Lagrangian Vector Quantization With Combined Entropy and Codebook Size Constraints

Robert M. Gray, *Fellow, IEEE*, Tamás Linder, *Senior Member, IEEE*, and John T. Gill, *Member, IEEE*

***Abstract*—In this paper, the Lagrangian formulation of variable-rate vector quantization is extended to quantization with simultaneous constraints on entropy and codebook size, including variable- and fixed-rate quantization as special cases. The formulation leads to a Lloyd quantizer design algorithm and generalizations of Gersho's approximations characterizing optimal performance for asymptotically large rate. A variation of Gersho's approach is shown to yield rigorous results partially characterizing the asymptotically optimal performance.**

***Index Terms*—Asymptotic, entropy constrained, high rate, Lagrangian, quantization.**

I. Introduction

The theory of quantization derives largely from Lloyd's work [19], which formalized the optimal performance and provided asymptotic approximations to the optimal performance for high-rate and fixed-rate scalar quantization. Zador [24] described extensions to vector quantization under one of either of two constraints on the "rate" of a code: $\log N$, the logarithm total number of codewords, and H, the Shannon entropy of the quantizer output. The goal was to characterize the smallest possible average distortion D given a constraint on the rate as measured by one of these two quantities under the assumption that the rate was asymptotically large. The first problem is generally known as *fixed-rate coding* because of the implied assumption of using an equal number of $\ln N$ nats to specify each codeword, while the second problem is generally known as *variable-rate coding* because of the implied assumption that a differing number of bits or nats will be used for each codeword by using an optimal lossless code to describe the codeword indices. The latter case is also known as entropy-constrained coding to reflect the constraint on entropy rather than on log codebook size, and in fact Zador introduced the constraint as a measure of required channel capacity to reliably transmit the codeword indices rather than as a measure of optimal lossless coding performance.

Zador's original proofs of the fixed-rate results for high-rate quantization were corrected and generalized in [4] and [11], and the entropy-constrained results were corrected and generalized in [12] using the Lagrangian formulation of [8]. The Lagrangian formulation replaces the traditional problem statement of minimizing the distortion subject to a rate constraint by an unconstrained minimization problem involving a Lagrangian distortion combining distortion and rate. In the variable-rate case, this provides a natural extension of the original Lloyd optimality properties and the resulting Lloyd algorithm from fixed-rate to variable-rate coding.

In Zador's development and in all developments since, the proofs for the two cases of fixed rate and variable rate have differed in significant ways as well as in the details. Zador [25] closed his paper with the observation that "It appears likely that the use of constraints of the type $C_1 \log N + C_2 H$ would help unify" the fixed- and variable-rate proofs. This combined rate constraint provides many potentially useful results in addition to that of the possible unification of the two traditional approaches. Such a linear combination of constraints can be viewed as the Lagrange dual problem (see, e.g., [3, Ch. 5]) to minimizing the average distortion subject to separate constraints on each definition of rate. By varying the positive Lagrangian multipliers C_1 and C_2, one can essentially consider all possible separate rate constraints. Thus, minimization of $D + C_1 \log N + C_2 H$ provides a Lagrangian dual to the minimization of average distortion subject to the combined constraints.

In addition, such a weighted linear combination of different definitions of "rate" is of practical interest as well as mathematical interest. The number of codewords can be important even for variable-length coding systems for several reasons. First, as a cost or penalty function, it makes explicit a quantity related to the storage needed for the codebooks required to synthesize the final reproduction and usually to encode the original signal. Second, without a penalty for the number of codewords, a codebook for a variable-rate coding scheme with only an entropy constraint might require an infinite number of codewords [17], which can cause both theoretical and practical problems with code design and implementation. Third, in some examples such as clustering of Gauss mixtures, a mixture having fewer Gaussian components is deemed superior to one with many Gaussian components because it is simpler. The Lloyd algorithm can be used for such designs [1], [13], and placing an explicit penalty on the number of codewords (in this case, Gaussian components of a mixture as represented by a covariance matrix and mean vector) forces the algorithm to trade off the number of codewords along with distortion and entropy [23]. Fourth, it is of interest to the traditional cases to consider an example where C_1/C_2 is close to, but not equal to, the extremes of 0 and 1. In these cases, one of the two definitions of rate is

Manuscript received August 6, 2006; revised July 9, 2007. This work was supported in part by the National Science Foundation under Grant CCR-0309701, by gifts from Hewlett-Packard Laboratories, and by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

R. M. Gray and J. T. Gill are with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: rmgray@stanford.edu; gill@stanford.edu).

T. Linder is with the Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: linder@mast.queensu.edu).

Communicated by V. A. Vaishampayan, Associate Editor At Large.

Digital Object Identifier 10.1109/TIT.2008.920205

the important one, but the other cannot be entirely discounted. For example, one can study the behavior of variable-rate codes with a mild additional constraint on the codebook size by using a small C_1/C_2.

As we are interested primarily in high-rate results where the cost of rate is small with respect to that of distortion, we will be interested in the relative behavior of C_1 and C_2 as both become small. Here, we make the assumption that it is the *relative* costs of entropy and log codebook size that remain fixed as the combined cost tends to zero. Thus, we focus on a Lagrangian minimization of the form $D + \lambda[(1-\eta)H + \eta \log N]$, where $\eta \in [0,1]$ reflects the relative importance of the two cost constraints and $\lambda > 0$ governs the overall rate cost. The high-rate regime will be considered by letting $\lambda \to 0$. Note that except for the asymptotic results, the original Zador choice of C_1 and C_2, and our choice of λ and η are equivalent in that either pair implies values for the other.

The Zador linear combination of the two definitions of rates is not the only possible general definition of average rate, which includes the traditional examples as special cases. For example, one could also consider a rate defined by a Renyi entropy as

$$R = \frac{1}{1-c} \ln \left(\sum_i \Pr(\alpha(X) = i)^c \right)$$

where $\alpha(X)$ is the codeword index assigned to input X by the quantizer's encoder, which will yield the fixed-rate definition if $c = 0$ and will converge to the Shannon entropy as $c \to 1$. The Zador form, however, has the advantage that it can be expressed as an average over the input distribution of an instantaneous rate (as will be seen), which leads to a Lagrangian distortion incorporating a cost of the rate per input symbol into the encoder. This in turn results in a characterization of a Lloyd-optimal encoder as one minimizing a Lagrangian distortion incorporating the instantaneous rate and hence leads to a Lloyd clustering algorithm for quantizer design. The Renyi entropy cannot be expressed as an expectation of an instantaneous rate and hence lacks these properties.

There is a very small literature considering combined rate constraints. Simultaneous constraints on entropy and codebook size were considered in the definition of the nth order Shannon rate-distortion functions by Rao and Pearlman [21]. The Shannon rate-distortion function is defined in terms of a mutual information constraint, which is weaker than constraints on output entropy and codebook size, so the Shannon rate-distortion function provides only a lower bound to the achievable distortion for fixed dimension, a lower bound which is guaranteed to be achievable only for asymptotically large dimension. Chan and Gersho introduced modifications of the Lloyd algorithm for tree-structured vector quantizer design, which explicitly incorporated codebook storage as a constraint by limiting the number of distinct node codebooks [6], [7].

This paper pursues Zador's proposal of a combined rate measure in some depth. Following a brief presentation of preliminaries, the development begins with an extension of the Lagrangian formulation for the variable-rate quantization case to a combined constraint case. An easy extension of the Lloyd algorithm provides necessary conditions for code optimality and hence an iterative design algorithm based on alternating optimization. Gersho's heuristics and approximations are used to develop formulas characterizing the optimal performance in the high-rate regime (of both large entropy and number of codewords). The results clarify the relations among three quantizer characteristics of interest at high rates: distortion, entropy, and codebook size. Traditional methods consider only the tradeoff between distortion and either entropy or codebook size. The approach provides new results on the behavior of asymptotically optimal sequences of quantizers, providing under certain conditions separate characterizations of the behavior of distortion, entropy, and codebook size in addition to Lagrangian combinations.

The rest of this paper is organized as follows. In Section II, vector quantization fundamentals are recalled, including some main results from the high-resolution theory, Gersho's heuristic derivations, and conjecture. Section III describes the combined codebook size and entropy constraint and states the high-rate equivalence of the traditional and Lagrangian formulations in this setting. Section IV provides a Gersho-type heuristic derivation of the high-rate quantizer performance with the combined constrained. In Section V, some auxiliary results are developed, which form the basis of our rigorous development. In Section VI, our first principal result, Theorem 1, provides the precise high-rate asymptotics of the best achievable performance under the combined rate constraint for the special case of the uniform distribution on the unit cube. Our second principal result, Theorem 2 in Section VII, proves that the asymptotic formula developed heuristically in Section IV is an upper bound on the best achievable performance for general source densities. Connections with Gersho's conjecture and a conjecture are also given.

Highly technical proofs are relegated to the appendices, which are terse in the interest of space. Detailed versions are available on request from the authors.

II. Quantization Fundamentals

A *quantizer* or *vector quantizer* q on $\Re^k$, k-dimensional Euclidean space, can be described by the following mappings and sets (all assumed to be measurable): an *encoder* $\alpha : \Re^k \to \mathcal{I}$, where $\mathcal{I} = \{0, 1, 2, \ldots, N(q) - 1\}$, an associated partition $\mathcal{S} = \{S_i;\ i \in \mathcal{I}\}$, $S_i \subset \Re^k$, such that $\alpha(x) = i$ if $x \in S_i$; a *decoder* $\beta : \mathcal{I} \to \Re^k$, an associated reproduction codebook $\mathcal{C} = \{\beta(i);\ i \in \mathcal{I}\}$ with $N(q) = |\mathcal{C}|$ of distinct elements; and a *length function* $\{\ell(i);\ i \in \mathcal{I}\}$, which is *admissible* in the sense that $\sum_{i \in \mathcal{I}} e^{-\ell(i)} \leq 1$. The condition for admissibility is Kraft's inequality in natural logarithms and it corresponds roughly to the quantity of nats required to communicate the index to the decoder using a uniquely decodable lossless code.

If $\ell(i) = \ln N(q)$ for all i with $N(q)$ finite, the quantizer is said to be *fixed rate*; otherwise, it is said to be *variable rate*. Let q denote both a shorthand for the collection of mappings (α, β, ℓ) and the overall mapping $q(x) = \beta(\alpha(x))$.

Let $d(x, \hat{x})$ denote the distortion between an input x and a quantized version $\hat{x} = \beta(\alpha(x))$, that is, a nonnegative measurable function. The basic optimality properties will be described for general distortion measures, but the high-rate results will focus on the squared error (squared Euclidean norm) distortion $\|x-\hat{x}\|^2$. If X is a random vector having density f, the *average distortion* is defined as $D_f(q) = E_f d(X, \beta(\alpha(X)))$, where E_f denotes expectation with respect to f. The *average rate* is defined by $R_f(q) = E_f \ell(\alpha(X))$; this reduces to $\ln|\mathcal{C}|$ in the fixed-rate case. The optimal performance is the minimum distortion achievable for a given rate, the operational distortion-rate function $\delta(f,R) = \inf_{q:R_f(q)\le R} D_f(q)$. To distinguish between the fixed-rate and variable-rate cases in a manner consistent with the later development, define

$$\begin{aligned}\delta_0(f,R) &= \inf_{q:E_f\ell(\alpha(X))\le R} D_f(q)\\ \delta_1(f,R) &= \inf_{q:\ln N(q)\le R} D_f(q).\end{aligned} \tag{1}$$

For fixed-rate quantizers, Lloyd's necessary conditions for a quantizer q to be optimal are as follows.

- For a given decoder β, the optimal encoder is $\alpha(x) = \mathrm{argmin}_i d(x,\beta(i))$. The minimum obviously exists since the index set is finite. Ties can be broken in an arbitrary fashion.
- For a given encoder α, the optimal decoder is $\beta(i) = \mathrm{argmin}_y E(d(X,y)|\alpha(X)=i)$ if the minimum exists. For squared error, $\beta(i) = E(X|\alpha(X)=i)$.
- $\Pr(\alpha(X)=i) \neq 0$ for $i \in \mathcal{I}$.

To obtain a similar set of properties in the variable-rate case, a Lagrangian approach is used. Define the Lagrangian distortion in terms of a Lagrangian multiplier $\lambda > 0$ by $\rho_\lambda(x,i) = d(x,\beta(i)) + \lambda\ell(i)$. The expected Lagrangian distortion for a quantizer q is $\rho(f,\lambda,q) = E_f(\rho_\lambda(X,\alpha(X)))= E_f(d(X,\beta(\alpha(X))) + \lambda\ell(\alpha(X)))$ and the optimal performance for a fixed λ is $\rho(f,\lambda) = \inf_q \rho(f,\lambda,q)$, where the infimum is over all quantizers q with admissible length functions. This notation where the optimization of a function over one of its arguments is denoted by the same function with that argument removed will often be used. The Lloyd conditions are then [8] as follow.

- For a given decoder β and length function ℓ, the optimal encoder is $\alpha(x) = \mathrm{argmin}_i\,(d(x,\beta(i)) + \lambda\ell(i))$. The minimum obviously exists when the index set is finite and can be shown to exist (for well-behaved d) if the index set is countably infinite. Again, ties can be broken arbitrarily.
- For a given encoder α, the optimal decoder satisfies $\beta(i) = \mathrm{argmin}_y E(d(X,y|\alpha(X)=i)$ if the minimum exists.
- For a given encoder α, the optimal length function is $\ell(i) = -\ln\Pr(\alpha(X)=i)$. Thus, for the optimal length function, $R_f(q) = H_f(q) = H_f(\mathcal{S})$, where $H(\mathcal{S}) = -\sum_{m=0}^{|\mathcal{S}|-1} P_f(S_m)\ln P_f(S_m)$ is the Shannon entropy of the encoder output (and, since the codewords are assumed distinct, of the decoder output).

Although the condition $\Pr(\alpha(X)=i) \neq 0$ for $i \in \mathcal{I}$ is not necessary for optimality of variable-length codes, it is often added as a requirement to avoid useless codewords. Given a partition $\mathcal{S}$ (or encoder α), the Lloyd properties determine the remaining components so optimizing over quantizers is equivalent to optimizing over encoders or partitions. Thus, we emphasize the quantizer q or the partition $\mathcal{S}$, as is convenient, and write $D_f(q) = D_f(\mathcal{S})$ and $R_f(q) = R_f(\mathcal{S})$ when we assume that the encoder and the length function are optimally matched to the partition. If the quantizer is a fixed-rate code, then the codebook $\mathcal{C}$ determines the encoder and hence the entire quantizer.

A. High-Rate Quantization

For the squared error distortion $d(x,\hat{x}) = \|x-\hat{x}\|^2$, the traditional form of Zador's high-rate (or high-resolution) result for fixed-rate quantizers is that if the probability density function (pdf) f satisfies the moment condition $E_f(\|X\|^{2+\delta}) < \infty$ for some $\delta > 0$, then [4], [11]

$$\lim_{N\to\infty} N^{\frac{2}{k}}\delta_1(f,\ln N) = a_k\|f\|_{\frac{k}{k+2}}$$

where a_k is a positive constant given by $a_k = \inf_{N\ge 1} N^{\frac{2}{k}}\delta_1(u,\ln N)$, u is the uniform pdf on the unit cube $[0,1]^k$, and $\|f\|_p = \left(\int f^p\right)^{1/p}$. The traditional form of Zador's result for variable-rate quantizers is

$$\lim_{R\to\infty} e^{\frac{2}{k}R}\delta_0(f,R) = b_k e^{\frac{2}{k}h(f)} \tag{2}$$

where $h(f) = -\int f\ln f$ is the differential entropy and the positive constant b_k is given by $b_k = \inf_{R>0} R^{\frac{2}{k}}\delta_0(u,R)$. The first rigorous proof of (2) used the Lagrangian approach. Assume that the pdf f is such that the $h(f)$ is finite and a uniform scalar quantized version of X with cubic cell volume 1 has finite entropy. Under these conditions [12]

$$\lim_{\lambda\to 0}\left(\frac{\rho(f,\lambda,0)}{\lambda} + \frac{k}{2}\ln\lambda\right) = \theta_k + h(f) \tag{3}$$

where

$$\theta_k \triangleq \inf_{\lambda>0}\left(\frac{\rho(u,\lambda,0)}{\lambda} + \frac{k}{2}\ln\lambda\right) = \frac{k}{2}\ln\frac{2eb_k}{k}. \tag{4}$$

The equivalence of the traditional Zador formulation and the Lagrangian formulation was shown in [12, Lemma 1]: (2) holds if and only if (iff) the Lagrangian form (3) holds. Similar arguments show that in the fixed-rate case, (2) holds for f iff

$$\lim_{\lambda\to 0}\left(\frac{\rho(f,\lambda,1)}{\lambda} + \frac{k}{2}\ln\lambda\right) = \psi_k + \ln\|f\|_{k/(k+2)}^{k/2} \tag{5}$$

where

$$\psi_k = (k/2)\ln(2ea_k/k). \tag{6}$$

This has the same form as the variable-rate Zador result with $\ln\|f\|_{k/(k+2)}^{k/2}$ replacing $h(f)$ and ψ_k replacing θ_k.

The details of the proofs of the high-rate results for the traditional cases differ significantly, but most proofs of these results follow the original Zador approach: 1) prove the result for u, the uniform pdf on the unit cube; 2) extend the result to pdfs that are piecewise constant on disjoint cubes of equal side a; 3) prove the result for a general pdf on a cube; and 4) prove the result for general pdfs. The first step is a key one both because it provides the primary building block for the subsequent results, and because it suffices to study Zador's constants.

B. Gersho's Conjecture and Heuristics

The rigorous proofs of the high-rate results are notoriously difficult, and the results had a limited audience until Gersho [10] provided a relatively simple heuristic development of the basic formulas. His approach, although not rigorous, provides insight into the results and a consistency check with the rigorous development. Furthermore, a goal of this work is to develop an approach that has some of the flavor of Gersho's as an aid to intuition, yet which is also amenable to rigorous analysis without assuming more than the traditional cases.

Gersho's conjecture involves two assumptions regarding asymptotically optimal sequences of fixed-rate and variable-rate quantizers. First, it is assumed that there exists a quantizer point density function $\Lambda(x)$ such that a sequence of optimal codes with N codewords, $N = 1, 2, \ldots$, will satisfy for all "reasonable" $S \subset \Re^k$

$$\lim_{N\to\infty} \frac{1}{N} \times (\#\text{ of reproduction vectors in a set } S) = \int_S \Lambda(x)dx.$$

By definition, $\int_{\Re^k} \Lambda(x)\,dx = 1$. A point density function is simply a nonnegative function that integrates to 1 and hence is mathematically equivalent to a pdf and we use the abbreviation pdf for both. Second, Gersho assumed that if $f(x)$ is smooth and R is large, then both the fixed-rate and variable-rate minimum distortion quantizers have cells S_i that are (approximately) scaled, rotated, and translated copies of S^*, the convex polytope that tessellates $\Re^k$ with minimum normalized moment of inertia

$$M(S) = \frac{1}{kV(S)^{1+2/k}} \int_S \|x - y(S)\|^2 dx$$

where $V(S)$ is the k-dimensional volume of S and $y(S) = V(S)^{-1}\int_S dx$ denotes the centroid of S with respect to the uniform distribution. Specifically, define c_k as the minimum of $M(S)$ over all tessellating convex polytopes S. Under these assumptions, Gersho argued that for large N

$$D_f(q) \approx c_k E_f\left((\frac{1}{N(q)\Lambda(X)})^{2/k} \right) \tag{7}$$

$$\begin{aligned} H_f(q(X)) &\approx h(X) - E_f\left(\log(\frac{1}{N(q)\Lambda(X)}) \right) \\ &= \ln N(q) - H(f\|\Lambda) \end{aligned} \tag{8}$$

where the continuous relative entropy $H(f\|\Lambda)$ is given by

$$H(f\|\Lambda) = \int f(x) \ln \frac{f(x)}{\Lambda(x)} dx.$$

The distortion approximation results from the approximation of integrals by Riemann sums. The entropy approximation, however, is known to hold only in very special cases such as those where all quantization cells have equal volume as in Voronoi regions of lattices [9], [18], and no general result along these lines is known. These approximations can be combined with the Holder and Jensen inequalities to obtain the Zador results [10], as will be seen as a special case in Section III.

