambda < \lambda_0$, then any quantizer $q_1$ that satisfies $\theta(g, \lambda, q_1) \leq \theta(g, \lambda) + \epsilon$ has a collection $\{S_i^1 : i \in \mathcal{J}_1\}$ of cells with total probability bounded as*

$$\frac{\epsilon}{2} \leq \sum_{i\in\mathcal{J}_1} P_g(S_i^1) \leq \epsilon. \tag{28}$$

Let $\lambda$ be small enough and choose $q_1$ so that it satisfies the conclusion of Lemma 11. Define

$$p^* = \sum_{i\in\mathcal{J}_1} P_g(S_i^1) \tag{29}$$

and choose the length function $\ell_1$ optimally with respect to $g$ and $q_1$, i.e.,

$$\ell_1(i) = -\ln P_g(S_i^1), \qquad \theta(g, \lambda, q_1, \ell_1) = \theta(g, \lambda, q_1). \tag{30}$$

A second quantizer $q_2$ is a uniform $k$-dimensional quantizer with side width $\Delta = 1/N$, where $N = \lfloor \lambda^{-1/2} \rfloor$, so that $N \leq \lambda^{-1/2}$, $\Delta \leq \sqrt{\lambda}/(1-\sqrt{\lambda})$, $\Delta^2 \leq \lambda/(1-2\sqrt{\lambda}) \leq 2\lambda$ if $\lambda \leq 1/16$, which we can assume without loss of generality in the asymptotic ($\lambda \to 0$) analysis. Then for all $x \in C_1$,

$$d(x, q_2(x)) \leq k\frac{\Delta^2}{4} \leq \frac{k}{2}\lambda.$$

Let $\alpha_2$ and $\mathcal{I}_2$ denote the encoder and index set of $q_2$, and define the (constant) length function $\ell_2$ by

$$\ell_2(i) = -\ln p^* + 1 - \frac{k}{2}\ln\lambda.$$

Note that $\ell_2$ is admissible since

$$\sum_{i\in\mathcal{I}_2} e^{-\ell_2(i)} = N^k e^{-1} p^* \lambda^{k/2} \leq e^{-1}p^* < 1. \tag{31}$$

The quantizer $q_1$ is designed to be good for a particular pdf while the quantizer $q_2$ is designed to provide a bound on the distortion and length which will be valid for any pdf. A composite quantizer $\bar{q}$ can be formed by merging $q_1$ and $q_2$ which will still be well-matched to a specified pdf, but will now also have a uniform bound on distortion and length over all pdfs. This bound will permit us to bound the performance resulting from applying the quantizer to distinct pdfs. The merging is accomplished by the universal coding technique of considering the union codebook and simply finding the minimum Lagrangian distortion codeword in the combined codebook: given an input vector $x$, to find the code yielding the smallest Lagrangian distortion, i.e, let

$$m(x) = \operatorname*{argmin}_l (d(x, q_l(x)) + \lambda\ell_l(\alpha_l(x)))$$

define the encoder of $\bar{q}$ by

$$\bar{\alpha}(x) = (m, i) = (m(x), \alpha_{m(x)}(x))$$

and define the decoder by $\bar{\beta}(m, i) = \beta_m(i)$. Recall the definition of the index subset $\mathcal{J}_1 \subset \mathcal{I}_1$ in (28), and define the length function for $\bar{q}$ as

$$\bar{\ell}(m,i) = \begin{cases} \ell_1(i) & \text{if } m = 1 \text{ and } \ i \in \mathcal{I}_1 \setminus \mathcal{J}_1 \\ \ell_1(i) + 1 & \text{if } m = 1 \text{ and } i \in \mathcal{J}_1 \\ \ell_2(i) & \text{if } m = 2. \end{cases}$$

Then $\bar{\ell}$ is admissible since by (30), (29), and (31),

$$\begin{aligned} \sum_{m,i} e^{-\bar{\ell}(m,i)} &= \sum_{i\in\mathcal{I}_1\setminus\mathcal{J}_1} e^{-\ell_1(i)} + \sum_{i\in\mathcal{J}_1} e^{-\ell_1(i)-1} + \sum_{i\in\mathcal{I}_2} e^{-\ell_2(i)} \\ &\leq \sum_{i\in\mathcal{I}_1\setminus\mathcal{J}_1} P_g(S_i^1) + \sum_{i\in\mathcal{J}_1} e^{-1}P_g(S_i^1) + e^{-1}p^* \\ &= 1 - p^* + e^{-1}p^* + e^{-1}p^* \\ &\leq 1. \end{aligned}$$

Set $B = \{x : m(x) = 2\}$ and $W = \bigcup_{i \in \mathcal{J}_1} S_i^1$. Then the definition of $\bar{\ell}$ implies

$$d(x, \bar{q}(x)) + \lambda \bar{\ell}(\bar{\alpha}(x)) = \min_l [d(x, q_l(x)) + \lambda \ell_l(\alpha_l(x))] + \lambda 1_{W \cap B^c}(x)$$

and hence

$$d(x, \bar{q}(x)) + \lambda \bar{\ell}(\bar{\alpha}(x)) \leq d(x, q_l(x)) + \lambda \ell_l(\alpha_l(x)) + \lambda 1_{W \cap B^c}(x); \quad l = 1, 2. \tag{32}$$

In particular, the upper bound for $l = 2$ implies

$$d(x, \bar{q}(x)) + \lambda \bar{\ell}(\bar{\alpha}(x)) \leq \left( \frac{k}{2} - \ln p^* + 1 \right) \lambda - \frac{k}{2} \lambda \ln \lambda. \tag{33}$$

The next lemma is proved in Appendix C.

