Quantization

Robert M. Gray, *Fellow, IEEE,* and David L. Neuhoff, *Fellow, IEEE*

(Invited Paper)

***Abstract*— The history of the theory and practice of quantization dates to 1948, although similar ideas had appeared in the literature as long ago as 1898. The fundamental role of quantization in modulation and analog-to-digital conversion was first recognized during the early development of pulse-code modulation systems, especially in the 1948 paper of Oliver, Pierce, and Shannon. Also in 1948, Bennett published the first high-resolution analysis of quantization and an exact analysis of quantization noise for Gaussian processes, and Shannon published the beginnings of rate distortion theory, which would provide a theory for quantization as analog-to-digital conversion and as data compression. Beginning with these three papers of fifty years ago, we trace the history of quantization from its origins through this decade, and we survey the fundamentals of the theory and many of the popular and promising techniques for quantization.**

***Index Terms*— High resolution theory, rate distortion theory, source coding, quantization.**

I. Introduction

THE dictionary (*Random House*) definition of quantization is the division of a quantity into a discrete number of small parts, often assumed to be integral multiples of a common quantity. The oldest example of quantization is rounding off, which was first analyzed by Sheppard [468] for the application of estimating densities by histograms. Any real number x can be rounded off to the nearest integer, say $q(x)$, with a resulting quantization error $e = q(x) - x$ so that $q(x) = x + e$. More generally, we can define a quantizer as consisting of a set of intervals or *cells* $\mathcal{S} = \{S_i;\ i \in \mathcal{I}\}$, where the index set $\mathcal{I}$ is ordinarily a collection of consecutive integers beginning with 0 or 1, together with a set of *reproduction values* or *points* or *levels* $\mathcal{C} = \{y_i;\ i \in \mathcal{I}\}$, so that the overall quantizer q is defined by $q(x) = y_i$ for $x \in S_i$, which can be expressed concisely as

$$q(x) = \sum_i y_i 1_{S_i}(x) \tag{1}$$

where the indicator function $1_S(x)$ is 1 if $x \in S$ and 0 otherwise. For this definition to make sense we assume that $\mathcal{S}$ is a partition of the real line. That is, the cells are disjoint and exhaustive. The general definition reduces to the rounding off example if $S_i = (i-1/2,\ i+1/2]$ and $y_i = i$ for all integers i. More generally, the cells might take the form $S_i = (a_{i-1},\ a_i]$ where the a_i's, which are called *thresholds*, form an increasing sequence. The width of a cell S_i is its length $a_i - a_{i-1}$. The function $q(x)$ is often called the *quantization rule*. A simple quantizer with five reproduction levels is depicted in Fig. 1 as a collection of intervals bordered by thresholds along with the levels for each interval.

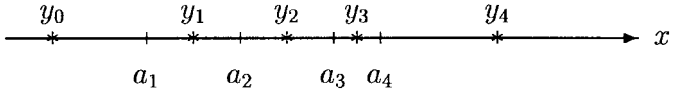


Fig. 1. A nonuniform quantizer: $a_0 = \infty$, $a_5 = \infty$.

A quantizer is said to be *uniform* if, as in the roundoff case, the levels y_i are equispaced, say Δ apart, and the thresholds a_i are midway between adjacent levels. If an infinite number of levels are allowed, then all cells S_i will have width equal to Δ, the separation between levels. If only a finite number of levels are allowed, then all but two cells will have width Δ and the outermost cells will be semi-infinite. An example of a uniform quantizer with cell width Δ and $N = 8$ levels is given in Fig. 2. Given a uniform quantizer with cell width Δ, the region of the input space within $\Delta/2$ of some quantizer level is called the *granular region* or simply the *support* and that outside (where the quantizer error is unbounded) is called the *overload* or *saturation* region. More generally, the support or granular region of a nonuniform quantizer is the region of the input space within a relatively small distance of some level, and the overload region is the complement of the granular region. To be concrete, "small" might be defined as half the width of the largest cell of finite width.

The quality of a quantizer can be measured by the goodness of the resulting reproduction in comparison to the original. One way of accomplishing this is to define a distortion measure $d(x, \hat{x})$ that quantifies cost or distortion resulting from reproducing x as $\hat{x}$ and to consider the average distortion as a measure of the quality of a system, with smaller average distortion meaning higher quality. The most common distortion measure is the squared error $d(x, \hat{x}) = |x - \hat{x}|^2$, but we shall encounter others later. In practice, the average will be a sample average when the quantizer is applied to a sequence of real data, but the theory views the data as sharing a common probability density function (pdf) $f(x)$ corresponding to a generic random variable X and the average distortion becomes an expectation

$$D(q) = E[d(X, q(X))] = \sum_i \int_{S_i} d(x, y_i) f(x)\, dx. \tag{2}$$

Manuscript received January 7, 1998; revised June 6, 1998. This work was supported in part by the National Science Foundation under Grants NCR-941574 and MIP-931190.

R. M. Gray is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA.

D. L. Neuhoff is with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109 USA.

Publisher Item Identifier S 0018-9448(98)06317-2.

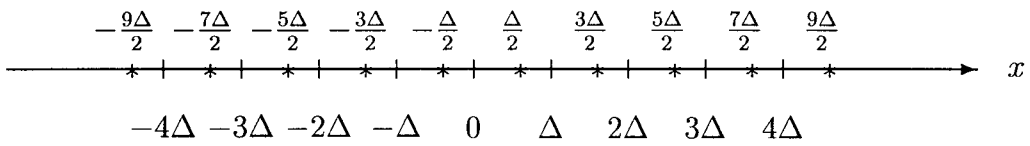


Fig. 2. A uniform quantizer.

If the distortion is measured by squared error, $D(q)$ becomes the mean squared error (MSE), a special case on which we shall mostly focus.

It is desirable to have the average distortion as small as possible, and in fact negligible average distortion is achievable by letting the cells become numerous and tiny. There is a cost in terms of the number of bits required to describe the quantizer output to a decoder, however, and arbitrarily reliable reproduction will not be possible for digital storage and communication media with finite capacity. A simple method for quantifying the cost for communications or storage is to assume that the quantizer "codes" an input x into a binary representation or channel codeword of the quantizer index i specifying which reproduction level should be used in the reconstruction. If there are N possible levels and all of the binary representations or binary codewords have equal length (a temporary assumption), the binary vectors will need $\log N$ (or the next larger integer, $\lceil \log N \rceil$, if $\log N$ is not an integer) components or bits. Thus one definition of the *rate* of the code in bits per input sample is

$$R(q) = \log N. \tag{3}$$

A quantizer with fixed-length binary codewords is said to have *fixed rate* because all quantizer levels are assumed to have binary codewords of equal length. Later this restriction will be weakened. Note that all logarithms in this paper will have base 2, unless explicitly specified otherwise.

In summary, the goal of quantization is to encode the data from a source, characterized by its probability density function, into as few bits as possible (i.e., with low rate) in such a way that a reproduction may be recovered from the bits with as high quality as possible (i.e., with small average distortion). Clearly, there is a tradeoff between the two primary performance measures: average distortion (or simply *distortion*, as we will often abbreviate) and rate. This tradeoff may be quantified as the *operational distortion-rate function* $\delta(R)$, which is defined to be the least distortion of any scalar quantizer with rate R or less. That is,

$$\delta(R) \equiv \inf_{q:\, R(q) \le R} D(q). \tag{4}$$

Alternatively, one can define the operational *rate-distortion function* $r(D)$ as the least rate of any fixed-rate scalar quantizer with distortion D or less, which is the inverse of $\delta(R)$.

We have so far described *scalar quantization with fixed-rate coding*, a technique whereby each data sample is independently encoded into a fixed number of bits and decoded into a reproduction. As we shall see, there are many alternative quantization techniques that permit a better tradeoff of distortion and rate; e.g., less distortion for the same rate, or vice versa. The purpose of this paper is to review the development of such techniques, and the theory of their design and performance. For example, for each type of technique we will be interested in its operational distortion-rate function, which is defined to be the least distortion of any quantizer of the given type with rate R or less. We will also be interested in the best possible performance among *all* quantizers. Both as a preview and as an occasional benchmark for comparison, we informally define the class of all quantizers as the class of quantizers that can 1) operate on scalars or vectors instead of only on scalars (vector quantizers), 2) have fixed or variable rate in the sense that the binary codeword describing the quantizer output can have length depending on the input, and 3) be memoryless or have memory, for example, using different sets of reproduction levels, depending on the past. In addition, we restrict attention to quantizers that do not change with time. That is, when confronted with the same input and the same past history, a quantizer will produce the same output regardless of the time. We occasionally use the term *lossy source code* or simply *code* as alternatives to *quantizer*. The rate is now defined as the average number of bits per source symbol required to describe the corresponding reproduction symbol. We informally generalize the operational distortion-rate function $\delta(R)$ providing the best performance for scalar quantizers, to $\overline{\delta}(R)$, which is defined as the infimum of the average distortion over all quantization techniques with rate R or less. Thus $\overline{\delta}(R)$ can be viewed as the best possible performance over all quantizers with no constraints on dimension, structure, or complexity.

Section II begins with a historical tour of the development of the theory and practice of quantization over the past fifty years, a period encompassing almost the entire literature on the subject. Two complementary approaches dominate the history and present state of the theory, and three of the key papers appeared in 1948, two of them in Volume 27 (1948) of the *Bell Systems Technical Journal*. Likely the approach best known to the readers of these TRANSACTIONS is that of rate-distortion theory or source coding with a fidelity criterion—Shannon's information-theoretic approach to source coding—which was first suggested in his 1948 paper [464] providing the foundations of information theory, but which was not fully developed until his 1959 source coding paper [465]. The second approach is that of high resolution (or high-rate or asymptotic) quantization theory, which had its origins in the 1948 paper on PCM by Oliver, Pierce, and Shannon [394], the 1948 paper on quantization error spectra by Bennett [43], and the 1951 paper by Panter and Dite [405]. Much of the history and state of the art of quantization derives from these seminal works.

In contrast to these two asymptotic theories, there is also a small but important collection of results that are not asymptotic in nature. The oldest such results are the exact analyses

for special nonasymptotic cases, such as Clavier, Panter, and Grieg's 1947 analysis of the spectra of the quantization error for uniformly quantized sinusoidal signals [99], [100], and Bennett's 1948 derivation of the power spectral density of a uniformly quantized Gaussian random process [43]. The most important nonasymptotic results, however, are the basic optimality conditions and iterative-descent algorithms for quantizer design, such as first developed by Steinhaus (1956) [480] and Lloyd (1957) [330], and later popularized by Max (1960) [349].

Our goal in the next section is to introduce in historical context many of the key ideas of quantization that originated in classical works and evolved over the past 50 years, and in the remaining sections to survey selectively and in more detail a variety of results which illustrate both the historical development and the state of the field. Section III will present basic background material that will be needed in the remainder of the paper, including the general definition of a quantizer and the basic forms of optimality criteria and descent algorithms. Some such material has already been introduced and more will be introduced in Section II. However, for completeness, Section III will be largely self-contained. Section IV reviews the development of quantization theories and compares the approaches. Finally, Section V describes a number of specific quantization techniques.

In any review of a large subject such as quantization there is no space to discuss or even mention all work on the subject. Though we have made an effort to select the most important work, no doubt we have missed some important work due to bias, misunderstanding, or ignorance. For this we apologize, both to the reader and to the researchers whose work we may have neglected.

II. History

The history of quantization often takes on several parallel paths, which causes some problems in our clustering of topics. We follow roughly a chronological order within each and order the paths as best we can. Specifically, we will first track the design and analysis of practical quantization techniques in three paths: fixed-rate scalar quantization, which leads directly from the discussion of Section I, predictive and transform coding, which adds linear processing to scalar quantization in order to exploit source redundancy, and variable-rate quantization, which uses Shannon's lossless source coding techniques [464] to reduce rate. (Lossless codes were originally called *noiseless*.) Next we follow early forward-looking work on vector quantization, including the seminal work of Shannon and Zador, in which vector quantization appears more to be a paradigm for analyzing the fundamental limits of quantizer performance than a practical coding technique. A surprising amount of such vector quantization theory was developed outside the conventional communications and signal processing literature. Subsequently, we review briefly the developments from the mid-1970's to the mid-1980's which mainly concern the emergence of vector quantization as a practical technique. Finally, we sketch briefly developments from the mid-1980's to the present. Except where stated otherwise, we presume squared error as the distortion measure.

A. Fixed-Rate Scalar Quantization: PCM and the Origins of Quantization Theory

Both quantization and source coding with a fidelity criterion have their origins in pulse-code modulation (PCM), a technique patented in 1938 by Reeves [432], who 25 years later wrote a historical perspective on and an appraisal of the future of PCM with Deloraine [120]. The predictions were surprisingly accurate as to the eventual ubiquity of digital speech and video. The technique was first successfully implemented in hardware by Black, who reported the principles and implementation in 1947 [51], as did another Bell Labs paper by Goodall [209]. PCM was subsequently analyzed in detail and popularized by Oliver, Pierce, and Shannon in 1948 [394]. PCM was the first *digital* technique for conveying an analog information signal (principally telephone speech) over an analog channel (typically, a wire or the atmosphere). In other words, it is a modulation technique, i.e., an alternative to AM, FM, and various other types of pulse modulation. It consists of three main components: a sampler (including a prefilter), a quantizer (with a fixed-rate binary encoder), and a binary pulse modulator. The sampler converts a continuous-time waveform $x(t)$ into a sequence of samples $x_n = x(n/f_s)$, where f_s is the sampling frequency. The sampler is ordinarily preceded by a lowpass filter with cutoff frequency $f_s/2$. If the filter is ideal, then the Shannon–Nyquist or Shannon–Whittaker–Kotelnikov sampling theorem ensures that the lowpass filtered signal can, in principle, be perfectly recovered by appropriately filtering the samples. Quantization of the samples renders this an approximation, with the MSE of the recovered waveform being, approximately, the sum of the MSE of the quantizer $D(q)$ and the high-frequency power removed by the lowpass filter. The binary pulse modulator typically uses the bits produced by the quantizer to determine the amplitude, frequency, or phase of a sinusoidal carrier waveform. In the evolutionary development of modulation techniques it was found that the performance of pulse-amplitude modulation in the presence of noise could be improved if the samples were quantized to the nearest of a set of N levels before modulating the carrier (64 equally spaced levels was typical). Though this introduces quantization error, deciding which of the N levels had been transmitted in the presence of noise could be done with such reliability that the overall MSE was substantially reduced. Reducing the number of quantization levels N made it even easier to decide which level had been transmitted, but came at the cost of a considerable increase in the MSE of the quantizer. A solution was to fix N at a value giving acceptably small quantizer MSE and to binary encode the levels, so that the receiver had only to make binary decisions, something it can do with great reliability. The resulting system, PCM, had the best resistance to noise of all modulations of the time.

As the digital era emerged, it was recognized that the sampling, quantizing, and encoding part of PCM performs an analog-to-digital (A/D) conversion, with uses extending much beyond communication over analog channels. Even in the communications field, it was recognized that the task of analog-to-digital conversion (and source coding) should be factored out of binary modulation as a separate task. Thus

PCM is now generally considered to just consist of sampling, quantizing, and encoding; i.e., it no longer includes the binary pulse modulation.

Although quantization in the information theory literature is generally considered as a form of data compression, its use for modulation or A/D conversion was originally viewed as data expansion or, more accurately, bandwidth expansion. For example, a speech waveform occupying roughly 4 kHz would have a Nyquist rate of 8 kHz. Sampling at the Nyquist rate and quantizing at 8 bits per sample and then modulating the resulting binary pulses using amplitude- or frequency-shift keying would yield a signal occupying roughly 64 kHz, a 16–fold increase in bandwidth! Mathematically this constitutes compression in the sense that a continuous waveform requiring an infinite number of bits is reduced to a finite number of bits, but for practical purposes PCM is not well interpreted as a compression scheme.

In an early contribution to the theory of quantization, Clavier, Panter, and Grieg (1947) [99], [100] applied Rice's characteristic function or transform method [434] to provide exact expressions for the quantization error and its moments resulting from uniform quantization for certain specific inputs, including constants and sinusoids. The complicated sums of Bessel functions resembled the early analyses of another nonlinear modulation technique, FM, and left little hope for general closed-form solutions for interesting signals.

The first general contributions to quantization theory came in 1948 with the papers of Oliver, Pierce, and Shannon [394] and Bennett [43]. As part of their analysis of PCM for communications, they developed the oft-quoted result that for large rate or resolution, a uniform quantizer with cell width Δ yields average distortion $D(q) \cong \Delta^2/12$. If the quantizer has N levels and rate $R = \log N$, and the source has input range (or *support*) of width A, so that $\Delta = A/N$ is the natural choice, then the $\Delta^2/12$ approximation yields the familiar form for the signal-to-noise ratio (SNR) of

$$10 \log_{10} \frac{\mathrm{var}(X)}{E[(q(X)-X)^2]} = c + 20R \log_{10} 2 \cong c + 6R \text{ dB}$$

showing that for large rate, the SNR of uniform quantization increases 6 dB for each one-bit increase of rate, which is often referred to as the "6-dB-per-bit rule." The $\Delta^2/12$ formula is considered a *high-resolution* formula; indeed, the first such formula, in that it applies to the situation where the cells and average distortion are small, and the rate is large, so that the reproduction produced by the quantizer is quite accurate. The $\Delta^2/12$ result also appeared many years earlier (albeit in somewhat disguised form) in Sheppard's 1898 treatment [468].

Bennett also developed several other fundamental results in quantization theory. He generalized the high-resolution approximation for uniform quantization to provide an approximation to $D(q)$ for companders, systems that preceded a uniform quantizer by a monotonic smooth nonlinearity called a "compressor," say G, and used the inverse nonlinearity when reconstructing the signal. Thus the output reproduction $\hat{x}$ given an input x was given by $\hat{x} = G^{-1}(q(G(x)))$, where q is a uniform quantizer. Bennett showed that in this case

$$D(q) \cong \frac{\Delta^2}{12} \int \frac{f(x)}{g^2(x)}\, dx \tag{5}$$

where $g(x) = dG(x)/dx$, Δ is the cellwidth of the uniform quantizer, and the integral is taken over the granular range of the input. (The constant $1/12$ in the above assumes that G maps to the unit interval $[0,1]$.) Since, as Bennett pointed out, any nonuniform quantizer can be implemented as a compander, this result, often referred to as "Bennett's integral," provides an asymptotic approximation for any quantizer. It is useful to jump ahead and point out that g can be interpreted, as Lloyd would explicitly point out in 1957 [330], as a constant times a "quantizer point-density function $\lambda(x)$," that is, a function with the property that for any region S

$$\text{number of quantizer levels in } S \approx N \int_S \lambda(x)\, dx. \tag{6}$$

Since integrating $\lambda(x)$ over a region gives the fraction of quantizer reproduction levels in the region, it is evident that $\lambda(x)$ is normalized so that $\int_{\Re} \lambda(x)\, dx = 1$. It will also prove useful to consider the unnormalized quantizer point density $\Lambda(x)$, which when integrated over S gives the total number of levels within S rather than the fraction. In the current situation $\Lambda(x) = N\lambda(x)$, but the unnormalized density will generalize to the case where N is infinite.

Rewriting Bennett's integral in terms of the point-density function yields its more common form

$$D(q) \cong \frac{1}{12} \frac{1}{N^2} \int \frac{f(x)}{\lambda^2(x)}\, dx. \tag{7}$$

The idea of a quantizer point-density function will generalize to vectors, while the compander approach will not in the sense that not all vector quantizers can be represented as companders [192].

Bennett also demonstrated that, under assumptions of high resolution and smooth densities, the quantization error behaved much like random "noise": it had small correlation with the signal and had approximately a flat ("white") spectrum. This led to an "additive-noise" model of quantizer error, since with these properties the formula $q(X) = X + [q(X) - X]$ could be interpreted as representing the quantizer output as the sum of a signal and white noise. This model was later popularized by Widrow [528], [529], but the viewpoint avoids the fact that the "noise" is in fact dependent on the signal and the approximations are valid only under certain conditions. Signal-independent quantization noise has generally been found to be perceptually desirable. This was the motivation for randomizing the action of quantization by the addition of a dither signal, a method introduced by Roberts [442] as a means of making quantized images look better by replacing the artifacts resulting from deterministic errors by random noise. We shall return to dithering in Section V, where it will be seen that suitable dithering can indeed make exact the Bennett approximations of uniform distribution and signal independence of the overall quantizer noise. Bennett also used a variation of Rice's method to derive an exact computation of the spectrum of quantizer noise when a Gaussian process

is uniformly quantized, providing one of the very few exact computations of quantization error spectra.

In 1951 Panter and Dite [405] developed a high-resolution formula for the distortion of a fixed-rate scalar quantizer using approximations similar to Bennett's, but without reference to Bennett. They then used variational techniques to minimize their formula and found the following formula for the operational distortion-rate function of fixed-rate scalar quantization: for large values of R

$$\delta(R) \cong \frac{1}{12}\left(\int f^{1/3}(x)\,dx\right)^3 2^{-2R} \tag{8}$$

which is now called the Panter and Dite formula.[1] As part of their derivation, they demonstrated that an optimal quantizer resulted in roughly equal contributions to total average distortion from each quantization cell, a result later called the "partial distortion theorem." Though they did not rederive Bennett's integral, they had in effect derived the optimal compressor function for a compander, or, equivalently, the optimal quantizer point density

$$\lambda(x) = \frac{f^{1/3}(x)}{\int f^{1/3}(x')\,dx'}. \tag{9}$$

Indeed, substituting this point density into Bennett's integral and using the fact that $R = \log N$ yields (8). As an example, if the input density is Gaussian with variance σ^2, then

$$\delta(R) \cong \frac{1}{12}\,6\pi\sqrt{3}\sigma^2 2^{-2R}. \tag{10}$$

The fact that for large rates $\delta(R)$ decreases with R as 2^{-2R} implies that the signal-to-noise ratio increases according to the 6-dB-per-bit rule. Virtually all other high resolution formulas to be given later will also obey this rule. However, the constant that adds to $6R$ will vary with the source and quantizer being considered.

The Panter–Dite formula for $\delta(R)$ can also be derived directly from Bennett's integral using variational methods, as did Lloyd (1957) [330], Smith (1957) [474], and, much later without apparent knowledge of earlier work, Roe (1964) [443]. It can also be derived without using variational methods by application of Hölder's inequality to Bennett's integral [222], with the additional benefit of demonstrating that the claimed minimum is indeed global. Though not known at the time, it turns out that for a Gaussian source with independent and identically distributed (i.i.d.) samples, the operational distortion-rate function given above is $\pi\sqrt{3}/2 = 2.72$ times larger than $\overline{\delta}(R)$, the least distortion achievable by any quantization technique with rate R or less. (It was not until Shannon's 1959 paper [465] that $\overline{\delta}(R)$ was known.) Equivalently, the induced signal-to-noise ratio is 4.35 dB less than the best possible, or for a fixed distortion D the rate is 0.72 bits/sample larger than that achievable by the best quantizers.

In 1957, Smith [474] re-examined companding and PCM. Among other things, he gave somewhat cleaner derivations of Bennett's integral, the optimal compressor function, and the Panter–Dite formula.

Also in 1957, Lloyd [330] made an important study of quantization with three main contributions. First, he found necessary and sufficient conditions for a fixed-rate quantizer to be locally optimal; i.e., conditions that if satisfied implied that small perturbations to the levels or thresholds would increase distortion. Any optimal quantizer (one with smallest distortion) will necessarily satisfy these conditions, and so they are often called the *optimality conditions* or the *necessary conditions*. Simply stated, Lloyd's optimality conditions are that for a fixed-rate quantizer to be optimal, the quantizer partition must be optimal for the set of reproduction levels, and the set of reproduction levels must be optimal for the partition. Lloyd derived these conditions straightforwardly from first principles, without recourse to variational concepts such as derivatives. For the case of mean-squared error, the first condition implies a minimum distance or nearest neighbor quantization rule, choosing the closest available reproduction level to the source sample being quantized, and the second condition implies that the reproduction level corresponding to a given cell is the conditional expectation or *centroid* of the source value given that it lies in the specified cell; i.e., it is the minimum mean-squared error estimate of the source sample. For some sources there are multiple locally optimal quantizers, not all of which are globally optimal.

Second, based on his optimality conditions, Lloyd developed an iterative descent algorithm for designing quantizers for a given source distribution: begin with an initial collection of reproduction levels; optimize the partition for these levels by using a minimum distortion mapping, which gives a partition of the real line into intervals; then optimize the set of levels for the partition by replacing the old levels by the centroids of the partition cells. The alternation is continued until convergence to a local, if not global, optimum. Lloyd referred to this design algorithm as "Method I." He also developed a Method II based on the optimality properties. First choose an initial smallest reproduction level. This determines the cell threshold to the right, which in turn implies the next larger reproduction level, and so on. This approach alternately produces a level and a threshold. Once the last level has been chosen, the initial level can then be rechosen to reduce distortion and the algorithm continues. Lloyd provided design examples for uniform, Gaussian, and Laplacian random variables and showed that the results were consistent with the high resolution approximations. Although Method II would initially gain more popularity when rediscovered in 1960 by Max [349], it is Method I that easily extends to vector quantizers and many types of quantizers with structural constraints.

Third, motivated by the work of Panter and Dite but apparently unaware of that of Bennett or Smith, Lloyd rederived Bennett's integral and the Panter–Dite formula based on the concept of point-density function. This was a critically important step for subsequent generalizations of Bennett's integral to vector quantizers. He also showed directly that in situations where the global optimum is the only local optimum, quantizers that satisfy the optimality conditions have, asymptotically, the optimal point density given by (9).

[1] They also indicated that it had been derived earlier by P. R. Aigrain.

Unfortunately, Lloyd's work was not published in an archival journal at the time. Instead, it was presented at the 1957 Institute of Mathematical Statistics (IMS) meeting and appeared in print only as a Bell Laboratories Technical Memorandum. As a result, its results were not widely known in the engineering literature for many years, and many were independently rediscovered. All of the independent rediscoveries, however, used variational derivations, rather than Lloyd's simple derivations. The latter were essential for later extensions to vector quantizers and to the development of many quantizer optimization procedures. To our knowledge, the first mention of Lloyd's work in the IEEE literature came in 1964 with Fleischer's [170] derivation of a sufficient condition (namely, that the log of the source density be concave) in order that the optimal quantizer be the only locally optimal quantizer, and consequently, that Lloyd's Method I yields a globally optimal quantizer. (The condition is satisfied for common densities such as Gaussian and Laplacian.) Zador [561] had referred to Lloyd a year earlier in his Ph.D. dissertation, to be discussed later.

Later in the same year in another Bell Telephone Laboratories Technical Memorandum, Goldstein [207] used variational methods to derive conditions for global optimality of a scalar quantizer in terms of second-order partial derivatives with respect to the quantizer levels and thresholds. He also provided a simple counterintuitive example of a symmetric density for which the optimal quantizer was asymmetric.

In 1959, Shtein [471] added terms representing overload distortion to the $\Delta^2/12$ formula and to Bennett's integral and used them to optimize uniform and nonuniform quantizers. Unaware of prior work, except for Bennett's, he rederived the optimal compressor characteristic and the Panter–Dite formula.

In 1960, Max [349] published a variational proof of the Lloyd optimality properties for rth-power distortion measures, rediscovered Lloyd's Method II, and numerically investigated the design of fixed-rate quantizers for a variety of input densities.

Also in 1960, Widrow [529] derived an exact formula for the characteristic function of a uniformly quantized signal when the quantizer has an infinite number of levels. His results showed that under the condition that the characteristic function of the input signal be zero when its argument is greater than π/Δ, the moments of the quantized random variable are the same as the moments of the signal plus an additive signal-independent random variable uniformly distributed on $(-\Delta/2, \Delta/2]$. This has often been misinterpreted as saying that the quantized random variable can be approximated as being the input plus signal-independent uniform noise, a clearly false statement since the quantizer error $q(X) - X$ is a deterministic function of the signal. The "bandlimited" property of the characteristic function implies from Fourier transform theory that the probability density function must have infinite support since a signal and its transform cannot both be perfectly bandlimited.

We conclude this subsection by mentioning early work that appeared in the mathematical and statistical literature and which, in hindsight, can be viewed as related to scalar quantization. Specifically, in 1950–1951 Dalenius *et al.* [118], [119] used variational techniques to consider optimal grouping of Gaussian data with respect to average squared error. Lukaszewicz and H. Steinhaus [336] (1955) developed what we now consider to be the Lloyd optimality conditions using variational techniques in a study of optimum go/no-go gauge sets (as acknowledged by Lloyd). Cox in 1957 [111] also derived similar conditions. Some additional early work, which can now be seen as relating to vector quantization, will be reviewed later [480], [159], [561].

B. Scalar Quantization with Memory

It was recognized early that common sources such as speech and images had considerable "redundancy" that scalar quantization could not exploit. The term "redundancy" was commonly used in the early days and is still popular in some of the quantization literature. Strictly speaking, it refers to the statistical correlation or dependence between the samples of such sources and is usually referred to as *memory* in the information theory literature. As our current emphasis is historical, we follow the traditional language. While not disrupting the performance of scalar quantizers, such redundancy could be exploited to attain substantially better rate-distortion performance. The early approaches toward this end combined linear processing with scalar quantization, thereby preserving the simplicity of scalar quantization while using intuition-based arguments and insights to improve performance by incorporating memory into the overall code. The two most important approaches of this variety were predictive coding and transform coding. A shared intuition was that a preprocessing operation intended to make scalar quantization more efficient should "remove the redundancy" in the data. Indeed, to this day there is a common belief that data compression is equivalent to redundancy removal and that data without redundancy cannot be further compressed. As will be discussed later, this belief is contradicted both by Shannon's work, which demonstrated strictly improved performance using vector quantizers even for memoryless sources, and by the early work of Fejes Toth (1959) [159]. Nevertheless, removing redundancy leads to much improved codes.

Predictive quantization appears to originate in the 1946 delta modulation patent of Derjavitch, Deloraine, and Van Mierlo [129], but the most commonly cited early references are Cutler's patent [117] 2 605 361 on "Differential quantization of communication signals" and on DeJager's Philips technical report on delta modulation [128]. Cutler stated in his patent that it "is the object of the present invention to improve the efficiency of communication systems by taking advantage of correlation in the signals of these systems" and Derjavitch *et al.* also cited the reduction of redundancy as the key to the reduction of quantization noise. In 1950, Elias [141] provided an information-theoretic development of the benefits of predictive coding, but the work was not published until 1955 [142]. Other early references include [395], [300], [237], [511], and [572]. In particular, [511] claims Bennett-style asymptotics for high-resolution quantization error, but as will be discussed later, such approximations have yet to be rigorously derived.