III. Combined Codebook Size and Entropy Constraints

Define a combined instantaneous rate by

$$r(i) = (1-\eta)\ell(i) + \eta \ln N(q) \tag{9}$$

with $\eta \in [0,1]$ and ℓ an admissible length function. The average combined rate with the optimal choice of admissible ℓ yields the Zador affine combined rate $Er(\alpha(X)) = \lambda[(1-\eta)H(q) + \eta \ln N]$. The choice of instantaneous rate implies a Lagrangian distortion

$$\rho_\lambda(x,i) = d(x, \beta(i)) + \lambda r(i) \tag{10}$$

for use in the encoder optimization step. This provides an encoder that takes into account the cost in nats or bits of transmitting the index as does the traditional variable-rate case.

Because the combined rate $r(i)$ will play the same role as played by the traditional notion of "rate" in the fixed-rate and variable-rate cases, it will be referred to as "rate." It should not, however, be considered as the transmission rate required since that term better applies to ℓ. The multiplier η can be viewed as a Lagrangian multiplier reflecting the relative importance of the length function and the codebook size, with the traditional cases of fixed-rate and variable-rate quantization corresponding to $\eta = 1$ and 0, respectively. The traditional operational distortion-rate functions of (1) immediately generalize to

$$\delta_\eta(f,R) = \inf_{q: E_f r(\alpha(X)) \le R} D_f(q)$$

where the infimum is over all quantizers q with instantaneous rate functions r as in (9) with admissible length functions ℓ. Define the Lagrangian distortion by (10) and define the average distortion and optimal performance by

$$\begin{aligned} \rho(f,\lambda,\eta,q) &= E_f \rho_\lambda(X, \alpha(X)) \\ &= E_f\big[d(X, \beta(\alpha(X))) \\ &\qquad + \lambda[(1-\eta)\ell(\alpha(X)) + \eta \ln N(q)]\big] \\ \rho(f,\lambda,\eta) &= \inf_q \rho(f,\lambda,\eta,q). \end{aligned}$$

For $\lambda > 0$ and $\eta \in [0,1]$, the optimization will have the necessary Lloyd conditions of the variable-rate case. Lloyd's requirement for the elimination of zero probability words can be generalized as a means of "pruning" codebooks. For example, if one uses a subpartition (superpartition or refinement) of a quantizer partition $\mathcal{S}$, then rate will go down (up) and distortion down (up). If $D + \lambda R$ decreases, however, then the original partition could not be optimal. This provides a means of testing subpartitions to see if the reduction in rate more than compensates for the increase in distortion [15] and to prune unneeded partition cells if the subpartition is better.

As before, given a partition $\mathcal{S}$ (or encoder α), the Lloyd properties determine the remaining components so optimizing over quantizers is equivalent to optimizing over partitions. Thus, for example, $\rho(f,\lambda,\eta) = \inf_{\mathcal{S}} \rho(f,\lambda,\eta,\mathcal{S})$, where $\rho(f,\lambda,\eta,\mathcal{S})$ is the minimum value of $\rho(f,\lambda,\eta,q)$ over all quantizers q having partition $\mathcal{S}$.

We will use the following notation for the performance for partitions, the optimal performance over partitions, and the asymptotically optimal performance:

$$\theta(f,\lambda,\eta,\mathcal{S}) = \frac{\rho(f,\lambda,\eta,\mathcal{S})}{\lambda} + \frac{k}{2}\ln\lambda \tag{11}$$

$$\begin{aligned} \theta(f,\lambda,\eta) &= \inf_{\mathcal{S}} \theta(f,\lambda,\eta,\mathcal{S}) = \frac{\rho(f,\lambda,\eta)}{\lambda} + \frac{k}{2}\ln\lambda \\ \overline{\theta}(f,\eta) &= \limsup_{\lambda\to 0} \theta(f,\lambda,\eta) \\ \underline{\theta}(f,\eta) &= \liminf_{\lambda\to 0} \theta(f,\lambda,\eta) \end{aligned} \tag{12}$$

and, if the limit exists, $\theta(f,\eta) = \lim_{n\to\infty} \theta(f,\lambda,\eta) = \overline{\theta}(f,\eta) = \underline{\theta}(f,\eta)$.

The equivalence of the high-rate results for the distortion-rate and Lagrangian formulation follows in the combined constraint case as it did in the traditional fixed- and variable-rate cases by a straightforward modification of the [12, proof of Lemma 1], as summarized in the following lemma.

Lemma 1: If

$$\lim_{R\to\infty} e^{\frac{2}{k}R}\delta_\eta(f,R)=\delta(f,\eta) \tag{13}$$

exists for a positive finite $\delta(f,\eta)$, then

$$\lim_{\lambda\to 0}\theta(f,\lambda,\eta)=\frac{k}{2}\ln\left(\frac{2e}{k}\delta(f,\eta)\right).$$

Conversely, if

$$\lim_{\lambda\to 0}\theta(f,\lambda,\eta)=\theta(f,\eta) \tag{14}$$

for a finite $\theta(f,\eta)$, then

$$\lim_{R\to\infty} e^{\frac{2}{k}R}\delta_\eta(f,R)=\frac{k}{2}e^{(2/k)\theta(f,\eta)-1}.$$

Thus, the traditional distortion rate and the Lagrangian forms of the high-rate results are equivalent in that the traditional style limit exists iff the Lagrangian limit exists, and the limits are related by

$$\theta(f,\eta)=\frac{k}{2}\ln\left(\frac{2e}{k}\delta(f,\eta)\right) \tag{15}$$

which includes (4) and (6) as special cases.

IV. Heuristic Derivation of High-Rate Performance

Gersho's approximations can be used to develop a solution for the combined constraint case for general η. Suppose that a quantizer q has a quantizer point density Λ and a total of N quantization levels for N large, then using (7), (8), and the $\ln r \le r-1$ inequality

$$\begin{aligned}
\theta(f,\lambda,\eta,q) &= \frac{D_f(q)}{\lambda}+(1-\eta)H_f(q)+\eta\ln N+\frac{k}{2}\ln\lambda\\
&\approx \frac{c_kE_f\left((N\Lambda(X))^{-2/k}\right)}{\lambda}\\
&\quad +(1-\eta)[\ln N-H(f\|\Lambda)]+\eta\ln N+\frac{k}{2}\ln\lambda\\
&=\frac{k}{2}\left[\frac{2c_kN^{-2/k}E_f\left((\Lambda(X))^{-2/k}\right)}{\lambda}\right.\\
&\qquad\left.-\ln\frac{2c_kN^{-2/k}E_f\left((\Lambda(X))^{-2/k}\right)}{\lambda}-1\right]\\
&\quad +\frac{k}{2}\ln\left[\frac{2ec_k}{k}E_f\left((\Lambda(X))^{-2/k}\right)\right]-(1-\eta)H(f\|\Lambda)\\
&\ge\frac{k}{2}\ln\left[\frac{2ec_k}{k}E_f\left((\Lambda(X))^{-2/k}\right)\right]-(1-\eta)H(f\|\Lambda)
\end{aligned}$$

with equality iff

$$N=\left[\frac{2}{k}\frac{c_kE_f\left((\Lambda(X))^{-2/k}\right)}{\lambda}\right]^{k/2}. \tag{16}$$

Since the goal is to minimize $\theta(f,\lambda,\eta,q)$, this is the optimal choice of N given λ. Thus, for small λ

$$\theta(f,\lambda,\eta,q)\approx\frac{k}{2}\ln\left(\frac{2ec_k}{k}\right)+\phi(f,\eta,\Lambda) \tag{17}$$

where

$$\phi(f,\eta,\Lambda)=\frac{k}{2}\ln\left(E_f\left((\Lambda(X))^{-2/k}\right)\right)-(1-\eta)H(f\|\Lambda). \tag{18}$$

The best possible performance will be the one that minimizes $\phi(f,\lambda,\eta,q)$ over all q and hence Gersho's arguments suggest that if

$$\phi(f,\eta)=\inf_{\Lambda}\phi(f,\eta,\Lambda) \tag{19}$$

and the infimum is over all pdfs Λ, for which $\phi(f,\eta,\Lambda)$ is well defined, then

$$\theta(f,\eta)=\lim_{\lambda\to 0}\theta(f,\lambda,\eta)=(k/2)\ln(2ec_k/k)+\phi(f,\eta). \tag{20}$$

Since the functionals $\phi(f,\eta,\Lambda)$ and $\phi(f,\eta)$ of (18) and (19) arise here in the context of Gersho's conjecture and heuristic development, we will refer to them as Gersho functionals.

The functional $\phi(f,\eta,\Lambda)$ can be expressed as

$$\begin{aligned}
\phi(f,\eta,\Lambda) &= (1-\eta)\phi(f,0,\Lambda)+\eta\phi(f,1,\Lambda) && (21)\\
&=\phi(f,0,\Lambda)+\eta(\phi(f,1,\Lambda)-\phi(f,0,\Lambda))\\
&=\phi(f,0,\Lambda)+\eta H(f\|\Lambda)\\
&=\phi(f,1,\Lambda)-(1-\eta)H(f\|\Lambda). && (22)
\end{aligned}$$

The nonnegativity of the relative entropy $H(f\|\Lambda)$ implies immediately that $\phi(f,0,\Lambda)\le\phi(f,\eta,\Lambda)\le\phi(f,1,\Lambda)$. If the derived approximations are valid, then (16) implies that for a given point density function Λ the codebook size will be given approximately in terms of the Lagrangian multiplier by

$$\ln N\approx\frac{k}{2}\ln\frac{2ec_k}{k}+\phi(f,1,\Lambda)+\ln\lambda^{-k/2} \tag{23}$$

and hence, from (8)

$$\ln N(q)-H_f(q)\approx H(f\|\Lambda) \tag{24}$$

and hence, the log codebook size and the quantizer output entropy differ by a constant as the codebook size grows. Interestingly enough, there is no explicit dependence on η here; the dependence is implicit through the selection of a Λ minimizing $\phi(f,\eta,\Lambda)$.

As a check on the combined constraint result derived using Gersho's heuristics, consider the traditional cases. If $\eta=1$, Holder's inequality yields the bound

$$\begin{aligned}
\phi(f,1,\Lambda) &= \frac{k}{2}\ln\left(\int f(x)\Lambda(x)^{-2/k}dx\right)\\
&\ge\frac{k}{2}\ln\|f\|_{k/(k+2)}=\phi(f,1) && (25)
\end{aligned}$$

with equality iff

$$\Lambda(x)=\frac{f(x)^{k/(k+2)}}{\|f\|_{k/(k+2)}^{k/(k+2)}} \tag{26}$$

the well-known solution for the fixed-rate case. From [11, Remark 6.3], the moment condition (47) ensures that $\|f\|_{k/(k+2)}$ is finite and the Λ minimizing $\phi(f,1,\Lambda)$ is given by (26). If $\eta=0$, then from Jensen's inequality

$$\begin{aligned}\phi(f,0,\Lambda) &= \frac{k}{2}\ln\left(\int f(x)\Lambda(x)^{-2/k}dx\right) - H(f\|\Lambda) \geq h(f)\\ &= \phi(f,0) \qquad (27)\end{aligned}$$

with equality iff $\Lambda(x)$ is constant for the support set of X, again agreeing with the classical result. (Here, equality requires that the distribution of X has bounded support.)

The general minimization of $\phi(f,\eta,\Lambda)$ for $\eta\in(0,1)$ does not seem to have such a nice form. Since the infimum of a sum of terms is bound below by the sum of the minima, it is easy to see from (22), (25), and (27) that

$$\begin{aligned}\phi(f,\eta,\Lambda) &\geq (1-\eta)\phi(f,0)+\eta\phi(f,1)\\ &= (1-\eta)h(f)+\eta\ln\|f\|_{k/(k+2)}^{\frac{2}{k}}\end{aligned}$$

but the inequality is strict except for the endpoints since, in general, distinct Λ yields those minima. The bound does hold, however, for $f=u$, in which case Λ uniform on the unit cube yields

$$\phi(u,\eta)=0. \qquad (28)$$

In this case, (20) implies that

$$\theta(u,\eta)=\lim_{\lambda\to 0}\theta(u,\lambda,\eta)=\frac{k}{2}\ln\left(\frac{2ec_k}{k}\right) \qquad (29)$$

which is independent of η (and hence $a_k=b_k$ if Gersho's conjectures and approximations are true), and hence, Gersho's conjecture and approximations lead to the general conjecture

$$\lim_{\lambda\to 0}\theta(f,\lambda,\eta)=\theta(f,\eta)=\theta(u,\eta)+\phi(f,\eta) \qquad (30)$$

which reduces to the known results in the traditional cases.

Unfortunately, $\phi(f,\eta,\Lambda)$ is not convex in Λ, and hence, it is not immediately obvious how to approach its minimization. The following lemma shows that a transformation yields an equivalent convex optimization problem, and hence, the optimization inherits all of the algorithms and properties of convex optimization theory (see, e.g., [3]). The proof is straightforward but long and is not essential to the paper, so it is omitted.

Lemma 2: $\phi(f,\eta)=\inf_\nu\psi(f,\eta,\nu)=\psi(f,\eta)$, where $\phi(f,\eta)$ is given by (18) and (19), where

$$\begin{aligned}\psi(f,\eta,\nu) &= \frac{k}{2}\left(\int f(x)e^{\nu(x)}dx-(1-\eta)\int f(x)\nu(x)dx-1\right)\\ &\quad+\eta\ln\left(\int e^{-k\nu(x)/2}dx\right)+(1-\eta)h(f) \qquad (31)\end{aligned}$$

where the integrals are over the support set of f and the infimum over ν is over all measurable functions ν, for which $\psi(f,\eta,\nu)$ is well defined. The functional $\psi(f,\eta,\nu)$ is (strictly) convex in ν.

The optimization over ν instead of λ can be viewed as a form of "geometric programming" (e.g., [3, Sec. 4.5]). Unfortunately, in this infinite-dimensional case, convexity does not guarantee the existence of a minimizing ν. Strict convexity does, however, guarantee that if a minimizing ν exists, it is unique (at least up to a set of measure zero). In particular, if there is a local minimum of $\phi(f,\eta,\nu)$ with respect to ν, then it is the unique global minimum.

An obvious problem with this heuristic approach is that it rests on the assumption of the existence of a quantizer point density function corresponding to a sequence of asymptotically optimal quantizers. In some cases, one can use this assumption along with others to derive traditional high-rate quantization theorems as has been done by Na and Neuhoff [20], but this raises the question of whether the existence of asymptotically optimal high-rate quantizers implies the existence of the point density functions. In fact, the existence of point density functions has been rigorously proved only for the fixed-rate ($\eta=1$) case. The result was first stated by Bucklew [5] and subsequently proved rigorously by Graf and Luschgy [11]. The existence of the density for the variable-rate case has not been similarly proved, although Gersho's heuristic arguments suggest that it is uniform (and this is often assumed). We will see that one can prove results bearing a close resemblance to those predicted by Gersho's arguments.

V. Lagrangian Distortion Inequalities and Asymptotics

The following lemma provides a lower bound to the Lagrangian average distortion that is independent of λ.

Lemma 3: Given a pdf f and $\eta\in[0,1]$ and partition $\mathcal{S}$

$$\begin{aligned}\theta(f,\lambda,\eta,\mathcal{S}) &\geq \frac{k}{2}\ln\left(\frac{2e}{k}D_f(\mathcal{S})e^{\frac{2}{k}((1-\eta)H_f(\mathcal{S})+\eta\ln|\mathcal{S}|)}\right)\\ &\triangleq \Theta(f,\eta,\mathcal{S}) \qquad (32)\end{aligned}$$

with equality iff $D_f(\mathcal{S})=k\lambda/2$.

Proof: Rearranging terms in the definition of $\theta(f,\lambda,\eta,\mathcal{S})$ and using the $\ln r\leq r-1$ inequality yields

$$\begin{aligned}\theta(f,\lambda,\eta,\mathcal{S}) &= \frac{k}{2}\left[\frac{2D_f(\mathcal{S})}{k\lambda}-\ln\frac{2D_f(\mathcal{S})}{k\lambda}-1\right]\\ &\quad+\frac{k}{2}\ln\left(\frac{2e}{k}D_f(\mathcal{S})e^{\frac{2}{k}((1-\eta)H_f(\mathcal{S})+\eta\ln|\mathcal{S}|)}\right)\\ &\geq \frac{k}{2}\ln\left(\frac{2e}{k}D_f(\mathcal{S})e^{\frac{2}{k}((1-\eta)H_f(\mathcal{S})+\eta\ln|\mathcal{S}|)}\right)\end{aligned}$$

with equality iff $D_f(\mathcal{S})=k\lambda/2$. □

Corollary 1: Given the assumptions of Lemma 3

a) $\theta(f,\lambda,\eta,\mathcal{S}) \geq \Theta(f,\eta,\mathcal{S}) = \inf_{\lambda>0}\theta(f,\lambda,\eta,\mathcal{S}) \geq \Theta(f,\eta) = \inf_{\mathcal{S}}\Theta(f,\eta,\mathcal{S})$;

b) $\inf_{\lambda>0}\theta(f,\lambda,\eta)=\Theta(f,\eta)$;

c) $\underline{\theta}(f,\eta)\geq\Theta(f,\eta)$.

Proof: The first statement follows from the previous lemma and the definitions. The second follows from $\inf_{\lambda>0}\theta(f,\lambda,\eta) = \inf_{\lambda>0}\inf_{\mathcal{S}}\theta(f,\lambda,\eta,\mathcal{S}) = \inf_{\mathcal{S}}\inf_{\lambda>0}\theta(f,\lambda,\eta,\mathcal{S}) = \Theta(f,\eta)$. The final statement follows from the first statement since $\theta(f,\lambda,\eta)\geq\Theta(f,\eta)$. □

The functions $\theta(f,\lambda,\eta,\mathcal{S})$ and $\Theta(f,\eta,\mathcal{S})$ both have the form of the function ϕ of (21), i.e., they are an affine combination of their endvalues as in

$$\Theta(f,\eta,\mathcal{S})=(1-\eta)\Theta(f,0,\mathcal{S})+\eta\Theta(f,1,\mathcal{S}). \qquad (33)$$

The functions $\theta(f,\lambda,\eta)$, $\Theta(f,\eta)$, and $\phi(f,\eta)$ are all infima of the previous affine functions. These functions share many useful properties, which are summarized in the following lemma. We use the ϕ notation as being typical, but the results apply to any functions of the form described.

Lemma 4: Suppose that $\phi(f,\eta,\Lambda)$, $\eta\in[0,1]$, has the form

$$\begin{aligned}\phi(f,\eta,\Lambda) &= (1-\eta)\phi(f,0,\Lambda)+\eta\phi(f,1,\Lambda)\\ &= \phi(f,0,\Lambda)+\eta\left(\phi(f,1,\Lambda)-\phi(f,0,\Lambda)\right) \quad (34)\end{aligned}$$

with

$$\phi(f,1,\Lambda)\geq\phi(f,0,\Lambda). \quad (35)$$

Then, the functionals $\phi(f,\eta,\Lambda)$ and $\phi(f,\eta)=\inf_\Lambda\phi(f,\eta,\Lambda)$ are monotonically nondecreasing, concave, and continuous functions of $\eta\in[0,1]$. Furthermore, the left derivative $\phi'(f,\eta_-)$ exists and is finite on $(0,1]$ and the right derivative $\phi'(f,\eta_+)$ exists on $[0,1)$ and is finite on $(0,1)$. Last, $\phi(f,\eta)$ is differentiable with respect to η [i.e., $\phi'(f,\eta_-)=\phi'(f,\eta_+)$] for all $\eta\in(0,1)$ except possibly for a countable set of points.