**Lemma 12** *The quantizer $\bar{q}$ satisfies*

$$\begin{aligned} &|\theta(f, \lambda, \bar{q}, \bar{\ell}) - \theta(\hat{f}^{(M)}, \lambda, \bar{q}, \bar{\ell})| \\ &\qquad \leq \left[ \left( \frac{k}{2} - \ln p^* + 1 \right) + \frac{k}{2} \ln \pi \right] \|f - \hat{f}^{(M)}\|_1 \\ &\qquad\quad + |h(f) - h(\hat{f}^{(M)})| + b(M) \end{aligned}$$

*where $b(M)$ depends only on $f$ and $M$, and $\lim_{M \to \infty} b(M) = 0$.*

Fix $M$ large enough such that

$$\left[ \frac{k}{2} - \ln \frac{\epsilon}{2} + 1 + \frac{k}{2} \ln \pi \right] \frac{1}{2} \|f - \hat{f}^{(M)}\|_1 + b(M) \leq \epsilon. \tag{34}$$

Use the design pdf $g = \hat{f}^{(M)}$ to construct $q_1$. Then $q_1$ satisfies Lemma 11 for all $\lambda$ small enough. Recall that by construction,

$$\theta(\hat{f}^{(M)}, \lambda, q_1, \ell_1) = \theta(\hat{f}^{(M)}, \lambda, q_1) \leq \theta(\hat{f}^{(M)}, \lambda) + \epsilon.$$

This and (32) imply

$$\begin{aligned} \theta(\hat{f}^{(M)}, \lambda, \bar{q}, \bar{\ell}) &\leq \int \left( \frac{d(x, q_1(x))}{\lambda} + \ell_1(\alpha_1(x)) + 1_{W \cap B^c}(x) \right) \hat{f}^{(M)}(x)\, dx \\ &\quad + \frac{k}{2} \ln \lambda - h(\hat{f}^{(M)}) \\ &= \theta(\hat{f}^{(M)}, \lambda, q_1) + P_{\hat{f}^{(M)}}(W \cap B^c) \\ &\leq \theta(\hat{f}^{(M)}, \lambda) + 2\epsilon \end{aligned}$$

where the last inequality holds since $P_{\hat{f}^{(M)}}(W \cap B^c) \leq p^* \leq \epsilon$. Combine this with the bound of Lemma 12 to obtain

$$\begin{aligned} \theta(f, \lambda) \leq \theta(f, \lambda, \bar{q}, \bar{\ell}) &\leq \theta(\hat{f}^{(M)}, \lambda, \bar{q}, \bar{\ell}) + \left[ \frac{k}{2} - \ln p^* + 1 + \frac{k}{2} \ln \pi \right] \|f - \hat{f}^{(M)}\|_1 \\ &\quad + |h(f) - h(\hat{f}^{(M)})| + b(M) \\ &\leq \theta(\hat{f}^{(M)}, \lambda) + 2\epsilon + \left[ \frac{k}{2} - \ln p^* + 1 + \frac{k}{2} \ln \pi \right] \|f - \hat{f}^{(M)}\|_1 \\ &\quad + |h(f) - h(\hat{f}^{(M)})| + b(M) \\ &\leq \theta(\hat{f}^{(M)}, \lambda) + 3\epsilon \end{aligned}$$

where the last inequality follows from (34) and the fact that $p^* \geq \epsilon/2$. Since this bound holds for all $\lambda$ small enough, we obtain $\overline{\theta}(f) \leq \overline{\theta}(\hat{f}^{(M)}) + 3\epsilon$. This is equivalent to $\overline{\theta}(f) \leq \theta_k + 3\epsilon$ since $\hat{f}^{(M)}$ has the Zador property. Thus $\overline{\theta}(f) \leq \theta_k$ since $\epsilon > 0$ was arbitrary. The converse inequality $\theta_k \leq \underline{\theta}(f)$ is proved in a similar fashion using the design pdf $g = f$.

### Final step: Proof of theorem

Carve $\Omega$ into disjoint unit cubes $C_1(m)$ and write the pdf $f$ as the disjoint mixture

$$f(x) = \sum_m P_f(C_1(m)) f_m(x), \quad f_m(x) = \frac{f(x)}{P_f(C_1(m))} 1_{C_1(m)}(x).$$

From Corollary 1,

$$\sum_m w_m \underline{\theta}(f_m) \leq \underline{\theta}(f) \leq \overline{\theta}(f) \leq \sum_m w_m \overline{\theta}(f_m). \tag{35}$$

From the previous step, $\underline{\theta}(f_m) = \overline{\theta}(f_m) = \theta_k$ for all $m$, and hence $\underline{\theta}(f) = \overline{\theta}(f) = \theta_k$, which proves the theorem.