From the point of view of least squares estimation theory, if one were to optimally predict a data sequence based on its past

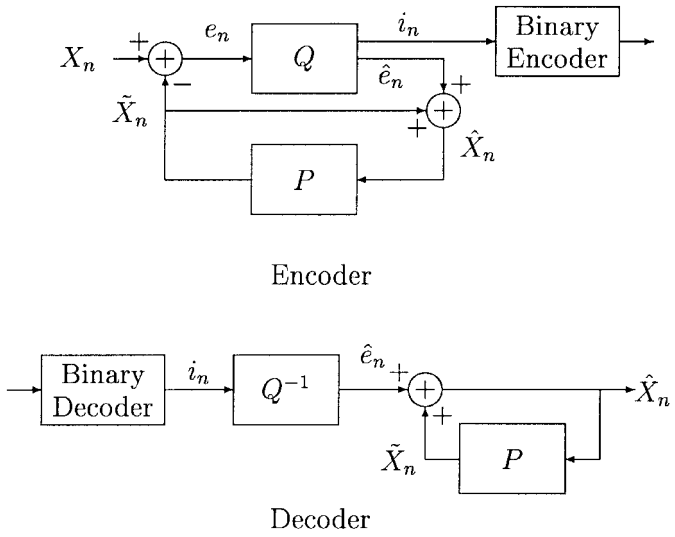


Fig. 3. Predictive quantizer encoder/decoder.

in the sense of minimizing the mean-squared error, then the resulting error or residual or innovations sequence would be uncorrelated and it would have the minimum possible variance. To permit reconstruction in a coded system, however, the prediction must be based on past reconstructed samples and not true samples. This is accomplished by placing a quantizer inside a prediction loop and using the same predictor to decode the signal. A simple predictive quantizer or differential pulse-coded modulator (DPCM) is depicted in Fig. 3. If the predictor is simply the last sample and the quantizer has only one bit, the system becomes a delta-modulator. Predictive quantizers are considered to have *memory* in that the quantization of a sample depends on previous samples, via the feedback loop.

Predictive quantizers have been extensively developed, for example there are many adaptive versions, and are widely used in speech and video coding, where a number of standards are based on them. In speech coding they form the basis of ITU-G.721, 722, 723, and 726, and in video coding they form the basis of the interframe coding schemes standardized in the MPEG and H.26X series. Comprehensive discussions may be found in books [265], [374], [196], [424], [50], and [458], as well as survey papers [264] and [198].

Though decorrelation was an early motivation for predictive quantization, the most common view at present is that the primary role of the predictor is to reduce the variance of the variable to be scalar-quantized. This view stems from the facts that a) it is the prediction errors rather than the source samples that are quantized, b) the overall quantization error precisely equals that of the scalar quantizer operating on the prediction errors, c) the operational distortion-rate function $\delta(R)$ for scalar quantization is proportional to variance (more precisely, a scaling of the random variable being quantized by a factor a results in a scaling of $\delta(R)$ by a^2), and d) the density of the prediction error is usually sufficiently similar in form to that of the source that its operational distortion-rate function is smaller than that of the original source by, approximately, the ratio of the variance of the source to that of the prediction error, a quantity that is often called a *prediction gain* [350], [396], [482], [397], [265]. Analyses of this form usually claim that under high-resolution conditions the distribution of the prediction error approaches that of the error when predictions are based on past source samples rather than past reproductions. However, it is not clear that the accuracy of this approximation increases sufficiently rapidly with finer resolution to ensure that the difference between the operational distortion-rate functions of the two types of prediction errors is small relative to their values, which are themselves decreasing as the resolution becomes finer. Indeed, it is still an open question whether this type of analysis, which typically uses Bennett and Panter–Dite formulas, is asymptotically correct. Nevertheless, the results of such high resolution approximations are widely accepted and often compare well with experimental results [156], [265]. Assuming that they give the correct answer, then for large rates and a stationary, Gaussian source with memory, the distortion of an optimized DPCM quantizer is less than that of a scalar quantizer by the factor σ_1^2/σ^2, where σ^2 is the variance of the source and σ_1^2 is the one-step prediction error; i.e., the smallest MSE of any prediction of one sample based on previous samples. It turns out that this exceeds $\overline{\delta}(R)$ by the same factor by which the distortion of optimal fixed-rate scalar quantization exceeds $\overline{\delta}(R)$ for a memoryless Gaussian source. Hence, it appears that DPCM does a good job of exploiting source memory given that it is based on scalar quantization, at least under the high-resolution assumption.

Because it has not been rigorously shown that one may apply Bennett's integral or the Panter–Dite formula directly to the prediction error, the analysis of such feedback quantization systems has proved to be notoriously difficult, with results limited to proofs of stability [191], [281], [284], i.e., asymptotic stationarity, to analyses of distortion via Hermite polynomial expansions for Gaussian processes [124], [473], [17], [346], [241], [262], [156], [189], [190], [367]–[369], [293], to analyses of distortion when the source is a Wiener process [163], [346], [240], and to exact solutions of the nonlinear difference equations describing the system and hence to descriptions of the output sequences and their moments, including power spectral densities, for constant and sinusoidal signals and finite sums of sinusoids using Rice's method, results which extend the work of Panter, Clavier, and Grieg to quantizers inside a feedback loop [260], [71], [215], [216], [72]. Conditions for use in code design resembling the Lloyd optimality conditions have been studied for feedback quantization [161], [203], [41], but the conditions are not optimality conditions in the Lloyd sense, i.e., they are not necessary conditions for a quantizer within a feedback loop to yield the minimum average distortion subject to a rate constraint. We will return to this issue when we consider finite-state vector quantizers. There has also been work on the optimality of certain causal coding structures somewhat akin to predictive or feedback quantization [331], [414], [148], [534], [178], [381], [521].

Transform coding is the second approach to exploiting redundancy by using scalar quantization with linear preprocessing. Here, the source samples are collected into a vector of, say, dimension k that is multiplied by an orthogonal matrix (an

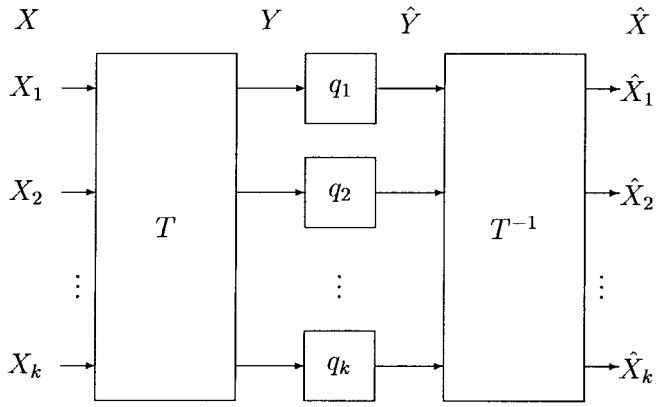


Fig. 4. Transform code.

orthogonal transform) and the resulting transform coefficients are scalar quantized, usually with a different quantizer for each coefficient. The operation is depicted in Fig. 4. This style of code was introduced in 1956 by Kramer and Mathews [299] and analyzed and popularized in 1962–1963 by Huang and Schultheiss [247], [248]. Kramer and Mathews simply assumed that the goal of the transform was to decorrelate the symbols, but Huang and Schultheiss proved that decorrelating does indeed lead to optimal transform code design, at least in the case of Gaussian sources and high resolution. Transform coding has been extensively developed for coding images and video, where the discrete cosine transform (DCT) [7], [429] is most commonly used because of its computational simplicity and its good performance. Indeed, DCT coding is the basic approach dominating current image and video coding standards, including H.261, H.263, JPEG, and MPEG. These codes combine uniform scalar quantization of the transform coefficients with an efficient lossless coding of the quantizer indices, as will be considered in the next section as a variable-rate quantizer. For discussions of transform coding for images see [533], [422], [375], [265], [98], [374], [261], [424], [196], [208], [408], [50], [458], and More recently, transform coding has also been widely used in high-fidelity audio coding [272], [200].

Unlike predictive quantizers, the transform coding approach lent itself quite well to the Bennett high-resolution approximations, the classical analysis being that of Huang and Schultheiss [247], [248] of the performance of optimized transform codes for fixed-rate scalar quantizers for Gaussian sources, a result which demonstrated that the Karhunen–Loève decorrelating transform was optimum for this application for the given assumptions. If the transform is the Karhunen–Loève transform, then the coefficients will be uncorrelated (and hence independent if the input vector is also Gaussian). The seminal work of Huang and Schultheiss showed that high-resolution approximation theory could provide analytical descriptions of optimal performance and design algorithms for optimizing codes of a given structure. In particular, they showed that under the high-resolution assumptions with Gaussian sources, the average distortion of the best transform code with a given rate is less than that of optimal scalar quantization by the factor $(\det K_k)^{1/k}/\sigma^2$, where σ^2 is the average of the variances of the components of the source vector and K_k is its $k \times k$ covariance matrix. Note that this reduction in distortion becomes larger for sources with more memory (more correlation) because the covariance matrices of such sources have smaller determinants. When k is large, it turns out that the distortion of optimized transform coding with a given rate exceeds $\overline{\delta}(R)$ by the same factor by which the distortion of optimal fixed-rate scalar quantization exceeds $\overline{\delta}(R)$ for a memoryless Gaussian source. Hence, like DPCM, transform coding does a good job of exploiting source memory given that it is a system based on scalar quantization.

C. Variable-Rate Quantization

Shannon's lossless source coding theory (1948) [464] made it clear that assigning equal numbers of bits to all quantization cells is wasteful if the cells have unequal probabilities. Instead, the number of bits produced by the quantizer will, on the average, be reduced if shorter binary codewords are assigned to higher probability cells. Of course, this means that longer codewords will need to be assigned to the less probable cells, but Shannon's theory shows that, in general, there is a net gain. This leads directly to *variable-rate quantization*, which has the partition into cells and codebook of levels as before, but now has binary codewords of varying lengths assigned to the cells (alternatively, the levels). Ordinarily, the set of binary codewords is chosen to satisfy the prefix condition that no member is a prefix of another member, in order to insure unique decodability. As will be made precise in the next section, one may view a variable-rate quantizer as consisting of a partition, a codebook, and a lossless binary code, i.e., an assignment of binary codewords.

For variable-rate quantizers the rate is no longer defined as the logarithm of the codebook size. Rather, the instantaneous rate for a given input is the number of binary symbols in the binary codeword (the length of the binary codeword) and the rate is the average length of the binary codewords, where the average is taken over the probability distribution of the source samples. The operational distortion-rate function $\delta(R)$ using this definition is the smallest average distortion over all (variable-rate) quantizers having rate R or less. Since we have weakened the constraint by expanding the allowed set of quantizers, this operational distortion-rate function will ordinarily be smaller than the fixed-rate optimum.

Huffman's algorithm [251] provides a systematic method of designing binary codes with the smallest possible average length for a given set of probabilities, such as those of the cells. Codes designed in this way are typically called Huffman codes. Unfortunately, there is no known expression for the resulting minimum average length in terms of the probabilities. However, Shannon's lossless source coding theorem implies that given a source and a quantizer partition, one can always find an assignment of binary codewords (indeed, a prefix set) with average length not more than $H(q(X)) + 1$, and that no uniquely decodable set of binary codewords can have average length less than $H(q(X))$, where

$$H(q(X)) = -\sum_i P_i \log P_i$$

is the Shannon *entropy* of the quantizer output and $P_i = \Pr(X \in S_i)$ is the probability that the source sample X lies in the ith cell S_i. Shannon also provided a simple way of attaining performance within the upper bound: if the quantizer index is i, then assign it a binary codeword with length $\lceil -\log P_i \rceil$ (the Kraft inequality ensures that this is always possible by simply choosing paths in a binary tree). Moreover, tighter bounds have been developed. For example, Gallager [181] has shown that the entropy can be at most $P_{\max}+0.0861$ smaller than the average length of the Huffman code, when $P_{\max}$, the largest of the P_i's, is less than $1/2$. See [73] for discussion of this and other bounds. Since $P_{\max}$ is ordinarily much smaller than $1/2$, this shows that $H(q(X))$ is generally a fairly accurate estimate of the average rate, especially in the high-resolution case.

Since there is no simple formula determining the rate of the Huffman code, but entropy provides a useful estimate, it is reasonable to simplify the variable-length quantizer design problem a little by redefining the instantaneous rate of a variable-rate quantizer as $-\log P_i$ for the ith quantizer level and hence to define the average rate as $H(q(X))$, the entropy of its output. As mentioned above, this underestimates the true rate by a small amount that in no case exceeds one. We could again define an operational distortion-rate function as the minimum average distortion over all variable-rate quantizers with output entropy $H(q(X)) \le R$. Since the quantizer output entropy is a lower bound to actual rate, this operational distortion-rate function may be optimistic; i.e., it falls below $\delta(R)$ defined using average length as rate. A quantizer designed to provide the smallest average distortion subject to an entropy constraint is called an *entropy-constrained scalar quantizer*.

Variable-rate quantization is also called *variable-length quantization* or *quantization with entropy coding*. We will not, except where critical, take pains to distinguish entropy-constrained quantizers and entropy-coded quantizers. And we will usually blur the distinction between average length and entropy as measures of the rate of such quantizers unless, again, it is important in some particular discussion. This is much the same sort of blurring as using $\log N$ instead of $\lceil \log N \rceil$ as the measure of rate in fixed-rate quantization.

It is important to note that the number of quantization cells or levels does not play a primary role in variable-rate quantization because, for example, there can be many levels in places where the source density is small with little effect on either distortion or rate. Indeed, the number of levels can be infinite, which has the advantage of eliminating the overload region and resulting overload distortion.

A potential drawback of variable-rate quantization is the necessity of dealing with the variable numbers of bits that it produces. For example, if the bits are to be communicated through a fixed-rate digital channel, one will have to use buffering and to take buffer overflows and underflows into account. Another drawback is the potential for error propagation when bits are received by the decoder in error.

The most basic and simple example of a variable-rate quantizer, and one which plays a fundamental role as a benchmark for comparison, is a uniform scalar quantizer with a variable-length binary lossless code.

The possibility of applying variable-length coding to quantization may well have occurred to any number of people who were familiar with both quantization and Shannon's 1948 paper. The earliest references to such that we have found are in the 1952 papers by Kretzmer [300] and Oliver [395]. In 1960, Max [349] had such in mind when he computed the entropy of nonuniform and uniform quantizers that had been designed to minimize distortion for a given number of levels. For a Gaussian source, his results showed that variable-length coding would yield rate reductions of about 0.5 bit/sample.

High-resolution analysis of variable-rate quantization developed in a handful of papers from 1958 to 1968. However, since these papers were widely scattered or unpublished, it was not until 1968 that the situation was well understood in the IEEE community.

The first high-resolution analysis was that of Schutzenberger (1958) [462] who showed that the distortion of optimized variable-rate quantization (both scalar and vector) decreases with rate as 2^{-2R}, just as with fixed-rate quantization. But he did not find the multiplicative factors, nor did he describe the nature of the partitions and codebooks that are best for variable-rate quantization.

In 1959, Renyi [433] showed that a uniform scalar quantizer with infinitely many levels and small cell width Δ has output entropy given approximately by

$$H(q(X)) \cong h(X) - \log \Delta \tag{11}$$

where

$$h(X) = -\int f(x) \log f(x)\, dx$$

is the *differential entropy* of the source variable X.

In 1963, Koshelev [579] discovered the very interesting fact that in the high-resolution case, the mean-squared error of uniform scalar quantization exceeds that of the least distortion achievable by any quantization scheme whatsoever, i.e., $\overline{\delta}(R)$, by a factor of only $\pi e/6 = 1.42$. Equivalently, the induced signal-to-noise ratio is only 1.53 dB less than the best possible, or for a fixed distortion D, the rate is only 0.255 bit/sample larger than that achievable by the best quantizers. (For the Gaussian source, it gains 2.82 dB or 0.47 bit/sample over the best fixed-rate scalar quantizer.) It is also of interest to note that this was the first paper to compare the performance of a specific quantization scheme to $\overline{\delta}(R)$. Unfortunately, Koshelev's paper was published in a journal that was not widely circulated.

In an unpublished 1966 Bell Telephone Laboratories Technical Memo [562], Zador also studied variable-rate (as well as fixed-rate) quantization. As his focus was on vector quantization, his work will be described later. Here we only point out that for variable-rate scalar quantization with large rate, his results showed that the operational distortion-rate function (i.e., the least distortion of such codes with a given rate) is

$$\delta(R) \cong \frac{1}{12} 2^{2h(X)} 2^{-2R}. \tag{12}$$

Though he was not aware of it, this turns out to be the formula found by Koshelev, therby demonstrating that in the high-

resolution case, uniform is the best type of scalar quantizer when variable-rate coding is applied.

Finally, in 1967 and 1968 two papers appeared in the IEEE literature (in fact in these TRANSACTIONS) on variable-rate quantization, without reference to any of the aforementioned work. The first, by Goblick and Holsinger [205], showed by numerical evaluation that uniform scalar quantization with variable-rate coding attains performance within about 1.5 dB (or 0.25 bit/sample) of the best possible for an i.i.d. Gaussian source. The second, by Gish and Pierce [204], demonstrated analytically what the first paper had found empirically. Specifically, it derived (11), and more generally, the fact that a high-resolution nonuniform scalar quantizer has output entropy

$$H(q(X)) \cong h(X) + \int f(x) \log \Lambda(x)\, dx \qquad (13)$$

where $\Lambda(x)$ is the unnormalized point density of the quantizer. They then used these approximations along with Bennett's integral to rederive (12) and to show that in the high-resolution case, uniform scalar quantizers achieve the operational distortion-rate function of variable-rate quantization. Next, by comparing to what is called the *Shannon lower bound* to $\overline{\delta}(R)$, they showed that for i.i.d. sources, the latter is only 1.53 dB (0.255 bit/sample) from the best possible performance $\overline{\delta}(R)$ of any quantization system whatsoever, which is what Koshelev [579] found earlier. Their results showed that such good performance was attainable for any source distribution, not just the Gaussian case checked by Goblick and Holsinger. They also generalized the results from squared-error distortion to nondecreasing functions of magnitude error.

Less well known is their proof of the fact that in the high resolution case, the entropy of k successive outputs of a uniformly scalar quantized stationary source, e.g., with memory, is

$$H(q(X_1), \cdots, q(X_k)) \cong h(X_1, \cdots, X_k) - \log \Delta. \qquad (14)$$

They used this, and the generalization of (13) to vectors, to show that when rate and k are large, uniform scalar quantization with variable-length coding of k successive quantizer outputs (*block entropy coding*) achieves performance that is 1.53 dB (0.255 bit/sample) from $\overline{\delta}(R)$, even for sources with memory. (They accomplished this by comparing to Shannon lower bounds.) This important result was not widely appreciated until rediscovered by Ziv (1985) [578], who also showed that a similar result holds for small rates. Note that although uniform scalar quantizers are quite simple, the lossless code capable of approaching the kth-order entropy of the quantized source can be quite complicated. In addition, Gish and Pierce observed that when coding vectors, performance could be improved by using quantizer cells other than the cube implicitly used by uniform scalar quantizers and noted that the hexagonal cell was superior in two dimensions, as originally demonstrated by Fejes Toth [159] and Newman [385].

Though uniform quantization is asymptotically best for entropy-constrained quantization, at lower rates nonuniform quantization can do better, and a series of papers explored algorithms for designing them. In 1969, Wood [539] provided a numerical descent algorithm for designing an entropy-constrained scalar quantizer, and showed, as predicted by Gish and Pierce, that the performance was only slightly superior to a uniform scalar quantizer followed by a lossless code.

In a 1972 paper dealing with a vector quantization technique to be discussed later, Berger [47] described Lloyd-like conditions for optimality of an entropy-constrained scalar quantizer for squared-error distortion. He formulated the optimization as an unconstrained Lagrangian minimization and developed an iterative algorithm for the design of entropy-constrained scalar quantizers. He showed that Gish and Pierce's demonstration of approximate optimality of uniform scalar quantization for variable-rate quantization holds approximately even when the rate is not large and holds exactly for exponential densities, provided the levels are placed at the centroids. In 1976, Netravali and Saigal introduced a fixed-point algorithm with the same goal of minimizing average distortion for a scalar quantizer with an entropy constraint [376]. Yet another approach was taken by Noll and Zelinski (1978) [391]. Berger refined his approach to entropy-constrained quantizer design in [48].

Variable-rate quantization was also extended to DPCM and transform coding, where high-resolution analysis shows that it gains the same relative to fixed-rate quantization as it does when applied to direct scalar quantizing [154], [398]. We note, however, that the variable-rate quantization analysis for DPCM suffers from the same flaws as the fixed-rate quantization analysis for DPCM.

Numerous extensions of the Bennett-style asymptotic approximations and the approximation of $r(D)$ or $\delta(R)$ and the characterizations of properties of optimal high-resolution quantization for both fixed- and variable-rate quantization for squared error and other error moments appeared during the 1960's, e.g., [497], [498], [55], [467], [8]. An excellent summary of the early work is contained in a 1970 paper by Elias [143].

We close this section with an important practical observation. The current JPEG and related standards can be viewed as a combination of transform coding and variable-length quantization. It is worth pointing out how the standard resembles and differs from the models considered thus far. As previously stated, the transform coefficients are separately quantized by possibly different uniform quantizers, the bin lengths of the quantizers being determined by a customizable quantization table. This typically produces a quantized transformed image with many zeros. The lossless, variable-length code then scans the image in a zig-zag (or Peano) fashion, producing a sequence of runlengths of the zeros and indices corresponding to nonzero values, which are then Huffman-coded (or arithmetic-coded). This procedure has the effect of coding only the transform coefficients with the largest magnitude, which are the ones most important for reconstruction. The early transform coders typically coded the first, say, K coefficients, and ignored the rest. In essence, the method adopted for the standards selectively coded the most important coefficients, i.e., those having the largest magnitude, rather than simply

the lowest frequency coefficients. The runlength coding step can in hindsight be viewed as a simple way of locating the most significant coefficients, which in turn are described the most accurately. This implicit "significance" map was an early version of an idea that would later be essential to wavelet coders.

D. The Beginnings of Vector Quantization

As described in the three previous subsections, the 1940's through the early 1970's produced a steady stream of advances in the design and analysis of practical quantization techniques, principally scalar, predictive, transform, and variable-rate quantization, with quantizer performance improving as these decades progressed. On the other hand, at roughly the same time there was a parallel series of developments that were more concerned with the fundamental limits of quantization than with practical quantization issues. We speak primarily of the remarkable work of Shannon and the very important work of Zador, though there were other important contributors as well. This work dealt with what is now called *vector quantization* (VQ) (or *block* or *multidimensional quantization*), which is just like scalar quantization except that all components of a vector, of say k successive source samples, are quantized simultaneously. As such they are characterized by a k-dimensional partition, a k-dimensional codebook (containing k-dimensional *points*, *reproduction codewords* or *codevectors*), and an assignment of binary codewords to the cells of the partition (equivalently, to the codevectors).

An immediate advantage of vector quantization is that it provides a model of a general quantization scheme operating on vectors without any structural constraints. It clearly includes transform coding as a special case and can also be considered to include predictive quantization operating locally within the vector. This lack of structural constraints makes the general model more amenable to analysis and optimization. In these early decades, vector quantization served primarily as a paradigm for exploring fundamental performance limits; it was not yet evident whether it would become a practical coding technique.

Shannon's Source Coding Theory: In his classic 1948 paper, Shannon [464] sketched the idea of the rate of a source as the minimum bit rate required to reconstruct the source to some degree of accuracy as measured by a fidelity criterion such as mean-squared error. The sketch was fully developed in his 1959 paper [465] for i.i.d. sources, additive measures of distortion, and block source codes, now called vector quantizers. In this later paper, Shannon showed that when coding at some rate R, the least distortion achievable by vector quantizers of any kind is equal to a function $D(R)$, subsequently called the *Shannon distortion-rate function*, that is determined by the statistics of the source and the measure of distortion.[2]

[2]Actually, Shannon described the solution to the equivalent problem of minimizing rate subject to a distortion constraint and found that the answer was given by a function $R(D)$, subsequently called the *Shannon rate-distortion function*, which is the inverse of $D(R)$. Accordingly, the theory is often called *rate-distortion theory*, cf. [46].

To elaborate on Shannon's theory, we note that one can immediately extend the quantizer notation of (1), the distortion and rate definitions of (2) and (3), and the operational distortion-rate functions to define the smallest distortion $\delta_k(R)$ possible for a k-dimensional fixed-rate vector quantizer that achieves rate R or less. (The distortion between two k-dimensional vectors is defined to be the numerical average of the distortions between their respective components. The rate is $1/k$ times the (average) number of bits to describe a k-dimensional source vector.) We will make the dimension k explicit in the notation when we are allowing it to vary and omit it when not. Furthermore, as with Shannon's channel coding and lossless source coding theories, one can consider the best possible performance over codes of *all* dimensions (assuming the data can be blocked into vectors of arbitrary size) and define an operational distortion-rate function

$$\overline{\delta}(R) = \inf_k \delta_k(R). \tag{15}$$

The operational rate-distortion functions $r_k(D)$ and $\overline{r}(D)$ are defined similarly. For finite dimension k, the function $\delta_k(R)$ will depend on the definition of rate, i.e., whether it is the log of the reproduction size, the average binary codeword length, or the quantizer output entropy. It turns out, however, that $\overline{\delta}(R)$ is not affected by this choice. That is, it is the same for all definitions of rate.

For an i.i.d. source $\{X_n\}$, the *Shannon distortion-rate function* $D(R)$ is defined as the minimum average distortion $E[d(X, Y)]$ over all conditional distributions of Y given X for which the mutual information $I(X; Y)$ is at most R, where we emphasize that X and Y are scalar variables here. In his principal result, the coding theorem for source coding with a fidelity criterion, Shannon showed that for every R, $\overline{\delta}(R) = D(R)$. That is, no VQ of any dimension k with rate R could yield smaller average distortion than $D(R)$, and that for some dimension—possibly very large—there exists a VQ with rate no greater than R and distortion very nearly $D(R)$. As an illustrative example, the Shannon distortion-rate function of an i.i.d. Gaussian source with variance σ^2 is

$$D(R) = \sigma^2 2^{-2R} \tag{16}$$

where σ^2 is the variance of the source. Equivalently, the Shannon rate-distortion function is $R(D) = \frac{1}{2}\log(\sigma^2/D)$, $0 \le D \le \sigma^2$. Since it is also known that this represents the best possible performance of any quantization scheme whatsoever, it is these formulas that we used previously when comparing the performance of scalar quantizers to that of the best quantization schemes. For example, comparing (10) and (16), one sees why we made earlier the statement that the operational distortion-rate function of scalar quantization is $\pi\sqrt{3}/2$ times larger than $\overline{\delta}(R)$. Notice that (16) shows that for this source the 2^{-2R} exponential rate of decay of distortion with rate, demonstrated by high resolution arguments for high rates, extends to all rates. This is not usually the case for other sources.

Shannon's approach was subsequently generalized to sources with memory, cf. [180], [45], [46], [218], [549], [127], [126], [282], [283], [138], and [479]. The general

definitions of distortion-rate and rate-distortion functions resemble those for operational distortion-rate and rate-distortion functions in that they are infima of kth-order functions. For example, the kth-order distortion-rate function $D_k(R)$ of a stationary random process $\{X_n\}$ is defined as an infimum of the average distortion $E[d(X, Y)]$ over all conditional probability distributions of $Y = (Y_1, Y_2, \cdots, Y_k)$ given $X = (X_1, X_2, \cdots, X_k)$ for which average mutual information $(1/k)I(X, Y) \leq R$. The distortion-rate function for the process is then given by $\overline{D}(R) = \inf_k D_k(R)$. For i.i.d. sources $\overline{D}(R) = D_1(R)$, where $D_1(R)$ is what we previously called $D(R)$ for i.i.d. sources. (The rate-distortion functions $R_k(D)$ and $\overline{R}(D)$ are defined similarly.) A source coding theorem then shows under appropriate conditions that, for sources with memory, $\overline{\delta}(R) = \overline{D}(R)$ for all rates R. In other words, Shannon's distortion-rate function represents an asymptotically achievable, but never beatable, lower bound to the performance of any VQ of any dimension. The *positive coding theorem* demonstrating that the Shannon distortion-rate function is in fact achievable if one allows codes of arbitrarily large dimension and complexity is difficult to prove, but the existence of good codes rests on the law of large numbers, suggesting that large dimensions might indeed be required for good codes, with consequently large demands on complexity, memory, and delay.

Shannon's results, like those of Panter and Dite, Zador, and Gish and Pierce provide benchmarks for comparison for quantizers. However, Shannon's results provide an interesting contrast with these early results on quantizer performance. Specifically, the early quantization theory had derived the limits of scalar quantizer performance based on the assumption of high resolution and showed that these bounds were achievable by a suitable choice of quantizer. Shannon, on the other hand, had fixed a finite, nonasymptotic rate, but had considered asymptotic limits as the dimension k of a vector quantizer was allowed to become arbitrarily large. The former asymptotics, high resolution for fixed dimension, are generally viewed as quantization theory, while the latter, fixed-rate and high dimension, are generally considered to be source coding theory or information theory. Prior to 1960, quantization had been viewed primarily as PCM, a form of analog-to-digital conversion or digital modulation, while Shannon's source coding theory was generally viewed as a mathematical approach to data compression. The first to explicitly apply Shannon's source coding theory to the problem of analog-to-digital conversion combined with digital transmission appear to be Goblick and Holsinger [205] in 1967, and the first to make explicit comparisons of quantizer performance to Shannon's rate-distortion function was Koshelev [579] in 1963.

A distinct variation on the Shannon approach was introduced to the English literature in 1956 by Kolmogorov [288], who described several results by Russian information theorists inspired by Shannon's 1948 treatment of coding with respect to a fidelity criterion. Kolmogorov considered two notions of the rate with respect to a fidelity criterion: His second notion was the same as Shannon's, where a mutual information was minimized subject to a constraint on the average distortion, in this case measured by squared error. The first peformed a similar minimization of mutual information, but with the requirement that maximum distortion between the input and reproduction did not exceed a specified level ϵ. Kolmogorov referred to both functions as the "ϵ-entropy" $H_\epsilon(X)$ of a random object X, but the name has subsequently been considered to apply to the maximum distortion being constrained to be less than ϵ, rather than the Shannon function, later called the rate-distortion function, which constrained the average distortion. Note that the maximum distortion with respect to a distortion measure d can be incorporated in the average distortion formulation if one considers a new distortion measure ρ defined by

$$\rho(x, \hat{x}) = \begin{cases} 0, & \text{if } d(x, y) \leq \epsilon \\ \infty, & \text{otherwise.} \end{cases} \tag{17}$$

As with Shannon's rate-distortion function, this was an information-theoretic definition. As with quantization, there are corresponding operational definitions. The operational epsilon entropy (ϵ-entropy) of a random variable X can be defined as the smallest entropy of a quantized output such that the reproduction is no further from the input than ϵ (at least with probability 1):

$$\mathcal{H}_\epsilon(X) = \inf_{q:\ \sup_x\ d(x, q(x)) \leq \epsilon} H(q(X)). \tag{18}$$

This is effectively a variable-rate definition since lossless coding would be required to achieve a bit rate near the entropy. Alternatively, one could define the operational epsilon entropy as $\log N_\epsilon$, where N_ϵ is the smallest number of reproduction codevectors for which all inputs are (with probability 1) within ϵ of a codevector. This quantity is clearly infinite if the random object X does not have finite support. As in the Shannon case, all these definitions can be made for k-dimensional vectors X^k and the limiting behavior can be studied. Results regarding the convergence of such limits and the equality of the information-theoretic and operational notions of epsilon entropy can be found, e.g., in [421], [420], [278], and [59]. Much of the theory is concerned with approximating epsilon entropy for small ϵ.