Proof: Equations (34) and (35) imply that $\phi(f,\eta,\Lambda)$ and hence also $\phi(f,\eta)$ are monotonically nondecreasing in η. Since $\phi(f,\eta,\Lambda)$ is affine in η, it is both convex and concave and it is continuous and its infimum over Λ, $\phi(f,\eta)$, is concave. Since $\phi(f,\eta)$ is nondecreasing and concave, it is continuous everywhere except possibly at the origin $\eta=0$. It is also continuous at 0 since

$$\begin{aligned}\inf_{\eta\in(0,1)}\phi(f,\eta) &= \inf_{\eta\in(0,1)}\inf_\Lambda\phi(f,\eta,\Lambda)\\ &= \inf_\Lambda\inf_{\eta\in(0,1)}\phi(f,\eta,\Lambda)\\ &= \inf_\Lambda\phi(f,0,\Lambda)\\ &= \phi(f,0).\end{aligned}$$

Since $\phi(f,\eta)$ is concave on $[0,1]$, its left derivative $\phi'(f,\eta_-)$ exists and is finite on $(0,1]$ and its right derivative $\phi'(f,\eta_+)$ exists on $[0,1)$ and is finite on $(0,1)$. Furthermore, $\phi(f,\eta)$ is differentiable [i.e., $\phi'(f,\eta_-)=\phi'(f,\eta_+)$] for all $\eta\in(0,1)$ except possibly for a countable set of points (see, e.g., [22]). □

It will be necessary during the development to tease apart the separate behavior of $\phi(f,0,\Lambda)$ and $\phi(f,1,\Lambda)$ when Λ is chosen to minimize the convex combination $\phi(f,\eta,\Lambda)$. This allows one to quantify separately the log codebook size when it is the combination of codebook size and entropy that is being controlled. The following corollary accomplishes this. The proof is in part A of the Appendix.

Corollary 2: Given a functional $\phi(f,\eta,\Lambda)$, $\eta\in[0,1]$, satisfying (34), suppose that Λ_n, $n=1,2,\ldots$, is chosen so that

$$\lim_{n\to\infty}\phi(f,\eta,\Lambda_n)=\phi(f,\eta)=\inf_\Lambda\phi(f,\eta,\Lambda). \quad (36)$$

Then

$$\begin{aligned}\phi'(f,\eta_-) &\geq \limsup_{n\to\infty}\left(\phi(f,1,\Lambda_n)-\phi(f,0,\Lambda_n)\right)\\ &\geq \liminf_{n\to\infty}\left(\phi(f,1,\Lambda_n)-\phi(f,0,\Lambda_n)\right)\\ &\geq \phi'(f,\eta_+) \quad (37)\end{aligned}$$

and for all except possibly a countable set of $\eta\in(0,1)$

$$\lim_{n\to\infty}\left(\phi(f,1,\Lambda_n)-\phi(f,0,\Lambda_n)\right)=\phi'(f,\eta). \quad (38)$$

In addition

$$\begin{aligned}\phi(f,\eta)+(1-\eta)\phi'(f,\eta_-) &\geq \limsup_{n\to\infty}\phi(f,1,\Lambda_n)\\ &\geq \liminf_{n\to\infty}\phi(f,1,\Lambda_n)\\ &\geq \phi(f,\eta)+(1-\eta)\phi'(f,\eta_+)\end{aligned}$$

and for all except possibly a countable set of $\eta\in(0,1)$

$$\lim_{n\to\infty}\phi(f,1,\Lambda_n)=\phi(f,\eta)+(1-\eta)\phi'(f,\eta).$$

The functions $\theta(f,\lambda,\eta,\mathcal{S})$ and $\theta(f,\lambda,\eta)$ of (11) and (12) have the form considered in Lemma 4, which with known results for the traditional cases $\eta=0,1$ yields the following.

Corollary 3:

$$\begin{aligned}\theta(f,\lambda,0,\mathcal{S}) \leq \theta(f,\lambda,\eta,\mathcal{S}) &\leq \theta(f,\lambda,1,\mathcal{S})\\ \theta(f,\lambda,0) \leq \theta(f,\lambda,\eta) &\leq \theta(f,\lambda,1).\end{aligned}$$

If f satisfies conditions for the traditional (i.e., fixed- and variable-rate) results, then

$$\begin{aligned}\theta_k+h(f) &= \theta(f,0) \leq \liminf_{\lambda\to0}\theta(f,\lambda,\eta)\\ &\leq \limsup_{\lambda\to0}\theta(f,\lambda,\eta)\leq\theta(f,1)\\ &= \psi_k+\ln\|f\|_{k/(k+2)}^{2/k}.\end{aligned}$$

A. Limiting Distortion and Rate

The following is a simple technical result that extends a property of the variable-rate and fixed-rate cases to the combined constraint case.

Lemma 5: Suppose that $\lambda_n>0$ converges to 0 and a sequence of quantizer partitions $\mathcal{S}_n$ satisfies $\limsup_{n\to\infty}\theta(f,\lambda_n,\eta,\mathcal{S}_n)=c$ for a finite constant c. Then

$$\begin{aligned}\lim_{n\to\infty}\lambda_n\left[(1-\eta)H_f(\mathcal{S}_n)+\eta\ln|\mathcal{S}_n|\right] &= 0\\ \lim_{n\to\infty}D_f(\mathcal{S}_n) &= 0.\end{aligned}$$

Proof: Since $\lambda_n>0$

$$\begin{aligned}&\limsup_{n\to\infty}\lambda_n\theta(f,\lambda_n,\eta,\mathcal{S}_n)\\ &\leq\left(\limsup_{n\to\infty}\lambda_n\right)\left(\limsup_{n\to\infty}\theta(f,\lambda_n,\eta,\mathcal{S}_n)\right)\\ &=0\end{aligned}$$

and hence

$$\begin{aligned}0 &\geq \limsup_{n\to\infty}\Bigg(D_f(\mathcal{S}_n)+\lambda_n\bigg((1-\eta)H_f(\mathcal{S}_n)+\eta\ln|\mathcal{S}_n|\bigg)\\ &\qquad +\frac{k}{2}\lambda_n\ln\lambda_n\Bigg)\\ &= \limsup_{n\to\infty}\left(D_f(\mathcal{S}_n)+\lambda_n\left((1-\eta)H_f(\mathcal{S}_n)+\eta\ln|\mathcal{S}_n|\right)\right)\end{aligned}$$

which implies the lemma since all terms in the last expression are all nonnegative. □

The next result applies the previous ideas to show that under certain conditions, the separate asymptotic behavior of distortion, codebook size, and entropy can be teased apart.

Lemma 6: Given a pdf f and $\lambda_n \to 0$, suppose that a sequence of partitions $\mathcal{S}_n$ satisfies

$$\lim_{n\to\infty} \theta(f, \lambda_n, \eta, \mathcal{S}_n) = \Theta(f, \eta) \tag{39}$$

then

$$\lim_{n\to\infty} \frac{2D_f(\mathcal{S}_n)}{k\lambda_n} = 1 \tag{40}$$

$$\begin{aligned}\lim_{n\to\infty} &\left((1-\eta)H_f(\mathcal{S}_n) + \eta \ln|\mathcal{S}_n| + \frac{k}{2}\ln\lambda_n\right)\\ &= \Theta(f,\eta) - \frac{k}{2}\end{aligned} \tag{41}$$

$$\begin{aligned}\Theta'&(f,\eta_-)\\ &\geq \limsup_{n\to\infty}\,(\ln|\mathcal{S}_n| - H_f(\mathcal{S}_n))\\ &\geq \liminf_{n\to\infty}\,(\ln|\mathcal{S}_n| - H_f(\mathcal{S}_n)) \geq \Theta'(u,\eta_+)\end{aligned} \tag{42}$$

$$\begin{aligned}\Theta(u,\eta) &- \eta\Theta'(u,\eta_+) - \frac{k}{2}\\ &\geq \limsup_{n\to\infty}\left(H_f(\mathcal{S}_n) + \frac{k}{2}\ln\lambda_n\right)\\ &\geq \liminf_{n\to\infty}\left(H_f(\mathcal{S}_n) + \frac{k}{2}\ln\lambda_n\right)\\ &\geq \Theta(f,\eta) - \eta\Theta_k'(f,\eta_-) - \frac{k}{2}\end{aligned} \tag{43}$$

$$\begin{aligned}\Theta(f,\eta) &+ (1-\eta)\Theta'(f,\eta_-) - \frac{k}{2}\\ &\geq \limsup_{n\to\infty}\left(\ln|S_n| + \frac{k}{2}\ln\lambda_n\right)\\ &\geq \liminf_{n\to\infty}\left(\ln|S_n| + \frac{k}{2}\ln\lambda_n\right)\\ &\geq \Theta(f,\eta) + (1-\eta)\Theta'(f,\eta_+) - \frac{k}{2}.\end{aligned} \tag{44}$$

For all except possibly a countable number of η, the inequalities become equalities and the left and right derivatives equal the derivatives. For $\eta = 0$, the rightmost inequalities of (42) and (44) and the leftmost inequality of (43) remain valid. For $\eta = 1$, the leftmost inequalities of (42) and (44) and the rightmost inequality of (43) remain valid.

Proof: Equation (39), Corollary 1, and the bound (32) imply that

$$\lim_{n\to\infty}\left[\frac{2D_{u_a}(\mathcal{S}_n)}{k\lambda_n} - \ln\frac{2D_{u_a}(\mathcal{S}_n)}{k\lambda_n} - 1\right] = 0$$

which from the continuity of $r - \ln r - 1$ implies (40), which in turn combined with (39) yields (41). Lemma 4 and Corollary 2 provide the means of accomplishing this separation. Since $\Theta(f,\eta,\mathcal{S})$ has the form of Lemma 4 and $\Theta(f,1,\mathcal{S}) - \Theta(f,0,\mathcal{S}) = \ln|\mathcal{S}| - H_f(\mathcal{S})$, Corollary 2 yields the third relation. The penultimate result follows from subtracting η times the third result from the second result. The final result comes from adding $1-\eta$ times the third result to the second result. □

The lemma implies that under the assumed conditions, λ_n controls the separate asymptotic behavior of distortion and rate and not just their Lagrangian combination: for small λ_n $D_f(\mathcal{S}_n) \approx \frac{k}{2}\lambda_n$ and $(1-\eta)H_f(\mathcal{S}_n) + \eta\ln|\mathcal{S}_n| \approx \Theta(f,\eta) - \frac{k}{2}\ln(e\lambda_n)$. The lemma also sandwiches the difference between the entropy and the log codebook size between lower and upper bounds, which are equal for all except possibly a countable number of η, in which case the difference between the entropy and the log codebook size converges to a finite constant.

VI. Uniform Density

Our first of two principal results extends Zador's results completely for the uniform distribution on the unit cube to the combined constraint case. The theorem is proved in part B of the Appendix.

Theorem 1: Let u denote the uniform density on the unit cube and let $\eta \in [0,1]$. Then

$$\lim_{\lambda\to 0}\theta(u,\lambda,\eta) = \Theta(u,\eta) \tag{45}$$

where

$$\Theta(u,\eta) = \inf_{\lambda>0}\theta(u,\lambda,\eta)$$

is a finite constant.

In the traditional cases of $\eta = 0$ and 1, we have $\Theta(u,0) = \theta_k$ and $\Theta(u,1) = \psi_k$, respectively. In general, the theorem does not explicitly identify the value of $\Theta(u,\eta)$, but only states that the limit in (45) exists and is finite.

From (13) and (14), the theorem implies the traditional Zador asymptotic form

$$\lim_{R\to\infty} e^{\frac{2}{k}R}\delta_\eta(u,R) = \frac{k}{2}e^{(2/k)\Theta(u,\eta)-1}.$$

Comparing (45) with the formula (20) derived using Gersho's conjecture and approximations shows that the theorem is consistent with Gersho's approach if we identify

$$\Theta(u,\eta) = (k/2)\ln(2ec_k/k). \tag{46}$$

If Gersho's conjecture were true for a certain k, we would have $a_k = b_k = c_k$, and hence $\Theta(u,\eta) = (k/2)\ln(2ec_k/k)$ would not depend on η. This is the case for $k = 1$ (scalar quantization), where it is known that $a_1 = b_1 = 1/12$.

Theorem 1 describes the asymptotics for the case of a uniform density on the unit cube and shows that $\lim_{\lambda\to 0}\theta(u,\lambda,\eta) = \Theta(u,\eta)$. It follows the general approach of Zador and the proofs of the fixed-rate and variable-rate cases. The theorem states that the infimum over λ of $\theta(u,\lambda,\eta)$ is in fact a limit as λ goes to zero. The result has the intuitive interpretation that the best values of λ in the sense of minimizing $\theta(u,\lambda,\eta)$ are the values near 0, that is, in the high-rate regime.

Theorem 1 demonstrates that the assumptions of Lemma 6 are met in the case of the uniform distribution on a unit cube. With $f = u$ and $\eta = 1$, Lemma 6 thus implies that

$$\begin{aligned}\frac{2}{k}\ln\frac{a_k}{b_k} = \frac{\Theta(u,1) - \Theta(u,0)}{1} &\geq \Theta'(u,1_-)\\ &\geq \limsup_{n\to\infty}\,(\ln|\mathcal{S}_n| - H_f(\mathcal{S}_n))\end{aligned}$$

which provides an improvement to the upper bound on the asymptotic difference of entropy and codebook size of [14]. If the Gersho approximations were valid, the left hand side would be 0 and the optimal codebooks would have the maximum possible entropy. With the possible exception of $\eta = 0$, the pure variable-rate case, regardless of η log codebook size cannot asymptotically exceed the entropy by more than a finite constant.

VII. Achievable Performance

Our second principal result demonstrates that the Gersho functional provides an upper bound to the optimal achievable performance in the high-rate regime. The theorem is proved in parts C–F of the Appendix.

Recall that the function $\Theta(u,\eta)$ is nondecreasing, concave, and continuous in $\eta \in [0,1]$. Its left derivative $\Theta(u,\eta_-)$ and right derivative $\Theta(u,\eta_+)$ with respect to η are defined, nonnegative, nonincreasing, and finite for $\eta \in (0,1)$, and $\Theta'(u,\eta_-) \geq \Theta'(u,\eta_+)$. The left and right derivatives are equal, and hence, the derivative exists for all except possibly a countable set of $\eta \in (0,1)$.

Theorem 2: Assume that X has an absolutely continuous distribution with pdf f such that

$$E_f(\|X\|^{2+\delta}) = \int f(x)\|x\|^{2+\delta}dx < \infty \tag{47}$$

for some $\delta > 0$ and

$$h(f) > -\infty \tag{48}$$

and that $\eta \in [0,1]$. Then

$$\limsup_{\lambda\to 0} \theta(f,\lambda,\eta) \leq \Theta(u,\eta) + h(f,\eta)$$

where $\Theta(u,\eta)$ is defined in Theorem 1, where

$$h(f,\eta) = \begin{cases} \phi(f,\eta)+\eta(1-\eta)\left(\Theta'(u,\eta_-)-\Theta'(u,\eta_+)\right), & \eta\in(0,1)\\ \phi(f,\eta), & \eta=0,1\end{cases}$$

where $\phi(f,\eta)$ is the Gersho functional given by (18) and (19), and where $h(f,\eta) = \phi(f,\eta)$ for all except possibly a countable set of $\eta \in [0,1]$.

If Gersho's conjecture and approximations were true, then $(k/2)\ln(2ec_k/k) = \Theta(u,\eta)$ and (46) would imply that $\Theta'(u,\eta) = 0$ for all η and hence $h(f,\eta) = \phi(f,\eta)$ for all η, which implies exact agreement with the optimal asymptotic performance derived using Gersho's conjecture. If Gersho's conjecture and approximations are not true, then the result shows that the formula (20) derived using Gersho's methods at least provides an upper bound for all except possibly a countable set of η in $(0,1)$ provided the $(k/2)\ln(2ec_k/k)$ term is replaced by $\Theta(u,\eta)$. Based on Gersho's heuristics and the known traditional special cases, we conjecture that the converse results are true, that is, that

$$\lim_{\lambda\to 0} \theta(f,\lambda,\eta) = \theta(f,\eta) = \Theta(u,\eta) + \phi(f,\eta). \tag{49}$$

The conjecture is known to be true for the traditional cases where $\eta = 0$ if f has finite differential entropy and if the partition of Euclidean space into unit cubes has finite entropy [12], and where $\eta = 1$ if for some $\delta > 0$ (47) holds [4], [11]. In these cases, $h(f,\eta) = \phi(f,\eta)$ and $\Theta(u,0) = \theta_k$, $\phi(f,0) = h(f)$, $\Theta(u,1) = \psi_k$, $\phi(f,1) = \ln\|f\|_{k/(k+2)}^{k/2}$. If X has finite second moment [as is the case if condition (47) for $\eta = 1$ holds], then $h(f) < \infty$ and the entropy of the uniformly quantized X with cell side 1 is finite. Henceforth, we assume that the pdf f satisfies the moment condition (47) and that (48) holds, and hence also $h(f)$ is finite. The conjecture will also be shown to be true for the uniform densities u_a on a cube of side a with $\phi(u_a,\eta) = k\ln a$.

VIII. Discussion

The Zador-style results for fixed- and variable-rate quantization have been extended to combined entropy and codebook size constraints for uniform distributions on a cube. It has also been shown for general source densities that formulas developed based on the uniform distribution case coupled with a rigorous version of Gersho's heuristic arguments characterize an achievable performance for all except possibly a countable set of $\eta \in (0,1)$.

The development can be viewed as a variation on Gersho's methods, which provides heuristics that can be rigorously demonstrated. Instead of using assumptions on the optimal cell shapes and a heuristic development of the asymptotic entropy, we follow Zador's methods and base the asymptotics on high-rate optimal quantizers for uniform densities on small cubes. The approach shows that in place of the assumption of an asymptotic quantizer point density function Λ, a function playing the same role follows from a convex optimization problem involving Lagrangian multipliers of the component codes used to prove the theorem. The approach provides a means of estimating the log codebook size even though only a weighted sum of the entropy and log codebook size is constrained. The log codebook bounding does not provide a useful characterization of the purely variable rate case because the left derivatives required for the upper bound are not defined. This is reflected in the fact that without the constraint of a finite codebook, an infinite number of codewords may be needed to achieve the optimal variable length code [17]. This makes it remarkable that in the case of a very small but nonzero η, the log codebook size can asymptotically be greater than the entropy by no more than a constant, and hence, the fractional difference goes to zero.

A natural question is how much the performance might suffer in the purely variable-rate case by the addition of a constraint on the log codebook size.