## Appendix A

### Proof of Lemma 1

The result was first stated in [14]. The proof here follows similar lines, but corrects several errors. Analogous to the $\theta$ notation introduced earlier for the Lagrangian formulation, we define similar quantities for the traditional form. Recall that the unit of entropy is usually nats, but bits will be used when entropies appear in an exponent of 2.

$$\begin{aligned}
\zeta(f, R, q) &= D_f(q) 2^{\frac{2}{k}(R - h(f))} \\
\zeta(f, R) &= \inf_{q: H_f(q) \leq R} \zeta(f, R, q) \\
\underline{\zeta}(f) &= \liminf_{R \to \infty} \zeta(f, R) \\
\overline{\zeta}(f) &= \limsup_{R \to \infty} \zeta(f, R).
\end{aligned}$$

Thus we have

$$\delta_f(R) 2^{\frac{2}{k}(R - h(f))} = \zeta(f, R).$$

The traditional form of Zador's property can now be described as $\underline{\zeta}(f) = \overline{\zeta}(f) = b(2, k)$ and the Lagrangian form as $\underline{\theta}(f) = \overline{\theta}(f) = \theta_k$.

The connection between the limits in one direction follows from the following equality, which is used repeatedly in the proof.

$$\begin{aligned}\theta(f,\lambda,q) &= \frac{k}{2}\left[\frac{2D_f(q)}{k\lambda} - \ln\frac{2D_f(q)}{k\lambda} - 1\right] \\ &\quad + \frac{k}{2}\ln\left(\frac{2e}{k}D_f(q)2^{\frac{2}{k}(H_f(q)-h(f))}\right). \end{aligned} \tag{36}$$

The term in the square brackets is nonnegative since $\ln r \le r-1$.

Since $\underline{\theta}(f)$ is finite by Lemma 3, just as in Lemma 6 we can choose $\lambda_n \to 0$ as $n \to \infty$ so that $\theta(f,\lambda_n) \to \underline{\theta}(f)$ and hence a sequence of quantizers $q_n$ exists such that $\theta(f,\lambda_n,q_n) \to \underline{\theta}(f)$. Thus by eq. (24),

$$D_f(q_n) \to 0. \tag{37}$$

Define $\lambda_n^* = 2D_f(q_n)/k$ and observe that by (36)

$$\begin{aligned}\theta(f,\lambda_n^*,q_n) &= \frac{k}{2}\ln\left(\frac{2e}{k}D_f(q_n)2^{\frac{2}{k}(H_f(q_n)-h(f))}\right) \\ &\le \theta(f,\lambda_n,q_n). \end{aligned} \tag{38}$$

From Lemma 3, $\theta(f,\lambda,q) \ge -\frac{k}{2}\ln\pi$. Since $D_f(q_n) \to 0$, this necessarily implies that

$$H_f(q_n) \to \infty. \tag{39}$$

Since $q_n$ has entropy $H_f(q_n)$ and distortion $D_f(q_n) \le \delta_f(H_f(q_n))$, the minimum average distortion over all quantizers having rate $H_f(q_n)$, we have that

$$\begin{aligned}\underline{\theta}(f) &= \lim_{n\to\infty}\theta(f,\lambda_n,q_n) \\ &\ge \liminf_{n\to\infty}\frac{k}{2}\ln\left(\frac{2e}{k}D_f(q_n)2^{\frac{2}{k}(H_f(q_n)-h(f))}\right) \\ &\ge \liminf_{n\to\infty}\frac{k}{2}\ln\left(\frac{2e}{k}\delta_f(H_f(q_n))2^{\frac{2}{k}(H_f(q_n)-h(f))}\right) \\ &= \frac{k}{2}\ln\left(\frac{2e}{k}\liminf_{n\to\infty}\delta_f(H_f(q_n))2^{\frac{2}{k}(H_f(q_n)-h(f))}\right) \\ &\ge \frac{k}{2}\ln\frac{2e}{k}\underline{\zeta}(f). \end{aligned}$$

Summarizing,

$$\underline{\theta}(f) \ge \frac{k}{2}\ln\frac{2e}{k}\underline{\zeta}(f). \tag{40}$$

Now suppose that Zador's traditional result holds, hence $\underline{\zeta}(f) = \overline{\zeta}(f) = b(2,k)$ and for any sequence $R_n \to \infty$ there is a sequence of quantizers $q_n$ with $H_f(q_n) \le R_n$ for which $\zeta(f,R_n,q_n) \to b(2,k)$ so that

$$D_f(q_n)2^{\frac{2}{k}(R_n-h(f))} \to b(2,k). \tag{41}$$

Choose $\lambda_n \to 0$ such that $\theta(f, \lambda_n) \to \overline{\theta}(f)$. For this sequence $\lambda_n$, define

$$R_n = h(f) + \frac{k}{2} \ln \frac{2b(2,k)}{k\lambda_n} \tag{42}$$

and construct $q_n$ as in (41) for this $R_n$. Then

$$\begin{aligned}
\frac{k}{2} \ln \frac{2e}{k} b(2,k) &= \lim_{n\to\infty} \frac{k}{2} \ln \frac{2e}{k} D_f(q_n) 2^{\frac{2}{k}(R_n - h(f))} \\
&\geq \liminf_{n\to\infty} \left( \frac{k}{2} \ln \frac{2e}{k} D_f(q_n) 2^{\frac{2}{k}(H_f(q_n) - h(f))} \right) \\
&= \liminf_{n\to\infty} \Bigg( \theta(f, \lambda_n, q_n) - \\
&\qquad \frac{k}{2} \left[ \frac{2D_f(q_n)}{k\lambda_n} - \ln \frac{2D_f(q_n)}{k\lambda_n} - 1 \right] \Bigg).
\end{aligned}$$