Epsilon entropy extends to function approximation theory with a slight change by removing the notion of probability. Here the epsilon entropy becomes the log of the smallest number of balls of radius ϵ required to cover a compact metric space (e.g., a function space—see, e.g., [520] and [420] for a discussion of various notions of epsilon entropy).

We mention epsilon entropy because of its close mathematical connection to rate-distortion theory. Our emphasis, however, is on codes that minimize average, not maximum, distortion.

The Earliest Vector Quantization Work: Outside of Shannon's sketch of rate-distortion theory in 1948, the earliest work with a definite vector quantization flavor appeared in the mathematical and statistical literature. Most important was the remarkable work of Steinhaus in 1956 [480], who considered a problem equivalent to a three-dimensional generalization of scalar quantization with a squared-error distortion measure.

Suppose that a mass density $m(x)$ is defined on Euclidean space. For any finite N, let $\mathcal{S} = \{S_i;\ i = 1, \cdots, N\}$ be a partition of Euclidean space into N disjoint bodies (cells) and let $\mathcal{C} = \{y_i;\ i = 1, \cdots, N\}$ be a collection of N vectors, one associated with each cell of the partition. What partition $\mathcal{S}$ and collection of vectors $\mathcal{C}$ minimizes

$$\sum_{i=1}^{N} \int_{S_i} m(x)\|x - y_i\|^2\, dx$$

the sum of the moments of inertia of the cells about the associated vectors? This problem is formally equivalent to a fixed-rate three-dimensional vector quantizer with a squared-error distortion measure and a probability density $m(x)/\int m(x')\,dx'$. Steinhaus derived what we now consider to be the Lloyd optimality conditions (centroid and nearest neighbor mapping) from fundamental principles (without variational techniques), proved the existence of a solution, and described the iterative descent algorithm for finding a good partition and vector collection. His derivation applies immediately to any finite-dimensional space and hence, like Lloyd's, extends immediately to vector quantization of any dimension. Steinhaus was aware of the problems with local optima, but stated that "generally" there would be a unique solution. No mention is made of "quantization," but this appears to be the first paper to both state the vector quantization problem and to provide necessary conditions for a solution, which yield a design algorithm.

In 1959, Fejes Toth described the specific application of Steinhaus' problem in two dimensions to a source with a uniform density on a bounded support region and to quantization with an asymptotically large number of points [159]. Using an earlier inequality of his [158], he showed that the optimal two-dimensional quantizer under these assumptions tessellated the support region with hexagons. This was the first evaluation of the performance of a genuinely multidimensional quantizer. It was rederived in a 1964 Bell Laboratories Technical Memorandum by Newman [385]; its first appearance in English. It made a particularly important point: even in the simple case of two independent uniform random variables, with no redundancy to remove, the performance achievable by quantizing vectors using a hexagonal-lattice encoding partition is strictly better than that achievable by uniform scalar quantization, which can be viewed as a two-dimensional quantizer with a square encoding lattice.

The first high-resolution approximations for vector quantization were published by Schutzenberger in 1958 [462], who found upper and lower bounds to the least distortion of k-dimensional variable-rate vector quantizers, both of the form $K2^{-2R}$. Unfortunately, the upper and lower bounds diverge as k increases.

In 1963, Zador [561] made a very large advance by using high-resolution methods to show that for large rates, the operational distortion-rate function of fixed-rate quantization has the form

$$\delta_k(R) \cong b_k \|f\|_{k/(k+2)} 2^{-2R} \tag{19}$$

where b_k is a term that is independent of the source, $f(x)$ is the k-dimensional source density, and

$$\|f\|_{k/(k+2)} = \left(\int f^{k/(k+2)}(x)\, dx \right)^{(k+2)/k}$$

is the term that depends on the source. This generalized the Panter–Dite formula to the vector case. While the formula for $\delta_k(R)$ obviously matches the Shannon distortion-rate function $D(R)$ when both dimension and rate are large (because in this case both are approximations to $\delta_k(R) \cong \overline{\delta}(R)$), Zador's formula has the advantage of being applicable for any dimension k while the Shannon theory is applicable only for large k. On the other hand, Shannon theory is applicable for any rate R while high resolution theory is applicable only for large rates. Thus the two theories are complementary. Zador also explicitly extended Lloyd's optimality properties to vectors with distortion measures that were integer powers of the Euclidean norm, thereby also generalizing Steinhaus' results to dimensions higher than three, but he did not specifically consider descent design algorithms. Unfortunately, the results of Zador's thesis were not published until 1982 [563] and were little known outside of Bell Laboratories until Gersho's important paper of 1979 [193], to be described later.

Zador's dissertation also dealt with the analysis of variable-rate vector quantization, but the asymptotic formula given there is not the correct one. Rather it was left to his subsequent unpublished 1966 memo [562] to derive the correct formula. (Curiously, his 1982 paper [563] reports the formula from the thesis rather than the memo.) Again using high-resolution methods, he showed that for large rates, the operational distortion-rate function of variable-rate vector quantization has the form

$$\delta_k(R) \cong c_k 2^{2h_k(X)} 2^{-2R} \tag{20}$$

where c_k is a term that is independent of the source and $h_k = (1/k)h(X_1, \cdots, X_k)$ is the dimension-normalized differential entropy of the source. This completed what he and Schutzenberger had begun.

In the mid-1960's, the optimality properties described by Steinhaus, Lloyd, and Zador and the design algorithm of Steinhaus and Lloyd were rediscovered in the statistical clustering literature. Similar algorithms were introduced in 1965 by Forgey [172], Ball and Hall [29], [230], Jancey [263], and in 1969 by MacQueen [341] (the "k-means" algorithm). These algorithms were developed for statistical clustering applications, the selection of a finite collection of templates that well represent a large collection of data in the MSE sense, i.e., a fixed-rate VQ with an MSE distortion measure in quantization terminology, cf. Anderberg [9], Diday and Simon [133], or Hartigan [238]. MacQueen used an incremental incorporation of successive samples of a training set to design the codes, each vector being first mapped into a minimum-distortion reproduction level representing a cluster, and then the level for that cluster being replaced by an adjusted centroid. Forgey and Jancey used simultaneous updates of all centroids, as did Steinhaus and Lloyd.

Unfortunately, many of these early results did not propagate among the diverse groups working on similar problems. Zador's extensions of Lloyd's results were little known outside of Bell Laboratories. The work of Steinhaus has been virtually unknown in the quantization community until recently. The work in the clustering community on what were effectively vector quantizer design algorithms in the context of statistical clustering was little known at the time in the quantization community, and it was not generally appreciated that Lloyd's algorithm was in fact a clustering algorithm. Part of the lack of interest through the 1950's was likely due to the fact that there had not yet appeared any strong motivation to consider the quantization of vectors instead of scalars. This motivation came as a result of Shannon's landmark 1959 paper on source coding with a fidelity criterion.

E. Implementable Vector Quantizers

As mentioned before, it was not evident from the earliest studies that vector quantization could be a practical technique. The only obvious encoding procedure is brute-force nearest neighbor encoding: compare the source vector to be quantized with all reproduction vectors in the codebook. Since a (fixed-rate) VQ with dimension k and rate R has 2^{kR} codevectors, the number of computations required to do this grows exponentially with the dimension-rate product kR, and gets quickly out of hand. For example, if $k = 10$ and $R = 2$, there are roughly one million codevectors. Moreover, these codevectors need to be stored, which also consumes costly resources. Finally, the proof of Shannon's source coding theorem relies on the dimension becoming large, suggesting that large dimension might be needed to attain good performance. As a point of reference, we note that in the development of channel codes, for which Shannon's theory had also suggested large dimension, it was common circa 1970 to consider channel codes with dimensions on the order of 100 or more. Thus it no doubt appeared to many that similarly large dimensions might be needed for effective quantization. Clearly, a brute-force implementation of VQ with such dimensions would be out of the question. On the other hand, the channel codes of this era with large dimension and good performance, e.g., BCH codes, were highly *structured* so that encoding and decoding need not be done by brute force.

From the above discussion, it should not be surprising that the first VQ intended as a practical technique had a reproduction codebook that was highly structured in order to reduce the complexity of encoding and decoding. Specifically, we speak of the fixed-rate vector quantizer introduced in 1965 by Dunn [137] for multidimensional i.i.d. Gaussian vectors. He argued that his code was effectively a permutation code as earlier used by Slepian [472] for channel coding, in that the reproduction codebook contains only codevectors that are permutations of each other. This leads to a quantizer with reduced (but still fairly large) complexity. Dunn compared numerical computations of the performance of this scheme to the Shannon rate-distortion function. As mentioned earlier, this was the first such comparison. In 1972, Berger, Jelinek, and Wolf [49], and Berger [47] introduced lower complexity encoding algorithms for permutation codes, and Berger [47] showed that for large dimensions, the operational distortion-rate function of permutation codes is approximately equal to that of optimal variable-rate scalar quantizers. While they do not attain performance beyond that of scalar quantization, permutation codes have the advantage of avoiding the buffering and error propagation problems of variable-rate quantization.

Notwithstanding the skepticism of some about the feasibility of brute-force unstructured vector quantization, serious studies of such began to appear in the mid-1970's, when several independent results were reported describing applications of clustering algorithms, usually k-means, to problems of vector quantization. In 1974–1975, Chaffee [76] and Chaffee and Omura [77] used clustering ideas to design a vector quantizer for very low rate speech vocoding. In 1977, Hilbert used clustering algorithms for joint image compression and image classification [242]. These papers appear to be the first applications of direct vector quantization for speech and image coding applications. Also in 1977, Chen used an algorithm equivalent to a two-dimensional Lloyd algorithm to design two-dimensional vector quantizers [87].

In 1978 and 1979, a vector extension of Lloyd's Method I was applied to linear predictive coded (LPC) speech parameters by Buzo and others [220],[67], [68], [223] with a weighted quadratic distortion measure on parameter vectors closely related to the Itakura–Saito spectral distortion measure [258], [259], [257]. Also in 1978, Adoul, Collin, and Dalle [3] used clustering ideas to design two-dimensional vector quantizers for speech coding. Caprio, Westin, and Esposito in 1978 [74] and Menez, Boeri, and Esteban in 1979 [353] also considered clustering algorithms for the design of vector quantizers with squared error and magnitude error distortion measures.

The most important paper on quantization during the 1970's was without a doubt Gersho's paper on "Asymptotically optimal block quantization" [193]. The paper popularized high resolution theory and the potential performance gains of vector quantization, provided new, simplified variations and proofs of Zador's results and vector extensions of Gish and Pierce's results with squared-error distortion, and introduced lattice vector quantization as a means of achieving the asymptotically optimal quantizer point density for entropy-constrained vector quantization for a random vector with bounded support. The simple derivations combined the vector quantizer point-density approximations with the use of Hölder's and Jensen's inequalities, generalizing a scalar quantizer technique introduced in 1977 [222]. One step of the development rested on a still unproved conjecture regarding the asymptotically optimal quantizer cell shapes and Zador's constants, a conjecture which since has borne Gersho's name and which will be considered at some length in Section IV. Portions of this work were extended to nondecreasing functions of norms in [554].

Gersho's work stimulated renewed interest in the theory and design of direct vector quantizers and demonstrated that, contrary to the common impression that very large dimensions were required, significant gains could be achieved over scalar quantization by quantizing vectors of modest dimension and,

as a result, such codes might be competitive with predictive and transform codes in some applications.

In 1980, Linde, Buzo, and Gray explicitly extended Lloyd's algorithm to vector quantizer design [318]. As we have seen, the clustering approach to vector quantizer design originated years earlier, but the Linde *et al.* paper introduced it as a direct extension to the original Lloyd optimal PCM design algorithm, extended it to more general distortion measures than had been previously considered (including an input-weighted quadratic distortion useful in speech coding), and succeeded in popularizing the algorithm to the point that it is often referred to as the "LBG algorithm." A "splitting" method for designing the quantizer from scratch was developed, wherein one first designs a quantizer with two words (2-means), then doubles the codebook size by adding a new codevector near each existing codevector, then runs Lloyd's algorithm again, and so on. The numerical examples of quantizer design complemented Gersho's high-resolution results much as Lloyd's had complemented Panter and Dite: it was shown that even with modest dimensions and modest rates, significant gains over scalar quantization could be achieved by direct vector quantization of modest complexity. Later in the same year, Buzo *et al.* [69] developed a tree-structured vector quantizer (TSVQ) for ten-dimensional LPC vectors that greatly reduced the encoder complexity from exponential growth with codebook size to linear growth by searching a sequence of small codebooks instead of a single large codebook. The result was an 800-bits/s LPC speech coder with intelligible quality comparable to that of scalar-quantized LPC speech coders of four times the rate. (See also [538].) In the same year, Adoul, Debray, and Dalle [4] also used a spectral distance measure to optimize predictors for DPCM and the first thorough study of vector quantization for image compression was published by Yamada, Fujita, and Tazaki [551].

In hindsight, the surprising effectiveness of low-dimensional VQ, e.g., $k = 2$ to 10, can be explained by the fact that in Shannon's theory large dimension is needed to attain performance arbitrarily close to the ideal. In channel coding at rates less than capacity, ideal performance means zero error probability, and large dimension is needed for codes to approach this. However, when quantizing at a given rate R, ideal performance means distortion equal to $\overline{\delta}(R)$. Since this is not zero, there is really no point to making the difference between actual and ideal performance arbitrarily small. For example, it might be enough to come within 5% to 20% (0.2 to 0.8 dB) of $\overline{\delta}(R)$, which does not require terribly large dimension. We will return to this in Section IV with estimates of the required dimension.

There followed an active period for all facets of quantization theory and design. Many of these results developed early in the decade were fortuitously grouped in the March 1982 special issue on Quantization of these TRANSACTIONS, which published the Bell Laboratories Technical Memos of Lloyd, Newman, and Zador along with Berger's extension of the optimality properties of entropy-constrained scalar quantization to rth-power distortion measures and his extensive comparison of minimum-entropy quantizers and fixed-rate permutation codes [48], generalizations by Trushkin of Fleischer's conditions for uniqueness of local optima [503], results on the asymptotic behavior of Lloyd's algorithm with training-sequence size based on the theory of k-means consistency by Pollard [418], two seminal papers on lattice quantization by Conway and Sloane [103], [104], rigorous developments of the Bennett theory for vector quantizers and rth-power distortion measures by Bucklew and Wise [64], Kieffer's demonstration of stochastic stability for a general class of feedback quantizers including the historic class of predictive quantizers and delta modulators along with adaptive generalizations [281], Kieffer's study of the convergence rate of Lloyd's algorithm [280], and the demonstration by Garey, Johnson, and Witsenhausen that the Lloyd–Max optimization was NP-hard [187].

Toward the middle of the 1980's, several tutorial articles on vector quantization appeared, which greatly increased the accessibility of the subject [195], [214], [342], [372].

F. The Mid-1980's to the Present

In the middle to late 1980's, a wide variety of vector quantizer design algorithms were developed and tested for speech, images, video, and other signal sources. Some of the quantizer design algorithms developed as alternatives to Lloyd's algorithm include simulated annealing [140], [507], [169], [289], deterministic annealing [445]–[447], pairwise nearest neighbor [146] (which had its origins in earlier clustering techniques [524]), stochastic relaxation [567], [571], self-organizing feature maps [290], [544], [545], and other neural nets [495], [301], [492], [337], [65]. A variety of quantization techniques were introduced by constraining the structure of the vector quantization to better balance complexity with performance and these methods were applied to real signals (especially speech and images) as well as to random sources, which permitted comparison to the theoretical high-resolution and Shannon bounds. The literature begins to grow too large to cite all works of possible interest, but several of the techniques will be considered in Section V. Here, we only mention several examples with references and leave further discussion to Section V.

As will be discussed in some depth in Section V, fast search algorithms were developed for unstructured reproduction codebooks, and even faster searches for reproduction codebooks constrained to have a simple structure, for example to be a subset of points of a regular lattice as in a lattice vector quantizer. Additional structure can be imposed for faster searches with virtually no loss of performance, as in Fisher's pyramid VQ [164], which takes advantage of the asymptotic equipartition property to choose a structured support region for the quantizer. Tree-structured VQ uses a tree-structured reproduction codebook with a matched tree-structured search algorithm. A tree-structured VQ with far less memory is provided by a multistage or residual VQ. A variety of product vector quantizers use a Cartesian product reproduction codebook, which often can be rapidly searched. Examples include polar vector quantizers, mean-removed vector quantizers, and shape-gain vector quantizers. Trellis encoders and trellis-coded quantizers use a Viterbi algorithm encoder matched to a reproduction codebook with a trellis structure. Hierarchical

table-lookup vector quantizers provide fixed-rate vector quantizers with minimal computational complexity. Many of the early quantization techniques, results, and applications can be found in original form in Swaszek's 1985 reprint collection on quantization [484] and Abut's 1990 IEEE Reprint Collection on Vector Quantization [2].

We close this section with a brief discussion of two specific works which deal with optimizing variable-rate scalar quantizers without additional structure, the problem that leads to the general formulation of optimal quantization in the next section. In 1984 Farvardin and Modestino [155] extended Berger's [47] necessary conditions for optimality of an entropy-constrained scalar quantizer to more general distortion measures and described two design algorithms: the first is similar to Berger's iterative algorithm, but the second was a fixed-point algorithm which can be considered as a natural extension of Lloyd's Method I from fixed-rate to variable-rate vector quantization. In 1989, Chou *et al.* [93] developed a generalized Lloyd algorithm for entropy-constrained vector quantization that generalized Berger's [47], [48] Lagrangian formulation for scalar quantization and Farvardin and Modestino's fixed-point design algorithm [155] to vectors. Optimality properties for minimizing a Lagrangian distortion $D(q) + \lambda R(q)$ were derived, where rate could be either average length or entropy. Lloyd's optimal decoder remained unchanged and the lossless code is easily seen to be an optimal lossless code for the encoded vectors, but this formulation shows that the optimal encoder must simultaneously consider both the distortion and rate resulting from the encoder. In other words, quantizers with variable rate should use an encoder that minimizes a sum of squared error and weighted bit rate, and not only the squared error. Another approach to entropy-constrained scalar quantization is described in [285].

This is a good place to again mention Gish and Pierce's result that if the rate is high, optimal entropy-constrained scalar or vector quantization can provide no more than roughly 1/4-bit improvement over uniform scalar quantization with block entropy coding. Berger [47] showed that permutation codes achieved roughly the same performance with a fixed-rate vector quantizer. Ziv [578] showed in 1985 that if subtractive dithering is allowed, dithered uniform quantization followed by block lossless encoding will be at most 0.754 bit worse than the optimal entropy-constrained vector quantizer with the same block size, even if the rate is not high. (Subtractive dithering, as will be discussed later, adds a random dither signal to the input and removes it from the decompressed output.) As previously discussed, these results do not eliminate the usefulness of fixed-rate quantizers, because they may be simpler and avoid the difficulties associated with variable-rate codes. These results do suggest, however, that uniform quantization and lossless coding is always a candidate and a benchmark for performance comparison. It is not known if the operational distortion-rate function of variable-rate quantization with dithering is better than that without dithering.

The present decade has seen continuing activity in developing high resolution theory and design algorithms for a variety of quantization structures, and in applying many of the principles of the theory to optimizing signal processing and communication systems incorporating quantizers. As the arrival of the present is a good place to close our historical tour, many results of the current decade will be sketched through the remaining sections. It is difficult to resist pointing out, however, that in 1990 Lloyd's algorithm was rediscovered in the statistical literature under the name of "principal points," which are distinguished from traditional k-means by the assumption of an absolutely continuous distribution instead of an empirical distribution [171], [496], a formulation included in the VQ formulation for a general distribution. Unfortunately, these works reflect no awareness of the rich quantization literature.

Most quantizers today are indeed uniform and scalar, but are combined with prediction or transforms. In many niche applications, however, the true vector quantizers, including lattices and other constrained code structures, exhibit advantages, including the coding of speech residuals in code excited linear predictive (CELP) speech coding systems and VXTreme/Microsoft streaming video in WebTheater. Vector quantization, unlike scalar quantization, is usually applied to digital signals, e.g., signals that have already been "finely" quantized by an A/D converter. In this case, quantization (vector or scalar) truly represents compression since it reduces the number of bits required to describe a signal and it reduces the bandwidth required to transmit the signal description if an analog link is used.

Modern video coding schemes often incorporate the Lagrangian distortion viewpoint for accomplishing rate control, while using predictive quantization in a general sense through motion compensation and uniform quantizers with optimized lossless coding of transform coefficients for the intraframe coding (cf. [201], [202]).

III. Quantization Basics: Encoding, Rate, Distortion, and Optimality

This section presents, in a self-contained manner, the basics of memoryless quantization, that is, vector quantizers which operate independently on successive vectors. For brevity, we omit the "memoryless" qualifier for most of the rest of this section. A key characteristic of any quantizer is its *dimension* k, a positive integer. Its input is a k-dimensional vector $x = (x_1, \cdots, x_k)$ from some alphabet $A \subset \Re^k$. (Abstract alphabets are also of interest in rate-distortion theory, but virtually all alphabets encountered in quantization are real-valued vector spaces, in which case the alphabet is often called the *support* of the source distribution.) If $k = 1$ the quantizer is *scalar*; otherwise, it is *vector*. In any case, the quantizer consists of three components—a *lossy encoder* $\alpha\colon A \to \mathcal{I}$, where the index set $\mathcal{I}$ is an arbitrary countable set, usually taken as a collection of consecutive integers, a *reproduction decoder* $\beta\colon \mathcal{I} \to \hat{A}$, where $\hat{A} \subset \Re^k$ is the *reproduction alphabet*, and a *lossless encoder* $\gamma\colon \mathcal{I} \to \mathcal{J}$, an invertible mapping (at least with probability 1) into a collection $\mathcal{J}$ of variable-length binary vectors that satisfies the prefix condition. Alternatively, a lossy encoder is specified by a partition $\mathcal{S} = \{S_i;\ i \in \mathcal{I}\}$ of A, where $S_i = \{x\colon \alpha(x) = i\}$; a reproduction decoder is specified by a *(reproduction) codebook*

$\mathcal{C} = \{\beta(i);\ i \in \mathcal{I}\}$ of *points*, *codevectors*, or *reproduction codewords*; and the lossless encoder γ can be described by its *binary codebook* $\mathcal{J} = \{\gamma(i);\ i \in \mathcal{I}\}$ containing *binary* or *channel codewords*. The *quantization rule* is the function $q(x) = \beta(\alpha(x))$ or, equivalently, $q(x) = \beta(i)$ whenever $x \in S_i$.

A k-dimensional quantizer is used by applying its lossy and lossless encoders, followed by the corresponding decoders, to a sequence of k-dimensional input vectors $\{\underline{x}_n;\ n = 1, 2, \cdots\}$ extracted from the data being encoded. There is not a unique way to do such vector extraction; and the design and performance of the quantizer usually depend significantly on the specific method that is used. For data that naturally forms a sequence $x_1, x_2, \cdots$ of scalar-valued samples, e.g., speech, vector extraction is almost always done by parsing the data into successive k-tuples of adjacent samples, i.e., $\underline{x}_n = (x_{(n-1)k+1}, \cdots, x_{nk})$. As an example of other possibilities, one could also extract the first k even samples, followed by the first k odd samples, the next k even samples, and so on. This subsampling could be useful for a multiresolution reconstruction, as in interpolative vector quantization [234], [194]. For other types of data there may be no canonical extraction method. For example, in stereo speech the k-dimensional vectors might consist just of left samples, or just of right samples, or half from each, or k from the left followed by k from the right, etc. Another example is grayscale imagery where the k-dimensional vectors might come from parsing the image into rectangular m-by-n blocks of pixels, where $mn = k$, or into other tiling polytopes, such as hexagons and other shapes aimed at taking advantage of the eye's insensitivity to noise along diagonals in comparison with along horizontal and vertical lines [226]. Or the vectors might come from some less regular parsing. If the image has color, with each pixel value represented by some three-dimensional vector, then k-dimensional vectors can be extracted in even more ways. And if the data is a sequence of color of images, e.g., digital video, the extraction possibilities increase immensely.[3]

There are two generic domains in which (memoryless) quantization theory, both analysis and design, can proceed. In the first, which we call the *random vector domain*, the input data, i.e., source, to be quantized is described by a fixed value of k, an alphabet $A \subset \Re^k$, and a probability distribution on A; and the quantizer must be k-dimensional. This is the case when the specific vector dimension and contents are not allowed to vary, e.g., when ten-dimensional speech parameter vectors of line spectral pairs or reflection coefficients are coded together. In the second, which we call the *random process domain*, the input data is characterized as a discrete parameter random process, i.e., a countable collection (usually infinite) of random variables; and different ways of extracting vectors from its component variables may be considered and compared, including different choices of the dimension k. As indicated above, there are in general many ways to do this. However, for concreteness and because it provides the opportunity to make some key points, whenever the random process domain is of interest in this and the next section, we focus exclusively on the canonical case where the data naturally forms a one-dimensional, scalar-valued sequence, and successive k-tuples of adjacent samples are extracted for quantization. We will also assume that the random process is stationary, unless a specific exception is made. Stationary models can easily be defined to include processes that exhibit distinct local and global stationarity properties (such as speech and images) by the use of models such as composite, hidden Markov, and mixture sources. In the random vector domain, there is no first-order stationarity assumption; e.g., the individual components within each vector need not be identically distributed. In either domain we presume that the quantizer operates on a k-dimensional random vector $X = (X_1, \cdots, X_k)$, usually assumed to be absolutely continuous so that it is described by a probability density function (pdf) $f(x)$. Densities are usually assumed to have finite variance in order to avoid technical difficulties.

Memoryless quantizers, as described here, are also referred to as "vanilla" vector quantizers or block-source codes. The alternative is a quantizer with *memory*. Memory can be incorporated in a variety of ways; it can be used separately for the lossy encoder (for example, different mappings can be used, conditional on the past) or for the lossless encoder (the index produced by a quantizer can be coded conditionally based on previous indices). We shall return to vector quantizers with memory in Section V, but our primary emphasis will remain on memoryless quantizers. We will occasionally use the term *code* as a generic substitute for *quantizer*.

The instantaneous rate of the quantizer applied to a particular input is the normalized length $r(x) = (1/k)l(\gamma(\alpha(x)))$ of the channel codeword, the number of bits per source symbol that must be sent to describe the reproduction. An important special case is when all binary codewords have the same length r, in which case the quantizer is referred to as *fixed-length* or *fixed-rate*.

To measure the quality of the reproduction, we assume the existence of a nonnegative distortion measure $d(x, \hat{x})$ which assigns a distortion or cost to the reproduction of input x by $\hat{x}$. Ideally, one would like a distortion measure that is easy to compute, useful in analysis, and perceptually meaningful in the sense that small (large) distortion means good (poor) perceived quality. No single distortion measure accomplishes all three goals, but the common squared-error distortion

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

satisfies the first two. Although much maligned for lack of perceptual meaningfulness, it often is a useful indicator of perceptual quality and, perhaps more importantly, it can be generalized to a class of distortion measures that have proved useful in perceptual coding, the input-weighted quadratic distortion measures of the form

$$d(x, \hat{x}) = (x - \hat{x})^t W_x (x - \hat{x}) \tag{21}$$

where W_x is a positive-definite matrix that depends on the input, cf. [258], [259], [257], [224], [387], [386], [150], [186], [316], [323], [325]. Most of the theory and design techniques

[3] For example, the video community has had a longstanding debate between progressive versus interlaced scanning—two different extraction methods.

considered here extend to such measures, as will be discussed later. We also assume that $d(x, \hat{x}) = 0$ if and only if $x = \hat{x}$, an assumption that involves no genuine loss of generality and allows us to consider a lossless code as a code for which $d(x, \beta(\alpha(x))) = 0$ for all inputs x.

There exists a considerable literature for various other distortion measures, including l_p and other norms of differences and convex or nondecreasing functions of norms of differences. These have rarely found application in real systems, however, so our emphasis will be on the MSE with comments on generalizations to input-weighted quadratic distortion measures.

The overall performance of a quantizer applied to a source is characterized by the normalized rate

$$R(\alpha, \gamma) = E[r(X)] = \frac{1}{k} E[l(\gamma(\alpha(X)))] = \frac{1}{k} \sum_i l(\gamma(i)) \int_{S_i} f(x)\, dx$$

and the normalized average distortion

$$D(\alpha, \beta) = \frac{1}{k} E[d(X, \beta(\alpha(X)))] = \frac{1}{k} \sum_i \int_{S_i} d(x, y_i) f(x)\, dx.$$

Every quantizer (α, γ, β) is thus described by a rate-distortion pair $(R(\alpha, \gamma), D(\alpha, \beta))$. The goal of compression system design is to optimize the rate-distortion tradeoff. Fixed-rate quantizers constrain this optimization by not allowing a code to assign fewer bits to inputs that might benefit from such, but they provide simpler codes that avoid the necessity of buffering in order to match variable-rate codewords to a possibly fixed-rate digital channel.

The optimal rate-distortion tradeoff for a fixed dimension k can be formalized in several ways: by optimizing distortion for a constrained rate, by optimizing rate for a constrained distortion, or by an unconstrained optimization using a Lagrange approach. These approaches lead, respectively, to the operational distortion-rate function

$$\delta(R) = \inf_{(\alpha, \gamma, \beta): R(\alpha, \gamma) \le R} D(\alpha, \beta)$$

the operational rate-distortion function

$$r(D) = \inf_{(\alpha, \gamma, \beta): D(\alpha, \beta) \le D} R(\alpha, \gamma)$$

and the operational Lagrangian or weighted distortion-rate function

$$L(\lambda) = \inf_{(\alpha, \gamma, \beta)} D(\alpha, \beta) + \lambda R(\alpha, \gamma)$$

where λ is a nonnegative number. A small value of λ leads to a low-distortion, high-rate solution and a large value leads to a low-rate, high-distortion solution. Note that

$$D(\alpha, \beta) + \lambda R(\alpha, \gamma) = E[d(X, \beta(\alpha(X)) + \lambda l(\gamma(\alpha(X)))]$$

so that the bracketed term can be considered to be a modified or Lagrangian distortion, and that $L(\lambda)$ is the smallest average Lagrangian distortion. All of these formalizations of optimal performance have their uses, and all are essentially equivalent: the distortion-rate and rate-distortion functions are duals and every distortion-rate pair on the convex hull of these curves corresponds to the Lagrangian for some value of λ. Note that if one constrains the problem to fixed-rate codes, then the Lagrangian approach reduces to the distortion-rate approach since $R(\alpha, \gamma)$ no longer depends on the code and γ can be considered as just a binary indexing of $\mathcal{I}$.