Let q be a quantizer that is optimal under the combined rate constraint and has large rate $r(q) = (1-\eta)H_f(q) + \eta N(q)$. Assuming that conjecture (49) holds, the arguments (13) and (14) imply

$$D_f(q) = \delta_\eta(f, r(q)) \approx \frac{k}{2}e^{\frac{2}{k}\theta(f,\eta)-1}e^{-\frac{2}{k}r(q)}.$$

It is easy to see that the conjecture and the argument of Lemmas 3–6 imply $\ln N(q) - H_f(q) \approx \theta'(f,\eta)$, so we can express the distortion in terms of the quantizer's entropy rather than its combined rate as

$$D_f(q) \approx \frac{k}{2}e^{\frac{2}{k}\theta(f,\eta)-1}e^{-\frac{2}{k}\eta\theta'(f,\eta)}e^{-\frac{2}{k}H_f(q)}.$$

Let $\hat{q}$ denote an optimal entropy-constrained quantizer with large entropy $H_f(\hat{q})$. The above with $\eta = 0$ reduces to Zador's result

$$D_f(\hat{q}) \approx \frac{k}{2} e^{\frac{2}{k}\theta(f,0)-1} e^{-\frac{2}{k}H_f(\hat{q})}.$$

We have $\theta(f,0) = \theta_k + h(f)$ and $\theta(f,\eta) = \Theta(u,\eta) + \phi(f,\eta)$, where $\phi(f,\eta)$ is the Gersho functional given by (18) and (19). If the two quantizers have equal entropy, the logarithm of the distortion loss suffered by q with respect to $\hat{q}$ due to the constraint on its codebook size is given by

$$\ln\frac{D_f(q)}{D_f(\hat{q})} \approx \frac{2}{k}\Bigg[(\Theta(u,\eta) - \theta_k - \eta\Theta'(u,\eta)) \\ +(\phi(f,\eta) - h(f) - \eta\phi'(f,\eta))\Bigg].$$

By definition, the loss is always nonnegative (this can also be seen from the fact that both $\Theta(u,\eta)$ and $\phi(f,\eta)$ are concave functions of η). If Gersho's conjecture holds true, the loss reduces to

$$\ln\frac{D_f(q)}{D_f(\hat{q})} \approx \frac{2}{k}\bigg(\phi(f,\eta) - h(f) - \eta\phi'(f,\eta)\bigg).$$

Since we do not have a closed-form expression for the Gersho functional, we cannot readily evaluate the loss, but numerical optimization methods can be used to give a good approximation for well-behaved source densities.

Appendix

A. *Proof of Corollary 2*

From (34), we have for all $\Delta\eta > 0$

$$\begin{aligned}\phi(f,1,\Lambda_n) - \phi(f,0,\Lambda_n) &= \frac{\phi(f,\eta+\Delta\eta,\Lambda_n) - \phi(f,\eta,\Lambda_n)}{\Delta\eta} \\ &\geq \frac{\phi(f,\eta+\Delta\eta) - \phi(f,\eta,\Lambda_n)}{\Delta\eta} \\ \phi(f,1,\Lambda_n) - \phi(f,0,\Lambda_n) &= \frac{\phi(f,\eta,\Lambda_n) - \phi(f,\eta-\Delta\eta,\Lambda_n)}{\Delta\eta} \\ &\leq \frac{\phi(f,\eta,\Lambda_n) - \phi(f,\eta-\Delta\eta)}{\Delta\eta}.\end{aligned}$$

Letting $n \to \infty$ yields

$$\begin{aligned}\frac{\phi(f,\eta) - \phi(f,\eta-\Delta\eta)}{\Delta\eta} &\geq \limsup_{n\to\infty}(\phi(f,1,\Lambda_n) - \phi(f,0,\Lambda_n)) \\ &\geq \liminf_{n\to\infty}(\phi(f,1,\Lambda_n) - \phi(f,0,\Lambda_n)) \\ &\geq \frac{\phi(f,\eta+\Delta\eta) - \phi(f,\eta)}{\Delta\eta}.\end{aligned}$$

Letting $\Delta\eta \to 0$ proves (37) and (38) follows from the previous lemma. The remainder of the corollary follows from $\phi(f,1,\Lambda_n) = \phi(f,\eta,\Lambda_n) + (1-\eta)(\phi(f,1,\Lambda_n) - \phi(f,0,\Lambda_n)$, (36), and (37). □

B. *Proof of Theorem 1*

Obviously $\underline{\theta}(u,\eta) \geq \Theta(u,\eta)$, so we need only to show that $\overline{\theta}(u,\eta) \leq \Theta(u,\eta)$. The proof mimics the first step of the corresponding result for the entropy-constrained case in [12] with some nontrivial changes. For any fixed integer M, carve the unit cube C_1 into a collection of disjoint cubes $\{C_{1/M,m}; m = 1,\ldots,M^k\}$ with sides $a = 1/M$. Let $u_{1/M,m}$ denote the uniform density on cube $C_{1/M,m}$. Suppose that $\mathcal{S}^{(1)}$ is an approximately optimal quantizer partition for $u_{1/M,1}$, that is, for an arbitrarily small $\epsilon > 0$, $\theta(u_{1/M,1},\lambda,\eta,\mathcal{S}^{(1)}) \leq \theta(u_{1/M,1},\lambda,\eta) + \epsilon$, and suppose that the quantizer associated with $\mathcal{S}^{(1)}$ has N_M words, entropy H_M, codebook $\mathcal{C}_M$. Let ℓ_M denote the length function. For all other subcubes $C_{1/M,m}$, use a translate of the partition $\mathcal{S}^{(1)}$ to form $\mathcal{S}^{(m)}$, $m = 2,\ldots,M^k$. All of the subcodes will have the same number of codewords, the same length function, the same entropy, and the same average distortion $D_{u_{1/M}}(\mathcal{S}^{(1)})$. Thus, $\theta(u_{1/M},\lambda,\eta,\mathcal{S}^{(1)}) = \theta(u_{1/M,m},\lambda,\eta,\mathcal{S}^{(m)})$, all m. Form a composite or union code q with partition $\mathcal{S}$ with atoms comprising all of the atoms of the subcode partitions $\mathcal{S}^{(m)}$. The composite codebook will have $N = M^k N_M$ words, an encoder $\alpha(x) = (\alpha_{M,m}(x), m)$ if $x \in C_{1/M,m}$, and a decoder $\beta(i,m) = \beta_{M,m}(i)$. The performance resulting from this composite code will be bound below by the best possible performance, $\theta(u,\lambda,\eta)$. We have, with $w_m = M^{-k}$, that

$$\begin{aligned}&\theta(u_{1/M},\lambda,\eta) + \epsilon \\ &\quad = \sum_{m=1}^{M^k} w_m \theta(u_{1/M,m},\lambda,\eta) + \epsilon \\ &\quad \geq \sum_{m=1}^{M^k} w_m \theta(u_{1/M,m},\lambda,\eta,\mathcal{S}^{(m)}) \\ &\quad = \sum_{m=1}^{M^k} w_m \Bigg(\frac{D_{u_{1/M,m}}(\mathcal{S}^{(m)})}{\lambda} + \frac{k}{2}\ln\lambda \\ &\qquad\qquad + \Big[(1-\eta)H_{u_{1/M,m}}(\mathcal{S}^{(m)}) + \eta\ln N_M\Big]\Bigg).\end{aligned}$$

Since $u_{1/M,m}$ is the conditional density of u given the cube $C_{M,m}$

$$\begin{aligned}D_u(\mathcal{S}) &= \sum_m w_m D_{u_{1/M,m}}(\mathcal{S}^{(m)}) \\ H_u(\mathcal{S}) &= -\sum_{m=1}^{M^k}\sum_{i=1}^{N_M} P_u(S_i^{(m)})\ln P_u(S_i^{(m)}) \\ &= \sum_{m=1}^{M^k} w_m H_{u_{1/M,m}}(\mathcal{S}^{(m)}) + H_u(Z)\end{aligned}$$

where we have defined the random variable $Z = m$ if $X \in C_{1/M,m}$ (so that $H_u(Z) = \ln M^k$). Therefore

$$\begin{aligned}&\theta(u_{1/M},\lambda,\eta) + \epsilon \\ &\quad \geq \sum_{m=1}^{M^k} w_m \Bigg(\frac{D_{u_{1/M,m}}(\mathcal{S}^{(m)})}{\lambda} + \frac{k}{2}\ln\lambda \\ &\qquad\qquad + \Big[(1-\eta)H_{u_{1/M,m}}(\mathcal{S}^{(m)}) + \eta\ln N_M\Big]\Bigg) \\ &\quad = \frac{D_u(\mathcal{S})}{\lambda} + \frac{k}{2}\ln\lambda + (1-\eta)H_u(\mathcal{S}) + \eta\ln N - k\ln M \\ &\quad \geq \theta(u,\lambda,\eta) - k\ln M\end{aligned}$$

which implies that $\theta(u_{1/M},\lambda,\eta) \geq \theta(u,\lambda,\eta) - k\ln M$, which with Lemma 7 in part C of the Appendix means that for any $\lambda > 0$ and any integer M, $\theta(u,\lambda,\eta) - k\ln M \leq \theta(u_{1/M},\lambda,\eta) = \theta(u,M^2\lambda,\eta) - k\ln M$ or $\theta(u,\lambda,\eta) \leq \theta(u,M^2\lambda,\eta)$. Replacing $M^2\lambda$ by λ, $\theta(u,\lambda,\eta) \geq \theta(u,M^{-2}\lambda,\eta)$, any $\lambda > 0$, and integer $M \geq 1$.

The remainder of this proof follows closely the proof of Lemma 9 for the entropy-constrained case in [12]. Fix λ and note that $(0,\lambda] = \bigcup_{m=1}^{\infty}(\lambda/(M+1)^2, \lambda/M^2]$, so for any $\lambda' \in (0,\lambda]$, there is an integer M such that $\lambda/(M+1)^2 < \lambda' \leq \lambda/M^2$. $\rho(f,\lambda,\eta)$ is easily seen to be nonincreasing with decreasing λ, hence

$$\begin{aligned}\theta(u,\lambda,\eta) \geq& \frac{\rho(u,\lambda',\eta)}{\left(\frac{M+1}{M}\right)^2\lambda'} + \frac{k}{2}\ln\lambda' \\ =& \left(\frac{M}{M+1}\right)^2\theta(u,\lambda',\eta) + \left(\frac{2M+1}{M^2+2M+1}\right)\frac{k}{2}\ln\lambda'.\end{aligned}$$

Choose any subsequence of λ' tending to zero. The largest possible value of the limit superior of the right-hand side is $\overline{\theta}(u,\eta)$, and hence, $\theta(u,\lambda,\eta) \geq \overline{\theta}(u,\eta)$, which means that $\Theta(u,\eta) \triangleq \inf_\lambda \theta(u,\lambda,\eta) \geq \overline{\theta}(u,\eta)$. Hence, $\underline{\theta}(u,\eta) \geq \Theta(u,\eta) \geq \overline{\theta}(u,\eta)$, and hence, the limit $\lim_{\lambda\to 0}\theta(u,\lambda,\eta)$ must exist and equal $\Theta(u,\eta)$. □

C. Uniform Densities on a Cube

Let $f = u_a$, the uniform density on a cube of side a and, in particular, on the unit cube. The traditional results for $\eta = 0$ and 1 are well known for this case. Define the uniform pdf $u_{a,r}$ on the cube $C_{a,r}$ by $u_{a,r}(x) = V(C_{a,r})^{-1}1_{C_{a,r}}(x)$, where $1_C(x)$ denotes the indicator function of the set $C \subset \Re^k$. We often abbreviate $u_{a,0}$ to u_a and u_1 to u. Then, $u_{a,r}(x) = a^{-k}1_{C_a}(x-r) = a^{-k}u((x-r)/a)$ and $V(C_{a,r}) = a^k$, $h(u_{a,r}) = \ln V(C_{a,r}) = k\ln a$, and $\ln\|u_{a,r}\|_{k/(k+2)}^{k/2} = k\ln a$.

Define a cube in $\Re^k$ with side a and location r as $C_{a,r} = \{x : r_i \leq x_i < r_i + a; i = 1,\ldots,k\} = \prod_{i=1}^{k}[r_i, r_i+a)$. 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. Suppose that X is a random variable with pdf f_1 on the unit cube C_1. Then, the scaled random variable $Y = aX + r$ for any $a > 0$ has a pdf $f_{a,r}(x) = a^{-k}f_1((x-r)/a)$ on the cube $C_{a,r}$. Any quantizer q_1 with encoder α_1 and decoder β_1 for X implies a corresponding quantizer for Y

$$\begin{aligned}\alpha_{a,r}(x) =& \alpha_1\left(\frac{x-r}{a}\right), \quad \beta_{a,r}(i) = a\beta_1(i) + r \\ q_{a,r}(x) =& \beta_{a,r}(\alpha_{a,r}(x)) = a\beta_1\left(\alpha_1\left(\frac{x-r}{a}\right)\right) + r \\ =& aq_1\left(\frac{x-r}{a}\right) + r. \end{aligned} \tag{50}$$

Conversely, given a quantizer $q_{a,r}$ for $f_{a,r}$, one can construct a corresponding quantizer for f_1.

Lemma 7: $\theta(f_{a,r},\lambda,\eta,q_{a,r}) = \theta(f_1,a^{-2}\lambda,\eta,q_1) + k\ln a$, $\theta(f_{a,r},\lambda,\eta) = \theta\left(f_1,a^{-2}\lambda,\eta\right) + k\ln a$.

The following result relates the performance for a given random variable with support on a cube to that of a shifted or scaled version of the random variable. The result is an extension of [12, Lemmas 7 and 8] to more general densities and combined rate constraints. The proof is essentially a change of variables and follows [12] closely. The details are in part G of the Appendix. The lemma allows us to focus on the particular case of densities on the unit cube to infer the properties of densities on any shifted and scaled cube.

Theorem 1 and Lemma 7 immediately imply the following.

Corollary 4: Equation (49) holds for a uniform density u_a on a cube of size $a > 0$ with $h(u_a,\eta) = \phi(u_a,\eta) = k\ln a$ and $\Delta_k(\eta) = 0$. Thus, $\theta(u_a,\eta) = \Theta(u_a,\eta) = \Theta(u,\eta) + k\ln a$.

The corollary shows that the conditions of Lemma 6 are again satisfied.

D. Piecewise Constant Pdfs on Cubes

In the variable-rate case, the result for Zador's second step is easy because of the nice behavior of the limiting functions on disjoint mixtures [12]. One constructs separate codes for all of the cubes with constant pdfs and then quantifies the behavior of the union codebook using an essentially linear decomposition of the conditional distortions and entropies. In the fixed-rate case, the corresponding step is much harder [4], [11], [24] because one must solve a bit allocation problem across the cubes in order to optimize the collection of codebooks overall by assigning to each an appropriate number of quantization points, which sum to the total available. Neither approach works alone in the constrained case. As in the traditional cases, we build a union codebook, but we choose local Lagrange multipliers so as to optimize an overall average.

We begin with a heuristic development that can be viewed as a variation on Gersho's heuristic approach wherein instead of making assumptions on the behavior of individual cells in asymptotically optimal quantizers, we focus on the provable behavior in Lemma 6 of asymptotically optimal quantizers on individual cubes and then combine the collection of quantizers into a single overall quantizer. For simplicity, the heuristic development focuses on η for which the derivatives in Lemma 6 exist and the inequalities are equalities, that is, for all except possibly a countable number of $\eta \in (0,1)$. The supporting rigorous arguments will use the more general bounds of Lemma 6.

Assume that a pdf f is zero outside the union of a finite number M of disjoint cubes $\{C(m); m = 1,2,\ldots,M\}$ of equal side a. In particular, consider a pdf of the form

$$f(x) = \sum_m w_m f^{(m)}(x) = \sum_m w_m a^{-k}1_{C(m)}(x) \tag{51}$$

where a^k is the volume of each $C(m)$, $w_m \geq 0$, and $\sum_m w_m = 1$, so that w_m is a probability mass function (pmf). For this section, we consider a fixed. In the next section, we will use pdfs of this form with small a to approximate more general pdfs.

In terms of the pdf f

$$h(f) = -\int f(x)\ln f(x)dx = H(w) + k\ln a$$

and

$$\ln \|f\|_{k/(k+2)}^{k/2} = \ln \|w\|_{k/(k+2)}^{k/2} + k \ln a$$

where $w = (w_1, \ldots, w_M)$, $H(w) = -\sum_m w_n \ln w_m$, and $\|w\|_\rho = (\sum_m w_m^\rho)^{1/\rho}$.

For each cube $C(m)$ with nonzero probability $w_m > 0$, we design a nearly optimal code with partition $\mathcal{S}^{(m)}$ for the conditionally uniform pdf $f^{(m)}$ using a Lagrange multiplier $\lambda_m > 0$ and a common value of η for all m. For the moment, we leave open the choice of the λ_m except for the assumptions that it is strictly positive (or the solution would yield infinite rate and zero distortion) and that for all m, for which $w_m > 0$, the λ_m are small enough to ensure from Lemma 6 that

$$D_{f^{(m)}}(\mathcal{S}^{(m)}) \approx \frac{k\lambda_m}{2} \tag{52}$$

$$\ln |S_m| - H_{f^{(m)}}(\mathcal{S}^{(m)}) \approx \Theta'(u,\eta) \tag{53}$$

$$H_{f^{(m)}}(\mathcal{S}^{(m)}) + \frac{k}{2}\ln \lambda_m \approx \Theta(u,\eta) - \eta\Theta'(u,\eta) + k\ln a - \frac{k}{2} \tag{54}$$

$$\ln |\mathcal{S}^{(m)}| + \frac{k}{2}\ln \lambda_m \approx \Theta(u,\eta) + (1-\eta)\Theta'(u,\eta) + k \ln a - \frac{k}{2}. \tag{55}$$

Observe that λ_m controls the number of quantization points in each cell $C(m)$, i.e.,

$$|S^{(m)}| \approx \lambda_m^{-k/2} e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta)+k\ln a - k/2}. \tag{56}$$

The constant of proportionality is complicated, but it disappears in the fraction of quantizer points falling in a single cell $C(m)$ given by

$$\frac{|\mathcal{S}^{(m)}|}{\sum_l |\mathcal{S}^{(l)}|} \approx \frac{\lambda_m^{-k/2}}{\sum_l \lambda_l^{-k/2}} \triangleq \Lambda_m \tag{57}$$

where we have defined the *quantizer pmf* Λ_m as the fraction of quantizer points within the mth cube. Thus, if we wish to find the fraction of quantization levels within any set consisting of the union of a disjoint collection of the small cubes, we need only to sum up the values of Λ_m over the m indexing the small cubes in the subset. Thus, integrating the function

$$\Lambda(x) \triangleq \sum_m \frac{\Lambda_m}{V(C(m))} 1_{C(m)}(x) = a^{-k} \sum_m \Lambda_m 1_{C(m)}(x) \tag{58}$$

over this set gives the fraction of quantizer points in the set. For sets consisting of unions of partition cells, this characterizes $\Lambda(x)$ as a quantizer point density function.

Each cube $C(m)$, for which $w_m = 0$, should be assigned zero rate, that is, it should have zero entropy and zero log codebook size so as to not waste bits.

Construct a composite code based on these subcodes. The composite partition $\mathcal{S} = \{S_i\}$ has as atoms all of the atoms of all the small cube partitions for those cubes with nonzero probability $\mathcal{S}^{(m)} = \{S_i^{(m)}\}$, together with a single atom, which we denote S_0, which is the union of all of the cells having zero probability. For this partition, we proceed to evaluate the performance using the composite code. Keep in mind that all sums are over the support set of w.