Since $\theta(f, \lambda_n, q_n) \geq \theta(f, \lambda_n)$ and the ratios in square brackets go to 1 (based on (41) and (42)), it follows that

$$\frac{k}{2} \ln \frac{2e}{k} b(2,k) \geq \liminf_{n\to\infty} \theta(f, \lambda_n) = \overline{\theta}(f).$$

This combined with (40) completes the proof that if the traditional Zador limit holds, then so does the Lagrangian form with

$$\theta_k = \frac{k}{2} \ln \frac{2e}{k} b(2,k). \tag{43}$$

Suppose instead that the Lagrangian form of Zador's theorem holds, so that $\underline{\theta}(f) = \overline{\theta}(f) = \theta_k$. Let $\hat{\delta}_f(R)$ denote the convex hull of $\delta_f(R)$ defined as the largest convex function on $(0, \infty)$ that is majorized by $\delta_f(R)$. First we show that

$$\lim_{R\to\infty} \hat{\delta}_f(R) 2^{\frac{2}{k}(R - h(f))} = b(2,k) \tag{44}$$

where

$$b(2,k) = \frac{k}{2} e^{\frac{2}{k}\theta_k - 1}$$

and then we prove that (44) implies

$$\underline{\zeta}(f) = \liminf_{R\to\infty} \delta_f(R) 2^{\frac{2}{k}(R - h(f))} \geq b(2,k) \tag{45}$$

and

$$\overline{\zeta}(f) = \limsup_{R\to\infty} \delta_f(R) 2^{\frac{2}{k}(R - h(f))} \leq b(2,k). \tag{46}$$

To show (44) note that $\rho(f, \lambda) - \lambda R$ is the largest affine function with slope $-\lambda$ that is majorized by $\delta_f(R)$. Since $\hat{\delta}_f(R)$ is the pointwise supremum of all affine functions that are majorized by $\delta_f(R)$ (see, e.g., [12]), and $\delta_f(R)$ is nonincreasing,

$$\hat{\delta}_f(R) = \sup_{\lambda > 0} \Big( \rho(f, \lambda) - \lambda R \Big).$$

By assumption $\theta(f,\lambda) = \theta_k + o(1)$, where $o(1) \to 0$ as $\lambda \to 0$; hence

$$\rho(f,\lambda) - \lambda R = \lambda\Big(\theta_k + o(1) + h(f) - R - \frac{k}{2}\ln\lambda\Big) \tag{47}$$

where the expression in parentheses is positive for all $\lambda > 0$ small enough. On the other hand,

$$\rho(f,\lambda) \le D_f(Q_1) + \lambda H_f(Q_1) \le \frac{k}{4} + \lambda H_f(Q_1)$$

hence for all $R > H_f(Q_1)$,

$$\rho(f,\lambda) - \lambda R \le 0 \quad \text{if} \quad \lambda \ge \lambda_R \triangleq \frac{k}{4(R - H_f(Q_1))}.$$

It follows that

$$\hat{\delta}_f(R) = \sup_{\lambda\in(0,\lambda_R]} \Big(\rho(f,\lambda) - \lambda R\Big).$$

Fix an arbitrary $\epsilon > 0$. Since $\lambda_R \to 0$ as $R \to \infty$, (47) implies that for all $R$ large enough

$$\hat{\delta}_f(R) \le \sup_{\lambda\in(0,\lambda_R]} \lambda\Big(\theta_k + \epsilon + h(f) - R - \frac{k}{2}\ln\lambda\Big)$$
$$\hat{\delta}_f(R) \ge \sup_{\lambda\in(0,\lambda_R]} \lambda\Big(\theta_k - \epsilon + h(f) - R - \frac{k}{2}\ln\lambda\Big).$$

The suprema are readily evaluated by differentiation taking advantage of the concavity of $-\lambda\ln\lambda$. More directly one can use the inequality $\ln r - r \le -1$ to show that for $c = \theta_k \pm \epsilon + h(f) - R$

$$\lambda(\frac{2}{k}c - \ln\lambda) = \lambda(\ln\frac{e^{\frac{2}{k}c-1}}{\lambda} + 1) \le e^{\frac{2}{k}c-1}$$

where equality holds if and only if $\lambda = e^{\frac{2}{k}c-1}$. Note that $e^{\frac{2}{k}c-1} \le \lambda_R$ for large enough $R$; hence

$$\hat{\delta}_f(R) \le \frac{k}{2} e^{\frac{2}{k}(\theta_k+\epsilon)-1} 2^{-\frac{2}{k}(R-h(f))}$$
$$\hat{\delta}_f(R) \ge \frac{k}{2} e^{\frac{2}{k}(\theta_k-\epsilon)-1} 2^{-\frac{2}{k}(R-h(f))}.$$

Since $\epsilon > 0$ was arbitrary, we obtain (44).