Formal definitions of quantizer optimality easily yield optimality conditions as direct vector extensions and variations on Lloyd's conditions. The conditions all have a common flavor: if two components of the code (α, γ, β) are fixed, then the third component must have a specific form for the code to be optimal. The resulting optimality properties are summarized below. The proofs are simple and require no calculus of variations or differentiation. Proofs may be found, e.g., in [94] and [196].

- For a fixed lossy encoder α, regardless of the lossless encoder γ, the optimal reproduction decoder β is given by
$$\beta(i) = \operatorname*{argmin}_y E[d(X, y)|\alpha(X) = i]$$
the output minimizing the conditional expectation of the distortion between the output and the input given that the encoder produced index i. These vectors are called the Lloyd centroids. Note that the optimal decoder output for a given encoder output i is simply the optimal estimate of the input vector X given $\alpha(X) = i$ in the sense of minimizing the conditional average distortion. If the distortion is squared-error, the reproduction decoder is simply the conditional expectation of X given it was encoded into i
$$\text{centroid}\,(S_i) = E[X|X \in S_i].$$
If the distortion measure is the input-weighted squared error of (21), then [318], [224]
$$\text{centroid}\,(S_i) = E[W_X|X \in S_i]^{-1} E[W_X X|X \in S_i].$$
- For a fixed lossy encoder α, regardless of the reproduction decoder β, the optimal lossless encoder γ is the optimal lossless code for the discrete source $\alpha(X)$, e.g., a Huffman code for the lossy encoded source.
- For a fixed reproduction decoder β, lossless code γ, and Lagrangian parameter λ, the optimal lossy encoder is a minimum-distortion (nearest neighbor) encoder for the modified Lagrangian distortion measure
$$\alpha(x) = \operatorname*{argmin}_{i \in \mathcal{I}} (d(x, \beta(i)) + \lambda l(\gamma(i))).$$

If the code is constrained to be fixed-rate, then the second property is irrelevant and the third property reduces to the familiar minimum distortion encoding with respect to d, as in the original formulation of Lloyd (and implicit in Shannon). (The resulting partition is often called a *Voronoi* partition.) In the general variable-rate case, the minimum distance (with respect to the distortion measure d) encoder is suboptimal;

the optimal rule takes into account both distortion and codeword length. Thus simply cascading a minimum MSE vector quantizer with a lossless code is suboptimal. Instead, in the general case, instantaneous rate should be considered in an optimal encoding, as the goal is to trade off distortion and rate in an optimal fashion. In all of these cases, the encoder can be viewed as a mechanism for controlling the output of the decoder so as to minimize the total Lagrangian distortion.

The optimality conditions imply a descent algorithm for code design: Given some λ, begin with an initial code (α, β, γ). Optimize the encoder α for the other two components, then optimize the reproduction decoder β for the remaining components, then optimize the lossless coder γ for the remaining components. Let T denote the overall transformation resulting from these three operations. One such iteration of T must decrease or leave unchanged the average Lagrangian distortion. Iterate until convergence or the improvement falls beneath some threshold. This algorithm is an extension and variation on the algorithm for optimal scalar quantizer design introduced for fixed-rate scalar quantization by Lloyd [330]. The algorithm is a fixed-point algorithm since if it converges to a code, the code must be a fixed point with respect to T. This generalized Lloyd algorithm applies to any distribution, including parametric models and empirical distributions formed from training sets of real data. There is no obvious means of choosing the "best" λ, so the design algorithm might sweep through several values to provide a choice of rate-distortion pairs. We also mention that Lloyd-style iterative algorithms have been used to design many structured forms of quantization. For example, when the codes are constrained to have fixed rate, the algorithm becomes k-means clustering, finding a fixed number of representative points that yield the minimum average distortion when a minimum distortion mapping is assumed.

As mentioned in Section I, a variety of other clustering algorithms exist that can be used to design vector quantizers (or solve any other clustering problems). Although each has found its adherents, none has convincingly yielded significant benefits over the Lloyd algorithm and its variations in terms of trading off rate and distortion, although some have proved much faster (and others much slower). Some algorithms such as simulated and deterministic annealing have been found experimentally to do a better job of avoiding local optima and finding globally optimal distortion-rate pairs than has the basic Lloyd algorithm, but repeated applications of the Lloyd algorithm with different initial conditions has also proved effective in avoiding local optima. We focus on the Lloyd algorithm because of its simplicity, its proven merit at designing codes, and because of the wealth of results regarding its convergence properties [451], [418], [108], [91], [101], [321], [335], [131], [36].

The centroid property of optimal reproduction decoders has interesting implications in the special case of a squared-error distortion measure, where it follows easily [137], [60], [193], [184], [196] that

- $E[q(X)] = E[X]$, so that the quantizer output can be considered as an unbiased estimator of the input.
- $E[q_i(X)(q_j(X) - X_j)] = 0$, for all i, j so that each component of the quantizer output is orthogonal to each component of the quantizer error. This is an example of the well-known fact that the minimum mean-squared error estimate of an unknown, X, given an observation, $\alpha(X)$, causes the estimate to be orthogonal to the error. In view of the previous property, this implies that the quantizer error is uncorrelated with the quantizer output rather than, as is often assumed, with the quantizer input.
- $E[||q(X) - X||^2] = E[||X||^2] - E[||q(X)||^2]$, which implies that the energy (or variance) of the quantized signal must be less than that in the original signal.
- $E[X^t(q(X) - X)] = -E[||q(X) - X||^2]$, which shows that the quantizer error is *not* uncorrelated with the input. In fact, the correlation is minus the mean-squared error.

It is instructive to consider the extreme points of the rate-distortion tradeoff, when the distortion is zero (or $\lambda = 0$) and the rate is 0 (when $\lambda = \infty$). First suppose that $\lambda = 0$. In this case, the rate does not affect the Lagrangian distortion at all, but MSE counts. If the source is discrete, then one can optimize this case by forcing zero distortion, that is, using a lossless code. In this case, Shannon's lossless coding theorem implies that for rate measured by average instantaneous codelength

$$H(X) \le r(0) < H(X) + 1$$

or, if rate is measured by entropy, then simply $r(0) = H(X)$, the entropy of the vector. In terms of the Lagrangian formulation, $L(0) = 0$. Conversely, suppose that $\lambda \to \infty$. In this case distortion costs a negligible amount and rate costs an enormous amount, so here the optimal is attained by using zero rate and simply tolerating whatever distortion one must suffer. The distortion for a zero-rate code is minimized by the centroid of the unconditional distribution,

$$D(0) = \min_y E[d(X, y)]$$

which is simply the mean $E[X]$ in the MSE case. Here the Lagrangian formulation becomes $L(\infty) = \min_y E[d(X, y)]$. Both of these extreme points are global optima, albeit the second is useless in practice.

So far, we have focused on the random vector domain and considered optimality for quantizers of a fixed dimension. In practice, however, and in source coding theory, the dimension k may be a parameter of choice, and it is of interest to consider how the optima depend on it. Accordingly, we now focus on the random process domain, assuming that the source is a one-dimensional, scalar-valued, stationary random process. In this situation, the various operational optima explicitly note the dimension, e.g., $\delta_k(R)$ denotes the operational distortion-rate function for dimension k and rate R and, similarly, $r_k(D)$ and $L_k(\lambda)$ denote the operational rate-distortion and Lagrange functions. Moreover, the overall optimal performance for all quantizers of rate less than or equal to R is defined by

$$\overline{\delta}(R) = \inf_k \delta_k(R). \tag{22}$$

Similar definitions hold for the rate-versus-distortion and the Lagrangian viewpoints.

Using stationarity, it can be shown (cf. [562], [577], [221], [217, Lemma 11.2.3]) that the operational distortion-rate function is *subadditive* in the sense that for any positive integers k and l

$$\delta_{k+l}(R) \le \frac{k}{k+l}\,\delta_k(R) + \frac{l}{k+l}\,\delta_l(R) \tag{23}$$

which shows the generally decreasing trend of the $\delta_k(R)$'s as k increases. It is not known whether or not $\delta_{k+1}(R)$ is always less than or equal to $\delta_k(R)$. However, it can be shown that subadditivity implies (cf. [180, p. 112])

$$\overline{\delta}(R) = \lim_{k\to\infty} \delta_k(R). \tag{24}$$

Hence high-dimensional quantizers can do as well as any quantizer. Note that (23) and (24) both hold for the special cases of fixed-rate quantizers as well as for variable-rate quantizers.

It is important to point out that for squared error and most other distortion measures, the "inf" in (22) is not a "min." Specifically, $\overline{\delta}(R)$ represents performance that cannot be achieved exactly, except in degenerate situations such as when $R = 0$ or the source distribution is discrete rather than continuous. Of course, by the infimum definition of $\overline{\delta}(R)$, there are always quantizers with performance arbitrarily close to it. We conclude that no quantizers are *truly* optimal. Thus it is essential to understand that whenever the word "optimal" is used in the random process domain, it is *always* in the context of some specific constraint or class of quantizers, such as eight-dimensional fixed-rate VQ or entropy-constrained uniform scalar quantization or pyramid coding with dimension 20, to name a few at random. Indeed, though desirable, "optimality" loses a bit of its lustre when one considers the fact that an optimal code in one class might not work as well as a suboptimal code in another. It should now be evident that the importance of the Lloyd-style optimality principles lies ultimately in their ability to guide the optimization of quantizers within specific constraints or classes.

IV. High Resolution Quantization Theory

This section presents an overview of high resolution theory and compares its results to those of Shannon rate-distortion theory. For simplicity, we will adopt squared error as the distortion measure until late in the section, where extensions to other distortion measures are discussed. There have been two styles of high resolution theory developments: informal, where simple approximations are made, and rigorous, where limiting formulas are rigorously derived. Here, we proceed with the informal style until later when the results of the rigorous approach are summarized. We will also presume the "random vector domain" of fixed dimension, as described in the previous section, until stated otherwise.

A. Asymptotic Distortion

As mentioned earlier, the first and most elementary result in high resolution theory is the $\Delta^2/12$ approximation to the mean-squared error of a uniform scalar quantizer with step size Δ [43], [394], [468], which we now derive. Consider an N-level uniform quantizer q whose levels are $y_1, \cdots, y_N$, with $y_i = y_{i-1} + \Delta$. When this quantizer is applied to a continuous random variable X with probability density $f(x)$, when Δ is small, and when overload distortion can be ignored, the mean-squared error (MSE) distortion may be approximated as follows:

$$\begin{aligned} D(q) &= E[(X - q(X))^2] \\ &\cong \sum_{i=1}^{N} \int_{y_i-\Delta/2}^{y_i+\Delta/2} (x - y_i)^2 f(x)\,dx \\ &\cong \sum_{i=1}^{N} f(y_i) \int_{y_i-\Delta/2}^{y_i+\Delta/2} (x - y_i)^2\,dx \\ &= \frac{\Delta^2}{12} \sum_{i=1}^{N} f(y_i)\Delta \\ &\cong \frac{\Delta^2}{12} \int_{y_1-\Delta/2}^{y_N+\Delta/2} f(x)\,dx \\ &\cong \frac{\Delta^2}{12}. \end{aligned}$$

The first approximation in the above derives from ignoring overload distortion. If the source density is entirely contained in the granular region of the quantizer, then this approximation is not needed. The second approximation derives from observing that the density may be approximated as a constant on a small interval. Usually, as in the mean value theorem of integration, one assumes the density is continuous, but as any measurable function is approximately continuous, when Δ is sufficiently small this approximation is valid even for discontinuous densities. The third approximation derives from recognizing that by the definition of a Riemann integral, $\sum_{i=1}^{N} f(y_i)\Delta$ is approximately equal to the integral of f. Finally, the last approximation derives from again ignoring the overload region. As mentioned in earlier sections, there are situations, such as variable-rate quantization, where an infinite number of levels are permitted. In such cases, if the support of the uniform scalar quantizer contains that of the source density, then there will be no overload distortion to ignore, and again we have $D \cong \Delta^2/12$.

It is important to mention the sense in which D is approximated by $\Delta^2/12$. After all, when Δ is small, both D and $\Delta^2/12$ will be small, so it is not saying much to assert that their difference is small. Rather, as discussed later in the context of the rigorous framework for high resolution theory, it can be shown that under ordinary conditions, the ratio of D and $\Delta^2/12$ tends to 1 as Δ decreases. Though we will not generally mention it, all future high-resolution approximations discussed in this paper will also hold in this ratio-tending-to-one sense.

Each of the assumptions and simple approximations made in deriving $\Delta^2/12$ reoccurs in some guise in the derivation of all subsequent high-resolution formulas, such as for nonuniform, vector, and variable-rate quantizers. Thus they might be said to be principal suppositions. Indeed, the small cell type of supposition is what gives the theory its "high resolution" name.

In uniform quantization, all cells have the same size and shape and the levels are in the center of each cell (except for the outermost cells which are ignored). Thus the cell size Δ is the key performance determining gross characteristic. In more advanced, e.g., vector, quantization, cells may differ in size and shape, and the codevectors need not be in the centers of the cells. Consequently, other gross characterizations are needed. These are the *point density* and the *inertial profile*.

The point density of a vector quantizer is the direct extension of the point density introduced in Section II. That is, it is a nonnegative, usually smooth function $\lambda(x)$ that, when integrated over a region, determines the approximate fraction of codevectors contained in that region. In fixed-rate coding, the point density is usually normalized by the number of codevectors so that its total integral is one. In variable-rate coding, where the number of codevectors is not a key performance-determining parameter and may even be infinite, the point density is usually left unnormalized. As we consider fixed-rate coding first, we will presume λ is normalized, until stated otherwise. There is clearly an inverse relationship between the point density and the volume of cells, namely, $\lambda(x) \cong (N \operatorname{vol}(S_x))^{-1}$, where, as before, N is the number of codevectors or cells and S_x denotes the cell containing x.

As with any density that describes a discrete set of points, there is no unique way to define it for a specific quantizer. Rather, the point density is intended as a high-level gross characterization, or a model or target to which a quantizer aspires. It describes the codevectors, in much the way that a probability density describes a set of data points—it does not say exactly where they are located, but roughly characterizes their distribution. Quantizers with different numbers of codevectors can be compared on the basis of their point density, and there is an ideal point density to which quantizers aspire—they cannot achieve it exactly, but may approximate it. Nevertheless, there are times when a concrete definition of the point density of a specific quantizer is needed. In such cases, the following is often used: the *specific point density* of a quantizer q is $\lambda_q(x) \equiv (N \operatorname{vol}(S_x))^{-1}$. This piecewise-constant function captures all the (fine) detail in the quantizer's partition, in contrast to the usual notion of a point density as a gross characterization. As an example of its use, we mention that for fixed-rate quantization, the ideal point density $\lambda(x)$ is usually a smooth function, closely related to the source density, and one may say that a quantizer has point density approximately $\lambda(x)$ if $\lambda_q(x) \cong \lambda(x)$ for all x in some set with high probability (relative to the source density). When a scalar quantizer is implemented as a compander, $\lambda(x)$ is proportional to the derivative of the compressor function applied to the input. Though the notion of point density would no doubt have been recognizable to the earliest contributors such as Bennett, Panter, and Dite, as mentioned earlier, it was not explicitly introduced until Lloyd's work [330].

In nonuniform scalar quantization and vector quantization, there is the additional issue of codevector placement within cells and, in the latter case, of cell shape. The effect of point placement and cell shape is exhibited in the following approximation to the contribution of a small cell S_i with codevector y_i to the MSE of a k-dimensional vector quantizer

$$D_i(q) = \frac{1}{k} \int_{S_i} \|x - y_i\|^2 f(x)\, dx \tag{25}$$

$$\cong f(y_i) M(S_i, y_i) \operatorname{vol}(S_i)^{1+2/k} \tag{26}$$

where $M(S_i, y_i)$ is the normalized moment of inertia of the cell S_i about the point y_i, defined by

$$M(S_i, y_i) \equiv \frac{1}{k} \frac{1}{\operatorname{vol}(S_i)^{1+2/k}} \int_{S_i} \|x - y_i\|^2\, dx.$$

Normalizing by volume makes M independent of the size of the cell. Normalizing by dimension yields a kind of invariance to dimension, namely, that $M(S_i \times S_i, (y_i, y_i)) = M(S_i, y_i)$. We often write $M(S_i)$ when y_i is clear from the context. The normalized moment of inertia, and the resulting contribution $D_i(q)$, is smaller for sphere-like cells with codevectors in the center than for cells that are oblong, have sharply pointed vertices, or have displaced codevectors. In the latter cases, there are more points farther from y_i that contribute substantially to normalized moment of inertia, especially when dimension is large.

In some quantizers, such as uniform scalar and lattice quantizers, all cells (with the exception of the outermost cells) have the same shape and the same placement of codevectors within cells. In other quantizers, however, cell shape or codevector placement varies with position. In such cases, it is useful to characterize the variation of cell normalized moment of inertia by a nonnegative, usually smooth function $m(x)$, called the *inertial profile*. That is, $m(x) \cong M(S_i, y_i)$ when $x \in S_i$. As with point densities, we do not define $m(x)$ to be equal to $M(S_x, q(x))$, because we want it to be a high-level gross characterization or model to which a quantizer aspires. Instead, we let $m_q(x) \equiv M(S_x, q(x))$ be called the *specific inertial profile* of the quantizer q. This is a piecewise-constant function that captures the fine details of cell normalized moment of inertia.

Returning to $D_i(q)$ expressed in (26), the effect of cell size is obviously in the term $\operatorname{vol}(S_i)$. Using the inverse relationship between point density and cell volume yields

$$D_i(q) \cong \frac{1}{N^{2/k}} f(y_i) \frac{M(S_i, y_i)}{\lambda^{2/k}(y_i)} \operatorname{vol}(S_i)$$

which shows how point density locally influences distortion. Summing the above over all cells and recognizing the sum as an approximation to an integral yields the following approximation to the distortion of a vector quantizer:

$$D(q) \cong \frac{1}{N^{2/k}} \int \frac{m(x)}{\lambda^{2/k}(x)} f(x)\, dx. \tag{27}$$

For scalar quantizers $(k = 1)$ with points in the middle of the cells, $m(x) = 1/12$ and the above reduces to

$$D(q) \cong \frac{1}{12} \frac{1}{N^2} \int \frac{1}{\lambda^2(x)} f(x)\, dx \tag{28}$$

which is what Bennett [43] found for companders, as restated in terms of point densities by Lloyd [330]. Both (28) and the more general formula (27) are called *Bennett's integral*. The

extension of Bennett's integral to vector quantizers was first made by Gersho (1979) [193] for quantizers with congruent cells for which the concept of inertial profile was not needed, and then to vector quantizers with varying cell shapes (and codevector placements) by Na and Neuhoff (1995) [365].

Bennett's integral (27) can be expected to be a good approximation under the following conditions: i) Most cells are small enough that $f(x)$ can be approximated as being constant over the cell. (There can be some large cells where $f(x)$ is very small.) Ordinarily, this requires N to be large. ii) The specific point density of the quantizer approximately equals $\lambda(x)$ on a high probability set of x's. iii) The specific inertial profile approximately equals $m(x)$ on a high probability set of x's. iv) Adjacent cells have similar volumes. The last condition rules out quantizers such as a scalar one whose cells have alternating lengths such as Δ, $\frac{1}{2}\Delta$, $\frac{1}{2}\Delta$, Δ, $\frac{1}{2}\Delta$, $\frac{1}{2}\Delta$, Δ, $\cdots$. The point density of such a quantizer is $\lambda(x) = 3/(2\Delta N)$, because there are three points in an interval of width 2Δ. Assuming, for simplicity, that the source density is uniform on $[0, 1]$, it is easy to compute $D = (5/96)\Delta^2$, whereas Bennett's integral equals $(1/27)\Delta^2$. One may obtain the correct distortion by separately applying Bennett's integral to the union of intervals of length Δ and to the union of intervals of length $\frac{1}{2}\Delta$. The problem is that Bennett's integral is not linear in the point density. So for it to be accurate, cell size must change slowly or only occasionally. Since Bennett's integral is linear in the inertial profile, it is not necessary to assume that adjacent cells have similar shapes, although one would normally expect this to be the case in situations where Bennett's integral is applied. Examples of the use of the vector extension of Bennett's integral will be given later.

Approximating the source density as a constant over each quantization cell, which is a key step in the derivations of (26) and (28), is like assuming that the effect of quantization is to add noise that is uniformly distributed. However, the range of noise values must match the size and shape of the cell. And so when the cells are not all of the same size and shape, such quantization noise is obviously correlated with the vector X being quantized. On the other hand, for uniform scalar and lattice vector quantizers, the error and X are approximately uncorrelated. A more general result, mentioned in Section III, is that the correlation between the input and the quantization error is approximately equal to the MSE of the quantizer when the codevectors are approximately centroids.

B. Performance of the Best k-Dimensional, Fixed-Rate Quantizers

Having Bennett's integral for distortion, one can hope to find a formula for $\delta_k(R)$, the operational distortion-rate function for k-dimensional, fixed-rate vector quantization, by choosing the key characteristics, point density and inertial profile, to minimize (27). Unfortunately, it is not known how to find the best inertial profile. Indeed, it is not even known what functions are allowable as inertial profiles. However, Gersho (1979) [193] made the now widely accepted conjecture that when rate is large, most cells of a k-dimensional quantizer with rate R and minimum or nearly minimum MSE are approximately congruent to some basic tessellating[4] k-dimensional cell shape T_k. In this case, the optimum inertial profile is a constant and Bennett's integral can be minimized by variational techniques or Hölder's inequality [193], [222], resulting in the optimal point density

$$\lambda_k^*(x) = \frac{f^{k/(k+2)}(x)}{\int f^{k/(k+2)}(x')\,dx'} \tag{29}$$

and the following approximation to the operational distortion-rate function: for large R

$$\delta_k(R) \cong M_k \beta_k \sigma^2 2^{-2R} \equiv Z_k(R) \tag{30}$$

where $M_k \equiv M(T_k)$, which is the least normalized moment of inertia of k-dimensional tessellating polytopes, and

$$\beta_k \equiv \frac{1}{\sigma^2}\left(\int f^{k/(k+2)}(x)\,dx\right)^{(k+2)/k}$$

is the term depending on the source distribution. Dividing by variance makes β_k invariant to a scaling of the source. We will refer to M_k, β_k, and $Z_k(R)$ as, respectively, Gersho's constant (in dimension k), Zador's factor (for k-dimensional, fixed-rate quantization), and the Zador–Gersho function (for k-dimensional, fixed-rate quantization). (Zador's role will be described later.) When $k = 1$, $Z_1(R)$ reduces to the Panter–Dite formula (8).

From the form of $\lambda_k^*(x)$ one may straightforwardly deduce that cells are smaller and have higher probability where $f(x)$ is larger, and that all cells contribute roughly the same to the distortion; i.e., $D_i(q)$ in (26) is approximately the same for all i, which is the "partial distortion theorem" first deduced for scalar quantization by Panter and Dite.

A number of properties of M_k and β_k are known; here, we mention just a few. Gersho's constant M_k is known only for $k = 1$ and 2, where T_k is, respectively, an interval and a regular hexagon. It is not known whether the M_k's are monotonically nonincreasing for all k, but it can be shown that they form a subadditive sequence, which is a property strong enough to imply that the infimum over k equals the limit as k tends to infinity. Though it has long been presumed, only recently has it been directly shown that the M_k's tend to $1/2\pi e$ as k increases (Zamir and Feder [564]), which is the limit of the normalized moment of inertia of k-dimensional spheres as k tends to infinity. Previously, the assertion that the M_k's tend to $1/2\pi e$ depended on Gersho's conjecture. Zador's factor β_k tends to be smaller for source densities that are more "compact" (lighter tails and more uniform) and have more dependence among the source variables.

Fortunately, high resolution theory need not rely solely on Gersho's conjecture, because Zador's dissertation [561] and subsequent memo [562] showed that for large rate $\delta(R)$ has the form $b_k \beta_k \sigma^2 2^{-2R}$, where b_k is independent of the

[4] A cell T "tessellates" if there exists a partition of $\Re^k$ whose cells are, entirely, translations and rotations of T. The Voronoi cell of any lattice tessellates, but not all tessellations are generated by lattices. Gersho also conjectured that T_k would be *admissible* in the sense that the Voronoi partition for the centroids of the tessellation would coincide with the tessellation. But this is not essential.

source distribution. Thus Gersho's conjecture is really just a conjecture about b_k.

In deriving the key result, Zador first showed that for a random vector that is uniformly distributed on the unit cube, $\delta(R)$ has the form $b_k 2^{-2R}$ when R is large, which effectively defines b_k. (In this case, $\beta_k \sigma^2 = 1$.) He then used this to prove the general result by showing that no quantizer with high rate could do better than one whose partition is hierarchically constructed by partitioning $\Re^k$ into small equally sized cubes and then subdividing each with the partition of the quantizer that is best for a uniform distribution on that cube, where the number of cells within each cube depends on the source density in that cube. In other words, the local structure of an asymptotically optimal quantizer can be that of the optimum quantizer for a uniform distribution.

In this light, Gersho's conjecture is true if and only if. at high rates. one may obtain an asymptotically optimal quantizer for a uniform distribution by tessellating with T_k. The latter statement has been proven for $k = 1$ (cf. [106, p. 59]) and for $k = 2$ by Fejes Toth (1959) [159]; see also [385]. For $k = 3$, it is known that the best lattice tessellation is the body-centered cubic lattice, which is generated by a truncated octahedron [35]. It has not been proven that this is the best tessellation, though one would suspect that it is. In summary, Gersho's conjecture is known to be true only for $k = 1$ and 2. Might it be false for $k \geq 3$? If it is, it might be that the best quantizers for a uniform source have a *periodic* tessellation in which two or more cell shapes alternate in a periodic fashion, like the hexagons and pentagons on the surface of a soccer ball. If the cells in one period of the tessellation have the same volumes, then one may apply Bennett's integral, and (30) holds with M_k replaced by the average of the normalized moment of inertia of the cells in one period. However, if the cells have unequal volumes, then as in the example given while discussing Condition iv) of Bennett's integral, the MSE will be the average of distortions computed by using Bennett's integral separately on the union of cells of each type, and a macrolevel definition of M_k will be needed. It might also be that the structure of optimal quantizers is aperiodic. However, it seems likely to us that, asymptotically, one could always find a quantizer with a periodic structure that is essentially as good as any aperiodic one.

It is an open question in dimensions three and above whether the best tessellation is a lattice. In most dimensions, the best known tessellation is a lattice. However, tessellations that are better than the best known lattices have recently been found for dimensions seven and nine by Agrell and Eriksson [149].

From now on, we shall proceed assuming Gersho's conjecture is correct, with the knowledge that if this is not the case, then analyses based on M_k will be wrong (for $k \geq 3$) by the factor M_k/b_k, which will be larger than 1 (but probably not much larger), and which in any case will converge to one as $k \to \infty$, as discussed later.

C. Performance of the Best k-Dimensional, Variable-Rate Quantizers

Extensions of high resolution theory to variable-rate quantization can also be based on Bennett's integral, as well as approximations, originally due to Gish and Pierce [204], to the entropy of the output of a quantizer. Two such approximations, which can be derived using approximations much like those used to derive Bennett's integral, were stated earlier for scalar quantizers in (11) and (13). However, the approximation (13), which says that for quantizers with mostly small cells $H(q) \cong h(X) + E[\log \Lambda(X)]$, where $\Lambda(x)$ is the unnormalized point density, holds equally well for vector quantizers, when X is interpreted as a vector rather than a scalar variable. As mentioned before, unnormalized point density is used because with variable-rate quantization, the number of codevectors is not a primary characteristic and may even be infinite. For example, one can always add levels in a way that has negligible impact on the distortion and entropy.

We could now proceed to use Bennett's integral and the entropy approximation to find the operational distortion-rate function for variable-rate, k-dimensional, memoryless VQ. However, we wish to consider a somewhat more general case. Just as Gish and Pierce found something quite interesting by examining the best possible performance of scalar quantization with *block* entropy coding, we will now consider the operational distortion-rate function for vector quantization with block entropy coding. Specifically, we seek $\delta_{k,L}(R)$, which is defined to be the infimum of the distortions of any quantizer with rate R or less, whose lossy encoder is k-dimensional and memoryless, and whose lossless encoder simultaneously codes a block of L successive quantization indices with a variable-length prefix code. In effect, the overall code is a kL-dimensional, memoryless VQ. However, we will refer to it as a k-dimensional (memoryless) quantizer with Lth-order variable-length coding (or Lth-order entropy coding). When $L = 1$, the code becomes a conventional memoryless, variable-rate vector quantizer. It is convenient to let $L = 0$ connote fixed-length coding, so that $\delta_{k,0}(R)$ means the same as $\delta_k(R)$ of the previous section. By finding high-resolution approximations to $\delta_{k,L}(R)$ for all values of $k \geq 1$ and $L \geq 0$, we will be able to compare the advantages of increasing the dimension k of the quantizer to those of increasing the order L of the entropy coder.

To find $\delta_{k,L}(R)$ we assume that the source produces a sequence $(\underline{X}_1, \cdots, \underline{X}_L)$ of identical, but not necessarily independent, k-dimensional random vectors, each with density $f(x)$. A straightforward generalization of (13) shows that under high-resolution conditions, the rate is given by

$$R \cong \frac{1}{kL} h(X_1, \cdots, X_{kL}) + \frac{1}{k} \int f(x) \log \Lambda(x)\, dx. \quad (31)$$

On the other hand, the distortion of such a code may be approximated using Bennett's integral (27), with $\Lambda(x)/N^{2/k}$ substituted for the normalized point density $\lambda(x)$. Then, as with fixed-rate vector quantization, one would like to find $\delta_{k,L}(R)$ by choosing the inertial profile m and the point density Λ to minimize Bennett's integral subject to a constraint on the rate that the right-hand side of (31) be at most R.

Once again, though it is not known how to find the best inertial profile, Gersho's conjecture suggests that when rate is large, the cells of the best rate-constrained quantizers are, mostly, congruent to T_k. Hence, from now on we shall assume

that the inertial profile of the best variable-rate quantizers is, approximately, $m(x) = M_k$. In this case, using variational techniques or simply Jensen's inequality, one can show that the best point density is uniform on all of $\Re^k$ (or at least over the support of the source density). In other words, all quantizer cells have the same size, as in a tessellation. Using this fact along with (27) and (31) yields

$$\delta_{k,\,L}(R) \cong M_k \gamma_{kL} \sigma^2 2^{-2R} \equiv Z_{k,\,L}(R) \tag{32}$$

where

$$\gamma_k \equiv \frac{1}{\sigma^2}\, 2^{2(1/k)h(X_1, \cdots, X_k)}$$

is the term depending on the source distribution. Dividing by variance makes it invariant to scale. We call γ_k the (kth-order) Zador entropy factor and $Z_{k,\,L}(R)$ a Zador–Gersho function for variable-rate coding. Since fixed-rate coding is a special case of variable-length coding, it must be that γ_k is less than or equal to β_k in (30). This can be directly verified using Jensen's inequality [193].