Consider each term in the sum $\theta(f,\lambda,\eta,\mathcal{S}) = \frac{D_f(\mathcal{S})}{\lambda} + \frac{k}{2}\ln\lambda + [(1-\eta)H_f(\mathcal{S}) + \eta \ln |\mathcal{S}|]$. Apply (52)–(55) to write the approximations

$$\begin{aligned}
D_f(\mathcal{S}) &= \sum_m D_{f^{(m)}}(\mathcal{S}^{(m)}) w_m \approx \frac{k}{2}\sum_m \lambda_m w_m \\
H_f(\mathcal{S}) &= -\sum_i P_f(S_i) \ln P_f(S_i) \\
&= -\sum_m \sum_i P_f\left(S_i^{(m)}\right) \ln P_f\left(S_i^{(m)}\right) \\
&\approx -\frac{k}{2}\sum_m w_m \ln \lambda_m + H(w) + \Theta(u,\eta) \\
&\quad - \eta\Theta'(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e} \\
\ln |\mathcal{S}| &= \ln\left(\sum_{m=1}^{M} |\mathcal{S}^{(m)}|\right) \\
&\approx \ln\left(\sum_m \exp\left(-\frac{k}{2}\ln\lambda_m + \Theta(u,\eta) \right.\right. \\
&\qquad \left.\left. + (1-\eta)\Theta'(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e}\right)\right) \\
&= \Theta(u,\eta) + (1-\eta)\Theta'(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e} + \ln\sum_m \lambda_m^{-k/2}.
\end{aligned}$$

The sum for evaluating the total number of quantization cells is over the support set of w_m and hence excludes the cell of $\mathcal{S}$ containing all the zero probability cubes.

Combining the preceding equalities and approximations

$$\begin{aligned}
&\theta(f,\lambda,\eta,\mathcal{S}) \\
&\quad\approx \frac{k}{2}\sum_m \frac{\lambda_m}{\lambda} w_m + \frac{k}{2}\ln\lambda + (1-\eta) \\
&\qquad \times \left(-\frac{k}{2}\sum_m w_m \ln\lambda_m + H(w) + \Theta(u,\eta)\right. \\
&\qquad\qquad \left. - \eta\Theta'(u,\eta) + \frac{k}{2}\ln a^2 - \frac{k}{2}\right) \\
&\qquad + \eta\left[\Theta(u,\eta) + (1-\eta)\theta_k'(\eta) + \frac{k}{2}\ln a^2\right. \\
&\qquad\qquad \left. - \frac{k}{2} + \ln\sum_m \left(\frac{1}{\lambda_m}\right)^{\frac{k}{2}}\right] \\
&\quad = \Theta(u,\eta) + k\ln a - \frac{k}{2} + (1-\eta)H(w) \\
&\qquad + \frac{k}{2}\sum_m \frac{\lambda_m}{\lambda} w_m - (1-\eta)\frac{k}{2}\sum_m w_m \ln\frac{\lambda_m}{\lambda} \\
&\qquad + \eta \ln \sum_m \left(\frac{\lambda_m}{\lambda}\right)^{-k/2}.
\end{aligned}$$

Interestingly, the derivative terms are canceled. Recall that λ is considered fixed (and very small) and that only the λ_m are free to optimize. For convenience, we normalize the Lagrangian parameters as $\mu_m = \lambda_m/\lambda$

$$\begin{aligned}&\theta(f,\lambda,\eta,\mathcal{S})\\ &\quad\approx \Theta(u,\eta) + \left[\frac{k}{2}\ln\frac{a^2}{e} + (1-\eta)H(w) + \frac{k}{2}\sum_m w_m\mu_m\right.\\ &\qquad\qquad \left. - (1-\eta)\frac{k}{2}\sum_m w_m\ln\mu_m + \eta\ln\sum_m \mu_m^{-k/2}\right]\\ &\quad\triangleq \Theta(u,\eta)+\Phi(w,\eta,\mu).\end{aligned}\tag{59}$$

If we fix the μ_m and let $\lambda \to 0$, then $\lambda_m = \mu_m\lambda \to 0$ for all m, satisfying our requirement for invoking the asymptotic results. The μ_m can be considered as the relative Lagrangian multipliers for the cells, which are held constant as λ becomes small.

The following lemma makes this development precise. It is proved in part H of the Appendix

Lemma 8: Given a piecewise constant pdf f of the form (51) and a positive vector $\mu = \{\mu_m\}$, then for any $\lambda_n \to 0$, there exists a sequence of partitions $\mathcal{S}_n$ such that

$$\begin{aligned}&\Theta(u,\eta)+\Phi(w,\eta,\mu)+\eta(1-\eta)\left(\Theta'(u,\eta_-)-\Theta'(u,\eta_+)\right)\\ &\quad\ge \limsup_{n\to\infty}\theta(f,\lambda_n,\eta,\mathcal{S}_n)\\ &\quad\ge \liminf_{n\to\infty}\theta(f,\lambda_n,\eta,\mathcal{S}_n)\\ &\quad\ge \Theta(u,\eta)+\Phi(w,\eta,\mu)-\eta(1-\eta)\left(\Theta'(u,\eta_-)-\Theta'(u,\eta_+)\right).\end{aligned}$$

For all except possibly a countable number of $\eta \in (0,1)$, for any $\lambda_n \to 0$, there exists a sequence of partitions $\mathcal{S}_n$ such that

$$\lim_{n\to\infty}\theta(f,\lambda_n,\eta,\mathcal{S}_n) = \Theta(u,\eta)+\Phi(w,\eta,\mu).$$

The proof of the following corollary to Lemma 8 is contained in the proof of Lemma 8 in (108). The result will be important for the third step of the proof of Theorem 2.

Corollary 5: Given the assumptions of Lemma 8 and the composite quantizer construction with partition $\mathcal{S}_n$ of the proof, the codebook size behaves as

$$\begin{aligned}&a^k e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_-)-k/2}\sum_{m=1}^{M}\mu_m^{-k/2}\\ &\quad\ge \limsup_{n\to\infty}|\mathcal{S}_n|\lambda_n^{k/2}\\ &\quad\ge \liminf_{n\to\infty}|\mathcal{S}_n|\lambda_n^{k/2}\\ &\quad\ge a^k e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_+)-k/2}\sum_{m=1}^{M}\mu_m^{-k/2}.\end{aligned}$$

For all except possibly a countable number of $\eta \in (0,1)$ the upper and lower bounds are identical with $\Theta'(u,\eta_-) = \Theta'(u,\eta_+) = \Theta'(u,\eta)$

The following corollary follows immediately from Lemma 8.

Corollary 6: Given the assumptions of the previous lemma, define

$$h(f,\eta) = \Phi(w,\eta)+\eta(1-\eta)\left(\Theta'(u,\eta_-)-\Theta'(u,\eta_+)\right)$$

where $\Phi(w,\eta) = \inf_\mu\Phi(w,\eta,\mu)$, where the infimum is over all positive μ. Then, $\overline{\theta}(f,\eta) \le \Theta(u,\eta)+h(f,\eta)$. For all except possibly a countable collection of $\eta \in (0,1)$, $h(f,\eta) = \Phi(w,\eta)$.

Useful alternative forms for Φ are given by

$$\begin{aligned}\Phi(w,\eta,\mu) =&\ k\ln a + H(w) + \frac{k}{2}\sum_m w_m\left(\mu_m - \ln\mu_m - 1\right)\\ &+\eta H(w\|\Lambda)\end{aligned}\tag{60}$$

$$\begin{aligned}=&\ k\ln a - \frac{k}{2} + \frac{k}{2}\sum_m w_m\mu_m\\ &+\ln\sum_\ell \mu_\ell^{-k/2} - (1-\eta)H(w\|\Lambda)\end{aligned}\tag{61}$$

where from (57) $\Lambda_m = \mu_m^{-k/2}/\sum_l \mu_l^{-k/2}$ and the discrete relative entropy is given by $H(w\|\Lambda) = \sum_m w_m\ln w_m/\Lambda_m$. Equation (61) expresses Φ as $k\ln a + H(w) = h(f)$ plus the sum of two nonnegative terms, hence Φ is always defined (although it may be infinite). Define the pmf $\bar{\mu}$ by $\bar{\mu}_m = w_m\mu_m/\sum_l w_l\mu_l$. Then, (61) can be written as

$$\begin{aligned}\Phi(w,\eta,\mu) =&\ k\ln a + H(w) + \frac{k}{2}\left(\sum_m w_m\mu_m - \ln\sum_m w_m\mu_m - 1\right)\\ &+\frac{k}{2}H(w\|\bar{\mu}) + \eta H(w\|\Lambda)\\ \ge&\ k\ln a + H(w) + \frac{k}{2}H(w\|\bar{\mu}) + \eta H(w\|\Lambda).\end{aligned}$$

Given a vector μ, the lower bound can be achieved by replacing μ by the pmf μ' defined by $\mu'_m = \mu_m/\sum_l w_l\mu_l$, that is, by normalizing the μ vector with respect to w. Neither of the relative entropy terms changes since the rescaling factors are all canceled, so this substitution provides a strict improvement in $\Phi(w,\eta,\mu)$ if μ is not already suitably scaled. Thus, the infimum can be restricted to only those μ, for which

$$\sum_m w_m\mu_m = 1.\tag{62}$$

The function $\Phi(w,\eta,\mu)$ can be related to the ψ functional of Theorem 2 by the transformation by $\nu_m = \ln\mu_m$ and $\nu(x) = \sum_m 1_{C(m)}(x)\nu_m$, which with (59) and (31) implies

$$\begin{aligned}\Phi(w,\eta,\mu) =&\ (1-\eta)\left(k\ln a + H(w)\right) - \frac{k}{2} + \frac{k}{2}\sum_m w_m e^{\nu_m}\\ &-(1-\eta)\frac{k}{2}\sum_m w_m\nu_m + \eta\ln\sum_m a^k e^{-k\nu_m/2}\\ =&\ (1-\eta)h(f) + \frac{k}{2}\left[\int f(x)e^{\nu(x)}dx - (1-\eta)\right.\\ &\qquad\left.\times\int f(x)\nu(x)dx - 1\right]\\ &+\eta\ln\left(\int e^{-k\nu(x)/2}dx\right) = \psi(f,\eta,\nu).\end{aligned}$$

Straightforward application of the discrete analogs of the arguments of Lemma 2 shows that $\Phi(w,\eta) = \psi(w,\eta) = \phi(w,\eta)$, where $\Phi(w,\eta) = \inf_\mu \Phi(w,\eta,\mu)$, $\psi(w,\eta) = \inf_\nu \psi(w,\eta,\nu)$, $\phi(w,\eta) = \inf_\Lambda \phi(w,\eta,\Lambda)$, and where

$$\phi(w,\eta,\Lambda) \triangleq k\ln a + \frac{k}{2}\ln\left(\sum_m w_m \Lambda_m^{-k/2}\right) - (1-\eta)H(w\|\Lambda)$$

where the infima are over positive vectors μ, vectors ν, and probability mass functions Λ, respectively, for which the functions are well defined.

As with the ψ function, the ϕ function becomes the continuous function of (18) if stated in terms of the induced densities. The transformations relating the various optimizations are

$$\begin{aligned}\nu_m &= \ln\mu_m\\ \Lambda_m &= \frac{\mu_m^{-k/2}}{\sum\limits_l \mu_l^{-k/2}} = \frac{e^{-k\nu_m/2}}{\sum\limits_l e^{-k\nu_l/2}}\\ \mu_m &= \frac{\Lambda_m^{-2/k}}{\sum\limits_m w_m\Lambda_m^{-2/k}}\\ \left(\sum_m w_m\Lambda_m^{-2/k}\right)^{k/2} &= \sum_m \mu_m^{-k/2}.\end{aligned}$$

In this finite-dimensional case, a continuity-compactness argument guarantees the existence of a minimum and the strict convexity of ψ in ν guarantees a unique minimizing value but no simple form for the minimum or for the optimizing ν is known except for the traditional cases of $\eta = 0, 1$. The existence of a minimizing ν, however, implies the existence of minimizing Λ and μ.

The optimization over Λ has the nice intuitive interpretation of optimizing over the fraction of quantizer levels contained in each of the small cubes, which relates it to the traditional bit allocation approach for the fixed rate case of Zador and to the fixed-rate and variable-rate heuristic developments of Gersho based on his conjecture. In the case considered, the code construction used shows how Λ relates to Lagrangian multipliers used to locally optimize codebooks.

The following corollary uses the existence of an optimum μ to modify the bound of Corollary 5 on the codebook size of the composite quantizers used in the construction. The corollary will be useful because the bound is dependent on M, the number of components in the piecewise constant model, only though the functional $\Phi(w,\eta)$. The lemma strongly resembles the heuristically derived Gersho approximation (23) if the identification of (46) of $\Theta(u,\eta)$ and c_k terms is made, which implies also that $\Theta'(u,\eta) = 0$.

Corollary 7: Given the assumptions of Lemma 8, assume in addition that μ minimizes $\Phi(w,\eta,\mu)$ so that $\Phi(w,\eta) = \Phi(w,\eta,\mu)$. Then

$$c_- \ge \limsup_{n\to\infty} |\mathcal{S}_n|\lambda_n^{k/2} \ge \liminf_{n\to\infty} |\mathcal{S}_n|\lambda_n^{k/2} \ge c_+ \tag{63}$$

where

$$\begin{aligned}c_- &= e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_-)-k/2}e^{\Phi'(w,\eta_-)} &(64)\\ c_+ &= e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_+)-k/2}e^{\Phi'(w,\eta_+)}. &(65)\end{aligned}$$

For all except possibly a countable collection of $\eta \in (0,1)$

$$\lim_{n\to\infty} |\mathcal{S}_n|\lambda_n^{k/2} = e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta)-k/2}e^{\Phi'(w,\eta)}. \tag{66}$$

Proof: From (61) and (62), $a^k\sum_m \mu_m^{-k/2} = \Phi(f,1,\mu)$, and hence, (63) follows from Corollary 5 combined with Corollary 2 applied to $\Phi(f,\eta,\mu)$ and $\Phi(f,\eta)$. Note that here the minimizing μ or Λ exists so the limits of Corollary 2 are not needed. □

E. General Densities on a Cube

To generalize from piecewise constant pdfs on a cube to more general pdfs on the unit cube, we approximate the latter by the former and use the codes of the previous section.

Recall that the distribution of X is assumed to be absolutely continuous with respect to Lebesgue measure with pdf f. We assume that f is zero outside the unit cube and that $h(f) > -\infty$ so that $h(f)$ is finite. In this case, the moment condition (47) and the condition that uniform quantization into unit cubes yields finite entropy are automatically satisfied. Given such a pdf f, let $Dom(\psi)$ be the domain of $\psi(f,\eta,\nu)$, that is, the collection of all ν, for which the integrals defining $\psi(f,\eta,\nu)$ exist. This is the collection of all ν, for which $\nu(x)$ and $e^{\nu(x)}$ are in the normed linear space $L_1(f)$ of f-integrable functions and $e^{-k\nu(x)/2}$ is integrable with respect to Lebesgue measure on the unit cube. $\mathrm{Dom}(\psi)$ is a convex subset of $L_1(f)$ from Holder's inequality. Thus, $\psi(f,\eta)$ is defined by a convex optimization problem.

We proceed to the limiting results needed for the case of general densities on the unit cube. For any positive integer, M partitions C_1 into M^k cubes of side length $a = 1/M$, say $\mathcal{S}_M = \{C(m);\ 1,2,\ldots,M^k\}$. Given a pdf f, form a piecewise constant approximation

$$f_M(x) = \sum_{m=1}^{M^k}\frac{P_f(C(m))}{V(C(m))}1_{C(m)}(x) = \sum_{m=1}^{M^k} w_m M^k 1_{C(m)}(x).$$

The use of the piecewise constant approximation to the original pdf follows that of [4] and [11]. This is a disjoint mixture source with $w_m = P_f(C(m))$ and component pdfs $M^k 1_{C(m)}(x)$. If P_M denotes the distribution induced by f_M, i.e., $P_M(F) = \int_F f_M(x)dx$, then $f_M = dP_M/dV(x)$. The following lemma showing some not surprising asymptotic properties is proved in part A of the Appendix.

Lemma 9: Assume that f and f_M are as defined in this section. Then

$$\begin{aligned}\lim_{M\to\infty}\|f_M - f\|_1 &= 0 &(67)\\ \lim_{M\to\infty} h(f_M) &= h(f) &(68)\\ \lim_{M\to\infty}\psi(f_M,\eta) &= \psi(f,\eta) &(69)\\ \lim_{M\to\infty} h(f_M,\eta) &= h(f,\eta) &(70)\end{aligned}$$

$$\begin{aligned}\psi'(f,\eta_-) &\ge \limsup_{M\to\infty}\psi'(f_M,\eta_-)\\ &\ge \liminf_{M\to\infty}\psi'(f_M,\eta_+)\\ &\ge \psi'(f,\eta_+). \end{aligned}\tag{71}$$

Construction of Quantizers: We proceed as in the variable length case [12] to modify the quantizer for the piecewise constant approximation in a way that does not affect the performance much, but allows us to bound the limiting behavior when the quantizer is actually applied to the true pdf f. The approach here differs, however, in that both entropy and log codebook size must be controlled, and we wish to construct a single sequence of partitions that asymptotically approaches the performance promised by the Gersho style approximations.

For the entire development, assume we are given positive sequences λ_n and ϵ_M such that

$$\lim_{n\to\infty} \lambda_n = \lim_{M\to\infty} \epsilon_M = 0. \tag{72}$$

For each integer M, choose $\mu = \mu^{(M)}$ in Lemma 8 as the value minimizing $\Phi(w,\eta,\mu)$ as $\Phi(w,\eta,\mu) = \Phi(w,\eta) = \psi(f_M,\eta)$. Then, given the sequence λ_n, the lemma and Corollary 7 imply the existence of a sequence of partitions $\{\mathcal{S}_{M,n}; n = 1,2,\ldots\}$, for which

$$\begin{aligned}\Theta(u,\eta) &+ \psi(f_M,\eta) + \eta(1-\eta)\left(\Theta'(u,\eta_-) - \Theta'(u,\eta_+)\right)\\ &\geq \limsup_{n\to\infty} \theta(f_M,\lambda_n,\eta,\mathcal{S}_{M,n})\\ &\geq \liminf_{n\to\infty} \theta(f_M,\lambda_n,\eta,\mathcal{S}_{M,n})\\ &\geq \Theta(u,\eta)+\psi(f_M,\eta)-\eta(1-\eta)\left(\Theta'(u,\eta_-)-\Theta'(u,\eta_+)\right)\end{aligned}$$

and

$$c_-^{(M)} \geq \limsup_{n\to\infty} |\mathcal{S}_{M,n}|\lambda_n^{k/2} \geq \liminf_{n\to\infty} |\mathcal{S}_{M,n}|\lambda_n^{k/2} \geq c_+^{(M)}$$

where

$$c_-^{(M)} = e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_-)-k/2} e^{\psi'(f_M,\eta_-)} \tag{73}$$

$$c_+^{(M)} = e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_+)-k/2} e^{\psi'(f_M,\eta_+)}. \tag{74}$$

Thus, given $\epsilon_M > 0$, we can choose an $n_0(f_M,\epsilon_M)$ sufficiently large to ensure that for $n \geq n_0(f_M,\epsilon_M)$

$$\begin{aligned}\Theta(u,\eta)&+\psi(f_M,\eta)+\eta(1-\eta)\left(\Theta'(u,\eta_-)-\Theta'(u,\eta_+)\right)+\epsilon_M\\ &\geq \theta(f_M,\lambda_n,\eta,\mathcal{S}_{M,n})\\ c_-^{(M)}(1+\epsilon_M) &\geq |\mathcal{S}_{M,n}|\lambda_n^{k/2} \geq c_+^{(M)}(1-\epsilon_M).\end{aligned} \tag{75}$$

We modify these sequences of quantizers $\{\mathcal{S}_{M,n}; n = 1,2,\ldots\}$ in a way that will permit necessary bounding of the inaccuracies resulting when computing averages with respect to f instead of f_M. The following technical lemma is proved in part J of the Appendix. It shows that the partitions $\{\mathcal{S}_{M,n}; n = 1,2,\ldots\}$ will have for sufficiently small λ_n (or large enough n) a collection of subcells with total probability between $\epsilon_M/2$ and ϵ_M. For simplicity, for the moment, the dependence on M is suppressed as M can be considered fixed. Let $g = f_M$ and $\epsilon = \epsilon_M$.