By definition, $\hat{\delta}_f(R) \le \delta_f(R)$; hence (44) immediately yields (45). To prove the converse inequality, observe that (46) will follow from (44) via normalization and scaling if we can show that for any nonincreasing function $\delta(t)$ on $(0,\infty)$ whose convex hull $\hat{\delta}(t)$ satisfies

$$\lim_{t\to\infty} \hat{\delta}(t)e^t = 1 \tag{48}$$

we have

$$\limsup_{t\to\infty} \delta(t)e^t \le 1. \tag{49}$$

The proof is by contradiction. Assume (49) does not hold and so there exist $a > 0$ and a sequence $t_n \to \infty$ such that $\delta(t_n) \geq (1+a)e^{-t_n}$ for all $n$. Fix $0 < \epsilon < 1$ such that $\epsilon < a$ and choose $n$ large enough such that for all $t \geq t_n - \ln \frac{1+a}{1-\epsilon}$,

$$(1-\epsilon)e^{-t} < \hat{\delta}(t) < (1+\epsilon)e^{-t}. \tag{50}$$

Let $t'_n$ be the unique solution of the equation

$$\hat{\delta}(t'_n) = (1+a)e^{-t_n}. \tag{51}$$

(For $n$ large enough a unique solution $t'_n$ always exists since (48) and the fact that $\hat{\delta}(t)$ is convex imply that $\hat{\delta}(t)$ is strictly decreasing and $\lim_{t\to\infty} \hat{\delta}(t) = 0$.) Note that by (50)

$$t_n - \ln \frac{1+a}{1-\epsilon} < t'_n < t_n - \ln \frac{1+a}{1+\epsilon} < t_n. \tag{52}$$

Since $\delta(t)$ is nonincreasing and $\delta(t_n) \geq (1+a)e^{-t_n}$, we have $\delta(t) \geq (1+a)e^{-t_n}$ for $t \leq t_n$. Therefore the line segment in the $(t,\delta)$-plane joining the points $(t'_n, \hat{\delta}(t'_n))$ and $(t_n, \hat{\delta}(t_n))$ lies below $\delta(t)$. Since $\hat{\delta}(t)$ is the largest convex function such that $\hat{\delta}(t) \leq \delta(t)$, this implies that for all $t > 0$,

$$\hat{\delta}(t) \geq -\frac{\hat{\delta}(t'_n) - \hat{\delta}(t_n)}{t_n - t'_n}(t - t_n) + \hat{\delta}(t_n).$$

Define $h = \ln \frac{1+a}{1-\epsilon}$. Then $t_n - t'_n < h$ by (52), so for $t_n - h \leq t \leq t_n$,

$$\begin{aligned}
\hat{\delta}(t) &\geq -\frac{\hat{\delta}(t'_n) - \hat{\delta}(t_n)}{t_n - t'_n}(t - t_n) + \hat{\delta}(t_n) \\
&\geq -\frac{\hat{\delta}(t'_n) - \hat{\delta}(t_n)}{h}(t - t_n) + \hat{\delta}(t_n) \\
&\geq -\frac{\hat{\delta}(t'_n) - (1-\epsilon)e^{-t_n}}{h}(t - t_n) + (1-\epsilon)e^{-t_n} \\
&= -\frac{(1-\epsilon)e^{-t_n+h} - (1-\epsilon)e^{-t_n}}{h}(t - t_n) + (1-\epsilon)e^{-t_n}
\end{aligned}$$

where the third inequality follows from (50) and the last equality from (51). We obtain that for $t_n - h \leq t \leq t_n$,

$$\begin{aligned}
\hat{\delta}(t) - (1+\epsilon)e^{-t} &\geq (1-\epsilon)e^{-t_n}\left(-\frac{e^h - 1}{h}(t - t_n) + 1 - \frac{1+\epsilon}{1-\epsilon}e^{-t+t_n}\right) \\
&\triangleq (1-\epsilon)e^{-t_n} g(t).
\end{aligned}$$

Again using either calculus or the $\ln r \leq r - 1$ inequality the maximum of $g(t)$ is seen to be achieved at

$$t^*_n = t_n + \ln \frac{1+\epsilon}{1-\epsilon} - \ln \frac{e^h - 1}{h}.$$

Note that since $(e^h - 1)/h \le e^h$ for all $h \ge 0$, we have $t_n^* \ge t_n - h$. Also, it is easy to see that $t_n^* \le t_n$ if $\epsilon$ is small enough. Choosing such $\epsilon$ and a corresponding large $t_n$ we obtain

$$\max_{t\in[t_n-h,t_n]} g(t) = g(t_n^*) = \frac{e^h-1}{h}\left(\ln\frac{e^h-1}{h} - 1\right) + 1 - \frac{e^h-1}{h}\ln\frac{1+\epsilon}{1-\epsilon}.$$

Note that $h \to \log(1+a) > 0$ as $\epsilon \to 0$, implying that the rightmost term converges to zero as $\epsilon \to 0$. Since $(e^h-1)/h > 1$ for all $h > 0$ and $y(\ln y - 1) + 1 > 0$ for all $y > 1$, we have

$$\lim_{\epsilon\to 0} \frac{e^h-1}{h}\left(\ln\frac{e^h-1}{h} - 1\right) + 1 > 0.$$

Therefore we can choose $\epsilon > 0$ small enough and $t_n$ large enough so that

$$\max_{t\in[t_n-h,t_n]} (\hat{\delta}(t) - (1+\epsilon)e^{-t}) \ge \max_{t\in[t_n-h,t_n]} (1-\epsilon)e^{-t_n} g(t) > 0$$

contradicting (50). We conclude that (48) implies (49) which completes the proof that the theorem implies the traditional Zador conclusions. □