In the case of scalar quantization ($k = 1$), the optimality of the uniform point density and the operational distortion-rate function $\delta_{1,\,L}(R)$ were found by Gish and Pierce (1968) [204]. Zador (1966) [562] considered the $L = 1$ case and showed that $\delta_{k,\,1}(R)$ has the form $c_k \gamma_k \sigma^2 2^{-2R}$ when R is large, where c_k is a constant that is independent of the source density and no larger than the constant b_k that he found for fixed-rate quantization. Gersho [193] used the argument given above to find the form of $\delta_{k,\,1}(R)$ given in (32).

As with fixed-rate quantization, we shall proceed under the assumption that Gersho's conjecture is correct, in which case $c_k = b_k = M_k$. If it is wrong, then our analyses will be off by the factor M_k/c_k, which, as before, will probably be just a little larger than one, and which in any case will converge to one as $k \to \infty$.

D. Fixed-Rate Quantization with Arbitrary Dimension

We now restrict attention to the random process domain wherein the source is assumed to be a one-dimensional, scalar-valued, stationary random process. We seek a high-resolution approximation to the operational distortion-rate function $\overline{\delta}(R) \equiv \inf_k \delta_k(R)$, which represents the best possible performance of any fixed-rate (memoryless) quantizer. As mentioned in Section III, for stationary sources $\overline{\delta}(R) = \lim_{k\to\infty} \delta_k(R)$. Therefore, taking the limit of the high-resolution approximation (30) for $\delta_k(R)$ yields the fact that for large R

$$\overline{\delta}(R) \cong \overline{M}\,\overline{\beta}\sigma^2 2^{-2R} \equiv \overline{Z}(R) \tag{33}$$

where

$$\overline{M} = \lim_{k\to\infty} M_k = \frac{1}{2}\pi e$$
$$\overline{\beta} = \lim_{k\to\infty} \beta_k$$

and

$$\overline{Z}(R) \equiv \lim_{k\to\infty} Z_k(R)$$

is another Zador–Gersho function. This operational distortion-rate function was also derived by Zador [561], who showed that his unknown factors b_k and c_k converged to $1/2\pi e$. The derivation given here is due to Gersho [193]. Notice that in this limiting case, there is no doubt about the constant $\overline{M}$.

As previously mentioned, the M_k's are subadditive, so that they are smallest when k is large. Similarly, for stationary sources it can be shown that the sequence $\{\log \beta_k\}$ is also subadditive [193], so that they too are smallest when k is large. Therefore, another expression for the above Zador–Gersho function is $\overline{Z}(R) = \inf_k Z_k(R)$.

E. The Benefits of Increasing Dimension in Fixed-Rate Quantization

Continuing in the random process domain (stationary sources), the generally decreasing natures of M_k and β_k directly quantify the benefits of increasing dimension in fixed-rate quantization. (Of course, there is also a cost to increasing dimension, namely, the increase in complexity.) For example, M_k decreases from $1/12 = 0.0833$ for $k = 1$ to the limit $1/2\pi e = 0.0586$. In decibels, this represents a 1.53-dB decrease in MSE. For an i.i.d. Gaussian source, β_k decreases from $6\sqrt{3}\pi = 32.6$ for $k = 1$ to the limit $2\pi e = 17.1$, which represents an additional 2.81-dB gain. In total, high-dimensional quantization gains 4.35 dB over scalar quantization for the i.i.d. Gaussian source. For a Gauss–Markov source with correlation coefficient $\rho = 0.9$, β_k decreases from $6\sqrt{3}\pi = 32.6$ for $k = 1$ to the limit $2\pi e(1 - \rho^2) = 3.25$ or a gain of 10.0 dB, yielding a total high-dimensional VQ gain of 11.5 dB over scalar quantization. Because of the 6-dB-per-bit rule, any gain stated in decibels can be translated to a reduction in rate (bits per sample) by dividing by 6.02.

On the other hand, it is also important to understand what specific characteristics of vector quantizers improve with dimension and by how much. Motivated by several prior explanations [342], [333], [365], we offer the following. We wish to compare an optimal quantizer q_k with dimension k to an optimal k'-dimensional quantizer $q_{k'}$ with $k' \gg k$. To simplify the discussion, assume k' is a multiple of k. Though these two quantizers have differing dimensions, their characteristics can be fairly compared by comparing $q_{k'}$ to the "product" VQ $q_{\mathrm{pr},\,k'}$ that is implicitly formed when q_k is used k'/k times in succession. Specifically, the product quantizer has quantization rule

$$q_{\mathrm{pr},\,k'}(x) = (q_k(\underline{x}_1), \cdots, q_k(\underline{x}_{k'/k}))$$

where $\underline{x}_1, \cdots, \underline{x}_{k'/k}$ are the successive k-tuples of x, and reproduction codebook $\mathcal{C}_{\mathrm{pr},\,k'}$ consisting of the concatenations of all possible sequences of k'/k codevectors from q_k's reproduction codebook $\mathcal{C}_k$. The subscripts "k" and "$\mathrm{pr},\,k'$" will be attached as needed to associate the appropriate features with the appropriate quantizer. The distortion and rate of the product quantizer are easily seen to be those of the k-dimensional VQ. Thus the shortcomings of an optimal k-dimensional quantizer relative to an optimal high-dimensional quantizer may be identified with those of the product quantizer—in particular,

with the latter's suboptimal point density and inertial profile, which we now find.

To simplify discussion, assume for now that $k = 1$, and let q_1 be a fixed-rate scalar quantizer, with large rate, levels in the middle of the cells, and point density $\lambda_{\rm sq}(x_1)$. The cells of the product quantizer $q_{{\rm pr},k'}$ are k'-dimensional rectangles formed by Cartesian products of cells from the scalar quantizer. When the scalar cells have the same width, a k'-dimensional cube is formed; otherwise, a rectangle is formed, i.e., an "oblong" cube. Since the widths of the cells are, approximately, determined by $\lambda_{\rm sq}(x_1)$, the point density and inertial profile of $q_{{\rm pr},k'}$ are determined by $\lambda_{\rm sq}$. Specifically, from the rectangular nature of the product cells one obtains [365], [378]

$$\lambda_{{\rm pr},k'}(x) = \prod_{i=1}^{k'} \lambda_{\rm sq}(x_i) \tag{34}$$

and

$$m_{{\rm pr},k'}(x) = \frac{1}{12} \frac{\frac{1}{k'}\sum_{i=1}^{k'} \frac{1}{\lambda_{\rm sq}^2(x_i)}}{\left(\prod_{i=1}^{k'} \frac{1}{\lambda_{\rm sq}^2(x_i)}\right)^{1/k'}} \tag{35}$$

which derive, respectively, from the facts that the volume of a rectangle is the product of its side lengths, that the normalized moment of inertia of a rectangle is that of a cube $(1/12)$ times the ratio of the arithmetic mean of the square of the side lengths to their geometric mean, and that the side lengths are determined by the scalar point density. Note that along the diagonal of the first "quadrant" (where $x_1 = x_2 = \cdots = x_{k'}$), the product cells are cubes and $m_{{\rm pr},k'}(x) = 1/12$, the minimum value. Off the diagonal, the cells are usually rectangular and, consequently, $m_{{\rm pr},k'}(x)$ is larger.

To quantify the suboptimality of the product quantizer's principal feature, we factor the ratio of the distortions of $q_{{\rm pr},k'}(x)$ and $q_{k'}$, which is a kind of loss, into terms that reflect the loss due to the inertial profile and point density [365], [378][5]

$$\begin{aligned} L &= \frac{D(q_{{\rm pr},k'})}{\delta_{k'}(R)} \cong \frac{B(k', m_{{\rm pr},k'}, \lambda_{{\rm pr},k'}, f)}{B(k', M_{k'}, \lambda_{k'}^*, f)} \\ &= \underbrace{\frac{B(k', m_{{\rm pr},k'}, \lambda_{{\rm pr},k'}, f)}{B(k', M_{k'}, \lambda_{{\rm pr},k'}, f)}}_{L_{\rm ce}} \times \underbrace{\frac{B(k', M_{k'}, \lambda_{{\rm pr},k'}, f)}{B(k', M_{k'}, \lambda_{k'}^*, f)}}_{L_{\rm pt}} \end{aligned} \tag{36}$$

where

$$B(k, m, \lambda, f) \equiv \int \frac{m(x)}{\lambda^{2/k}(x)} f(x)\, dx$$

is the part of Bennett's integral that does not depend on N, where the *cell-shape loss*, $L_{\rm ce}$, is the ratio of the distortion of the product quantizer to that of a hypothetical quantizer with same point density and an optimal inertial profile, and where the *point-density loss*, $L_{\rm pt}$, is the ratio of the distortion of a hypothetical quantizer with the point density of the product quantizer and a constant (e.g., optimal) inertial profile to that of a hypothetical quantizer with an optimal point density and the same (constant) inertial profile. Substituting (35) into (36) and using the fact that for large k', $M_{k'} \cong 1/2\pi e$, one finds

$$\begin{aligned} L &= \frac{\pi e}{6} \times \frac{\int \frac{1}{k'}\sum_{i=1}^{k'} \frac{1}{\lambda_{\rm sq}^2(x_i)} f_{k'}(x)\, dx}{\int \frac{1}{\left(\prod_{i=1}^{k'} \lambda_{\rm sq}^2(x_i)\right)^{1/k'}} f_{k'}(x)\, dx} \\ &\quad \times \frac{\int \frac{1}{\prod_{i=1}^{k'} \lambda_{\rm sq}^{2/k'}(x_i)} f_{k'}(x)\, dx}{\int \frac{1}{\lambda_{k'}^{*2/k'}(x)} f_{k'}(x)\, dx} \\ &= L_{\rm sp} \times L_{\rm ob} \times L_{\rm pt} \end{aligned} \tag{37}$$

where the cell shape loss has been factored into the product of a *space-filling loss* [333],[6] $L_{\rm sp}$, which is the ratio of the normalized moment of inertia of a cube to that of a high-dimensional sphere, and an *oblongitis* loss, $L_{\rm ob}$, which is the factor by which the rectangularity of the cells makes the cell shape loss larger than the space-filling loss.

To proceed further, consider first an i.i.d. source (stationary and memoryless) and consider how to choose the scalar point density $\lambda_{\rm sq}(x_1)$ in order to minimize L. On the one hand, choosing $\lambda_{\rm sq}(x_1)$ to be uniform on the set where the one-dimensional density[7] $f_1(x_1)$ is not small causes the product cells in the region where the k'-dimensional density $f_{k'}(x)$ is not small to be cubes and, consequently, makes $L_{\rm ob} \cong 1$, which is the smallest possible value. However, it causes the product point density to be poorly matched to the source density and, as a result, $L_{\rm pt}$ is large. On the other hand, choosing $\lambda_{\rm sq}(x_1) \cong f_1(x_1)$ causes the product quantizer to have, approximately, the optimal point density[8]

$$\lambda_{{\rm pr},k'}(x) \cong \prod_{i=1}^{k'} f_1(x_i) = f_{k'}(x) \cong \lambda_{k'}^*(x)$$

where the last step uses the fact that k' is large. However, this choice causes $L_{\rm ob}$ to be infinite.[9] The best point density, as implicitly found by Panter and Dite, is the compromise

$$\lambda_1^*(x_1) = \frac{f_1^{1/3}(x_1)}{\int f_1^{1/3}(u)\, du}$$

as given in (29). In the region where $f_1(x_1)$ is not small, $\lambda_1^*(x_1)$ is "more uniform" than $\lambda_1(x_1) = f_1(x_1)$ that causes

[5] Na and Neuhoff considered the ratio of the product code distortion to that of an optimal k-dimensional VQ for arbitrary k, not just for large k.

[6] Actually, Lookabaugh and Gray defined the inverse as a vector quantizer *advantage*. The space-filling loss was called a *cubic* loss in [365].

[7] Dimension will be added as a subscript to f in places where the dimension of X needs to be emphasized.

[8] The fact that product quantizers can have the optimal point density is often overlooked.

[9] This implies that distortion will not decrease as 2^{-2R}.

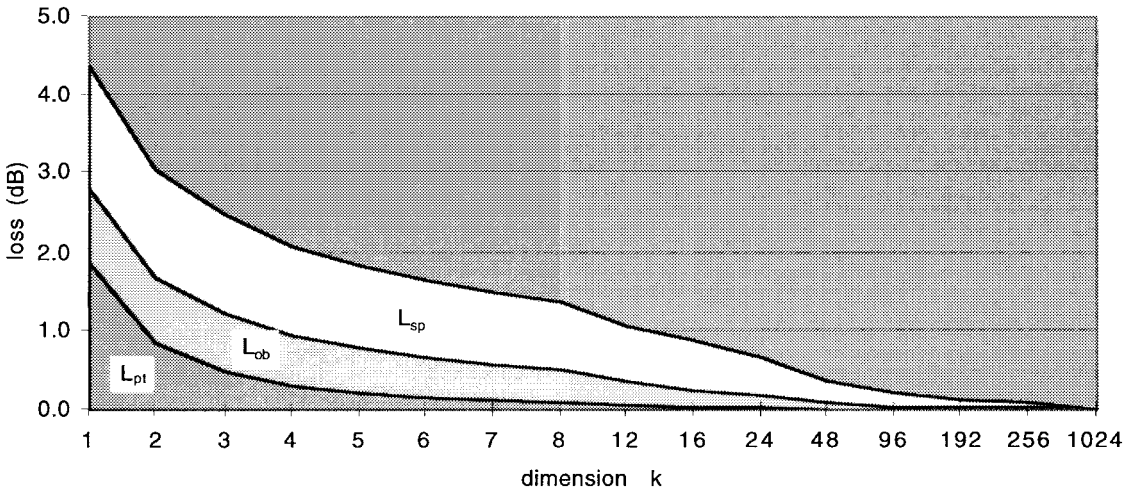


Fig. 5. Losses of optimal k-dimensional quantization relative to optimal high-dimensional quantization for an i.i.d. Gaussian source. The bottom curve is point-density loss; above that is point-density loss plus oblongitis loss; and the top curve is the total loss. For $k \geq 4$, the space-filling losses are estimates.

the product quantizer to have the optimum point density. Therefore, it generates a product quantizer whose cells in the region where $f_{k'}(x)$ is largest are more cubic, which explains why it has less oblongitis loss.

As an example, for an i.i.d. Gaussian source, the optimal choice of scalar quantizer causes the product quantizer to have 0.94-dB oblongitis loss and 1.88-dB point-density loss. The sum of these, 2.81 dB, which equals $10 \log_{10} \beta_1/\overline{\beta}$, has been called the "shape loss" [333] because it is determined by the shape of the density—the more uniform the density the less need for compromise because the scalar point densities leading to best product cell shapes and best point density are more similar. Indeed, for a uniform source density, there is no shape loss. In summary, for an i.i.d. source, in comparison to high-dimensional quantization, the shortcomings of scalar quantization with fixed-rate coding are 1) the $L_{\rm sp} = 1.53$-dB space-filling loss and 2) the lack of sufficient degrees of freedom to simultaneously attain good inertial profile (small $L_{\rm ob}$) and good point density (small $L_{\rm pt}$). On the other hand, it is often surprising to newcomers that vector quantization gains anything at all over scalar quantizers for i.i.d. sources, and secondly, that the gain is more than just the recovery of the space-filling loss.

A similar comparison can be made between k-dimensional $(k \geq 2)$ and high-dimensional VQ, by comparing the product quantizer formed by k'/k uses of a k-dimensional VQ to an optimal k'-dimensional quantizer, for large k'. The results are that as k increases 1) the space-filling loss $L_{\rm sp} = M_k/(1/2\pi e)$ decreases, and 2) there are more degrees of freedom so that less compromise is needed between the k-dimensional point density that minimizes oblongitis and the one that gives the optimal point density. As a result, the oblongitis, point density, and shape losses decrease to zero, along with the space-filling loss. For the i.i.d. Gaussian source, these losses are plotted in Fig. 5.

For sources with memory, scalar quantization $(k = 1)$ engenders an additional loss due to its inability to exploit the dependence between source samples. Specifically, when there is dependence/correlation between source samples, the product point density cannot match the ideal point density, not even approximately. See [333] and [365] for a definition of memory loss. (One can factor both the point density and oblongitis losses into two terms, one of which is due to the quantizer's inability to exploit memory.) There is also a memory loss for k-dimensional quantization, which decreases to 1 as k increases. The value of k for which the memory loss becomes close to unity (i.e., negligible) can be viewed as kind of "effective memory or correlation length" of the source. It is closely related to the decorrelation/independence length of the process, i.e., the smallest value of k such that source samples are approximately uncorrelated when separated by more than k.

F. Variable-Rate Quantization with Arbitrary Quantizer Dimension and Entropy Coding Order

We continue in the random process domain (stationary sources). To find the best possible performance of vector quantizers with block entropy coding over all possible choices of the dimension k of the lossy encoder and the order L of the entropy coder, we examine the high-resolution approximation (32), which shows that $\delta_{k,L}(R) \cong M_k \gamma_{kL} \sigma^2 2^{-2R}$. As mentioned previously, the M_k's are subadditive, so choosing k large makes M_k as small as possible, namely, as small as $\overline{M}$. Next, for stationary sources, it is well known that kth-order differential entropy $h_k \equiv (1/k)h(X_1, \cdots, X_k)$ is monotonically nonincreasing in k. Therefore, choosing either k or L large makes $\gamma_{kL} = 2^{2h_{kL}}$ as small as possible, namely, as small as $\overline{\gamma} \equiv \lim_{k\to\infty} \gamma_k$. Interestingly, $\overline{\gamma} = \overline{\beta} \equiv \lim_{k\to\infty} \beta_k$, as shown by Gersho [193], who credits Thomas Liggett. It follows immediately that the best possible performance of vector quantizers with block entropy coding is given by $\overline{\delta}(R) = \overline{M}\,\overline{\beta}\sigma^2 2^{-2R}$, which is the operational distortion-rate function of fixed-rate quantizers. In other words, entropy coding does not permit performance better than high-dimensional fixed-rate quantization.

Let us now re-examine the situation a bit more carefully. We may summarize the various high-resolution approximations to

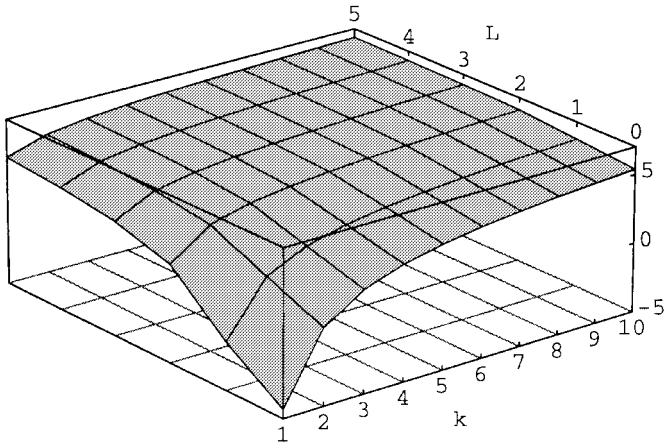


Fig. 6. $10 \log_{10} \alpha_{k,L}$ for a Gauss–Markov source with correlation coefficient 0.9.

operational distortion-rate functions as

$$\delta_{k,L}(R) \cong M_k \alpha_{k,L} \sigma^2 2^{-2R}, \qquad k \geq 1, L \geq 0 \tag{38}$$

where by convention $L = 0$ refers to fixed-rate coding, $L \geq 1$ refers to Lth-order entropy coding, and

$$\alpha_{k,L} \equiv \begin{cases} \beta_k, & L = 0 \\ \gamma_{kL}, & L \geq 1. \end{cases}$$

Note that both M_k's and $\alpha_{k,L}$'s tend to decrease as k or L increase. (The M_k's and the $\log \beta_k$'s are subadditive. The γ_k's are nonincreasing.) As an illustration, Fig. 6 plots $10 \log_{10} \alpha_{k,L}$ (in decibels) versus k and L for a Gauss–Markov source with correlation coefficient $\rho = 0.9$.

Consider how $\delta_{k,L}(R)$ decreases, i.e., improves, with k and L increasing. On the one hand, for fixed k, it decreases with increasing L (actually, it is monotonically nonincreasing) to

$$\delta_{k,\infty}(R) = M_k \overline{\beta} \sigma^2 2^{-2R} = \frac{M_k}{\overline{M}} \overline{\delta}(R). \tag{39}$$

Thus k-dimensional quantization with high-order entropy coding suffers only the k-dimensional space-filling loss. On the other hand, for fixed L, $\delta_{k,L}(R)$ decreases with k (actually it is subadditive) to

$$\delta_{\infty,L}(R) = \overline{M\beta} \sigma^2 2^{-2R} = \overline{\delta}(R). \tag{40}$$

Hence, high-dimensional quantization suffers no loss relative to the best possible performance, no matter the order or absence of an entropy coder.

From the above, we see that to attain performance close to $\overline{\delta}(R)$, k must be large enough that the space-filling loss $M_k/\overline{M}$ is approximately one, and the combination of k and L must be large enough that $\alpha_{k,L}/\overline{\beta}$ is also approximately one. Regarding the first of these, even $k = 1$ (scalar quantization) yields $M_1/\overline{M} = \pi e/6 = 1.42$, representing only a 1.53-dB loss, which may be acceptable in many situations. When it is not acceptable, k needs to be increased. Unfortunately, as evident in Fig. 5, the space-filling loss decreases slowly with increasing k. Regarding the second, we note that one has considerable freedom. There are two extreme cases: 1) k large and $L = 0$, i.e., fixed-rate high-dimensional quantization, or 2) L large and $k = 1$, i.e., scalar quantization with high-order entropy coding. In fact, uniform scalar quantization will suffice in the second case. Alternatively, one may choose moderate values for both k and L. Roughly speaking, kL must be approximately equal to the effective memory length of the source plus the value needed for a memoryless source. In effect, if the source has considerable memory, such memory can be exploited either by the lossy encoder (k large), or the lossless encoder (L large), or both (moderate values of k and L). Moreover, in such cases the potential reductions in $\alpha_{k,L}$ due to increasing k or L tend to be much larger than the potential reductions in the space-filling loss. For example, for the Gauss–Markov source of Fig. 6, $\alpha_{k,0} = \beta_k$ decreases 10.0 dB as k increases from one to infinity, and has already decreased 8.1 dB when $k = 6$.

From the point of view of the lossy encoder, the benefit of entropy coding is that it reduces the dimension required of the lossy encoder. Similarly, from the point of view of the lossless encoder, the benefit of increasing the dimension of the vector quantizer is that it decreases the order required of the lossless encoder. Stated another way, the benefits of entropy coding decrease with increasing quantizer dimension, and the benefits of increasing quantizer dimension decrease with increasing entropy coding order. In summary (cf. [377]), optimal performance is attainable with and only with a high-dimensional lossy encoder, and with or without entropy coding. However, good performance (within 1.53 dB of the best) is attainable with uniform scalar quantizer and high-order entropy coding. Both of these extreme approaches are quite complex, and so practical systems tend to be compromises with moderate quantizer dimension and entropy coding order.

As with fixed-rate quantization, it is important to understand what specific characteristics of variable-rate quantizers cause them to perform the way they do. Consequently, we will take another look at variable-rate quantization, this time from the point of view of the point density and inertial profile of the high-dimensional product quantizer induced by an optimal low-dimensional variable-rate quantizer. The situation is simpler than it was for fixed-rate quantization. As mentioned earlier, when rate is large, an optimal k-dimensional variable-rate quantizer has a uniform point density and a partition and codebook formed by tessellating T_k. Suppose k is small and k' is a large multiple of k. From the structure of optimal variable-rate quantizers, one sees that using an optimal k-dimensional quantizer k'/k times yields a k'-dimensional quantizer having the same (uniform) point density as the optimal k'-dimensional quantizer and differing, mainly, in that its inertial profile equals the constant M_k, whereas that of the optimal k'-dimensional quantizer equals $M_{k'} \cong \overline{M}$. Thus the loss due to k-dimensional quantization is only the space-filling loss $M_k/\overline{M}$, which explains what Gish and Pierce found for scalar quantizers in 1968 [204]. We emphasize that there is no point density, oblongitis, or memory loss, even for sources with memory. In effect, the entropy code has eliminated the need to shape the point density, and as a result, there is no need to compromise cell shapes.

Finally, let us compare the structure of the fixed-rate and variable-rate approaches when dimension is large. On the one

hand, optimal quantizers of each type have the same constant inertial profile, namely, $m(x) \cong M_k$. On the other hand, they have markedly different point densities: an optimal fixed-rate quantizer has point density $\lambda_k^*(x) \cong f_k(x)$, whereas an optimal variable-rate quantizer has point density that is uniform over all of $\Re^k$. How is it that two such disparate point densities do in fact yield the same distortion? The answer is provided by the asymptotic equipartition property (AEP) [110], which is the key fact upon which most of information theory rests. For a stationary, ergodic source with continuous random variables, the AEP says that when dimension k is large, the k-dimensional probability density is approximately constant, except on a set with small probability. More specifically, it shows $\Pr(X \in \mathcal{T}_k) \cong 1$, where

$$\mathcal{T}_k \equiv \left\{ x \in \Re^k : -\frac{1}{k} \log f_k(x) \cong h_\infty \right\}$$

is a set of *typical sequences*, where $h_\infty \equiv \lim_{k\to\infty} h_k$ is the *differential entropy rate* of the source. It follows immediately from the AEP and the fact that $\lambda_k^*(x) \cong f_k(x)$ that the point density of an optimal fixed-rate quantizer is approximately uniform on $\mathcal{T}_k$ and zero elsewhere. Moreover, for an optimal variable-rate quantizer, whose point density is uniform over all of $\Re^k$, we see that the cells not in $\mathcal{T}_k$ can be ignored, because they have negligible probability, and that the cells in $\mathcal{T}_k$ all have the same probability and, consequently, can be assigned codewords of equal length. Thus both approaches lead to quantizers that are identical on $\mathcal{T}_k$ (uniform point density and fixed-length codewords) and differ only in what they do on the complement of $\mathcal{T}_k$, a set of negligible probability.

It is worthwhile emphasizing that in all of the discussion in this section we have restricted attention to quantizers with memoryless lossy encoders and either fixed-rate, memoryless or block lossless encoders. Though there are many lossy and lossless encoders that are not of this form, such as DPCM or finite-state, predictive or address vector VQ, and Lempel–Ziv or arithmetic lossless coding, we believe that the easily analyzed case studied here shows, representatively, the effects of increasing memory in the lossy and lossless encoders.

G. Other Distortion Measures

By far the most commonly assumed distortion measure is squared error, which for scalars is defined by $d(x,y)=|x-y|^2$ and for vectors is defined by

$$d_k(x, y) = \sum_{i=1}^{k} |x_i - y_i|^2, \qquad \text{where } x = (x_1, \cdots, x_k).$$

Often the distortion is normalized by $1/k$. A variety of more general distortion measures have been considered in the literature, but the simplicity and tractability of squared error has long given it a central role. Intuitively, the average squared error is the average energy or power in the quantization noise. The most common extension of distortion measures for scalars is the rth-power distortion $d(x,y) = |x-y|^r$. For example, Roe [443] generalized Max's formulation to distortion measures of this form. Gish and Pierce [204] considered a more general distortion measure of the form $d(x, y) = L(x-y)$, where L is a monotone increasing function of the magnitude of its argument and $L(0) = 0$ with the added property that

$$M(v) \equiv \frac{1}{v} \int_{-v/2}^{v/2} L(u)\, du$$

has the property that $vM'(v)$ is monotone. None of these distortion measures has been widely used, although the magnitude error (rth power with $r = 1$) has been used in some studies, primarily because of its simple computation in comparison with the squared error (no multiplications).

The scalar distortion measures have various generalizations to vectors. If the dimension is fixed, then one needs only a distortion measure, say $d_k(x, y)$, defined for all $x, y \in \Re^k$. If the dimension is allowed to vary, however, then one requires a family of distortion measures $d_k(x, y)$, $k = 1, 2, \cdots$, which collection is called a *fidelity criterion* in source coding theory. Most commonly it is assumed that the fidelity criterion is *additive* or *single letter* in the sense that

$$\begin{aligned} &d_k((x_1, \cdots, x_k), (y_1, \cdots, y_k)) \\ &\quad = d_l((x_1, \cdots, x_l), (y_1, \cdots, y_l)) \\ &\qquad + d_{k-l}((x_{l+1}, \cdots, x_k), (y_{l+1}, \cdots, y_k)) \end{aligned} \tag{41}$$

for $l = 1, 2, \cdots, k-1$, or, equivalently,

$$d_k((x_1, \cdots, x_k), (y_1, \cdots, y_k)) = \sum_{i=1}^{k} d_1(x_i, y_i). \tag{42}$$

Additive distortion measures are particularly useful for proving source coding theorems since the normalized distortion will converge under appropriate conditions as the dimension grows large, thanks to the ergodic theorem. One can also assume more generally that the distortion measure is subbadditive in the sense that

$$\begin{aligned} &d_k((x_1, \cdots, x_k), (y_1, \cdots, y_k)) \\ &\quad \le d_l((x_1, \cdots, x_l), (y_1, \cdots, y_l)) \\ &\qquad + d_{k-l}((x_{l+1}, \cdots, x_k), (y_{l+1}, \cdots, y_k)) \end{aligned} \tag{43}$$

and the subadditive ergodic theorem will still lead to positive and negative coding theorems [218], [340].[10] An example of a subadditive distortion measure is the Levenshtein distance [314] which counts the number of insertions and deletions along with the number of changes that it takes to convert one sequence into another. Originally developed for studying error-correcting codes, the Levenshtein distance was rediscovered in the computer science community as the "edit distance."

For a fixed dimension k one can observe that the squared-error distortion measure can be written as $\|x-y\|^2$, where $\|x-y\|$ is the l_2 norm

$$\|x-y\| = \left(\sum_{i=1}^{k} |x_i - y_i|^2 \right)^{1/2}.$$

[10] This differs slightly from the previous definition of subadditive because the d_k are not assumed to be normalized. The previous definition applied to d_k/k is equivalent to this definition.

This idea can be extended by using any power of any l_p norm, e.g.,

$$d(x, y) = ||x - y||_p^r$$

where

$$||x - y||_p = \left(\sum_{i=1}^{k} |x_i - y_i|^p \right)^{1/p}.$$

(In this notation the l_2 norm is $|| \cdot ||_2$.) If we choose $p = r$, then this distortion measure (sometimes referred to simply as the rth-power distortion) is additive. Zador [562] defined a very general *rth-power distortion measure* as any distortion measure of the form $d(x, y) = \rho(x - y)$ where for any $a > 0$, $\rho(ax) = a^r \rho(|x_1|, \cdots, |x_k|)$, for some $r > 0$. This includes rth-power distortion in the narrow sense $||x - y||_2^r$, as well as the additive distortion measures of the form

$$||x - y||_r^r = \sum_{i=1}^{k} |x_i - y_i|^r$$

and even weighted average distortions such as

$$\left(\sum_{i=1}^{k} w_i |x_i - y_i|^2 \right)^r$$

and

$$\sum_{i=1}^{k} w_i |x_i - y_i|^r$$

where the w_i's are nonnegative.