Lemma 10: Let g be a pdf on C_1, for which $h(g)$ is finite and $\{\mathcal{S}_n; n = 1,2,\ldots\}$ is a sequence of partitions with corresponding quantizers q_n such that $\limsup_{n\to\infty}\theta(g,\lambda_n,\eta,\mathcal{S}_n) = c < \infty$. Then, for any $\epsilon \in (0,1)$, there is an $n_0 = n_0(g,\epsilon)$ such that if $n \geq n_0$ then all partitions $\mathcal{S}_n$ that satisfy $\theta(g,\lambda_n,\eta,\mathcal{S}_n) \leq c+\epsilon$ will have a collection $\{S_{n,i} : i \in \mathcal{J}_1\}$ of cells with total probability bounded as

$$\frac{\epsilon}{2} \leq \sum_{i\in\mathcal{J}_1} P_g(S_{n,i}) \leq \epsilon. \tag{76}$$

We continue to suppress the dependence on M until the modification of the quantizer is complete. Again, to simplify notation, we also suppress the n and will assume for the moment that $n \geq n_0$ is fixed as in the lemma. Abbreviate λ_n to λ. Consider the partition $\mathcal{S}^1 = \mathcal{S}_{M,n}$ and the corresponding quantizer q^1 with the optimal length function $\ell_1(i) - \ln P_g(S_i^1)$, where the single superscript 1 denotes that this is a quantizer designed for the distribution g as distinct from a second quantizer q^2, which will be designed to provide a worst case bond on Lagrangian distortion. Using Lemma 10 define

$$p^* = \sum_{i\in\mathcal{J}_1} P_g(S_i^1) \in [\epsilon/2,\epsilon]. \tag{77}$$

Choose a large constant $\gamma > 1$ to be specified later and define $\lambda' = \gamma\lambda$. Construct a second quantizer q^2 as a uniform k-dimensional (cubic lattice) quantizer with side width $\Delta = 1/K$, where $K = \lfloor \lambda'^{-1/2} \rfloor$ so that $K \leq \lambda'^{-1/2}$, $\Delta \leq \sqrt{\lambda'}/(1-\sqrt{\lambda'})$, $\Delta^2 \leq \lambda'/(1-2\sqrt{\lambda'}) \leq 2\lambda'$ if $\lambda' \leq 1/16$, which we can assume without loss of generality in the asymptotic ($\lambda \to 0$ and hence $\lambda' \to 0$) analysis, e.g., just redefine n_0. Then, for all $x \in C_1$

$$d(x,\beta_2(\alpha_2(x))) \leq k\frac{\Delta^2}{4} \leq \frac{k}{2}\lambda'. \tag{78}$$

Let α_2 and $\mathcal{I}_2$ denote the encoder and index set of q^2. This quantizer has

$$N_2 = K^k \leq \lambda'^{-\frac{k}{2}} \tag{79}$$

codewords. Define the (constant) length function ℓ_2 by

$$\ell_2(i) = -\ln p^* + 1 - \frac{k}{2}\ln\lambda'. \tag{80}$$

Note that ℓ_2 is admissible since

$$\sum_{i\in\mathcal{I}_2} e^{-\ell_2(i)} = K^k e^{-1} p^* \lambda'^{k/2} \leq e^{-1}p^* < 1. \tag{81}$$

A composite quantizer $\bar{q}$ is formed by merging the quantizers 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. The merging is accomplished by the universal coding technique of finding the minimum Lagrangian distortion codeword in the combined codebook: given an input vector x, define

$$m(x) = \operatorname*{argmin}_l \left(d(x,\beta_l(\alpha_l(x))) + \lambda(1-\eta)\ell_l(\alpha_l(x)) \right)$$

and define the encoder of $\bar{q}$ by $\bar{\alpha}(x) = (m,i) = (m(x),\alpha_{m(x)}(x))$. The minimum distortion rule uses the total number of codewords for the composite quantizer $\bar{N} = N_1 + N_2$, and hence the codebook sizes N_1 and N_2 do not affect the encoder.

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 (76), 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 from the choice of ℓ_1, (77), and (81)

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

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

$$\begin{aligned}&d(x, \bar{\beta}(\bar{\alpha}(x))) + \lambda\left((1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \eta \ln \bar{N}\right) \\ &\quad = \min_l \left[d(x, \beta_l(\alpha_l(x))) + \lambda\left((1-\eta)\ell_l(\alpha_l(x))\right)\right] \\ &\qquad + \lambda(1-\eta) 1_{W \cap B^c}(x) + \lambda \eta \ln \bar{N}\end{aligned}$$

and hence

$$\begin{aligned}&d(x, \bar{\beta}(\bar{\alpha}(x))) + \lambda\left((1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \eta \ln \bar{N}\right) \\ &\quad \leq d(x, \beta_l(\alpha_l(x))) \\ &\qquad + \lambda\left((1-\eta)[\ell_l(\alpha_l(x)) + 1_{W \cap B^c}(x)] + \eta \ln \bar{N}\right), \\ &\qquad\qquad l = 1, 2. \qquad (82)\end{aligned}$$

In particular, the upper bound for $l = 2$ with (78) and (80) implies

$$\begin{aligned}&d(x, \bar{\beta}(\bar{\alpha}(x))) + \lambda\left((1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \eta \ln \bar{N}\right) \\ &\quad \leq \frac{k}{2}\lambda' + \lambda\left((1-\eta)\left[-\ln p^* + 2 - \frac{k}{2}\ln \lambda'\right] + \eta \ln \bar{N}\right)\end{aligned}$$

and, therefore

$$\begin{aligned}&\frac{d(x, \bar{\beta}(\bar{\alpha}(x)))}{\lambda} + \left((1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \eta \ln \bar{N}\right) + \frac{k}{2}\ln \lambda \\ &\quad \leq \frac{k}{2}\frac{\lambda'}{\lambda} + (1-\eta)\left[-\ln p^* + 2 - \frac{k}{2}\ln \lambda'\right] + \eta \ln \bar{N} + \frac{k}{2}\ln \lambda \\ &\quad \leq \frac{k}{2}\frac{\lambda'}{\lambda} + \frac{k}{2}\ln \lambda + (1-\eta)\left[-\ln p^* + 2\right] + \eta \ln \bar{N}. \qquad (83)\end{aligned}$$

This completes the construction of the modified quantizer. Reintroducing the dependence on M and n, note that the construction depends on M, $g = f_M$, constants γ_M and ϵ_M, and $n \geq n_0(f_M, \epsilon_M)$. We choose γ_M so that it will grow to infinity, but not too fast in the sense that

$$\lim_{M \to \infty} \gamma_M \|f - f_M\|_1 = 0. \qquad (84)$$

Also, we now specify that $\epsilon_M \to 0$ is chosen such that

$$\lim_{M \to \infty} \|f - f_M\|_1 \ln \epsilon_M = 0. \qquad (85)$$

From Lemma 9, these requirements are met if, e.g., we let $\gamma_M = \|f - f_M\|_1^{-1/2}$ and $\epsilon_M = \|f - f_M\|_1$.

Let $q_{M,n}$ denote the original quantizer for f_M with encoder $\alpha_{M,n}$, length function $\ell_{M,n}$, and decoder $\beta_{M,n}$. Denote by $W_{M,n}$ and $B_{M,n}$ the sets used in the construction of the merged quantizers $\bar{q}_{M,n}$ with associated partitions $\bar{\mathcal{S}}_{M,n}$. We also need to control the behavior of the codebook sizes $N_1 = N_1(M,n) = |\mathcal{S}_{M,n}|$, $N_2 = N_2(M,n)$, and $\bar{N} = \bar{N}_{M,n} = |\bar{\mathcal{S}}_{M,n}| = N_1(M,n) + N_2(M,n)$. From (83), we have the upper bound

$$\begin{aligned}&\frac{d(x, \bar{\beta}_{M,n}(\bar{\alpha}_{M,n}(x)))}{\lambda_n} \\ &\quad + \left((1-\eta)\bar{\ell}_{M,n}(\bar{\alpha}_{M,n}(x)) + \eta \ln \bar{N}_{M,n}\right) + \frac{k}{2}\ln \lambda_n \\ &\quad \leq \frac{k}{2}\frac{\lambda_n'}{\lambda_n} + \frac{k}{2}\ln \lambda_n + (1-\eta)\left[-\ln p_{M,n}^* + 2\right] + \eta \ln \bar{N}_{M,n}\end{aligned} \qquad (86)$$

where $p_{M,n}^* \in [\epsilon_M/2, \epsilon_M]$. Equations (75) and (79) yield the following bounds on codebook sizes:

$$N_2(M,n) \leq \gamma_M^{-k/2} \lambda_n^{-k/2} \qquad (87)$$

$$N_1(M,n) \geq c_+^{(M)} \lambda^{-k/2}(1 - \epsilon_M) \qquad (88)$$

$$N_1(M,n) \leq c_-^{(M)}(1 + \epsilon_M)\lambda_n^{-k/2} \qquad (89)$$

$$N_2(M,n)/N_1(M,n) \leq \gamma_M^{-k/2} / \left(c_+^{(M)}(1 - \epsilon_M)\right) \qquad (90)$$

and

$$\hat{c}_M \lambda_n^{-k/2} \leq \bar{N}_{M,n} \leq c_M \lambda_n^{-k/2} \qquad (91)$$

where

$$\begin{aligned}\hat{c}_M &\triangleq \gamma_M^{-k/2} + c_+^{(M)}(1 - \epsilon_M), \\ c_M &\triangleq \gamma_M^{-k/2} + c_-^{(M)}(1 + \epsilon_M).\end{aligned} \qquad (92)$$

From (73), Lemma 9, and the assumed properties of γ_M and ϵ_M

$$\liminf_{M \to \infty} \hat{c}_M \geq e^{\theta_k(\eta) + (1-\eta)\Theta'(u, \eta_-) - k/2} e^{\psi'(f, \eta_+)} \qquad (93)$$

and

$$\limsup_{M \to \infty} c_M \leq e^{\theta_k(\eta) + (1-\eta)\Theta'(u, \eta_-) - k/2} e^{\psi'(f, \eta_-)}. \qquad (94)$$

Incorporating the bound of (91) into (86) yields

$$\begin{aligned}&\frac{d(x, \bar{\beta}_{M,n}(\bar{\alpha}_{M,n}(x)))}{\lambda_n} \\ &\quad + \left((1-\eta)\bar{\ell}_{M,n}(\bar{\alpha}_{M,n}(x)) + \eta \ln \bar{N}_{M,n}\right) + \frac{k}{2}\ln \lambda_n \\ &\quad \leq \frac{k}{2}\gamma_M + (1-\eta)\left[-\ln p_{M,n}^* + 2\right] + \eta \ln c_M. \qquad (95)\end{aligned}$$

The next lemma extends [12, Lemma 12] to the combined constraint case. It is proved in part K of the Appendix. The lemma uses the upper bound to the Lagrangian distortion of the merged quantizers $\bar{q}_{M,n}$ to provide an upper bound to the mismatch resulting from applying the quantizers designed for f_M to f.

Lemma 11: For $n \geq n_0(f_M, \epsilon_M)$, the quantizer $\bar{q}_{M,n}$ satisfies

$$\begin{aligned}&|\theta(f, \lambda_n, \eta, \bar{q}_{M,n}) - \theta(f_M, \lambda_n, \eta, \bar{q}_{M,n})| \\ &\quad \leq \left[\left(\frac{k}{2}\gamma_M + (1-\eta)\left[-\ln \epsilon_M + 2\right] + \eta \ln c_M\right) + \frac{k}{2}\ln \pi\right] \\ &\qquad \times \|f - f_M\|_1 + |h(f) - h(f_M)| + b(M)\end{aligned}$$

where $b(M)$ depends only on f and M, where $\lim_{M\to\infty} b(M) = 0$, and where c_M is defined by (92) and has a finite upper limit, and γ_M is defined in (84).

From the bounds of (82), (75), and (90), we have for $n \ge n_0(f_M, \epsilon_M)$ that

$$\begin{aligned}
&\theta(f_M, \lambda, \eta, \bar{q}_{M,n}) \\
&\quad \le \int \Bigg(\frac{d(x, q_{M,n}(x))}{\lambda} + (1-\eta)\big[\ell_1(\alpha_{M,n}(x)) \\
&\qquad\qquad + 1_{W_{M,n}\cap B^c_{M,n}}(x)\big] + \eta \ln \bar{N}_{M,n} \Bigg) \\
&\qquad \times f_M(x)dx + \frac{k}{2}\ln\lambda \\
&\quad = \theta(f_M, \lambda_n, \eta, q_{M,n}) + P_{f_M}(W_{M,n} \cap B^c_{M,n}) \\
&\qquad - \eta \ln N_1(M,n) + \eta \ln \bar{N}_{M,n} \\
&\quad \le h(f_M, \eta) + 2\epsilon_M + \eta \ln(1 + N_2(M,n)/N_1(M,n)) \\
&\quad \le h(f_M, \eta) + 2\epsilon_M + \eta\gamma_M^{-k/2} \Big/ \left(c_+^{(M)}(1-\epsilon_M)\right).
\end{aligned}$$

Combining this bound with the previous lemma implies that for $n \ge n_0(f_M, \epsilon_M)$

$$\begin{aligned}
&\theta(f, \lambda_n, \eta, \bar{q}_{M,n}) \\
&\quad \le h(f_M, \eta) + 2\epsilon_M + \eta \ln \\
&\qquad \times \left(1 + \gamma_M^{-k/2} \Big/ \left(c_+^{(M)}(1-\epsilon_M)\right)\right) \\
&\qquad + \left[\left(\frac{k}{2}\gamma_M + (1-\eta)\left[-\ln\epsilon_M + 2\right] + \eta \ln c_M\right) + \frac{k}{2}\ln\pi\right] \\
&\qquad \times \|f - f_M\|_1 + |h(f) - h(f_M)| + b(M).
\end{aligned} \tag{96}$$

Given the original $\lambda_n \to 0$ sequence, construct a final sequence of quantizers q_n with partitions $\mathcal{S}_n$ from the sequences $\mathcal{S}_{M,n}$ as follows. First, note that without loss of generality, we can assume that $n_0(f_M, \epsilon_M)$ are strictly increasing in M.

Let

$$M(n) = M, \qquad \text{if } n_0(f_M, \epsilon_M) \le n < n_0(f_{M+1}, \epsilon_{M+1})$$

and define the quantizer q_n as $q_n = \bar{q}_{M(n),n}$. By construction, $M(n)$ is monotone nondecreasing in n and grows without bound and $n \ge n_0(f_{M(n)}, \epsilon_{M(n)})$ once $n \ge n_0(f_1, \epsilon_1)$ (the sequence q_n can be initialized in an arbitrary manner). From (96)

$$\begin{aligned}
&\theta(f, \lambda_n, \eta, q_n) \\
&\quad \le h(f_{M(n)}, \eta) + 2\epsilon_{M(n)} \\
&\qquad + \eta \ln\left(1 + \gamma_{M(n)}^{-k/2} \Big/ \left(c_+^{M(n)}(1-\epsilon_{M(n)})\right)\right) \\
&\qquad + \Bigg[\left(\frac{k}{2}\gamma_{M(n)} + (1-\eta)\left[-\ln\epsilon_{M(n)} + 2\right] + \eta \ln c_{M(n)}\right) \\
&\qquad\qquad + \frac{k}{2}\ln\pi\Bigg] \|f - f_{M(n)}\|_1 \\
&\qquad + |h(f) - h(f_{M(n)})| + b(M(n)).
\end{aligned}$$

Since $M(n) \to \infty$ as $n \to \infty$, in view of Lemma 9, (84), and (85), we have proved the first part of the following lemma. The second part follows from (91), (93), and (94).

Lemma 12: Let X have an absolutely continuous distribution with pdf f, which is zero outside the unit cube C_1 and assume $h(f) > -\infty$. Given $\lambda_n \to 0$, there exists a sequence of partitions $\mathcal{S}_n$ such that $\limsup_{n\to\infty} \theta(f, \lambda_n, \eta, \mathcal{S}_n) \le \Theta(u, 0) + h(f, \eta)$ where $h(f, \eta) = \psi(f, \eta) + \eta(1-\eta)\,(\Theta'(u, \eta_-) - \Theta'(u, \eta_+))$ where $\psi(f, \eta) = \inf_\nu \psi(f, \eta, \nu)$ and $\psi(f, \eta, \nu)$ is defined in (31). Furthermore, for all except possibly a countable number of $\eta \in (0,1)$ $h(f, \eta) = \psi(f, \eta)$. The sizes of the codebooks satisfy

$$\begin{aligned}
&e^{\theta_k(\eta) + (1-\eta)\Theta'(u,\eta_-) - k/2} e^{\psi'(f,\eta_-)} \\
&\quad \ge \limsup_{n\to\infty} |\mathcal{S}_n| \lambda_n^{k/2} \\
&\quad \ge \liminf_{n\to\infty} |\mathcal{S}_n| \lambda_n^{k/2} \\
&\quad \ge e^{\theta_k(\eta) + (1-\eta)\Theta'(u,\eta_+) - k/2} e^{\psi'(f,\eta_+)}.
\end{aligned} \tag{97}$$

Since $\psi(f, \eta) = \phi(f, \eta)$, the Gersho functional describes the asymptotic behavior of the constructed sequence of quantizers.

The lemma implies the following corollary.

Corollary 8: Given the assumptions and definitions of the previous lemma

$$\overline{\theta}(f, \eta) \le \Theta(u, \eta) + h(f, \eta).$$

The development for general densities on a cube was done for the unit cube to keep things simple. Lemma 7 can be used to extend the result to a cube of arbitrary size.

Corollary 9: The results of Lemma 12 hold if the unit cube C_1 of the assumptions is replaced by the cube $C_{a,r}$ for finite $a > 0$ and r.

Proof: From Lemma 7, for any quantizer $q_{a,r}$ on $C_{a,r}$, there is an equivalent quantizer q_1 on C_1 with performance related by $\theta(f_{a,r}, \lambda, \eta, q_{a,r}) = \theta\left(f_1, a^{-2}\lambda, \eta, q_1\right) + k \ln a$. Furthermore, defining $\Lambda_{a,r}(x) = a^{-k}\Lambda_1((x-r)/a)$ yields by a change of variables $\phi(f_{a,r}, \eta, \Lambda_{a,r}) = \phi(f_1, \eta, \Lambda_1) + k \ln a$ so that $\phi(f_{a,r}, \eta) = \phi(f_1, \eta) + k \ln a$ and hence $h(f_{a,r}, \eta) = h(f_1, \eta) + k \ln a$. Thus, the results of Lemma 12 hold with the transformed quantizers and the addition of the scaling term $k \ln a$ to both target performance and actual performance. □

F. General Densities

Assume that X has a pdf satisfying the conditions of Theorem 2, which ensures that $h(f)$ is finite. Our proof for this case draws on results and bounds from Graf and Luschgy [11] for the fixed-rate case. With a slight change of notation, denote for any integer M the cube $C_M = [-M, M]^k$ of side $2M$ centered at the origin. Corollary 9 and Lemma 12 imply the existence of quantizers satisfying the bounds of Lemma 12 for the induced absolutely continuous distributions on C_M. Let f_M denote the conditional pdf given C_M, f_M^c the conditional pdf given C_{M^c}, and $p_M = \Pr(X \in C_M) = \int_{C_M} f(x)dx$ so that we have the disjoint mixture $f(x) = f_M(x)p_M + f_M^c(x)(1-p_M)$, where $p_M \to 1$ as $M \to \infty$. The following convergence properties are proved in part L of the Appendix.

Lemma 13: Given the previous definitions, $\lim_{M\to\infty} h(f_M) = h(f)$ and $\lim_{M\to\infty} h(f_M, \eta) = h(f, \eta)$.