# Appendix B

## Proof of Lemma 11

First we show that if for all $\lambda > 0$ we choose $q_\lambda$ to satisfy $\theta(g,\lambda,q_\lambda) \le \theta(g,\lambda) + \epsilon$, then the cells $\{S_{i,\lambda}\}$ of $q_\lambda$ satisfy

$$\lim_{\lambda\to 0} \max_i P_g(S_{i,\lambda}) = 0. \tag{53}$$

Choose $\lambda$ small enough so that $\theta(g,\lambda) \le \overline{\theta}(g) + \epsilon$, and let $q_\lambda$ be such that $\theta(g,\lambda,q_\lambda) \le \theta(g,\lambda) + \epsilon$. Then

$$D_g(q_\lambda) \le \lambda\overline{\theta}(g) + \lambda 2\epsilon - \lambda h(g) - \frac{k}{2}\lambda\ln\lambda = o(1) \tag{54}$$

where $o(1) \to 0$ as $\lambda \to 0$. Denote the codevector associated with $S_{i,\lambda}$ by $y_{i,\lambda}$, and define

$$d_g(S_{i,\lambda}) = \int_{S_{i,\lambda}} \|x - y_{i,\lambda}\|^2 g(x)\,dx.$$

Fix $c > 0$ and let $A_c = \{x : g(x) > c\}$. Then

$$\begin{aligned} d_g(S_{i,\lambda}) &\ge c\int_{S_{i,\lambda}\cap A_c} \|x - y_{i,\lambda}\|^2\,dx \\ &\ge cV(S_{i,\lambda}\cap A_c)^{1+2/k} G_k \end{aligned}$$

where $G_k$ is the normalized second moment of a $k$-dimensional sphere (see, e.g., [10]). Since $D_g(q_\lambda) = \sum_i d_g(S_{i,\lambda})$, this and (54) imply

$$\lim_{\lambda\to 0} \max_i V(S_{i,\lambda}\cap A_c) = 0$$

from which it follows that $\lim_{\lambda\to 0}\max_i P_g(S_{i,\lambda}\cap A_c)=0$ by the absolute continuity of $P_g$ with respect to the Lebesgue measure (see, e.g., [1]). Note that

$$P_g(S_{i,\lambda})\le P_g(S_{i,\lambda}\cap A_c)+P_g(\Re^k\setminus A_c)$$

and so

$$\lim_{\lambda\to 0}\max_i P_g(S_{i,\lambda})\le P_g(\Re^k\setminus A_c).$$

Since $\lim_{c\to 0}P_g(\Re^k\setminus A_c)=\lim_{c\to 0}P_g(\{x:g(x)\le c\})=0$, (53) follows.

The statement of the lemma follows by noticing that if $\max_i P_g(S_{i,\lambda})<\epsilon/2$, then there must exist a collection of partition cells with total probability between $\epsilon/2$ and $\epsilon$. Note in the above proof that the upper bounds on $\max_i P_g(S_{i,\lambda})$ depend only on $g$, $\lambda$, and $\epsilon$, and not on the particular choice of $q_\lambda$. Therefore the conclusion holds for any $q_1$ with $\theta(g,\lambda,q_1)\le\theta(g,\lambda)+\epsilon$ if $\lambda$ is less than a threshold depending only on $\epsilon$ and $g$. □

# Appendix C

## Proof of Lemma 12

By definition,

$$\begin{aligned}
&|\theta(f,\lambda,\bar q,\bar\ell)-\theta(\hat f^{(M)},\lambda,\bar q,\bar\ell)|\\
&\quad= \Big|\int[\frac{d(x,\bar q(x))}{\lambda}+\bar\ell(\bar\alpha(x))+\frac{k}{2}\ln\lambda]f(x)\,dx-h(f)\\
&\qquad-\int[\frac{d(x,\bar q(x))}{\lambda}+\bar\ell(\bar\alpha(x))+\frac{k}{2}\ln\lambda]\hat f^{(M)}(x)\,dx+h(\hat f^{(M)})\Big|\\
&\quad\le \Big|\int[\frac{d(x,\bar q(x))}{\lambda}+\bar\ell(\bar\alpha(x))+\frac{k}{2}\ln\lambda][f(x)-\hat f^{(M)}(x)]\,dx\Big|\\
&\qquad+|h(f)-h(\hat f^{(M)})|. \qquad (55)
\end{aligned}$$

For any real number $y$, let $y^+=\max(y,0)$ and $y^-=\max(-y,0)$, so that

$$y=y^+-y^-\ ,\quad |y|=y^++y^-. \qquad (56)$$

Then

$$\begin{aligned}
&\Big|\int[\frac{d(x,\bar q(x))}{\lambda}+\bar\ell(\bar\alpha(x))+\frac{k}{2}\ln\lambda][f(x)-\hat f^{(M)}(x)]\,dx\Big|\\
&\quad= \Big|\int[\frac{d(x,\bar q(x))}{\lambda}+\bar\ell(\bar\alpha(x))+\frac{k}{2}\ln\lambda][f(x)-\hat f^{(M)}(x)]^+\,dx\\
&\qquad-\int[\frac{d(x,\bar q(x))}{\lambda}+\bar\ell(\bar\alpha(x))+\frac{k}{2}\ln\lambda][f(x)-\hat f^{(M)}(x)]^-\,dx\Big|. \qquad (57)
\end{aligned}$$

The upper bound (33) implies

$$\begin{aligned}
&\int[\frac{d(x,\bar q(x))}{\lambda}+\bar\ell(\bar\alpha(x))+\frac{k}{2}\ln\lambda][f(x)-\hat f^{(M)}(x)]^+\,dx\\
&\qquad\le \left(\frac{k}{2}-\ln p^*+1\right)\int[f(x)-\hat f^{(M)}(x)]^+\,dx. \qquad (58)
\end{aligned}$$