A variation on the l_p norm is the l_∞ norm defined by $||x - y||_\infty = \max_i |x_i - y_i|$, which has been proposed as a candidate for a perceptually meaningful norm. Quantizer design algorithms exist for this case, but to date no high-resolution quantization theory or rate-distortion theory has been developed for this distortion measure (cf. [347], [231], and [348]).

High resolution theory usually considers a fixed dimension k, so neither additivity nor a family of distortion measures is required. However, high resolution theory has tended to concentrate on difference distortion measures, i.e., distortion measures that have the form $d(x, y) = L(x - y)$, where $x - y$ is the usual Euclidean difference and L is usually assumed to have nice properties, such as being monotonic in some norm of its argument. The rth-power distortion measures (of all types) fall into this category.

Recently, the basic results of high resolution theory have been extended to a family of nondifference distortion measures that are locally quadratic in the sense that provided $x \cong y$, the distortion measure is given approximately by a Taylor series expansion as $(x - y)^t B(y)(x - y)$, where $B(y)$ is a positive definite weighting matrix that depends on the output. This form is ensured by assuming that the distortion measure $d(x, y)$ has continuous partial derivatives of third order almost everywhere and that the matrix $B(y)$ defined as a k by k dimensional matrix with the $)j, n)$th element

$$B_{j,n}(y) = \frac{1}{2} \left. \frac{\partial^2 d(x, y)}{\partial x_j \, \partial x_n} \right|_{x=y} \tag{44}$$

is positive definite almost everywhere. The basic idea for this distortion measure was introduced by Gardner and Rao [186] to model a perceptual distortion measure for speech, where the matrix $B(y)$ is referred to as the "sensitivity matrix." The requirement for the existence of the derivatives of third order and for the $B(y)$ to be positive definite were added in [316] as necessary for the analysis. Examples of distortion measures meeting these conditions are the time-domain form of the Itakura–Saito distortion [258], [259], [257], [224], which has the form of an input-weighted quadratic distortion measure of the form of (21). For this case, the input weighting matrix W_x is related to the partial derivative matrix by $B(x) = \frac{1}{2}(W_x + W_x^t)$, so that positive definiteness of W_x assures that of $B(x)$ and the derivative conditions are transferred to W_x. Other distortion measures satisfying the assumptions are the image distortion measures of Eskicioglu and Fisher [150] and Nill [386], [387]. The Bennett integral has been extended to this type of distortion, and approximations for both fixed-rate and variable-rate operational distortion-rate functions have been developed [186], [316]. For the fixed-rate case, the result is that

$$D(q) \cong \frac{1}{N^{2/k}} \int f(x) (\det(B(x)))^{1/k} \frac{m(x)}{\lambda^{2/k}(x)} \, dx \tag{45}$$

where the modified inertial profile $m(x)$ is assumed to be the limit of

$$M(S_i, y_i) = (\det(B(y_i)))^{-(1/k)} \cdot \frac{\int_{S_i} (x_i - y_i)^t B(y_i)(x_i - y_i) \, dx}{[V(S_i)]^{(k+2)/k}}.$$

A natural extension of Gersho's conjecture to the nondifference distortion measures under consideration implies that, as in the squared-error case, the optimal inertial profile is assumed to be constant (which in any case will yield a bound) and minimizing the above (for example, using Hölder's inequality) yields the optimal point density

$$\lambda(x) = \frac{(f(x)(\det(B(x)))^{1/k})^{k/(k+2)}}{\int (f(x')(\det(B(x')))^{1/k})^{k/(k+2)} \, dx'} \tag{46}$$

and the operational distortion-rate function (analogous to (30))

$$\delta(R) \cong M_k \beta_k \sigma^2 2^{-2R} \tag{47}$$

where now

$$\beta_k = \frac{1}{\sigma^2} \left\{ \int (f(x)(\det(B(x)))^{1/k})^{k/(k+2)} \, dx \right\}^{(k+2)/k} \tag{48}$$

generalizes Zador's factor to the given distortion measure. As shown later in (58), M_k can be bounded below by the moment of inertia of a sphere. Similarly, in the variable-rate case

$$\delta(R) \cong M_k 2^{(2/k)(h(X) + (1/2) \int \log(\det(B(x))) f(x) \, dx)} 2^{-2R} \tag{49}$$

with optimal inertial profile $m(x) = M_k$ and optimal point density

$$\lambda(x) = \frac{(\det(B(x)))^{1/2}}{\int (\det(B(x')))^{1/2}\, dx'}. \tag{50}$$

Both results reduce to the previous results for the special case of a squared-error distortion measure since then $\det(B(x)) = 1$. Note in particular that the optimal point density for the entropy-constrained case is not in general a uniform density.

Parallel results for Shannon lower bounds to the rate-distortion function have been developed for this family of distortion measures by Linder and Zamir [323] and results for multidimensional companding with lattice codes for similar distortion measures have been developed by Linder, Zamir, and Zeger [325].

H. Rigorous Approaches to High Resolution Theory

Over the years, high-resolution analyses have been presented in several styles. Informal analyses of distortion, such as those used in this paper to obtain $\Delta^2/12$ and Bennett's integral (25), generally ignore overload distortion and estimate granular distortion by approximating the density as being constant within each quantization cell. In contrast, rigorous analyses generally focus on sequences of ever finer quantizers, for which they demonstrate that, in the limit, overload distortion becomes negligible in comparison to granular distortion and the ratio of granular distortion to some function of the fineness parameter tends to a constant. Though informal analyses generally lead to the same basic results as rigorous ones, the latter make it clear that the approximations are good enough that their percentage errors decrease to zero as the quantizers become finer, whereas the former do not. Moreover, the rigorous derivations provide explicit conditions under which the assumption of negligible overload distortion is valid. Some analyses (informal and rigorous) provide corrections for overload distortion, and some even give examples where the overload distortion cannot be asymptotically ignored but can be estimated nevertheless. Similar comments apply to informal versus rigorous analyses of asymptotic entropy. In the following we review the development of rigorous theory.

Many analyses—informal and rigorous—explicitly assume the source has finite range (i.e., a probability distribution with bounded support); so there is no overload distortion to be ignored [43], [405], [474]. In some cases, the source really does have finite range. In others, for example speech and images, the source samples have infinite range, but the measurement device has finite range. In such cases, the truncation by the measurement device creates an implicit overload distortion that is not affected by the design of the quantizer. It makes little sense, then, to choose a quantizer so fine that its (granular) distortion is significantly less than this implicit overload distortion. This means there is an upper limit to the fineness of quantizers that need be considered, and consequently, one must question whether such fineness is small enough that the source density can be approximated as constant within cells. Some analyses do not explicitly assume the source density has finite support, but merely assert that overload distortion can be ignored. We view that this differs only stylistically from an explicit assumption of finite support, for both approaches ignore overload distortion. However, assuming finite support is, arguably, humbler and mathematically more honest.

The earliest quantizer distortion analyses to appear in the open literature [43], [405], [474] assumed finite range and used the density-approximately-constant-in-cells assumption. Several papers avoided the latter by using a Taylor series expansion of the source density. For example, Lloyd [330] used this approach to show that, ignoring overload distortion, the approximation error in the Panter–Dite formula is $o(1/N^2)$, which means that it tends to zero, even when multiplied by N^2. Algazi [8], Roe [443], and Wood [539] also used Taylor series.

Overload distortion was first explicitly considered in the work of Shtein (1959) [471], who optimized the cell size of uniform scalar quantization using an explicit formula for the overload distortion (as well as $\Delta^2/12$ for the granular distortion) and while rederiving the Panter–Dite formula, added an overload distortion term.

The earliest rigorous analysis[11] is contained in Schutzenberger's 1958 paper [462], which showed that for k-dimensional variable-rate quantization ($L = 1$), rth-power distortion ($\|x - y\|^r$), and a source with finite differential entropy and $E[\|X\|^{r'}] < \infty$ for some $r' > r$, there is a $K_{k,r} > 0$, depending on the source and the dimension, such that any k-dimensional quantizer with finitely or infinitely many cells, and output entropy H, has distortion at least $K_{k,r}2^{-(r/k)H}$. Moreover, there exists $K'_{k,r} > K_{k,r}$ and a sequence of quantizers with increasing output entropies H and distortion no more than $K'_{k,r}2^{-(r/k)H}$. In essence, these results show that

$$K_{k,r}2^{(-r/k)R} \le \delta_{k,1}(R) \le K'_{k,r}2^{(-r/k)R}, \qquad \text{for all } R.$$

Unfortunately, as Schutzenberger notes, the ratio of $K'_{k,r}$ to $K_{k,r}$ tends to infinity as dimension increases. As he indicates, the problem is that in demonstrating the upper bound, he constructs a sequence of quantizers with cubic cells of equal size and then bounds from above the distortion in each cell by something proportional to its diameter to the rth power. If instead one were to bound the distortion by the moment of inertia of the cell times the maximum value of the density within it, then $K'_{k,r}/K_{k,r}$ would not tend to infinity.

Next, two papers appeared in the same issue of *Acta Math. Acad. Sci. Hungar.* in 1959. The paper by Renyi [433] gave, in effect, a rigorous derivation of (11) for a uniform quantizer with infinitely many levels. Specifically, it showed that $H(q_n(X)) = h(X) + \log n + o(1)$, provided that the source distribution is absolutely continuous and that $H(q_n(X))$ and $h(X)$ are finite, where q_n denotes a uniform quantizer with step size $1/n$ and $o(1)$ denotes a quantity that approaches zero as n goes to ∞. They paper also explores what happens when the distribution is not absolutely continuous.

[11] Though Lloyd [330] gave a fairly rigorous analysis of distortion, we do not include his paper in this category because it ignored overload distortion.

In the second paper, Fejes Toth [159] showed that for a two-dimensional random vector that is uniformly distributed on the unit square, the mean-squared error of any N-point quantizer is bounded from below by $M(\text{hexagon})/N$. This result was independently rederived in a simpler fashion by Newman (1964) [385]. Clearly, the lower bound is asymptotically achievable by a lattice with hexagonal cells. It follows then that the ratio of $\delta_2(R)$ to $M(\text{hexagon})\sigma^2 2^{-2R}$ tends to one, and also, that Gersho's conjecture holds for dimension two.

Zador's thesis (1963) [561] was the next rigorous work. As mentioned earlier, it contains two principal results. For fixed-rate quantization, rth-power distortion measures of the form $||x-y||^r$ and a source that is uniformly distributed on the unit cube, it first shows ([561, Lemma 2.3]) that the operational distortion-rate function[12] $\delta_k(N)$ multiplied by $N^{r/k}$ approaches a limit $b_{k,r}$ as $N \to \infty$. The basic idea, which Zador attributes to J. M. Hammersley, is the following: For any positive integers N and n, divide the unit cube into n^k subcubes, each with sides of length $1/n$. Clearly, the best code with $\tilde{N} = n^k N$ codevectors is at least as good as the code constructed by using the best code with N points for each subcube. It follows then that $\delta_k(\tilde{N}) \le \delta_k(n, N) = (1/n^r)\delta_k(N)$, where $\delta_k(n, N)$ is the operational distortion-rate function of a source that is uniformly distributed on a subcube and where the second relation follows from the fact that this "sub" source is just a scaling of the original source. Multiplying both sides by $\tilde{N}^{r/k}$ yields

$$\tilde{N}^{r/k}\delta_k(\tilde{N}) \le N^{r/k}\delta_k(N).$$

Thus we see that increasing the number of codevectors from N to $\tilde{N} = n^k N$ does not increase $N^{r/k}\delta_k(N)$. A somewhat more elaborate argument shows that this is approximately true for any sufficiently large $\tilde{N}$ and, as a result, that

$$\limsup_{N\to\infty} N^{r/k}\delta_k(N) \le \liminf_{N\to\infty} N^{r/k}\delta_k(N)$$

i.e., $N^{r/k}\delta_k(N)$ has a limit. One can see how the selfsimilarity of the uniform density (it is divisible into similar subdensities) plays a key role in this argument. Notice also that nowhere do the shapes of the cells or the point density enter into it.

Zador next addresses nonuniform densities. With $||f||_s$ denoting $(\int f^s(x)\,dx)^{1/s}$, his Theorem 2.2 shows that if the k-dimensional source density satisfies $||f||_{k/(k+r)} < \infty$ and $E[||X||^{k-1+r+\epsilon}] < \infty$ for some $\epsilon > 0$, then

$$N^{r/k}\delta_k(N) \to b_{k,r}||f||_{k/(k+r)}$$

as $N \to \infty$. The positive part, namely, that

$$\limsup_{N\to\infty} N^{r/k}\delta_k(N) \le b_{k,r}||f||_{k/(k+r)}$$

is established by constructing codes in, approximately, the following manner: Given N, one chooses a sufficiently large support cube (large enough that overload distortion contributes little), subdivides the cube into n^k equally sized subcubes, and places within each subcube a set of codevectors that are optimal for the uniform distribution on that subcube, where the number of codevectors in a subcube is carefully chosen so that the point density in that subcube approximates the optimal point density for the original source distribution. One then shows that the distortion of this code, multiplied by $N^{r/k}$, is approximately $b_{k,r}||f||_{k/(k+r)}$. The best codes are at least this good and it follows that

$$\limsup_{k\to\infty} N^{r/k}\delta_k(N) \le b_{k,r}||f||_{k/(k+r)}.$$

One can easily see how this construction creates codes with essentially optimal point density and cell shape. We will not describe the converse.

Zador's 1966 Bell Labs Memorandum [562] reproves these two main results under weaker conditions. The distortion measure is rth power in the general sense, which includes as special cases the narrow sense of the rth power of the Euclidean norm considered by Schutzenberger [462]. The requirement on the source density is only that each of its marginals has the property that it is bounded from above by $|x|^{r+\epsilon}$, for some $\epsilon > 0$ and all x of sufficiently large magnitude. This is a pure tail condition, as opposed to the finite moment condition of the thesis, which constrains both the tail and the peak of the density. Note also that it no longer requires that $||f||_{k/(k+r)}$ be finite.

As indicated earlier, Zador's memorandum also derives the asymptotic form of the operational distortion-rate function of variable-rate quantization. In other words, it finishes what his thesis and Schutzenberger [462] started, though he was apparently unaware of the latter. Specifically, it shows that

$$2^{rR}\delta_{k,1}(R) \to c_{k,r}2^{r(1/k)h(X_1,\cdots,X_k)} \text{ as } R \to \infty$$

where $c_{k,r}$ is some constant no larger than $b_{k,r}$, assuming the same conditions as the fixed-rate result, plus the additional requirement that for any $\epsilon > 0$ there is a bounded set containing all points x such that $f(x) \ge \epsilon$.

Gish and Pierce (1968) [204], who discovered that uniform is the asymptotically best type of scalar quantizer for variable-rate coding, presented both informal and rigorous derivations—the latter being the first to appear in these TRANSACTIONS. Specifically, they showed rigorously that for uniform scalar quantization with infinitely many cells of width Δ, the distortion D_Δ and the output entropy H_Δ behave as follows:

$$\lim_{\Delta\to 0} \frac{D_\Delta}{\Delta^2/12} = 1 \tag{51}$$

$$\lim_{\Delta\to 0} (H_\Delta + \log\Delta) = h(X) \tag{52}$$

which makes rigorous the $\Delta^2/12$ formula and (11), respectively. For this result, they required the density to be continuous except at finitely many points, and to satisfy a tail condition similar to Zador's and another condition about the behavior at points of discontinuity. The paper also outlined a rigorous proof of (32) in the scalar case, i.e., that $\delta_{1,1}(R)/Z_{1,1}(R) \to 1$ as $R \to \infty$. But as to the details it offered only that: "The complete proof is surprisingly long and will not be given here." Though Gish and Pierce were the first to informally derive (13), neither this paper nor any paper to date has provided a rigorous derivation.

[12] We abuse notation slightly and let $\delta_k(N)$ denote the least distortion of k-dimensional quantizers with N codevectors.

Elias (1970) [143] also made a rigorous analysis of scalar quantization, giving asymptotic bounds to the distortion of scalar quantizers with a rather singularly defined measure of distortion, namely, the rth root of the average of the rth power of the cell widths. A companion paper [144] considers similar bounds to the performance of vector quantizers with an analogous average-cell-size distortion measure.

In 1973, Csiszàr [114] presented a rigorous generalization of (52) to higher dimensional quantizers. Of most interest here is the following special case of his principal result ([114, Theorem 1]): Consider a k-dimensional source and a sequence of k-dimensional quantizers $q_1, q_2, \cdots$, where q_n has a countably infinite number of cells, each with volume v_n, where the v_n's and also the maximum of the cell diameters tends to zero. Then under certain conditions, including the condition that there be at least some quantizer with finite output entropy, the output entropy H_n satisfies

$$\lim_{n\to\infty}(H_n+\log v_n)=h(X). \tag{53}$$

Clearly, this result applies to quantizers generated by lattices and, more generally, tessellations. It also applies to quantizers with finitely many cells for sources with compact support. But it does not apply to quantizers with finitely many cells and sources with infinite support, because it does not deal with the overload region of such quantizers.

In 1977, Babkin *et al.* [580] obtained results indicating how rapidly the distortion of fixed-rate lattice quantizers approach $\overline{\delta}(R)$ as rate R and dimension k increase, for difference distortion measures. In 1978, these same authors [581] studied uniform scalar quantization with variable-rate coding, and extended Koshelev's results to rth power distortion measures.

The next contribution is that of Bucklew and Gallagher (1980) [63], who studied asymptotic properties of fixed-rate uniform scalar quantization. With Δ_N denoting the cell width that minimizes distortion among N cell uniform scalar quantizers and D_N denoting the resulting minimum mean-squared error, they showed that for a source with a Riemann integrable density $f(x)$

$$\lim_{N\to\infty} N\Delta_N=\operatorname{supp}(f)$$

and

$$\lim_{N\to\infty} N^2D_N=\frac{\operatorname{supp}(f)^2}{12}$$

where $\operatorname{supp}(f)$ is the length of the shortest interval (a,b) with probability one. When the support is finite, i.e., a and b are finite, the above implies $D_N/(\Delta_N^2/12)\to 1$ as $N\to\infty$, and so D_N decreases as $1/N^2$. This makes the $\Delta^2/12$ formula rigorous in the finite N case, at least when Δ is chosen optimally. However, when the support is infinite, e.g., a Gaussian density, D_N decreases at a rate slower than $1/N^2$, and the resulting signal-to-noise ratio versus rate curve separates from any line of slope 6 dB/bit. Consequently, the ratio of the operational distortion-rate functions of uniform and nonuniform scalar quantizers increases without bound as the rate increases; i.e., uniform quantization is asymptotically bad. Moreover, they showed that $D_N/(\Delta_N^2/12)$ does not always converge to 1. Instead, $\liminf_{N\to\infty} D_N/(\Delta_N^2/12)\geq 1$, and they exhibited densities where the inequality is strict. In such cases, the $\Delta^2/12$ formula is invalidated by the heavy tails of the density. It was not until much later that the asymptotic form of Δ_N and D_N were found, as will be described later.

Formal theory advanced further in papers by Bucklew and Wise, Cambanis and Gerr, and Bucklew. The first of these (1982) [64] demonstrated Zador's fixed-rate result for rth-power distortion $\|x-y\|^r$, assuming only that $E[\|X\|^{r+\delta}]<\infty$ for some $\delta>0$. It also contained a generalization to random vectors without probability densities, i.e., with distributions that are not absolutely continuous or even continuous. The paper also gave the first rigorous approach to the derivation of Bennett's integral for scalar quantization via companding. However, as pointed out by Linder (1991) [320], there was "a gap in the proof concerning the convergence of Riemann sums with increasing support to a Riemann integral." Linder fixed this and presented a correct derivation with weaker assumptions. Cambanis and Gerr (1983) [70] claimed a similar result, but it had more restrictive conditions and suffered from the same sort of problems as [64]. A subsequent paper by Bucklew (1984) [58] derived a result for vector quantizers that lies between Bennett's integral and Zador's formula. Specifically, it showed that when a sequence of quantizers is asymptotically optimal for one probability density $f^{(1)}(x)$, then its rth-power distortion on a source with density $f^{(2)}(x)$ is asymptotically given by $N^{-r/k}b_{k,r}\int\lambda^{-r/k}(x)f^{(2)}(x)\,dx$, where $\lambda(x)$ is the optimal point density for $f^{(1)}(x)$. On the one hand, this is like Bennett's integral in that $f^{(1)}(x)$, and consequently $\lambda(x)$, can be arbitrary. On the other hand, it is like Zador's result (or Gersho's generalization of Bennett's integral [193]) in that, in essence, it is assumed that the quantizers have optimal cell shapes.

In 1994, Linder and Zeger [326] rigorously derived the asymptotic distortion of quantizers generated by tessellations by showing that the quantizer q_α formed by tessellating with some basic cell shape S scaled by a positive number α has average (narrow-sense) rth-power distortion D_α satisfying

$$\lim_{\alpha\to 0}\frac{D_\alpha}{\alpha^r\operatorname{vol}(S)^{r/k}M(S)}=1.$$

They then combined the above with Csiszàr's result (53) to show that under fairly weak conditions (finite differential entropy and finite output entropy for some $\alpha>0$) the output entropy H_α and the distortion D_α are asymptotically related via

$$\lim_{\alpha\to 0}\frac{D_\alpha}{M(S)2^{(r/k)(h(X)-H_\alpha)}}=1$$

which is what Gersho derived informally [193].

The generalization of Bennett's integral to fixed-rate vector quantizers with rather arbitrary cell shapes was accomplished by Na and Neuhoff (1995) [365], who presented both informal and rigorous derivations. In the rigorous derivations, it was shown that if a sequence of quantizers $\{q_N\}$, parameterized by the number of codevectors, has specific point density and specific inertial profile converging in probability to a model point density and a model inertial profile,

respectively, then $N^{r/k}D(q_N)$ converges to Bennett's integral $\int m(x)\lambda^{-r/k}(x)f(x)\,dx$, where distortion is rth power $||x-y||^r$. A couple of additional conditions were also required, including one that is, implicitly, a tail condition.

Though uniform scalar quantization with finitely many levels is the oldest and most elementary form of quantization, the asymptotic form of the optimal step size Δ_N and resulting mean-squared error D_N has only recently been found for Gaussian and other densities with infinite support. Specifically, Hui and Neuhoff [253]–[255] have found that for a Gaussian density with variance σ^2

$$\lim_{N\to\infty}\frac{\Delta_N}{4\sigma\frac{1}{N}\sqrt{\ln N}}=1 \text{ and } \lim_{N\to\infty}\frac{D_N}{\frac{4}{3}\sigma^2\frac{1}{N^2}\ln N}=1.$$

This result was independently found by Eriksson and Agrell [149]. Moreover, it was shown that overload distortion is asymptotically negligible and that $D_N/(\Delta_N^2/12)\to 1$, which is the first time this has been proved for a source with infinite support. It follows from the above that the signal-to-noise ratio increases as $6.02R-10\log_{10}R$, which shows concretely how uniform scalar quantization is asymptotically bad. Hui and Neuhoff also considered non-Gaussian sources and provided a fairly general characterization of the asymptotic form of Δ_N and D_N. It turned out that the overload distortion is asymptotically negligible when and only when the tail parameter $\tau\equiv\lim_{y\to\infty}(E[X|X>y])/y$ equals one, which is the case for all generalized Gaussian densities. For such cases, more accurate approximations to Δ_N and D_N can be given. For densities with $\tau>1$, the ratio of overload to granular distortion is $(2\tau-2)/(2-\tau)$, and $D_N/(\Delta_N^2/12)\to\tau/(2-\tau)$. There are even densities with tails so heavy that $\tau=2$ and the granular distortion becomes negligible in comparison to the overload distortion. In a related result, the asymptotic form of the optimal scaling factor for lattice quantizers has also been found recently for an i.i.d. Gaussian source [359], [149].

We conclude this subsection by mentioning some gaps in rigorous high resolution theory. One, of course, is a proof or counterproof of Gersho's conjecture in dimensions three and higher. Another is the open question of whether the best tessellation in three or more dimensions is a lattice. Both of these are apparently difficult questions. There have been no rigorous derivations of (11), or its extension to higher dimensional tesselations, where the quantizers have finitely many levels, and overload distortion must be dealt with. Likewise, there have been no rigorous derivations of (13), or its higher dimensional generalization, except in the case where the point density is constant. Even assuming Gersho's conjecture is correct, there is no rigorous derivation of the Zador–Gersho formulas (30) and (32) along the lines of the informal derivations that start with Bennett's integral. We also mention that the tail conditions given in some of the rigorous results (e.g., [58], [365]) are very difficult to check. Simpler ones are needed. Finally, as discussed in Section II there are no convincing (let alone rigorous) asymptotic analyses of the operational distortion-rate function of DPCM.

I. Comparing High Resolution Theory and Shannon Rate Distortion Theory

It is interesting to compare and contrast the two principal theories of quantization, and we shall do so in a number of different domains.

Applicability: Sources—Shannon rate-distortion theory applies, fundamentally, to infinite sequences of random variables, i.e., to sources modeled as random processes. Its results derive from the frequencies with which events repeat, as expressed in a law of large numbers, such as the weak law or an ergodic theorem. As such, it applies to sources that are stationary in either the strict sense or some weaker sense, such as asymptotic mean stationarity (cf. [218, p. 16]). Though originally derived for ergodic sources, it has been extended to nonergodic sources [221], [469], [126], [138], [479]. In contrast, high resolution theory applies, fundamentally, to finite-dimensional random vectors. However, for stationary (or asymptotically stationary) sources, taking limits yields results for random processes. For example, the operational distortion-rate function $\overline{\delta}(R)$ was found to equal $\overline{Z}(R)$ in this way; see (33). Rate distortion theory also has one result relevant to finite-dimensional random vectors, namely, that the operational distortion-rate functions for fixed- and variable-rate quantization, $\delta_k(R)$ and $\delta_{k,1}(R)$, are (strictly) bounded from below by the kth-order Shannon distortion-rate function.

Both theories have been extended to continuous-time random processes. However, the high-resolution results are somewhat sketchy [43], [330], [204]. Both can be applied to two- or higher dimensional sources such as images or video. Both have been developed the most for Gaussian sources in the context of squared-error distortion, which is not surprising in view of the tractability of squared error and Gaussianity.

Applicability: Distortion Measures—Shannon rate distortion theory applies primarily to additive distortion measures; i.e., distortion measures of the form

$$d(x,y)=\sum_{i=1}^{k}d_1(x_i,y_i)$$

(or a normalized version), though there are some results for subadditive distortion measures [218], [340] and some for distortion measures such as $(x-y)^tB_x(x-y)$ [323]. High resolution theory has the most results for rth-power difference distortion measures, and as mentioned previously, some of its results have recently been extended to nondifference distortion measures such as $(x-y)^tB_x(x-y)$ [186], [316], [325]. In any event, both theories are the most fully developed for the squared-error distortion measure, especially for Gaussian sources. In addition, both theories require a finite moment condition, specific to the distortion measure. For squared-error distortion, it is simply that the variance of the source be finite. More generally, it is that $E[d(X,y)]<\infty$ for some y. In addition, as discussed previously, rigorous high resolution theory results require tail conditions on the source density, for example, $E[X^{2+\delta}]<\infty$ for some $\delta>0$.

Complementarity—The two theories are complementary in the sense that Shannon rate distortion theory prescribes the best possible performance of quantizers with a given rate and

asymptotically large dimension, while high resolution theory prescribes the best possible performance of codes with a given dimension and asymptotically large rate. That is, for fixed-rate codes

$$\delta_k(R) \cong \overline{D}(R), \qquad \text{for large } k \text{ and any } R \tag{54}$$

$$\delta_k(R) \cong Z_k(R), \qquad \text{for large } R \text{ and any } k \tag{55}$$

and, similarly, for variable-rate codes

$$\delta_{k,L}(R) \cong \overline{D}(R), \qquad \text{for large } k \text{ and any } L, R \tag{56}$$

$$\delta_{k,L}(R) \cong Z_{k,L}(R), \qquad \text{for large } R \text{ and any } k, L. \tag{57}$$

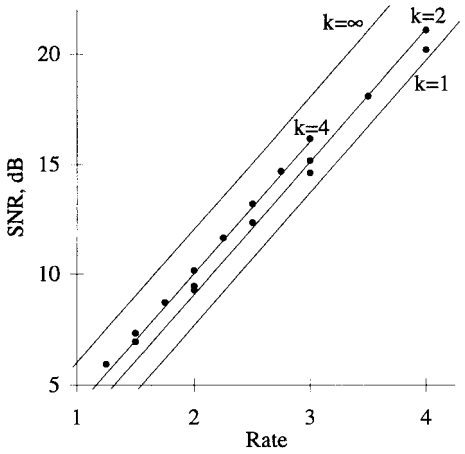
When both dimension and rate are large, they all give the same result, i.e.,

$$\delta_k(R) \cong \delta_{k,L}(R) \cong \overline{\delta}(R) \cong \overline{D}(R).$$

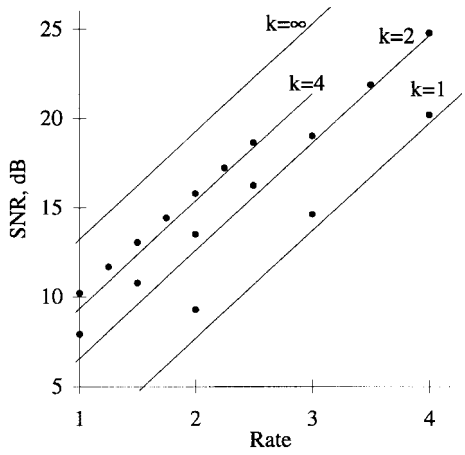
Rates of Convergence—It is useful to know how large R and k must be, respectively, for high resolution and rate distortion theory formulas to be accurate. As a rule of thumb, high resolution theory is fairly accurate for rates greater than or equal to about 3. And it is sufficiently accurate at rates about 2 for it to be useful when comparing different sources and codes. For example, Fig. 7 shows signal-to-noise ratios for fixed-rate quantizers produced by conventional design algorithms and predictions thereof based on the Zador–Gersho function $Z_k(R)$, for two Gaussian sources: i.i.d. and Markov with correlation coefficient 0.9. It is apparent from data such as this that the accuracy of the Zador–Gersho function approximation to $\delta_k(R)$ increases with dimension.