For any fixed M and positive sequence $\lambda_n \to 0$, let $\mathcal{S}_n$ be a sequence of partitions of C_M and q_n the corresponding quantizers satisfying the properties of Lemma 12 (from Corollary 9) for f_M. In particular

$$\limsup_{n\to\infty} \theta(f_M, \lambda_n, \eta, \mathcal{S}_n) \le \Theta(u, 0) + h(f_M, \eta) \tag{98}$$

and

$$c_+^{(M)} \le \liminf_{n\to\infty} |\mathcal{S}_n| \lambda_n^{k/2} \le \limsup_{n\to\infty} |\mathcal{S}_n| \lambda_n^{k/2} \le c_-^{(M)} \tag{99}$$

where $c_+^{(M)}$ and $c_-^{(M)}$ are finite constants, which depend only on f_M; see (97).

As in the development of Lemma 12, we modify these quantizers by merging them with other quantizers, but this time the second quantizers will be simpler than in the previous case because the two quantizers partition disjoint regions and can be handled separately.

Pick $\epsilon \in (0, 1/2)$ and construct a quantizer for C_M^c using partition $\hat{\mathcal{S}}_n$ with codebook size $\hat{N}_n = |\hat{\mathcal{S}}_n|$ constrained as

$$\frac{\epsilon}{1-\epsilon}|\mathcal{S}_n| - 1 \le \hat{N}_n = \left\lfloor \frac{\epsilon}{1-\epsilon}|\mathcal{S}_n| \right\rfloor \le \frac{\epsilon}{1-\epsilon}|\mathcal{S}_n| \tag{100}$$

so that if ϵ is small, then $\hat{\mathcal{S}}_n$ has only a small fraction of the number of codewords given to f_M. Choose $\hat{\mathcal{S}}_n$ to achieve $\delta_1(f_M^c, \ln \hat{N}_n)$, that is, we overbound the performance by considering an optimal fixed-rate quantizer on C_M^c.

Form the composite quantizer for $\Re^k$ as the partition $\bar{\mathcal{S}}_n$ consisting of the union of the atoms of $\mathcal{S}_n$ and $\hat{\mathcal{S}}_n$. The resulting performance will be

$$\begin{aligned}
&\theta(f, \lambda_n, \eta, \bar{\mathcal{S}}_n) \\
&\quad = \frac{D_f(\bar{\mathcal{S}}_n)}{\lambda_n} + (1-\eta) H_f(\bar{\mathcal{S}}_n) + \eta \ln |\bar{\mathcal{S}}_n| + \frac{k}{2}\ln \lambda_n \\
&\quad = (1-p_M)\frac{D_{f_M}(\mathcal{S}_n)}{\lambda_n} + p_M \frac{D_{f_M^c}(\hat{\mathcal{S}}_n)}{\lambda_n} + (1-\eta) \\
&\qquad \times \left[(1-p_M) H_{f_M}(\mathcal{S}_n) + p_M H_{f_M^c}(\hat{\mathcal{S}}_n) + h_2(p_M)\right] \\
&\qquad + \eta \ln\left(|\mathcal{S}_n| + |\hat{\mathcal{S}}_n|\right) + \frac{k}{2}\ln\lambda_n
\end{aligned}$$

where $h_2(p) \triangleq -p\ln p - (1-p)\ln(1-p)$. Regrouping terms

$$\begin{aligned}
&\theta(f, \lambda_n, \eta, \bar{\mathcal{S}}_n) \\
&\quad = (1-p_M)\Bigg[\frac{D_{f_M}(\mathcal{S}_n)}{\lambda_n} + (1-\eta) H_{f_M}(\mathcal{S}_n) \\
&\qquad\qquad + \eta \ln |\mathcal{S}_n| + \frac{k}{2}\ln \lambda_n\Bigg] \\
&\qquad + p_M \Bigg[\frac{D_{f_M^c}(\hat{\mathcal{S}}_n)}{\lambda_n} + (1-\eta) H_{f_M^c}(\hat{\mathcal{S}}_n) \\
&\qquad\qquad + \eta \ln |\hat{\mathcal{S}}_n| + \frac{k}{2}\ln\lambda_n\Bigg] + h_2(p_M) \\
&\qquad + \eta\left[\ln\left(|\mathcal{S}_n| + |\hat{\mathcal{S}}_n|\right) - (1-p_M)\ln|\mathcal{S}_n| - p_M \ln |\hat{\mathcal{S}}_n|\right] \\
&\quad = (1-p_M)\theta(f_M, \lambda_n, \eta, \mathcal{S}_n) + p_M \theta(f_M^c, \lambda_n, \eta, \hat{\mathcal{S}}_n) \\
&\qquad + h_2(p_M) + \eta\left[\ln\left(1 + \frac{|\hat{\mathcal{S}}_n|}{|\mathcal{S}_n|}\right) + p_M \ln \frac{|\mathcal{S}_n|}{|\hat{\mathcal{S}}_n|}\right] \\
&\quad \le (1-p_M)\theta(f_M, \lambda_n, \eta, \mathcal{S}_n) + p_M \theta(f_M^c, \lambda_n, 1, \hat{\mathcal{S}}_n) \\
&\qquad + h_2(p_M) + \eta\left[\ln\frac{1}{1-\epsilon} + p_M \ln\frac{2(1-\epsilon)}{\epsilon}\right]
\end{aligned} \tag{101}$$

where in the last inequality we used the fact that from (100) we have $\hat{N}_n \ge \frac{\epsilon}{2(1-\epsilon)}|\mathcal{S}_n|$ if $|\mathcal{S}_n| \ge 2(1-\epsilon)/\epsilon$ (i.e., for all n large enough since $\lim_{n\to\infty} |\mathcal{S}_n| = \infty$). To bound $\theta(f_M^c, \lambda_n, 1, \hat{\mathcal{S}}_n)$, we use Corollary 6.7 of Graf and Luschgy [11], which in our notation becomes $\delta_1(f_M^c, \ln \hat{N}_n) \le \hat{N}_n^{-2/k}\left(C_1 E_{f_M^c}(\|X\|^{2+\delta}) + C_2\right)$ for $\hat{N}_n \ge C_3$, where C_1, C_2, and C_3 depend only on δ and k, but not on f_M^c. Using (99) to bound relate $\hat{N}_n$ and $|\mathcal{S}_n|$, we have that

$$\begin{aligned}
&\theta(f_M^c, \lambda_n, 1, \hat{\mathcal{S}}_n) \\
&\quad = \frac{\delta_1(f_M^c, \ln N_n)}{\lambda_n} + \ln |\hat{\mathcal{S}}_n| + \frac{k}{2}\ln\lambda_n \\
&\quad \le \frac{\hat{N}_n^{-2/k}\left(C_1 E_{f_M^c}(\|X\|^{2+\delta}) + C_2\right)}{\lambda_n} + \ln|\mathcal{S}_n| + \frac{k}{2}\ln\lambda_n \\
&\quad = \left(\frac{2(1-\epsilon)}{\epsilon}\right)^{\frac{2}{k}} \left(C_1 E_{f_M^c}(\|X\|^{2+\delta}) + C_2\right) \\
&\qquad \times \left(|\mathcal{S}_n|\lambda_n^{k/2}\right)^{-2/k} + \ln|\mathcal{S}_n|\lambda_n^{k/2}.
\end{aligned}$$

Invoking (99) results in $\limsup_{n\to\infty}\theta(f_M^c, \lambda_n, 1, \hat{\mathcal{S}}_n) \le c(M, \epsilon)$, where $\lim_{M\to\infty} p_M c(M, \epsilon) = 0$ since $p_M E_{f_M^c}(\|X\|^{2+\delta}) = \int_{C_M^c} \|x\|^{2+\delta} f(x)dx \to 0$ as $M \to \infty$. Combining this with (101) and (98) shows that

$$\begin{aligned}
&\limsup_{n\to\infty}\theta(f, \lambda_n, \eta, \bar{\mathcal{S}}_n) \\
&\quad \le (1-p_M)(\Theta(u,0) + h(f_M, \eta)) + p_M c(M, \epsilon) + h(p_M) \\
&\qquad + \eta\left[\ln\frac{1}{1-\epsilon} + p_M\ln\frac{2(1-\epsilon)}{\epsilon}\right].
\end{aligned}$$

Since $p_M \to 0$, $p_M c(M, \epsilon) \to 0$, and $h(f_M, \eta) \to h(f, \eta)$ as $M \to \infty$ (by Lemma 13), the right-hand side can be made to be arbitrarily close to $\Theta(u, \eta) + h(f, \eta)$ by choosing first ϵ small enough and then M large enough. This proves that

$$\limsup_{\lambda\to 0}\theta(f, \lambda, \eta) \le \Theta(u, \eta) + h(f, \eta)$$

which completes the proof of Theorem 2 for the general case of pdfs satisfying the assumptions of the theorem.

G. *Proof of Lemma 7*

Let $\mathcal{S}_{a,r}$ and $\mathcal{S}_1$ denote the partitions corresponding to $\alpha_{a,r}$ and α_1, respectively. Then, $\Pr(\alpha_{a,r}(Y) = i) = \Pr(\alpha_{a,r}(aX + r) = i) = \Pr(\alpha_1(X) = i)$, and hence, $H_{f_{a,r}}(\mathcal{S}_{a,r}) = H_{f_1}(\mathcal{S}_1)$. The number of codewords for the two quantizers is identical by construction. A change of variables $y = ax + r$ yields average distortion

$$\int \|y - q_{a,r}(y)\|^2 f_{a,r}(y)\,dy = a^2 \int \|x - q_1(x)\|^2 f_1(x)dx$$

so that

$$\begin{aligned}
&\theta(f_{a,r}, \lambda, \eta, q_{a,r}) \\
&\quad = \frac{a^2}{\lambda} D_{f_1}(q_1) + \frac{k}{2}\ln\lambda + [(1-\eta)H_{f_1}(q_1) + \eta\ln N(q_1)]
\end{aligned}$$

which proves the first part of the lemma. The second follows by taking the infimum over q_1. □

H. Proof of Lemma 8

From Lemma 6, we can construct for $\lambda_n \to 0$ for each cube $C(m)$, for which $w_m > 0$ a sequence of partitions $\mathcal{S}_n^{(m)}$, for which

$$\lim_{n\to\infty} \frac{2D_{f^{(m)}}(\mathcal{S}_n^{(m)})}{k\mu_m\lambda_n} = 1 \tag{102}$$

$$\begin{aligned}&\lim_{n\to\infty}\left((1-\eta)H_{f^{(m)}}(\mathcal{S}_n^{(m)}) + \eta\ln|\mathcal{S}_n^{(m)}| + \frac{k}{2}\ln\mu_m\lambda_n\right)\\&\quad = \Theta(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e}\\&\Theta'(u,\eta_-)\\&\quad \geq \limsup_{n\to\infty}\left(\ln|\mathcal{S}_n^{(m)}| - H_{f^{(m)}}(\mathcal{S}_n^{(m)})\right)\\&\quad \geq \liminf_{n\to\infty}\left(\ln|\mathcal{S}_n^{(m)}| - H_{f^{(m)}}(\mathcal{S}_n^{(m)})\right) \geq \Theta'(u,\eta_+)\\&\Theta(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e} - \eta\Theta'(u,\eta_+)\\&\quad \geq \limsup_{n\to\infty} H_{f^{(m)}}(\mathcal{S}_n^{(m)}) + \frac{k}{2}\ln\mu_m\lambda_n\\&\quad \geq \liminf_{n\to\infty} H_{f^{(m)}}(\mathcal{S}_n^{(m)}) + \frac{k}{2}\ln\mu_m\lambda_n\\&\quad \geq \Theta(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e} - \eta\Theta'(u,\eta_-)\\&\Theta(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e} + (1-\eta)\Theta'(u,\eta_-)\\&\quad \geq \limsup_{n\to\infty}\ln\left|\mathcal{S}_n^{(m)}\right| + \frac{k}{2}\ln\mu_m\lambda_n\\&\quad \geq \liminf_{n\to\infty}\ln\left|\mathcal{S}_n^{(m)}\right| + \frac{k}{2}\ln\mu_m\lambda_n\\&\quad \geq \Theta(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e} + (1-\eta)\Theta'(u,\eta).\end{aligned} \tag{103}$$

The final equation is equivalent to

$$\begin{aligned}&\mu_m^{-k/2}a^k e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_-)-k/2}\\&\quad \geq \limsup_{n\to\infty}\left|\mathcal{S}_n^{(m)}\right|\lambda_n^{k/2}\\&\quad \geq \liminf_{n\to\infty}\left|\mathcal{S}_n^{(m)}\right|\lambda_n^{k/2}\\&\quad \geq \mu_m^{-k/2}a^k e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_+)-k/2}.\end{aligned} \tag{104}$$

The composite partition $\mathcal{S}_n$ is defined as the partition having as atoms all of the atoms of the individual cell partitions $\mathcal{S}_n^{(m)}$

$$\begin{aligned}\theta(f,\lambda_n,\eta,\mathcal{S}_n) &= \frac{D_f(\mathcal{S}_n)}{\lambda_n} + (1-\eta)H_f(\mathcal{S}_n) + \eta\ln|\mathcal{S}_n| + \frac{k}{2}\ln\lambda_n\\&= \frac{\sum_m w_m D_{f^{(m)}}(\mathcal{S}_n^{(m)})}{\lambda_n}\\&\quad + (1-\eta)\left(\sum_m w_m H_{f^{(m)}}(\mathcal{S}_n^{(m)}) + \frac{k}{2}\ln\lambda_n\right)\\&\quad + \eta\ln\left(\lambda_n^{k/2}\sum_m|\mathcal{S}_n^{(m)}|\right) + (1-\eta)H(w).\end{aligned} \tag{105}$$

From (102)

$$\lim_{n\to\infty}\frac{\sum_m w_m D_{f^{(m)}}(\mathcal{S}_n^{(m)})}{\lambda_n} = \frac{k}{2}\sum_m w_m\mu_m. \tag{106}$$

From (103)

$$\begin{aligned}&\Theta(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e} - \eta\Theta'(u,\eta_+) + \sum_m w_m\ln\mu_m^{-k/2}\\&\quad \geq \limsup_{n\to\infty}\left(\sum_m w_m H_{f^{(m)}}(\mathcal{S}_n^{(m)}) + \frac{k}{2}\ln\lambda_n\right)\\&\quad \geq \liminf_{n\to\infty}\left(\sum_m w_m H_{f^{(m)}}(\mathcal{S}_n^{(m)}) + \frac{k}{2}\ln\mu_m\lambda_n\right)\\&\quad \geq \Theta(u,\eta) + \frac{k}{2}\ln\frac{a^2}{e} - \eta\Theta'(u,\eta_-) + \sum_m w_m\ln\mu_m^{-k/2}\end{aligned} \tag{107}$$

and from (104)

$$\begin{aligned}&a^k e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_-)-k/2}\sum_m\mu_m^{-k/2}\\&\quad \geq \limsup_{n\to\infty}\sum_m|\mathcal{S}_n^{(m)}|\lambda_n^{k/2}\\&\quad \geq \liminf_{n\to\infty}\sum_m|\mathcal{S}_n^{(m)}|\lambda_n^{k/2}\\&\quad \geq a^k e^{\Theta(u,\eta)+(1-\eta)\Theta'(u,\eta_+)-k/2}\sum_m\mu_m^{-k/2}\end{aligned}$$

so that

$$\begin{aligned}&k\ln a - \frac{k}{2} + \Theta(u,\eta) + (1-\eta)\Theta'(u,\eta_-) + \ln\left(\sum_m\mu_m^{-k/2}\right)\\&\quad \geq \limsup_{n\to\infty}\ln\left(\sum_m|\mathcal{S}_n^{(m)}|\lambda_n^{k/2}\right)\\&\quad \geq \liminf_{n\to\infty}\ln\left(\sum_m|\mathcal{S}_n^{(m)}|\lambda_n^{k/2}\right)\\&\quad \geq k\ln a - \frac{k}{2} + \Theta(u,\eta) + (1-\eta)\Theta'(u,\eta_+) + \ln\left(\sum_m\mu_m^{-k/2}\right).\end{aligned} \tag{108}$$

The inequalities of the lemma follow from (105) and the sum of (106), $(1-\eta)$ times (107), and η times (108). The equality follows from the inequalities and Lemma 4 □

I. Proof of Lemma 9

The first two statements (67) and (68) are [12, Lemma 10], and (70) follows from (69), which we now prove. Keep in mind that all integrals in the proof are over C_1. In particular, we write $\int e^{-k\nu(x)/2}dx$ for $\int_{C_1} e^{-k\nu(x)(x)/2}dx$.

Suppose that $\nu^{(M)}$ is the minimizer for the piecewise constant pdf f_M satisfying $\psi\left(f_M,\eta,\nu^{(M)}\right) = \psi(f_M,\eta)$. Then, applying $\nu^{(M)}$ to the actual pdf f

$$\psi(f,\eta) \leq \psi(f,\eta,\nu^{(M)}) = \psi(f_M,\eta) - (1-\eta)\left[h(f_M) - h(f)\right]$$

and hence, from the second statement of the lemma, $\liminf_{M\to\infty}\psi(f_M,\eta) \geq \psi(f,\eta)$.

Conversely, suppose that ν yields a value of $\psi(f,\eta,\nu)$ within $\epsilon > 0$ of the infimum so that $\psi(f,\eta,\nu) \le \psi(f,\eta)+\epsilon$. Let $L > 0$ and define

$$\nu_L(x) = \begin{cases} \nu(x), & \text{if } |\nu(x)| \le L \\ -L, & \text{if } \nu(x) < -L \\ L, & \text{if } \nu(x) > L. \end{cases}$$

We have $|\nu_L(x)|f(x) \le Lf(x)$, $e^{\nu_L(x)}f(x) \le \max(e^{\nu(x)},1)f(x)$, and $e^{-k\nu_L(x)/2} \le \max(e^{-k\nu(x)/2},1)$. Since each upper bound is integrable on C_1 and $\lim_{L\to\infty}\nu_L(x) = \nu(x)$ for all x, the dominated convergence theorem implies $\lim_{L\to\infty}\int e^{\nu_L(x)}f(x)dx = \int e^{\nu(x)}f(x)dx$, $\lim_{L\to\infty}\int \nu_L(x)f(x)dx = \int \nu(x)f(x)dx$, and $\lim_{L\to\infty}\int e^{-k\nu_L(x)/2}dx = \int e^{-k\nu(x)/2}dx$. Thus, from the definition (31), $\lim_{L\to\infty}\psi(f,\eta,\nu_L) = \psi(f,\eta,\nu)$, and hence we can choose L such that $\psi(f,\eta,\nu_L) < \psi(f,\eta,\nu)+\epsilon < \psi(f,\eta)+2\epsilon$. Since $e^{\nu_L(x)} \le e^L$ and $|\nu_L(x)| \le L$, we have

$$\begin{aligned}
&|\psi(f,\eta,\nu_L) - \psi(f_M,\eta,\nu_L)| \\
&= \Bigg|\frac{k}{2}\left(\int e^{\nu_L(x)}f(x)dx - (1-\eta)\int \nu_L(x)f(x)dx\right) \\
&\qquad + (1-\eta)h(f) \\
&\qquad - \frac{k}{2}\left(\int e^{\nu_L(x)}f_M(x)dx - (1-\eta)\int \nu_L(x)f_M(x)dx\right) \\
&\qquad + (1-\eta)h(f_M)\Bigg| \\
&\le C_L\|f_M - f\|_1 + (1-\eta)|h(f_M) - h(f)|
\end{aligned}$$

for some constant C_L depending on L only. Hence, from the first and second statements of the lemma

$$\begin{aligned}
\liminf_{M\to\infty}\psi(f_M,\eta) &= \liminf_{M\to\infty}\inf_{\nu}\psi(f_M,\eta,\nu) \\
&\le \liminf_{M\to\infty}\psi(f_M,\eta,\nu_L) \\
&= \psi(f,\eta,\nu_L) \le \psi(f_M,\eta) + 2\epsilon
\end{aligned}$$

which proves the converse.