Note that (56) and the fact that $\int [f(x) - f^{(M)}(x)]\, dx = 0$ imply

$$\int [f(x) - \hat{f}^{(M)}(x)]^+ \, dx = \int [f(x) - \hat{f}^{(M)}(x)]^- \, dx = \frac{1}{2}\|f - \hat{f}^{(M)}\|_1$$

and so the function

$$F_M(x) = \frac{[f(x) - \hat{f}^{(M)}(x)]^+}{\frac{1}{2}\|f - \hat{f}^{(M)}\|_1}$$

is a pdf. Thus

$$\begin{aligned}
&\int [\frac{d(x, \bar{q}(x))}{\lambda} + \bar{\ell}(\bar{\alpha}(x)) + \frac{k}{2}\ln\lambda][f(x) - \hat{f}^{(M)}(x)]^+ \, dx \\
&\qquad = \Big(\theta(F_M, \lambda, \bar{q}, \bar{\ell}) + h(F_M)\Big)\frac{1}{2}\|f - \hat{f}^{(M)}\|_1 \\
&\qquad \geq \Big(\theta(F_M, \lambda, \bar{q}) + h(F_M)\Big)\frac{1}{2}\|f - \hat{f}^{(M)}\|_1 \\
&\qquad \geq \Big(-\frac{k}{2}\ln\pi + h(F_M)\Big)\frac{1}{2}\|f - \hat{f}^{(M)}\|_1 \qquad (59)
\end{aligned}$$

where in the last step we used the bound of Lemma 3. We have

$$\begin{aligned}
&\frac{1}{2}\|f - \hat{f}^{(M)}\|_1 h(F_M) \\
&\quad = \frac{1}{2}\|f - \hat{f}^{(M)}\|_1 \ln\Big(\frac{1}{2}\|f - \hat{f}^{(M)}\|_1\Big) - \int [f(x) - \hat{f}^{(M)}(x)]^+ \ln[f(x) - \hat{f}^{(M)}(x)]^+ \, dx.
\end{aligned}$$

By Lemma 10, $\lim_{M\to\infty} \hat{f}^{(M)}(x) = f(x)$ $V$-almost everywhere. Since $[f(x) - \hat{f}^{(M)}(x)]^+ \leq f(x)$,

$$|[f(x) - \hat{f}^{(M)}(x)]^+ \ln[f(x) - \hat{f}^{(M)}(x)]^+| \leq \max(e^{-1}, |f(x)\ln f(x)|).$$

Since $\int |f(x)\ln f(x)| < \infty$ and $[f(x) - \hat{f}^{(M)}(x)]^+$ is supported in the closure of $C_1$, the dominated convergence theorem implies

$$\lim_{M\to 0} \int [f(x) - \hat{f}^{(M)}(x)]^+ \ln[f(x) - \hat{f}^{(M)}(x)]^+ \, dx = 0.$$

Hence from Lemma 10

$$\lim_{M\to\infty} \frac{1}{2}\|f - \hat{f}^{(M)}\|_1 h(F_M) = 0. \qquad (60)$$

Letting $b_1(M) = \frac{1}{2}\|f - \hat{f}^{(M)}\|_1 |h(F_M)|$ and combining (58) and (59) we obtain

$$\begin{aligned}
&\Big|\int [\frac{d(x, \bar{q}(x))}{\lambda} + \bar{\ell}(\bar{\alpha}(x)) + \frac{k}{2}\ln\lambda][f(x) - \hat{f}^{(M)}(x)]^+ \, dx\Big| \\
&\qquad \leq \left[\frac{k}{2} - \ln p^* + 1 + \frac{k}{2}\ln\pi\right]\frac{1}{2}\|f - \hat{f}^{(M)}\|_1 + b_1(M)
\end{aligned}$$

where $b_1(M) \to 0$ as $M \to \infty$ by (60). A similar argument shows that

$$\left| \int [\frac{d(x, \bar{q}(x))}{\lambda} + \bar{\ell}(\bar{\alpha}(x)) + \frac{k}{2} \ln \lambda][f(x) - \hat{f}^{(M)}(x)]^- \, dx \right|$$
$$\leq \left[ \frac{k}{2} - \ln p^* + 1 + \frac{k}{2} \ln \pi \right] \frac{1}{2} \|f - \hat{f}^{(M)}\|_1 + b_2(M)$$

where $b_2(M) \to 0$ as $M \to \infty$. Let $b(M) = b_1(M) + b_2(M)$ and combine these bounds with (55) and (57) to obtain the bound of the lemma. □

*Acknowledgements*

The authors acknowledge the many helpful comments of Dave Neuhoff, Ken Zeger, András György, and of the students and colleagues in the Compression and Classification Group of Stanford University.

# References

[1] R. B. Ash, *Real analysis and probability.* New York: Academic Press, 1972.

[2] J. A. Bucklew and G. L. Wise, "Multidimensional asymptotic quantization theory with $r$th power distortion measures," *IEEE Trans. Inform. Theory*, vol. 28, pp. 239–247, March 1982.

[3] P. Billingsley, *Convergence of Probability Measures.* New York: Wiley, 1968.

[4] 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, January 1989.