The convergence rate of $\delta_k(R)$ to $\overline{\delta}(R)$ as k tends to infinity has also been studied [413], [548], [321], [576]. Roughly speaking these results show that for memoryless sources, the convergence rate is between $\sqrt{(\log k)/k}$ and $(\log k)/k$. Unfortunately, this theory does not enable one to actually predict how large the dimension must be in order that $\delta_k(R)$ is within some specified percentage, e.g., 10%, of $\overline{\delta}(R)$. However, one may use high resolution theory to do this, by comparing $M_k\beta_k$ (or $M_k\gamma_{kL}$ in the variable-rate case) to $\overline{M}\,\overline{\beta}$. For example, for the i.i.d. Gaussian source Fig. 5 shows that $\delta_k(R)$ yields distortions within 1 and 0.2 dB of that predicted by $\overline{\delta}(R)$ at dimensions 12 and 100, respectively. For sources with memory, the dimension needs to be larger, by roughly the effective memory length. One may conclude that the Shannon distortion-rate function approximation to $\delta_k(R)$ is applicable for moderate to large dimensions k.

Quantitative Relationships—For squared-error distortion, the Zador–Gersho function $\overline{Z}(R)$ is precisely equal to the well-known Shannon lower bound $\overline{D}_{\mathrm{slb}}(R)$ to the Shannon distortion-rate function. It follows that when rate is not large, $\overline{Z}(R)$ is, at least, a lower bound to $\overline{\delta}(R)$. Similarly, the Shannon lower bound $D_{\mathrm{slb,k}}(R)$ to the kth-order Shannon distortion-rate function equals $Z_{k,1}(R)(\overline{M}/M_k)$, from which it follows that $D_{\mathrm{slb,k}}(R)$ may be thought of as the distortion of a fictional quantizer having the distortion of an optimal k-dimensional variable-rate quantizer with first-order entropy coding, except that its cells have the normalized moment of inertia of a high-dimensional sphere instead of M_k. It is well known that $\overline{D}_{\mathrm{slb}}(R)/\overline{D}(R)$ approaches one as R increases



(a)



(b)

Fig. 7. Signal-to-noise ratios for optimal VQ's (dots) and predictions thereof based on the Zador–Gersho formula (straight lines). (a) i.i.d. Gaussian. (b) Gauss–Markov, correlation coefficient 0.9.

[327], [267], [46], [322], which is entirely consistent with the fact that $\overline{Z}(R)/\overline{\delta}(R)$ approaches one as R increases. The relationships among the various distortion-rate functions are summarized below. Inequalities marked with a "•" become tight as dimension k increases, and those marked with a "+" become tight as R increases.

$$\begin{array}{ccccc} \overline{D}(R) & \overset{+}{\geq} & \overline{D}_{\mathrm{slb}}(R) & = & \overline{Z}(R) \\ \wedge\vert\bullet & & \wedge\vert\bullet & & \wedge\vert\bullet \\ D_k(R) & \geq & D_{\mathrm{slb,k}}(R) = Z_{k,1}(R)\dfrac{\overline{M}}{M_k} & \overset{\bullet}{\leq} & Z_k(R). \end{array}$$

Applicability: Quantizer Types—Rate distortion theory finds the performance of the best quantizers of any type for stationary sources. It has nothing to say about suboptimal, structured or dimension-constrained quantizers except, as mentioned earlier, that quantizers of dimension k have distortion bounded from below by the kth-order Shannon distortion-rate function. In contrast, high resolution theory can be used to analyze and optimize the performance of a number of families of structured quantizers, such as transform, lattice, product, polar, two-stage, and, most directly, dimension-constrained quantizers. Such analyses are typically based on Bennett's integral. Indeed,

the ability to analyze structured or dimension-constrained quantizers is the true forte of high resolution theory.

Performance versus Complexity: Assessing performance versus complexity should be a major goal of quantization theory. On the one hand, rate distortion theory specifies the fundamental limits to performance without regard to complexity. On the other hand, because high resolution theory can analyze the performance of families of quantizers with complexity-reducing structure, one can learn much from it about how complexity relates to performance. In recent work, Hui and Neuhoff [256] have combined high resolution theory and Turing complexity theory to show that asymptotically optimal quantization can be implemented with complexity increasing at most polynomially with the rate.

Computability: First-order Shannon distortion-rate functions can be computed analytically for squared error and magnitude error and several source densites, such as Gaussian and Laplacian, and for some discrete sources, cf. [46], [494], [560], [217]. For other sources it can be computed with Blahut's algorithm [52]. And in the case of squared error, it can be computed with simpler algorithms [168], [444]. For sources with memory, complete analytical formulas for kth-order distortion-rate functions are known only for Gaussian sources. For other cases, the Blahut algorithm [52] can be used to compute $D_k(R)$, though its computational complexity becomes overwhelming unless k is small. Due to the difficulty of computing it, many (mostly lower) bounds to the Shannon distortion-rate function have been developed which for reasonably general cases yield the distortion-rate function exactly for a region of small distortion (cf. [465], [327], [267], [239], [46], [212], [550], [559], [217]). An important upper bound derives from the fact that with respect to squared error, the Gaussian source has the largest Shannon distortion-rate function (kth-order or in the limit) of any source with the same covariance function.

To compute a Zador–Gersho function, one needs to find M_k and either β_k or γ_k in the fixed- and variable-rate cases, respectively. Though M_k is known only for $k \leq 2$, there are bounds for other values of k. One lower bound is the normalized moment of inertia of a sphere of the same dimension

$$M_k \geq \frac{k}{k+2}\left(\frac{2\pi^{k/2}}{k\Gamma(k/2)}\right)^{-(2/k)}. \tag{58}$$

Another bound is given in [106]. One upper bound was developed by Zador; others derive from the currently best known tessellations (cf. [5] and [106]). The Zador factors β_k and γ_k can be computed straightforwardly for $k = 1$ and, also, for $k \geq 2$ for i.i.d. sources. In some cases, simple closed-form expressions can be found, e.g., for Gaussian, Laplacian, gamma densities. In other cases, numerical integration can be used. Upper bounds to β_1 are given in [294]. To the authors' knowledge, for sources with memory, simple expressions for the Zador factors have been found only for Gaussian sources; they depend on the covariance matrix.

Underlying Principles: Rate distortion theory is a deep and elegant theory based on the law of large numbers and the key information-theoretic property that derives from it, namely, the AEP. High resolution theory is a simpler, less elegant theory based on geometric characterizations and integral approximations over fine partitions.

Siblings: Lossless source coding and channel coding are sibling branches of information theory, also based on the law of large numbers and the asymptotic equipartition property. Siblings of high resolution theory include error probability analyses in digital modulation and channel coding based on minimum distance and a high signal-to-noise ratio assumption, and the average power analyses for the additive Gaussian channel based on the continuous approximation.

Code Design Philosophy: Neither theory is ordinarily considered to be constructive, yet each leads to its own design philosophy. Rate distortion theory shows that, with high probability, a good high-dimensional quantizer can be constructed by randomly choosing codevectors according to the output distribution of the test channel that achieves the Shannon rate-distortion function. As a construction technique, this leaves much to be desired because the dimension of such codes is large enough that the codes so constructed are completely impractical. On the other hand, the AEP indicates that such codevectors will be roughly uniformly distributed over a "typical" set, and this leads to the design philosophy that a good code has its codevectors uniformly distributed throughout this set. In the special case of squared-error distortion and an i.i.d. Gaussian source with variance σ^2, the output distribution is i.i.d. Gaussian with variance $\sigma^2 - D(R)$; the typical set is a thin shell near the surface of a sphere of radius $\sqrt{k(\sigma^2 - D(R))}$; and a good code has its codevectors uniformly distributed on this shell. Since the interior volume of such a (high-dimensional) sphere is negligible, it is equally valid for the codevectors to be uniformly distributed throughout the sphere. For other sources, the codevectors will be uniformly distributed over some subset of the shell.

High resolution theory indicates that for large rate and arbitrary dimension k, the quantization cells should be as spherical as possible—preferably shaped like T_k, with normalized moment of inertia M_k. Moreover, the codevectors should be distributed according to the optimal point density λ_k^*. Thus high resolution theory yields a very clear design philosophy. In the scalar case, one can use this philosophy directly to construct a good quantizer, by designing a compander whose nonlinearity $c(x)$ has derivative $\lambda_1^*(x)$, and extracting the resulting reconstruction levels and thresholds to obtain an approximately optimal point quantizer. This was first mentioned in Panter–Dite [405] and rediscovered several times. Unfortunately, at higher dimensions, companders cannot implement an optimal point density without creating large oblongitis [193], [56], [57]. So there is no direct way to construct optimal vector quantizers with the high resolution philosophy.

When dimension as well as rate is large, the two philosophies merge because the output distribution that achieves the Shannon distortion-rate function converges to the source density itself, as does the optimal point density. However, for small to moderate values of k, λ_k^* specifies a better distribution of points than the rate distortion philosophy of uniformly distributing codevectors over the typical set. For example, in

the i.i.d. Gaussian case it indicates that the point density should be a Gaussian hill with somewhat larger variance than that of the source density. Which design philosophy is more useful? At low rates (say 1 bit per sample or less), one has no choice but to look to rate distortion theory. But at moderate to high rates, it appears that the high-resolution design philosophy is the better choice. To see this consider an i.i.d. Gaussian source, a target rate R, and a k-dimensional quantizer with 2^{kR} points uniformly distributed throughout a spherical support region. This is the ideal code suggested by rate distortion theory. One obtains a lower bound to its distortion by assuming that source vectors outside the support region are quantized to the closest point on the surface of the sphere, and by assuming that the cells within the support region are k-dimensional spheres. In this case, at moderate to large rates (say rate ten), after choosing the diameter of the support region to minimize this lower bound, it has been found that the dimension k must be larger than 250 in order that the resulting signal-to-noise ratio be within 1 dB of that predicted by the Shannon distortion-rate function [25]. Similar results were reported by Pepin *et al*. [409]. On the other hand, as mentioned earlier, a quantizer with dimension 12 can achieve this same distortion. It is clear then that the ability to come fairly close to $\overline{\delta}(R)$ with moderately large dimension is not due to the rate distortion theory design philosophy, the AEP, nor the use of spherical codes. Rather, it is due to the fact that good codes with small to moderate dimension have appropriately tapered point densities, as suggested by high resolution theory.

Finally, it is interesting to note that high resolution theory actually contains some analyses of the Shannon random coding approach. For example, Zador's thesis [561] gives an upper bound on the distortion of a randomly generated vector quantizer.

Nature of the Error Process: Both theories have something to say about the distribution of quantization errors. Generally speaking, what rate distortion theory has to say comes from assuming that the error distribution caused by a quantizer whose performance is close to $\overline{\delta}(R)$ is similar to that caused by a test channel that comes close to achieving the Shannon distortion-rate function. This is reasonable because Shannon's random coding argument shows that using such a test channel to randomly generate high-dimensional codevectors leads, with very high probability, to a code whose distortion is close to $\overline{\delta}(R)$. For example, one may use this sort of argument to deduce that the quantization error of a good high-dimensional quantizer is approximately white and Gaussian when the source is memoryless, the distortion is squared error, and the rate is large, cf. [404], which shows Gaussian-like histograms for the quantization error of VQ's with dimensions 8 to 32. As another example, for a Gaussian source with memory and squared-error distortion, rate distortion theory shows there is a simple relation between the spectra of the source and the spectra of the error produced by an optimal high-dimensional quantizer, cf. [46].

High resolution theory also has a long tradition of analyzing the error process, beginning with Clavier *et al*. [95], [100], and Bennett [43], and focusing on the distribution of the error, its spectrum, and its correlation with the input. Bennett showed that in the high-resolution case, the power spectral density of the quantizer error with uniform quantization is approximately white (and uniformly distributed) provided the assumptions of the high resolution theory are met and the joint density of sample pairs is smooth. (See also [196, Sec. 5.6].) Bennett also found exact expressions for the power spectral density of a uniformly quantized Gaussian process. Sripad and Snyder [477] and Claasen and Jongepier [97] derived conditions under which the quantization error is white in terms of the joint characteristic functions of pairs of samples, two-dimensional analogs of Widrow's [529] condition. Zador [562] found high-resolution expressions for the characteristic function of the error produced by randomly chosen vector quantizers. Lee and Neuhoff [312], [379] found high-resolution expressions for the density of the error produced by fairly general (deterministic) scalar and vector quantizers in terms of their point density and their *shape profile*, which is a function that conveys more cell shape information than the inertial profile. As a side benefit, these expressions indicate that much can be deduced about the point density and cell shapes of a quantizer from a histogram of the lengths of the errors. Zamir and Feder [564] showed that the error produced by an optimal lattice quantizer with infinitely many small cells is asymptotically white in the sense that its components are uncorrelated with zero means and identical variances. Moreover, they showed that it becomes Gaussian as the dimension increases. The basic ideas are that as dimension increases good lattices have nearly spherical cells and that a uniform distribution over a high-dimensional sphere is approximately Gaussian, cf. [525]. Since optimal high-dimensional, high-rate VQ's can also be expected to have nearly spherical cells and since the AEP implies that most cells will have the same size, we reach the same conclusion as from rate distortion theory, namely, that good high-rate high-dimensional codes cause the quantization error to be approximately white and Gaussian.

Successive Approximation: Many vector quantizers operate in a successive approximation or progressive fashion, whereby a low-rate coarse quantization is followed by a sequence of finer and finer quantizations, which add to the rate. Tree-structured, multistage and hierarchical quantizers, to be discussed in the next section, are examples of such. Other methods can be used to design progressive indexing into given codebooks, as in Yamada and Tazaki (1991) [553] and Riskin *et al.* (1994) [440].

Successive approximation is useful in situations where the decoder needs to produce rough approximations of the data from the first bits it receives and, subsequently, to refine the approximation as more bits are received. Moreover, successive approximation quantizers are often structured in a way that makes them simpler than unstructured ones. Indeed, the three examples just cited are known more for their good performance with low complexity than for their progressive nature. An important question is whether the performance of a successive refinement quantizer will be better than one that does quantization in one step. On the one hand, rate distortion theory analysis [228], [291], [292], [557], [147], [437], [96] has shown that there are situations where successive approximation can be done without loss of optimality. On the other

hand, high-resolution analyses of TSVQ [383] and two-stage VQ [311] have quantified the loss of these particular codes, and in the latter case shown ways of modifying the quantizer to eliminate the loss. Thus both theories have something to say about successive refinement.

V. Quantization Techniques

This section presents an overview of quantization techniques (mainly vector) that have been introduced, beginning in the 1980's, with the goal of attaining rate/distortion performance better than that attainable by scalar-based techniques such as direct scalar quantization, DPCM, and transform coding, but without the inordinately large complexity of brute-force vector quantization methods. Recall that if the dimension of the source vector is fixed, say at k, then the goal is to attain performance close to the optimal performance as expressed by $\delta_k(R)$ in the fixed-rate case, or $\delta_{k,L}(R)$ (usually $\delta_{k,1}(R)$) in the general case where variable-rate codes are permitted. However, if, as in the case of a stationary source, the dimension k can be chosen arbitrarily, then in both the fixed- and variable-rate cases, the goal is to attain performance close to $\overline{\delta}(R)$. In this case, all quantizers with $R > 0$ are suboptimal, and quantizers with various dimensions and even memory (which blurs the notion of dimension) can be considered.

We would have liked to make a carefully categorized, ordered, and ranked presentation of the various methods. However, the literature and variety of such techniques is quite large; there are a number of competing ways in which to categorize the techniques; complexity is itself a difficult thing to quantify; there are several special cases (e.g., fixed or variable rate, and fixed or choosable dimension); and there has not been much theoretical or even quantitative comparison among them. Consequently, much work is still needed in sorting the wheat from the chaff, i.e., determining which methods give the best performance versus complexity tradeoff in which situations, and in gaining an understanding of why certain complexity-reducing approaches are better than others. Nevertheless, we have attempted to choose a reasonable set of techniques and an ordering of them for discussion. Where possible we will make comments about the efficacies of the techniques. In all cases, we include references.

We begin with a brief discussion of complexity. Roughly speaking, it has two aspects: arithmetic (or computational) complexity, which is the number of arithmetic operations per sample that must be performed when encoding or decoding, and storage (or memory or space) complexity, which is the amount of auxiliary storage (for example, of codebooks) that is required for encoding or decoding. Rather than trying to combine them, it makes sense to keep separate track, because their associated costs vary with implementation venue, e.g., a PC, UNIX platform, generic DSP chip, specially designed VLSI chip, etc. In some venues, storage is of such low cost that one is tempted to ignore it. However, there are techniques that benefit sufficiently from increased memory that even though the per-unit cost is trivial, to obtain the best performance–complexity tradeoff, memory usage should be increased until the marginal gain-to-cost ratio of further increases is small, at which point the total cost of memory may be signficant. As a result, one might think of a quantizer as being characterized by a four-tuple (R, D, A, M); i.e., arithmetic complexity A and storage complexity M have been added to the usual rate R and distortion D.

As a reminder, given a k-dimensional fixed-rate VQ with codebook $\mathcal{C}$ containing 2^{kR} codevectors, brute-force *full-search encoding* finds the closest codevector in $\mathcal{C}$ by computing the distortion between x and each codevector. In other words, it uses the optimal lossy encoder for the given codebook, creating the Voronoi partition. In the case of squared error, this requires computing approximately $A = 3\times 2^{kR}$ operations per sample and storing approximately $M = k \times 2^{kR}$ vector components. For example, a codebook with rate 0.25 bits per pixel (bpp) and vector dimension $8 \times 8 = 64$ has $2^{kR} = 2^{16}$ codevectors, an impractical number for, say, real-time video coding. This exponential explosion of complexity and memory can cause serious problems even for modest dimension and rate, but it can in general make codes completely impractical in either the high-resolution or high-dimension extremes. A brute-force variable-rate scheme of the same rate will be even more complex—typically involving a much greater number of codevectors, a Lagrangian distortion computation, and an entropy coding scheme as well. It is the high complexity of such brute-force techniques that motivates the reduced complexity techniques to be discussed later in this section.

Simple measures such as arithmetic complexity and storage need a number of qualifications. One must decide whether encoding and decoding complexities need to be counted separately or summed, or, indeed, whether only one of them is important. For example, in record-once-play-many situations, it is the decoder that must have low complexity. Having no particular application in mind, we will focus on the sum of encoder and decoder complexities. For some techniques (perhaps most) it is possible to trade computations for storage by the use of precomputed tables. In such cases a quantizer is characterized, not by a single A and M but by a curve of such. In some cases, a given set of precomputed tables is the heart of the method. Another issue is the cost of memory accesses. Such operations are usually signficantly less expensive than arithmetic operations. However, some methods do such a good job of reducing arithmetic operations that the cost of memory accesses becomes significant. Techniques that attain smaller values of distortion need higher precision in their arithmetic and storage, which though not usually accounted for in assessments of complexity may sometimes be of significance. For example, a recent study of VQ codebook storage has shown that in routine cases one needs to store codevector components with only about $R+4$ bits per component, where R is the rate of the quantizer [252]. Though this study did not assess the required arithmetic precision, one would guess that it need not be more than a little larger than that of the storage; e.g., R plus 5- or 6-bit arithmetic should suffice. Finally, variable-rate coding raises additional issues such as the costs associated with buffering, with storing and accessing variable-length codewords, and with the decoder having to parse binary sequences into variable-length codewords.

When assessing complexity of a quantization technique, it is interesting to compare the complexity invested in the lossy

encoder/decoder versus that in the lossless encoder/decoder. (Recall that good performance can theoretically be attained with either a simple lossy encoder, such as a uniform scalar quantizer, and a sophisticated lossless encoder or, vice versa, as in high-dimensional fixed-rate VQ.) A quantizer is considered to have low complexity only when both encoders have low complexity. In the discussion that follows we focus mainly on quantization techniques where the lossless encoder is conceptually if not quantitatively simple. We wish, however, to mention the indexing problem, which may be considered to lie between the lossless and the lossy encoder. There are certain fixed-rate techniques, such as lattice quantization, pyramid VQ, and scalar-vector quantization, where it is fairly easy to find the cell in which the source vector lies, but the cells are associated with some set of N indices that are not simply the integers from 1 to N, where N is the number of cells, and converting the identity of the cell into a sequence of $\log N$ bits is nontrivial. This is referred to as an *indexing* problem.

Finally, we mention two additional issues. The first is that there are some VQ techniques whose implementation complexities are not prohibitive, but which have sufficiently many codevectors that designing them is inordinately complex or requires an inordinate amount of training data. A second issue is that in some applications it is desirable that the output of the encoder be progressively decodable in the sense that a rough reproduction can be made from the first bits that it receives, and improved reproductions are made as more bits are received. Such quantizers are said to be *progressive* or embedded. Now it is true that a progressive decoder can be designed for any encoder (for example, it can compute the expected value of the source vector given whatever bits it has received so far). However, a "good" progressive code is one for which the intermediate distortions achieved at the intermediate rates are relatively good (though not usually as good as those of quantizers designed for one specific rate) and that rather than restarting from scratch every time the decoder receives a new bit (or group of bits), it uses some simple method to update the current reproduction. It is also desirable in some applications for the encoding to be progressive, as well. Though not designed with them in mind, it turns out that a number of the reduced-complexity VQ approaches also address these last two issues. That is, they are easier to design, as well as progressive.

A. Fast Searches of Unstructured Codebooks

Many techniques have been developed for speeding the full (minimum-distortion) search of an arbitrary codebook C containing N k-dimensional codevectors, for example, one generated by a Lloyd algorithm. In contrast to codebooks to be considered later these will be called *unstructured*. As a group these techniques use substantial amounts of additional memory in order to significantly reduce arithmetic complexity. A variety of such techniques are mentioned in [196, Sec. 12.16].

A number of fast-search techniques are similar in spirit to the following: the Euclidean distances between all pairs of codevectors are precomputed and stored in a table. Now, given a source vector x to quantize, some initial codevector $\tilde{y}$ is chosen. Then all codevectors y_i whose distance from $\tilde{y}$ is greater than $2||x-\tilde{y}||$ are eliminated from further consideration because they cannot be closer than $\tilde{y}$. Those not eliminated are successively compared to x until one that is closer than $\tilde{y}$ is found, which then replaces $\tilde{y}$, and the process continues. In this way, the set of potential codevectors is gradually narrowed. Techniques in this category, with different ways of narrowing the search, may be found in [362], [517], [475], [476], [363], [426], [249], [399], [273], [245], [229], [332], [307], [547], [308], and [493].

A number of other fast-search techniques begin with a "coarse" prequantization with some very low-complexity technique. It is called "coarse" because it typically has larger cells than the Voronoi regions of the codebook C that is being searched. The coarse prequantization often involves scalar quantization of some type or a tree-structuring of binary quantizers, such as what are called K-d trees. Associated with each coarse cell is a *bucket* containing the indices of each codevector that is the nearest codevector to some source vector in the cell. These buckets are determined in advance and saved as tables. Then to encode a source vector x, one applies the prequantization, finds the index of the prequantization cell in which x is contained, and performs a full search on the corresponding bucket for the closest codevector to x. Techniques of this type may be found in [44], [176], [88], [89], [334], [146], [532], [423], [415], [500], and [84]. In some of these, the coarse prequantization is one-dimensional; for example, the length of the source vector may be quantized, and then the bucket of all codevectors having similar lengths is searched for the closest codevector.

Another class of techniques is like the previous except that the low-complexity prequantization has much smaller cells than the Voronoi cells of C, i.e., it is finer. In this case, the buckets associated with most "fine" prequantization cells contain just one codevector, i.e., the same codevector in C is the closest codevector to each point in the fine cell. The indices of these codevectors, one for each fine cell, are stored in a precomputed table. For each of those relatively few fine cells that have buckets containing more than one codevector, one member of the bucket is chosen and its index is placed in the table as the entry for that fine cell. Quantization of x then proceeds by applying the fine prequantizer and then using the index of the fine cell in which x lies to address the table containing codevectors from C, which then outputs the index of a codeword in C. Due to the fact that not every bucket contains only one codevector, such techniques, which may be found in [86], [358], [357], [518], [75], and [219], do not do a perfect full search. Some quantitative analysis of the increased distortion is given in [356] for a case where the prequantization is a lattice quantizer. Other fast-search methods include the *partial distortion* method of [88], [39], [402] and the transform subspace-domain approach of [78].

Consideration of methods based on prequantization leads to the question of how fine the prequantization cells should be. Our experience is that the best tradeoffs come when the prequantization cells are finer rather than coarser, the explanation being that if one has prequantized coarsely and

now has to determine which codevector in a bucket is closest to x, it is more efficient to use some fast search method than to do full search. Dividing the coarse cells into finer ones is a way of doing just this. Another question that arises for all fast search techniques is whether it is worth the effort to perform a full search or whether one should instead stop short of this, as in the methods with fine prequantization cells. Our experience is that it is usually not worth the effort to do a full search, because by suffering only a very small increase in MSE one can achieve a significant reduction in arithmetic complexity and storage. Moreover, in the case of stationary sources where the dimension is subject to choice, for a given amount of arithmetic complexity and storage, one almost always gets better performance by doing a suboptimal search of a higher dimensional codebook than a full search of a lower dimensional one.

Fast search methods based on fine prequantization can be improved by optimizing the codebook for the given prequantizer. Each cell of the partition corresponding to C induced by prequantization followed by table lookup is the union of some number of fine cells of the prequantizer. Thus the question becomes: what is the best partition into N cells, each of which is the union of some number of fine cells. The codevectors in C should then be the centroids of these cells. Such techniques have been exploited in [86] and [358]. One technique worth particular mention is called *hierarchical table lookup* VQ [86], [518], [75], [219]. In this case, the prequantizer is itself an unstructured codebook that is searched with a fine prequantizer that is in turn searched with an even finer prequantizer, and so on. Specifically, the first prequantizer uses a high-rate scalar quantizer k times. The next level of prequantization applies a two-dimensional VQ to each of $k/2$ pairs of scalar quantizer outputs. The next level applies a four-dimensional VQ to each of $k/4$ pairs of outputs from the two-dimensional quantizers, and so on. Hence the method is hierarchical. Because each of the quantizers can be implemented entirely with table lookup, this method eliminates all arithmetic complexity except memory accesses. It has been successfully used for video coding [518], [75].

B. Structured Quantizers

We now turn to quantizers with structured partitions or reproduction codebooks, which in turn lend themselves to fast searching techniques and, in some cases, to greatly reduced storage. Many of these techniques are discussed in [196] and [458].

Lattice Quantizers: Lattice quantization can be viewed as a vector generalization of uniform scalar quantization. It constrains the reproduction codebook to be a subset of a regular lattice, where a lattice is the set of all vectors of the form $\sum_{i=1}^{n} m_i u_i$, where m_i are integers and the u_i are linearly independent (usually nondegenerate, i.e., $n = k$). The resulting Voronoi partition is a tessellation with all cells (except for those overlapping the overload region) having the same shape, size, and orientation. Lattice quantization was proposed by Gersho [193] because of its near optimality for high-resolution variable-rate quantization and, also, its near optimality for high-resolution fixed-rate quantization of uniformly distributed sources. (These assume that Gersho's conjecture holds and that the best lattice quantizer is approximately as good as the best tessellation.) Especially important is the fact that their highly structured nature has led to algorithms for implementing their lossy encoders with very low arithmetic and storage complexity [103]–[105], [459], [106], [199]. These find the integers m_i associated with the closest lattice point. Conway and Sloane [104], [106] have reported the best known lattices for several dimensions, as well as fast quantizing and decoding algorithms. Some important n-dimensional lattices are the root lattices $A_n (n \geq 1)$, $D_n (n \geq 2)$, and $E_n (n = 6, 7, 8)$, the Barnes–Wall lattice Λ_{16} in dimension 16, and the Leech lattice Λ_{24} in 24 dimensions. These latter give the best sphere packings and coverings in their respective dimensions. Recently, Agrell and Eriksson [5] have found improved lattices in dimensions 9 and 10.

Though low complexity algorithms have been found for the lossy encoder, there are other issues that affect the performance and complexity of lattice quantizers. For variable-rate coding, one must scale the lattice to obtain the desired distortion and rate, and one must implement an algorithm for mapping the m_i's to the variable-length binary codewords. The latter could potentially add much complexity. For fixed-rate coding with rate R, the lattice must be scaled and a subset 2^{kR} lattice points must be identified as the codevectors. This induces a support region. If the source has finite support, the lattice quantizer will ordinarily be chosen to have the same support. If not, then the scaling factor and lattice subset are usually chosen so that the resulting quantizer support region has large probability. In either case, a low complexity method is needed for assigning binary sequences to the chosen codevectors; i.e., for indexing. Conway and Sloane [105] found such a method for the important case that the support has the shape of an enlarged cell. For sources with infinite support, such as i.i.d. Gaussian, there is also the difficult question of how to quantize a source vector x lying outside the support region. For example, one might scale x so that it lies on or just inside the boundary of the support region, and then quantize the scaled vector in the usual way. Unfortunately, this simple method does not always find the closest codevector to x. Indeed, it often increases overload distortion substantially over that of the minimum-distance quantization rule. To date, there is apparently no low complexity method that does not substantially increase overload distortion.

High resolution theory applies immediately to lattice VQ when the entire lattice is considered to be the codebook. The theory becomes more difficult if, as is usually the case, only a bounded portion of the lattice is used as the codebook and one must separately consider granular and overload distortion. There are a variety of ways of considering the tradeoffs involved, cf. [580], [151], [359], [149], [409]. In any case, the essence of a lattice code is its uniform point density and nicely shaped cells with low normalized moment of inertia. For fixed-rate coding, they work well for uniform sources or other sources with bounded support. But as discussed earlier, for sources with unbounded support such as i.i.d. Gaussian, they require very large dimensions to achieve performance close to $\overline{\delta}(R)$.

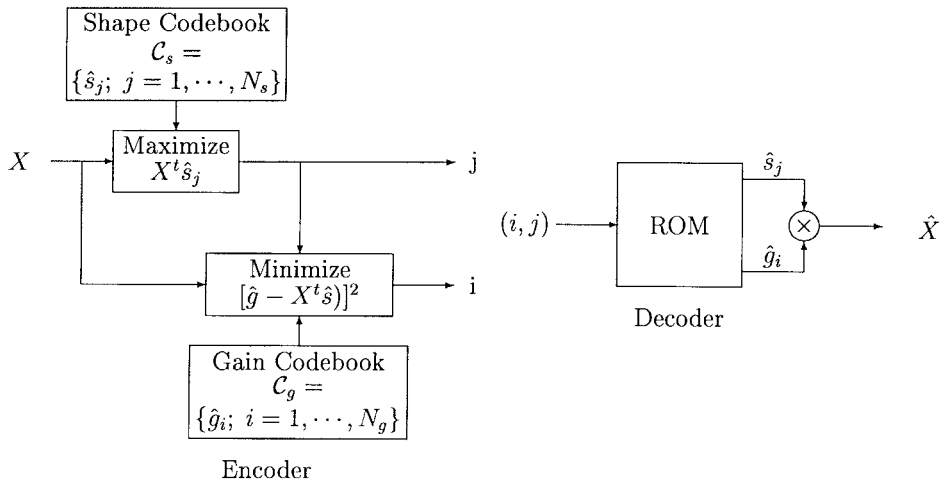


Fig. 8. Shape-gain VQ.