To prove (71), from the properties of left and right derivatives

$$\begin{aligned}
\psi'(f,\eta_-) &= \inf_{\Delta\eta>0}\lim_{M\to\infty}\frac{\psi(f_M,\eta)-\psi(f_M,\eta-\Delta\eta)}{\Delta\eta} \\
&\ge \limsup_{M\to\infty}\inf_{\Delta\eta>0}\frac{\psi(f_M,\eta)-\psi(f_M,\eta-\Delta\eta)}{\Delta\eta} \\
&= \limsup_{M\to\infty}\psi(f_M,\eta_-) \\
\psi'(f,\eta_+) &= \sup_{\Delta\eta>0}\lim_{M\to\infty}\frac{\psi(f_M,\eta+\Delta\eta)-\psi(f_M,\eta)}{\Delta\eta} \\
&\le \liminf_{M\to\infty}\sup_{\Delta\eta>0}\frac{\psi(f_M,\eta+\Delta\eta)-\psi(f_M,\eta)}{\Delta\eta} \\
&= \liminf_{M\to\infty}\psi(f_M,\eta_+).
\end{aligned}$$

□

J. Proof of Lemma 10

Let $S_{n,i}$ denote the ith atom of $\mathcal{S}_n$ and $y_{n,i}$ the associated reproduction codeword. First, we show that

$$\lim_{\lambda\to 0}\max_i P_g(S_{n,i}) = 0. \tag{109}$$

The assumptions of the lemma imply from Lemma 5 that $\lim_{n\to\infty}D_g(\mathcal{S}_n) = 0$. Define $d_g(S_{n,i}) = \int_{S_{n,i}}\|x - y_{n,i}\|^2 g(x)dx$. Fix $c > 0$ and let $A_c = \{x : g(x) > c\}$. Then

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

where G_k is the normalized second moment of a k-dimensional sphere (see, e.g., [16]). Since $D_g(\mathcal{S}_n) = \sum_i d_g(S_{n,i})$, this and the fact that $\lim_{n\to\infty}D_g(\mathcal{S}_n) = 0$ imply $\lim_{n\to\infty}\max_i V(S_{n,i}\cap A_c) = 0$, from which it follows that $\lim_{n\to\infty}\max_i P_g(S_{n,i}\cap A_c) = 0$ by the absolute continuity of P_g with respect to the Lebesgue measure (see, e.g., [2]). Note that $P_g(S_{n,i}) \le P_g(S_{n,i}\cap A_c) + P_g\left(\Re^k\setminus A_c\right)$ and so $\lim_{n\to\infty}\max_i P_g(S_{n,i}) \le P_g\left(\Re^k\setminus A_c\right)$. Since $\lim_{c\to 0}P_g\left(\Re^k\setminus A_c\right) = \lim_{c\to 0}P_g(\{x : g(x) \le c\}) = 0$, (109) follows.

The statement of the lemma follows by noticing that if $\max_i P_g(S_{n,i}) < \epsilon/2$, then there must exist a collection of partition cells with total probability between $\epsilon/2$ and ϵ. □

K. Proof of Lemma 11

The proof is very similar to [12, proof of Lemma 12] and hence only details that are distinct from those in [12] are given. For simplicity, we suppress the subscripts M, n in the quantizer, encoder, and decoder and the subscript n in λ_n.

By definition

$$\begin{aligned}
&|\theta(f,\lambda,\eta,\bar{q},\bar{\ell}) - \theta(f_M,\lambda,\eta,\bar{q},\bar{\ell})| \\
&= \left|\int\left[\frac{d(x,\bar{\beta}(\bar{\alpha}(x)))}{\lambda}+(1-\eta)\bar{\ell}(\bar{\alpha}(x))+\eta\ln\bar{N}+\frac{k}{2}\ln\lambda\right]\right. \\
&\qquad \left.\times\left[f(x)-f_M(x)\right]dx\right|.
\end{aligned}$$

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

$$y = y^+ - y^-, \quad |y| = y^+ + y^-. \tag{110}$$

Then, (111), shown at the bottom of the next page, holds. The pointwise upper bound (95) implies

$$\begin{aligned}
&\int\left[\frac{d(x,\bar{\beta}(\bar{\alpha}(x)))}{\lambda} + (1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \eta\ln\bar{N} + \frac{k}{2}\ln\lambda\right] \\
&\quad\times\left[f(x)-f_M(x)\right]^+dx \\
&\qquad\le\left(\frac{k}{2}\gamma + (1-\eta)\left[-\ln p^* + 2\right] + \eta\ln C_M\right) \\
&\qquad\quad\times\int\left[f(x)-f_M(x)\right]^+dx.
\end{aligned}\tag{112}$$

Note that (110) and the fact that $\int[f(x)-f_M(x)]dx = 0$ imply

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

and so the function $F_M(x) = [f(x) - f_M(x)]^+/\frac{1}{2}\|f - f_M\|_1$ is a pdf. Thus

$$\begin{aligned}\int & \left[\frac{d(x,\bar{\beta}(\bar{\alpha}(x)))}{\lambda} + (1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \eta \ln \bar{N} + \frac{k}{2}\ln\lambda\right] \\ &\times \left[f(x) - f_M(x)\right]^+ dx \\ &= \left(\theta(F_M, \lambda, \eta, \bar{q}, \bar{\ell})\right)\frac{1}{2}\|f - f_M\|_1 \\ &\geq \left(\theta(F_M, \lambda, \eta, \bar{q})\right)\frac{1}{2}\|f - f_M\|_1 \\ &\geq \left(-\frac{k}{2}\ln\pi + h(F_M)\right)\frac{1}{2}\|f - f_M\|_1 \end{aligned} \tag{113}$$

where in the last step we used Lemma 4 to infer that $\theta(F_M, \lambda, \eta, \bar{q}) \geq \theta(F_M, \lambda, 0, \bar{q})$ together with the bound of [12, Lemma 3]. From the proof of Lemma 12 in [12]

$$\lim_{M\to\infty} \frac{1}{2}\|f - f_M\|_1 h(F_M) = 0. \tag{114}$$

Letting $b_1(M) = \frac{1}{2}\|f - f_M\|_1 |h(F_M)|$ and combining the upper bound of (112) with the lower bound of (113), we obtain

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

where $b_1(M) \to 0$ as $M \to \infty$ by (114). A similar argument shows that

$$\begin{aligned}&\left|\int \left[\frac{d(x,\bar{\beta}(\bar{\alpha}(x)))}{\lambda} + (1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \frac{k}{2}\ln\lambda\right]\right. \\ &\qquad \left.\times \left[f(x) - f_M(x)\right]^- dx\right| \\ &\leq \left[\frac{k}{2}\gamma + (1-\eta)\left[-\ln p^* + 2\right] + \eta\ln C_M + \frac{k}{2}\ln\pi\right] \\ &\qquad \times \frac{1}{2}\|f - f_M\|_1 + b_2(M)\end{aligned}$$

where $b_2(M) \to 0$ as $M \to \infty$. Let $b(M) = b_1(M) + b_2(M)$ and combine these bounds with (110) and (111) to obtain the bound of the lemma. □

L. Proof of Lemma 13

To prove the first claim of the lemma, note that

$$\int_{C_M} f(x)\ln\frac{1}{f(x)}dx = (1-p_M)h(f_M) + (1-p_M)\ln\frac{1}{1-p_M}.$$

Since $-f(x)\ln f(x)$ is integrable by assumption, $C_M \subset C_{M+1}$ and $\bigcup_{M\geq 1} C_M = \Re^k$, the integral on the left converges to $h(f)$ as $M \to \infty$. This, together with $\lim_{M\to\infty} p_M = 0$, implies $\lim_{M\to\infty} h(f_M) = h(f)$, which is the first statement of the lemma.

For the next part, recall that $h(f_M, \eta) = \phi(f_M, \eta) + \eta(1-\eta)\left(\Theta'(u, \eta_-) - \Theta'(u, \eta_+)\right)$, where $\phi(f_M, \eta) = \inf_\Lambda \phi(f_M, \eta, \Lambda)$ and from (18)

$$\begin{aligned}\phi(f_M, \eta, \Lambda) = \frac{k}{2}\ln\int_{C_M} & f_M(x)\Lambda(x)^{-2/k}dx + (1-\eta) \\ &\times \int_{C_M} f_M(x)\ln\Lambda(x)dx + (1-\eta)h(f_M)\end{aligned}$$

for any pdf Λ, for which the integrals are finite. Thus, we need to prove that $\lim_{M\to\infty}\phi(f_M, \eta) = \phi(f, \eta)$.

Suppose that Λ_M is approximately optimal for f_M in the sense that $\phi(f_M, \eta, \Lambda) \leq \inf_\Lambda \phi(f_M, \eta, \Lambda) + \epsilon$. The density function Λ_M can be thought of as a conditional density function on C_M. Following (26) define $\tilde{\Lambda}(x) = f(x)^{k/(k+2)}/\int f(y)^{k/(k+2)}dy$, which is the optimal Λ for the full pdf f in the fixed-rate case. Define a point density function (pdf) $\hat{\Lambda}_M$ on $\Re^k$ by $\hat{\Lambda}_M(x) = 1_{C_M}(x)\Lambda_M(x)\pi_M + 1_{C_M^c}(x)\tilde{\Lambda}(x)$, where $\pi_M = \int_{C_M}\tilde{\Lambda}(x)dx$ so that

$$\begin{aligned}&\phi(f, \hat{\Lambda}_M, \eta) \\ &\quad = \frac{k}{2}\ln\left(\int_{C_M} f(x)\hat{\Lambda}_M(x)^{-2/k}dx + \int_{C_M^c} f(x)\hat{\Lambda}_M(x)^{-2/k}dx\right) \\ &\qquad + (1-\eta)\left(\int_{C_M} f(x)\ln\hat{\Lambda}_M(x)dx\right. \\ &\qquad\qquad \left. + \int_{C_M^c} f(x)\ln\hat{\Lambda}_M(x)dx\right) + (1-\eta)h(f).\end{aligned} \tag{115}$$

$$\begin{aligned}&\left|\int\left[\frac{d(x,\bar{\beta}(\bar{\alpha}(x)))}{\lambda} + (1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \eta\ln\bar{N} + \frac{k}{2}\ln\lambda\right]\left[f(x) - f_M(x)\right]dx\right| \\ &= \left|\int\left[\frac{d(x,\bar{\beta}(\bar{\alpha}(x)))}{\lambda} + (1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \eta\ln\bar{N} + \frac{k}{2}\ln\lambda\right]\left[f(x) - f_M(x)\right]^+ dx\right. \\ &\qquad \left. - \int\left[\frac{d(x,\bar{\beta}(\bar{\alpha}(x)))}{\lambda} + (1-\eta)\bar{\ell}(\bar{\alpha}(x)) + \eta\ln\bar{N} + \frac{k}{2}\ln\lambda\right]\left[f(x) - f_M(x)\right]^- dx\right|.\end{aligned} \tag{111}$$

$$
\begin{aligned}
&\ln\left(\int_{C_M} f(x)\hat{\Lambda}_M(x)^{-2/k}dx + \int_{C_M^c} f(x)\hat{\Lambda}_M(x)^{-2/k}dx\right)\\
&\quad = \ln\left((1-p_M)\pi_M^{-2/k}\int_{C_M} f_M(x)\Lambda_M(x)^{-2/k}dx + \int_{C_M^c} f(x)\tilde{\Lambda}(x)^{-2/k}dx\right)\\
&\quad = \ln\left(\int_{C_M} f_M(x)\Lambda_M(x)^{-2/k}dx\right) + \ln\left((1-p_M)\pi_M^{-2/k} + \frac{\int_{C_M^c} f(x)\tilde{\Lambda}(x)^{-2/k}dx}{\int_{C_M} f_M(x)\Lambda_M(x)^{-2/k}dx}\right)\\
&\quad = \ln\left(\int f_M(x)\Lambda_M(x)^{-2/k}dx\right) + \epsilon_M
\end{aligned}
$$

Then, from the definition of $\hat{\Lambda}_M$, the equation shown at the top of the page holds, where $\lim_{M\to\infty}\epsilon_M = 0$ as $M\to\infty$, since $\lim_{M\to\infty} p_M = 1$, $\lim_{M\to\infty}\pi_M = 1$, $\lim_{M\to\infty}\int_{C_M^c} f(x)\tilde{\Lambda}(x)^{-2/k}dx = 0$, and

$$
\begin{aligned}
\int_{C_M} f_M(x)\Lambda_M(x)^{-2/k}dx &\geq e^{\|f_M\|_{k/(k+2)}}\\
&\to e^{\|f\|_{k/(k+2)}} \quad \text{as } M\to\infty.
\end{aligned}
$$

Furthermore

$$
\begin{aligned}
&\int_{C_M} f(x)\ln\hat{\Lambda}_M(x)dx + \int_{C_M^c} f(x)\ln\hat{\Lambda}_M(x)dx\\
&\quad = (1-p_M)\int f_M(x)\ln\Lambda_M(x)dx + (1-p_M)\ln\pi_M\\
&\qquad + \int_{C_M^c} f(x)\ln\tilde{\Lambda}(x)dx\\
&\quad = (1-p_M)\int f_M(x)\ln\Lambda_M(x)dx + \epsilon'_M \qquad (116)
\end{aligned}
$$

where $\lim_{M\to\infty}\epsilon'_M = 0$ since $f(x)\ln\tilde{\Lambda}(x)$ is integrable, which ensures that $\int_{C_M^c} f(x)\ln\tilde{\Lambda}(x)dx$ goes to zero as $M\to\infty$.

Combining (115) and (116) with the first statement of the lemma, for all large enough M

$$
\begin{aligned}
\inf_{\Lambda}\phi(f,\eta,\Lambda) &\leq \phi(f,\eta,\hat{\Lambda}_M)\\
&\leq \phi(f_M,\eta,\Lambda_M) + \epsilon\\
&\leq \inf_{\Lambda}\phi(f_M,\eta,\Lambda) + 2\epsilon
\end{aligned}
$$

where the last inequality follows since Λ_M is asymptotically optimal. This proves that $\liminf_{M\to\infty}\phi(f_M,\eta) \geq \phi(f,\eta)$. Conversely, suppose that Λ is approximately optimal for f so that $\phi(f,\eta,\Lambda) \leq \inf_{\Lambda'}\phi(f,\eta,\Lambda') + \epsilon$. To form a candidate Λ for f_M, truncate Λ to C_M and then renormalize, that is, form $\hat{\Lambda}(x) = \Lambda(x)1_{C_M}(x)/\int_{C_M}\Lambda(y)dy$. Then

$$
\begin{aligned}
&\inf_{\Lambda}\phi(f_M,\eta,\Lambda)\\
&\quad \leq \phi(f_M,\eta,\hat{\Lambda})\\
&\quad = \frac{k}{2}\ln\left(E_{f_M}\left((\hat{\Lambda}(X))^{-2/k}\right)\right)\\
&\qquad + (1-\eta)h(f_M) + (1-\eta)E_{f_M}\left(\ln\hat{\Lambda}(X)\right)\\
&\quad = \frac{k}{2}\ln\left((1-p_M)\int_{C_M} d(x)f(x)\Lambda(x)^{-2/k}\right)\\
&\qquad + (1-\eta)\ln\left(\int_{C_M}\Lambda(y)dy\right)\\
&\qquad + (1-\eta)h(f_M) + (1-\eta)(1-p_M)\int_{C_M} f(x)\ln\Lambda(x)dx.
\end{aligned}
$$

As $M\to\infty$, the right-hand side converges to $\phi(f,\eta,\Lambda)$, whence $\limsup_{M\to\infty}\phi(f_M,\eta) \leq \phi(f,\eta)$, which finishes the proof of the second statement. □

Acknowledgment

The authors would like to thank S. Boyd for discussions on convex optimization and M. Effros and D. O'Brien for the comments.

References

[1] A. Aiyer, K.-S. Pyun, Y.-Z. Huang, D. B. O'Brien, and R. M. Gray, "Lloyd clustering of Gauss mixture models for image compression and classification," *Signal Process.: Image Commun.*, vol. 20, pp. 459–485, Jun. 2005.

[2] R. B. Ash*, Real Analysis and Probability*. New York: Academic, 1972.

[3] S. Boyd and L. Vandenberghe*, Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[4] J. A. Bucklew and G. L. Wise, "Multidimensional asymptotic quantization theory with rth power distortion measures," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 239–247, Mar. 1982.

[5] J. A. Bucklew, "Two results on the asymptotic performance of quantizers," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 2, pp. 341–348, Mar. 1984.

[6] W.-Y. Chan and A. Gersho, "Constrained-storage vector quantization in high fidelity audio transform coding," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Toronto, ON, Canada, May 1991, pp. 3597–3600.

[7] W.-Y. Chan and A. Gersho, "Enhanced multistage vector quantization by joint codebook design," *IEEE Trans. Commun.*, vol. 40, no. 11, pp. 1693–1697, Nov. 1992.

[8] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 1, pp. 31–42, Jan. 1989.

[9] I. Csiszár, "Generalized entropy and quantization problems," in *Proc. 6th Prague Conf. Inf. Theory, Statist. Decision Functions, Random Processes*, 1973, pp. 159–174.

[10] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 4, pp. 373–380, Jul. 1979.

[11] S. Graf and H. Luschgy, "Foundations of quantization for probability distributions," in *Lecture Notes in Mathematics*. Berlin, Germany: Springer-Verlag, 2000, vol. 1730.

[12] R. M. Gray, T. Linder, and J. Li, "A Lagrangian formulation of Zador's entropy-constrained quantization theorem," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 695–707, Mar. 2002.

[13] R. M. Gray and T. Linder, "Mismatch in high rate entropy constrained vector quantization," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1204–1217, May 2003.

[14] R. M. Gray and T. Linder, "Results and conjectures on high rate quantization," in *Proc. Data Compress. Conf.*, Snowbird, UT, 2004, pp. 3–12.

[15] R. M. Gray and J. T. Gill, III, "A Lagrangian formulation of fixed rate and entropy/memory constrained quantization," in *Proc. IEEE Data Compress. Conf.*, Mar. 23–31, 2005, pp. 223–235 [Online]. Available: http://ee.stanford.edu/~gray/dcc05.pdf

[16] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 10, pp. 2325–2384, Oct. 1998.

[17] A. György, T. Linder, P. A. Chou, and B. J. Betts, "Do optimal entropy-constrained quantizers have a finite or infinite number of codewords?," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 3031–3037, Nov. 2003.

[18] T. Linder and K. Zeger, "Asymptotic entropy constrained performance of tessellating and universal randomized lattice quantization," *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 575–579, Mar. 1994.

[19] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[20] S. Na and D. L. Neuhoff, "Bennett's integral for vector quantizers," *IEEE Trans. Inf. Theory*, vol. 41, no. 4, pp. 886–900, Jul. 1995.

[21] R. P. Rao and W. A. Pearlman, "Alphabet-constrained vector quantization," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1167–1179, Jul. 1993.

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

[23] S. Yoon and R. M. Gray, "Clustering and finding the number of clusters by unsupervised learning of mixture models using vector quantization," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Honolulu, HI, Apr. 2007, vol. 3, pp. 1081–1084.

[24] P. L. Zador, "Topics in the asymptotic quantization of continuous random variables," Bell Labs. Tech. Memorandum, 1966.

[25] P. L. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 139–148, Mar. 1982.