[5] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[6] I. Csiszár, "Generalized entropy and quantization problems," *Proc. Sixth Prague Conf.*, pp. 159–174, 1973.

[7] I. Csiszár. "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, pp. 57–70, 1974.

[8] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Springer, Lecture Notes in Mathematics, 1730, Berlin, 2000.

[9] R. M. Gray, *Entropy and Information Theory*, Springer–Verlag, 1990.

[10] R.M. Gray and D.L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2325–2384, October 1998.

[11] R.M. Gray and J. Li, "On Zador's Entropy-Constrained Quantization Theorem," *Proceedings Date Compression Conference 2001*, IEEE Computer Science Press, Los Alimitos, pp. 3–12, March 2001.

[12] R. T. Rockafellar, *Convex Analysis.* Princeton, NJ: Princeton University Press, 1970.

[13] D. J. Sakrison, "Worst sources and robust codes for difference distortion measures," *IEEE Trans. Inform. Theory*, vol. 21, pp. 301–309, May 1975.

[14] J. Shih, A. Aiyer, and R.M. Gray, "A Lagrangian formulation of high rate quantization," *Proceedings ICASSP 2001.*

[15] L. Vandenberghe, S. Boyd, and S.-P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 19, pp. 499-533, 1998.

[16] R. L. Wheeden and A. Zygmund, *Measure and Integral.* New York: Marcel Dekker, 1977.

[17] P. L. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. Dissertation, Stanford University, 1963. (Also Stanford University Department of Statistics Technical Report.)

[18] P. L. Zador, "Topics in the asymptotic quantization of continuous random variables," Bell Laboratories Technical Memorandum, 1966.

[19] P. L. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Trans. Inform. Theory*, vol. 28, pp. 139–148, March 1982.

**Biography of Robert M. Gray**

Robert M. Gray (S'68-M'69-SM'77-F'80) was born in San Diego, Calif., on November 1, 1943. He received the B.S. and M.S. degrees from M.I.T. in 1966 and the Ph.D. degree from U.S.C. in 1969, all in Electrical Engineering. Since 1969 he has been with Stanford University, where he is currently a Professor and Vice Chair of the Department of Electrical Engineering. His research interests are the theory and design of signal compression and classification systems.

He was a member of the Board of Governors of the IEEE Information Theory Group (1974–1980,1985–1988) as well as an Associate Editor (1977–1980) and Editor-in-Chief (1980–1983) of the *IEEE Transactions on Information Theory.* He is currently a member of the Board of Governors of the IEEE Signal Processing Society. He was Co-Chair of the 1993 IEEE International Symposium on Information Theory and Program Co-Chair of the 1997 IEEE International Conference on Image Processing. He is a member of the Image and Multidimensional Signal Processing Technical Committee of the IEEE Signal Processing Society and served as Chair 2000-2001.

He was corecipient with L.D. Davisson of the 1976 IEEE Information Theory Group Paper Award and corecipient with A. Buzo, A.H. Gray, and J.D. Markel of the 1983 IEEE ASSP Senior Award. He was awarded an IEEE Centennial medal in 1984, the 1993 Society Award from the IEEE Signal Processing Society, and the Technical Achievement Award from the Signal Processing Society and a Golden Jubilee Award for Technological Innovation from the IEEE Information Theory Society in 1998, and an IEEE Third Millennium Medal in 2000. He is a Fellow of the Institute of Mathematical Statistics and has held fellowships from the Japan Society for the Promotion of Science at the University of Osaka (1981), the Guggenheim Foundation at the University of Paris XI (1982), and NATO/Consiglio Nazionale delle Ricerche at the University of Naples (1990). During spring 1995 he was a Vinton Hayes Visiting Scholar at the Division of Applied Sciences of Harvard University.

He is a member of Sigma Xi, Eta Kappa Nu, AAAS, and the Société des Ingénieurs et Scientifiques de France. He holds an Advanced Class Amateur Radio License (KB6XQ).

**Biography of Tamás Linder**

Tamás Linder (S'92-M'93-SM'00) was born in Budapest, Hungary, in 1964. He received the M.S. degree in electrical engineering from the Technical University of Budapest in 1988, and the Ph.D degree from the Hungarian Academy of Sciences in electrical engineering in 1992.

He was a post-doctoral fellow at the University of Hawaii in 1992, and a Visiting Fulbright Scholar at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign in 1993-94. From 1994 to 1998 he was a faculty member of the Technical University of Budapest in the Department of Computer Science and Information Theory. He was a visiting research scholar in the Department of Electrical and Computer Engineering, University of California, San Diego from 1996 to 1998. He is now an Associate Professor of Mathematics and Engineering in the Department of Mathematics and Statistics, Queen's University, Kingston, Ont., Canada. His research interests include communications and information theory, source coding and vector quantization, machine learning, and statistical pattern recognition.

**Biography of Jia Li**

Jia Li is Assistant Professor of Statistics at The Pennsylvania State University. She was born in Hunan, China, in 1974. She received the BS degree in Electrical Engineering from Xi'an JiaoTong University, China, in 1993, the MSc degree in Electrical Engineering in 1995, the MSc degree in Statistics in 1998, and the PhD degree in Electrical Engineering in 1999, all from Stanford University. She worked as Research Associate in the Computer Science Department at Stanford University in 1999. She was a researcher at the Xerox Palo Alto Research Center (PARC) from 1999 to 2000. Her research interests include statistical classification and modeling, data compression, image processing, and image retrieval.