Product Quantizers: A product quantizer uses a reproduction codebook that is the Cartesian product of lower dimensional reproduction codebooks. For example, the application of a scalar quantizer to k successive samples $X_1, X_2, \cdots, X_k$ can be viewed as a product quantizer operating on the k-dimensional vector $X = (X_1, X_2, \cdots, X_k)$. The product structure makes searching easier and, unlike the special case of a sequence of scalar quantizers, the search need not be comprised of k independent searches. Products of vector quantizers are also possible. Typically, the product quantizer is applied, not to the original vector of samples, but to some functions or features extracted from the vector. The complexities of a product quantizer (arithmetic and storage, encoding and decoding) are the sums of those of the component quantizers. As such, they are ordinarily much less than the complexities of an unstructured quantizer with the same number of codevectors, whose complexities equal the product of those of the components of a product quantizer.

A *shape-gain* vector quantizer [449], [450] is an example of a product quantizer. It uses a product reproduction codebook consisting of a gain codebook $\mathcal{C}_g = \{\hat{g}_i; i = 1, \cdots, N_g\}$ of positive scalars and a shape codebook $\mathcal{C}_s = \{\hat{s}_j; j = 1, \cdots, N_s\}$ of unit norm k-dimensional vectors, and the overall reproduction vector is defined by $\hat{x} = \hat{g}\hat{s}$. It is easy to see the minimum-squared-error reproduction codeword $\hat{g}_i\hat{s}_j$ for an input vector x is found by the following encoding algorithm: First choose the index j that maximizes the correlation $x^t\hat{s}_j$, then for this chosen j choose the index i minimizing $|\hat{g}_i - x^t\hat{s}_j|$. This sequential rule gives the minimum-squared-error reproduction codeword without explicitly normalizing the input vector (which would be computationally expensive). The encoder and decoder are depicted in Fig. 8.

A potential advantage of such a system is that by separating these two "features," one is able to use a scalar quantizer for the gain feature and a lower rate codebook for the shape feature, which can then have a higher dimension, for the same search complexity. A major issue arises here: given a total rate constraint, how does one best divide the bits between the two codebooks? This is an example of a rate-allocation problem that arises in all product codebooks and about which more will be said shortly.

It is important to notice that the use of a product quantizer does not mean the use of independent quantizers for each component. As with shape-gain VQ, the optimal lossy encoder will in general not view only one coordinate at a time. Separate and independent quantization of the components provides a low-complexity but generally suboptimal encoder. In the case of the shape-gain VQ, the optimal lossy encoder is happily a simple sequential operation, where the gain quantizer is scalar, but the selection of one of its quantization levels depends on the result of another quantizer, the shape quantizer. Similar ideas can be used for mean-removed VQ [20], [21] and mean/gain/shape VQ [392]. The most general formulation of product codes has been given by Chan and Gersho [82]. It includes a number of schemes with dependent quantization, even tree-structured and multistage quantization, to be discussed later.

Fischer's *pyramid VQ* [164] is also a kind of shape-gain VQ. In this case, the codevectors of the shape codebook are constrained to lie on the surface of a k-dimensional pyramid, namely, the set of all vectors whose components have magnitudes summing to one. Pyramid VQ's are very well suited to i.i.d. Laplacian sources. An efficient method for indexing the shape codevectors is needed and a suitable method is included in pyramid VQ.

Two-dimensional shape-gain product quantizers, usually called *polar quantizers*, have been extensively developed [182], [183], [407], [406], [61], [62], [530], [489], [490], [483], [485], [488], [360]. Here, a two-dimensional source vector is represented in polar coordinates and, in the basic scheme, the codebook consists of the Cartesian product of a nonuniform scalar codebook for the magnitude and a uniform scalar codebook for the phase. Early versions of polar quantization used independent quantization of the magnitude and phase information, but later versions used the better method described above, and some even allowed the phase quantizers to have a resolution that depends on the outcome of the magnitude quantizer. Such polar quantizers

are called "unrestricted" [488], [530]. High-resolution analysis can be used to study the rate-distortion performance of these quantizers [61], [62], [483], [485], [488], [360]. Among other things, such analyses find the optimal point density for the magnitude quantizer and the optimal bit allocation between magnitude and phase. Originally, methods were developed specifically for polar quantizers. However, recently it has been shown that Bennett's integral can be applied to analyze polar quantization in a straightforward way [380]. It turns out that for an i.i.d. Gaussian source, optimized conventional polar quantization gains about 0.41 dB over direct scalar quantization, and optimized unrestricted polar quantization gains another 0.73 dB. Indeed, the latter has, asymptotically, square cells and the optimal two-dimensional point density, and loses only 0.17 dB relative to optimal two-dimensional vector quantization, but is still 3.11 dB from $\overline{\delta}(R)$.

Product quantizers can be used for any set of features deemed natural for decomposing a vector. Perhaps the most famous example is one we have seen already and now revisit: transform coding.

Transform Coding: Though the goal of this section is mainly to discuss techniques beyond scalar quantization, DPCM and transform coding, we discuss the latter here because of its relationships to other techniques and because we wish to discuss work on the bit-allocation problem.

Traditional transform coding can be viewed as a product quantizer operating on the transform coefficients resulting from a linear transform on the original vector. We have already mentioned the traditional high-resolution fixed-rate analysis and the more recent high-resolution entropy-constrained analysis for separate lossless coding of each quantized transform coefficient. An asymptotic low-resolution analysis [338], [339] has also been performed. In almost all actual implementations, however, scalar quantizers are combined with a block lossless code, where the lossless code is allowed to effectively operate on an entire block of quantized coefficients at once, usually by combining run-length coding with Huffman or arithmetic coding. As a result, the usual high-resolution analyses are not directly applicable.

Although high resolution theory shows that the Karhunen–Loève transform is optimal for Gaussian sources, and the asymptotic low-resolution analysis does likewise, the dominant transform for many years has been the discrete cosine transform (DCT) used in most current image and video coding standards. The primary competition for future standards comes from discrete wavelet transforms, which will be considered shortly. One reason for the use of the DCT is its lower complexity. An "unstructured" transform like the Karhunen–Loève requires approximately $2k$ operations per sample, which is small compared to the arithmetic complexity of unstructured VQ, but large compared to the approximately $\log k$ operations per sample for a DCT. Another motivation for the DCT is that in some sense it approximates the behavior of the Karhunen–Loève transform for certain sources. And a final motivation is that the frequency decomposition done by the DCT mimics, to some extent, that done by the human visual system and so one may quantize the DCT coefficients taking perception into account. We will not delve into the large literature of transforms, but will observe that bit allocation becomes an important issue, and one can either use the high-resolution approximations or a variety of nonasymptotic allocation algorithms such as the "fixed-slope" or Pareto-optimality considered in [526], [470], [94], [439], [438], and [463]. The method involves operating all quantizers at points on their operational distortion-rate curves of equal slopes. For a survey of some of these methods, see [107] or [196, Ch. 10]. A combinatorial optimization method is given in [546].

As a final comment on traditional transform coding, the code can be considered as being suboptimal as a k-dimensional quantizer because of the constrained structure (transform and product code). It gains, however, in having a low complexity, and transform codes remain among the most popular compression systems because of their balance of performance and complexity.

Subband/Wavelet/Pyramid Quantization: Subband codes, wavelet codes, and pyramid codes are intimately related and all are cousins of a transform code. The oldest of these methods (so far as quantization is concerned) is the pyramid code of Burt and Adelsen [66] (which is quite different from Fischer's pyramid VQ). The Burt and Adelsen pyramid is constructed from an image first by forming a Gaussian pyramid by successively lowpass filtering and downsampling, and then by forming a Laplacian pyramid which replaces each layer of the Gaussian pyramid by a residual image formed by subtracting a prediction of that layer based on the lower resolution layers. The resulting pyramid of images can then be quantized, e.g., by scalar quantizers. The approximation for any layer can be reconstructed by using the inverse quantizers (reproduction decoders) and upsampling and combining the reconstructed layer and all lower resolution reconstructed layers. Note that as one descends the pyramid, one easily combines the new bits for that layer with the bits already used to produce a higher resolution spatially and in amplitude. The pyramid code can be viewed as one of the original multiresolution codes. It can be viewed as a transform code because the entire original structure can be viewed as a linear transform of the original image, but observe that the number of pixels has been roughly doubled.

Subband codes decompose an image into separate images by using a bank of linear filters, hence once again performing a linear transformation on the data prior to quantizing it. Traditional subband coding used filters of equal or roughly equal bandwidth. Wavelet codes can be viewed as subband codes of logarithmically varying bandwidths instead of equal bandwidths, where the filters used satisfy certain properties. Since the introduction of subband codes in the late 1980's and wavelet codes in the early 1990's, the field has blossomed and produced several of the major contenders for the best speech and image compression systems. The literature is beyond the scope of this article to survey, and much is far more concerned with the transforms, filters, or basis functions used and the lossless coding used following quantization than with the quantization itself. Hence we content ourselves with the mention of a few highlights. The interested reader is referred to the book by Vetterli and Kovačević on wavelets and subband coding [516].

Subband coding was introduced in the context of speech coding in 1976 by Crochiere *et al.* [113]. The extension of subband filtering from 1-D to 2-D was made by Vetterli [515] and 2-D subband filtering was first applied to image coding by Woods *et al.* [541], [527], [540]. Early wavelet-coding techniques emphasized scalar or lattice vector quantization [12], [13], [130], [463], [14], [30], [185], and other vector quantization techniques have also been applied to wavelet coefficients, including tree encoding [366], residual vector quantization [295], and other methods [107]. A major breakthrough in performance and complexity came with the introduction of zerotrees [315], [466], [457], which provided an extremely efficient embedded representation of scalar quantized wavelet coefficients, called *embedded zerotree wavelet* (EZW) coding. As done by JPEG in a primitive way, the zerotree approach led to a code which first sent bits about the transform coefficients with the largest magnitude, and then sent subsequent bits describing these significant coefficients to greater accuracy as well as bits about originally less significant coefficients that became significant as the accuracy improved. The zerotree approach has been extended to vector quantization (e.g., [109]), but the slight improvement comes at a significant cost in added complexity. Rate-distortion ideas have been used to optimize the rate-distortion tradeoffs using wavelet packets by minimizing a Lagrangian distortion over code trees and bit assignments [427]. Recently, competitive schemes have demonstrated that separate scalar quantization of individual subbands coupled with a sophisticated but low-complexity lossless coding algorithm called stack-run coding can provide performance nearly as good as EZW [504].

The best wavelet codes tend to use very smart lossless codes, lossless codes which effectively code very large vectors. While wavelet advocates may credit the decomposition itself for the gains in compression, the theory suggests that rather it is the fact that vector entropy coding for very large vectors is feasible.

Scalar-Vector Quantization: Like permutation vector quantization and Fischer's pyramid vector quantizer, Laroia and Farvardin's [305] *scalar-vector quantization* attempts to match the performance of an optimal entropy-constrained scalar quantizer with a low-complexity fixed-rate structured vector quantizer. A derivative technique called *block-constrained quantization* [24], [27], [23], [28] is simpler and easier to describe. Here the reproduction codebook is a subset of the k-fold product of some scalar codebook. Variable-length binary codewords are associated with the scalar levels, and given some target rate R, the k-dimensional codebook contains only those sequences of k quantization levels for which the sum of the lengths of the binary codewords associated with the levels is at most kR. The minimum distortion codevector can be found using dynamic programming. Alternatively, an essentially optimal search can be performed with very low complexity using a knapsack packing or Lagrangian approach. The output of the encoder is the sequence of binary codewords corresponding to the codevector that was found, plus some padded bits if the total does not equal kR. The simplest method requires approximately $20N^2/k + 20$ operations per sample and storage for approximately N^2 numbers, where N is the number of scalar quantization levels. The original scalar-vector method differs in that rational lengths rather than binary codewords are assigned to the scalar quantizer levels, dynamic programming is used to find the best codevector, and the resulting codevectors are losslessly encoded with a kind of lexicographic encoding. For i.i.d. Gaussian sources these methods attain SNR within about 2 dB of $\overline{\delta}(R)$ with k on the order of 100, which is about 0.5 dB from the goal of 1.53 dB larger than $\overline{\delta}(R)$. A high-resolution analysis is given in [26] and [23]. The scalar-vector method extends to sources with memory by combining it with transform coding using a decorrelating or approximately decorrelating transform [305].

Tree-Structured Quantization: In its original and simplest form, a k-dimensional tree-structured vector quantizer (TSVQ) [69] is a fixed-rate quantizer with, say, rate R whose encoding is guided by a balanced (fixed-depth) binary tree of depth kR. There is a codevector associated with each of its 2^{kR} terminal nodes (leaves), and a k-dimensional testvector associated with each of its $2^{kR} - 1$ internal nodes. Quantization of a source vector x proceeds in a tree-structured search by finding which of the two nodes stemming from the root node has the closer testvector to x, then finding which of the two nodes stemming from this node has the closer testvector, and so on, until a terminal node and codevector are found. The binary encoding of this codevector consists of the sequence of kR binary decisions that lead to it. Decoding is done by table lookup as in unstructured VQ. As in successive approximation scalar quantization, TSVQ yields an embedded code with a naturally progressive structure.

With this method, encoding requires storing the tree of testvectors and codevectors, demanding approximately twice the storage of an unstructured codebook. However, encoding requires only $2kR$ distortion calculations, which is a tremendous decrease over the 2^{kR} required by full search of an unstructured codebook. In the case of squared-error distortion, instead of storing testvectors and computing the distortion between x and each of them, at each internal node one may store the normal to the hyperplane bisecting the testvectors at the two nodes stemming from it, and determine on which side of the hyperplane x lies by comparing an inner product of x with the normal to a threshold that is also stored. This reduces the arithmetic complexity and storage roughly in half to approximately kR operations per sample and 2^{kR} vectors. Further reductions in storage are possible, as described in [252].

The usual (but not necessarily optimal) *greedy* method for designing a balanced TSVQ [69], [225] is first to design the testvectors stemming from the root node using the Lloyd algorithm on a training set. Then design the two testvectors stemming from, say, the left one of these by running the Lloyd algorithm on the training vectors that were mapped to the left one, and so on.

In the scalar case, a tree can be found that implements any quantizer, indeed, the optimal quantizer. So tree-structuring loses nothing, though the above design algorithm does not necessarily generate the best possible quantizers. In the multidimensional case, one cannot expect that the greedy algorithm will produce a TSVQ that is as good as the best unstructured

VQ or even the best possible TSVQ. Nevertheless, it seems to work pretty well. It has been observed that in the high-resolution case, the cells of the resulting TSVQ's are mostly a mixture of cubes, cubes cut in half, the latter cut in half again, and so on until smaller cubes are formed. And it has been found for i.i.d. Gauss and Gauss–Markov sources that the performances of TSVQ's with moderate to high rates designed by the greedy algorithm are fairly well predicted by Bennett's integral, assuming the point density is optimum and the cells are an equal mixture of cubes, cubes cut in half, and so on. This sort of analysis indicates that the primary weakness of TSVQ is in the shapes of the cells that it produces. Specifically, its loss relative to optimal k-dimensional fixed-rate VQ ranges from 0.7 dB for $k = 2$ to 2.2 dB for very large dimensions. Part of the loss is $(1/12)/M_k$, the ratio of the normalized moment of inertia of a cube to that of the best k-dimensional cell shape, which approaches 1.53 dB for large k, and the remainder, about 0.5 to 0.7 dB, is due to the oblongitis caused by the cubes being cut into pieces [383]. A paper investigating the nature of TSVQ cells is [569].

Our experience has been that when taking both performance and complexity into account, TSVQ is a very competitive VQ method. For example, we assert that for most of the fast search methods, one can find a TSVQ (with quite possibly a different dimension) that dominates it in the sense that D, R, A, and M are all at least as good. Indeed, many of the fast-search approaches use a tree-structured prequantization. However, in TSVQ the searching tree and codebook are matched in size and character in a way that makes them work well together. A notable exception is the hierarchical table lookup VQ which attains a considerably smaller arithmetic complexity than attainable with TSVQ, at the expense of higher storage. The TSVQ will still be competitive in terms of throughput, however, as the tree-structured search is amenable to pipelining.

TSVQ's can be generalized to unbalanced trees (with variable depth as opposed to the fixed depth discussed above) [342], [94], [439], [196] and with larger branching factors than two or even variable branching factors [460]. However, it should be recalled that the goodness of the original TSVQ means that the gains of such are not likely to be substantial except in the low-resolution case or if variable-rate coding is used or if the source has some complex structure that the usual greedy algorithm cannot exploit.

A tree-structured quantizer is analogous to a classification or regression tree, and as such unbalanced TSVQ's can be designed by algorithms based on a gardening metaphor of *growing* and *pruning*. The most well known is the CART algorithm of Breiman, Friedman, Olshen, and Stone [53], and the variation of CART for designing TSVQ's bears their initials: the BFOS algorithm [94], [439], [196]. In this method, a balanced or unbalanced tree with more leaves than needed is first grown and then pruned. One can grow a balanced tree by splitting all nodes in each level of the tree, or by splitting one node at a time, e.g., by splitting the node with the largest contribution to the distortion [342] or in a greedy fashion to maximize the decrease in distortion for the increase in rate [439]. Once grown, the tree can be pruned by removing all descendants of any internal node, thereby making it a leaf. This will increase average distortion, but will also decrease the rate. Once again, one can select for pruning the node that offers the best tradeoff in terms of the least increase in distortion per decrease in bits. It can be shown that, for quite general measures of distortion, pruning can be done in an optimal fashion and the optimal subtrees of decreasing rate are nested [94] (see also [355]). It seems likely that in the moderate- to high-rate case, pruning removes leaves corresponding to cells that are oblong such as cubes cut in half, leaving mainly cubic cells. We also wish to emphasize that if variable-rate quantization is desired, the pruning can be done so as to optimize the tradeoff between distortion and leaf entropy.

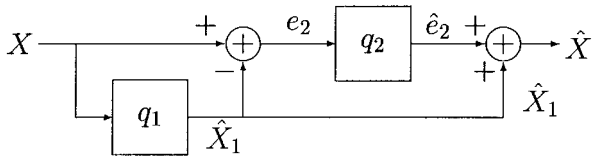


Fig. 9. Two-stage VQ.

There has been a flurry of recent work on the theory of tree-growing algorithms for vector quantizers, which are a form of recursive partitioning. See, for example, the work of Nobel and Olshen [390], [388], [389]. For other work on tree growing and pruning see [393], [439], [276], [22], and [355].

Multistage Vector Quantization: Multistage (or multistep, or cascade, or residual) vector quantization was introduced by Juang and Gray [274] as a form of tree-structured quantization with much reduced arithmetic complexity and storage. Instead of having a separate reproduction codebook for each branch in the tree, a single codebook could be used for all branches of a common length by coding the residual error accumulated to that point instead of coding the input vector directly. In other words, the quantization error (or residual) from the previous stage is quantized in the usual way by the following stage, and a reproduction is formed by summing the previous reproduction and the newly quantized residual. An example of a two-stage quantizer is depicted in Fig. 9. The rate of the multistage quantizer is the sum of the rates of the stages, and the distortion is simply that of the last stage. (It is easily seen that the overall error is just that of the last stage.) A multistage quantizer has a *direct sum* reproduction codebook in the sense that it contains all codevectors formed by summing codevectors from the reproduction codebooks used at each stage. One may also view it as a kind of product code in the sense that the reproduction codebook is determined by the Cartesian product of the stage codebooks. And like product quantization, its complexities (arithmetic and storage, encoding and decoding) are the sum of those of the stage quantizers plus a small amount for computing the residuals at the encoder or the sums at the decoder. In contrast, a conventional single-stage quantizer with the same rate and dimension has complexities equal to the product of those of the stage quantizers.

Since the total rate is the sum of the stage rates, a bit-allocation problem arises. In two-stage quantization using fixed-rate, unstructured, k-dimensional VQ's in both stages,

it usually happens that choosing both stages to have the same rate leads to the best performance versus complexity tradeoff. In this case, the complexities are approximately the square root of what they would be for a single-stage quantizer.

Though we restrict attention here to the case where all stages are fixed-rate vector quantizers with the same dimension, there is no reason why they need have the same dimension, have fixed rate, or have any similarity whatsoever. In other words, multistage quantization can be used (and often is) with very different kinds of quantizers in its stages (different dimensions and much different structures, e.g., DPCM or wavelet coding). For example, structuring the stage quantizers leads to good performance and further substantial reductions in complexity, e.g., [243], [79].

Of course, the multistage structuring leads to a suboptimal VQ for its given dimension. In particular, the direct-sum form of the codebook is not usually optimal, and the greedy-search algorithm described above, in which the residual from one stage is quantized by the next, does not find the closest codevector in the direct-sum codebook. Moreover, the usual greedy design method, which uses a Lloyd algorithm to design the first stage in the usual way and then to design the second stage to minimize distortion when operating on the errors of the first, and so on, does not, in general, design an optimal multistage VQ, even for greedy search. However, two-stage VQ's designed in this way work fairly well.

A high-resolution analysis of two-stage VQ using Bennett's integral on the second stage can be found in [311] and [309]. In order to apply Bennett's integral, it was necessary to find the form of the probability density of the quantization error produced by the first stage. This motivated the asymptotic error-density analysis of vector quantization in [312] and [379].

Multistage quantizers have been improved in a number of ways. More sophisticated (than greedy) encoding algorithms can take advantage of the direct sum nature of the codebook to make optimal or nearly optimal searches, though with some (and sometimes a great deal of) increased complexity. And more sophisticated design algorithms (than the greedy one) can also have benefits [32], [177], [81], [31], [33]. Variable-rate multistage quantizers have been developed [243], [297], [298], [441], [296].

Another way of improving multistage VQ is to adapt each stage to the outcome of the previous. One such scheme, introduced by Lee and Neuhoff [310], [309], was motivated by the observation that if the first stage quantizer has high rate, say R_1, then by Gersho's conjecture, the first stage cells all have approximately the shape of T_k, the tesselating polytope with least normalized moment of inertia, and the source density is approximately constant on them. This implies that the conditional distribution of the residual given that the source vector lies in the ith cell differs from that for the jth only by a scaling and rotation, because cell S_j differs from S_i by just a scaling and rotation. Therefore, if first-stage-dependent scaling and rotation are done prior to second-stage quantization, the conditional distribution of the residual will be the same for all cells, and the second stage can be designed for this distribution, rather than having to be a compromise, as is otherwise the case in two-stage VQ. Moreover, since this distribution is essentially uniform on a support region shaped like T_k, the second stage can itself be a uniform tesselation. The net effect is a quantizer that inherits the optimal point density of the first stage[13] and the optimal cell shapes of the second. Therefore, in the high-resolution case, this *cell-conditioned* two-stage VQ works essentially as well as an optimal (single-stage) VQ, but with much less complexity.

Direct implementation of cell-conditioned two-stage VQ, requires the storing of a scale factor and a rotation for each first stage cell, which operate on the first stage residual before quantization by the second stage. Their inverses are applied subsequently. However, since the first stage cells are so nearly spherical, the rotations gain only a small amount, typically about 0.1 dB, and may be omitted. Moreover, since the best known lattice tesselations are so close to the best known tesselations, one may use lattice VQ as the second stage, which further reduces complexity. Good schemes of this sort have even been developed for low to moderate rates by Gibson [270], [271] and Pan and Fischer [403], [404].

Cell-conditioned two-stage quantizers can be viewed as having a piecewise-constant point density of the sort proposed earlier by Kuhlmann and Bucklew [302] as a means of circumventing the fact that optimal vector quantizers cannot be implemented with companders. This approach was further developed by Swaszek in [487].

Another scheme for adapting each stage to the previous is called codebook sharing, as introduced by Chan and Gersho [80], [82]. With this approach, each stage has a finite set of reproduction codebooks, one of which is used to quantize the residual, depending on the sequence of outcomes from the previous stages. Thus each codebook is shared among some subset of the possible sequences of outcomes from the previous stages. This method lies between conventional multistage VQ in which each stage has one codebook that is shared among all sequences of outcomes from previous stages, and TSVQ in which, in effect, a different codebook is used for each sequence of outcomes from the previous stages. Chan and Gersho introduced a Lloyd-style iterative design algorithm for designing shared codebooks; they showed that by controlling the number and rate of the codebooks one could optimize multistage VQ with a constraint on storage; and they used this method to good effect in audio coding [80]. In the larger scheme of things, TSVQ, multistage VQ, and codebook sharing all fit within the broad family of generalized product codes that they introduced in [82].

Feedback Vector Quantization: Just as with scalar quantizers, a vector quantizer can be predictive; simply replace scalars with vectors in the predictive quantization structure depicted in Fig. 3 [235], [116], [85], [417]. Alternatively, the encoder and decoder can share a finite set of states and a quantizer custom designed for each state. Both encoder and decoder must be able to track the state in the absence of channel errors, so that the state must be determinable from knowledge of an initial state combined with the binary codewords transmitted to the decoder. The result is a finite-state version of a predictive

[13] Since the second stage uniformly refines the first stage cells, the overall point density is approximately that of the first stage.

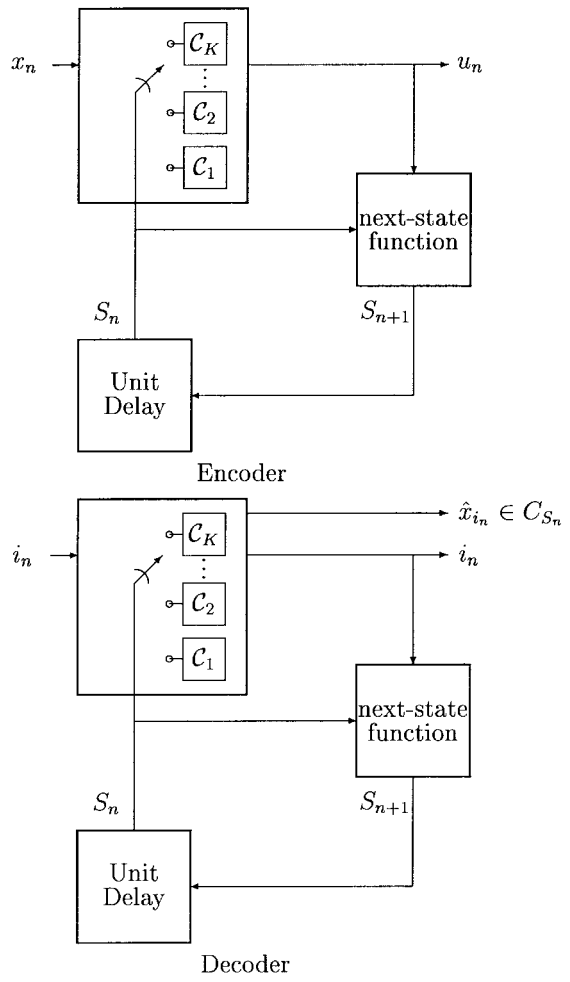


Fig. 10. Finite-state vector quantizer.

quantizer, referred to as a finite-state vector quantizer and depicted in Fig. 10. Although little theory has been developed for finite-state quantizers [161], [178], [179], a variety of design methods exist [174], [175], [136], [236], [15], [16], [286], [196], Lloyd's optimal decoder extends in a natural way to finite-state vector quantizers, the optimal reproduction decoder is a conditional expectation of the input vector given the binary codeword *and* the state. The optimal lossy encoder is not easily described, however, as the next state must be chosen in a way that ensures good future behavior, and not just in a greedy fashion that minimizes the current squared error. If look-ahead is allowed, however, then a tree or trellis search can be used to pick a long-term minimum distortion path, as will be considered in the next subsection.

Both predictive and finite-state vector quantizers typically use memory in the lossy encoder, but use a memoryless lossless code independently applied to each successive binary codeword. One can, of course, also make the lossless code depend on the state, or be conditional on the previous binary codeword. One can also use a memoryless VQ combined with a conditional lossless code (conditioned on the previous binary codeword) designed with a conditional entropy constraint [95], [188]. A simple approach that works for TSVQ is to code the binary path to the codevector for the present source vector relative to the binary path to that of the previous source vector, which is usually very similar. This is a kind of interblock lossless coding [384], [410], [428].

Address-vector quantization, introduced by Nasrabadi and Feng [371] (see also [160] and [373]), is another way to introduce memory into the lossy encoder of a vector quantizer with the goal of attaining higher dimensional performance with lower dimensional complexity. With this approach, in addition to the usual reproduction codebook C, there is an address codebook C_a containing permissible sequences of indices of codevectors in C. The address codebook plays the same role as the outer code in a concatenated channel code (or the trellis in trellis-encoded quantization discussed below), namely, it limits the allowable sequences of codewords from the inner code, which in this case is C. In this way, address-vector quantization can exploit the property that certain sequences of codevectors are much more probable than others; these will be the ones contained in C_a.

As with DPCM, the introduction of memory into the lossy encoder seriously complicates the theory of such codes, which likely explains why there is so little.

Tree/Trellis-Encoded Quantization: Channel coding has often inspired source coding or quantization structures. Channel coding matured much earlier and the dual nature of channel and source coding suggests that a good channel code can be turned into a good source code by reversing the order of encoder and decoder. This role reversal was natural for the codes which eased search requirements by imposition of a tree or trellis structure. Unlike the tree-structured vector quantizers, these earlier systems imposed the tree structure on the sequence of symbols instead of on a single vector of symbols. For the channel coding case, the encoder was a convolutional code, input symbols shifted into a shift register as output symbols, formed by linear combinations (in some field) of the shift-register contents, shifted out. Sequences of output symbols produced in this fashion could be depicted with a tree structure, where each node of the tree corresponded to the state of the shift register (all but the final or oldest symbol) and the branches connecting nodes were determined by the most recent symbol to enter the shift register and were labeled by the corresponding output, the output symbol resulting if that branch is taken. The goal of a channel decoder is to take such a sequence of tree branch labels that has been corrupted by noise, and find a minimum-distance valid sequence of branch labels. This could be accomplished by a tree-search algorithm such as the Fano, stack, or M-algorithm. Since the shift register is finite, the tree becomes redundant and new nodes will correspond to previously seen states so that the tree diagram becomes a merged tree or trellis, which can be searched by a dynamic programming algorithm, the Viterbi algorithm, cf. [173]. In the early 1970's, the algorithms for tree-decoding channel codes were inverted to form tree-encoding algorithms for sources by Jelinek, Anderson, and others [268], [269], [11], [132], [123], [10]. Later, trellis channel-decoding algorithms were modified to trellis-encoding algorithms for sources by Viterbi and Omura [519]. While linear encoders sufficed for channel coding, nonlinear decoders were required for the source coding application, and a variety of design algorithms were developed for designing the decoder to populate the

trellis searched by the encoder [319], [531], [481], [18], [40]. Observe that the reproduction decoder of a finite-state VQ can be used as the decoder in a trellis-encoding system, where the finite-state encoder is replaced by a minimum-distortion search of the decoder trellis implied by the finite-state VQ decoder, which is an optimal encoding for a sequence of inputs.

Tree- and trellis-encoded quantizers can both be considered as a VQ with large blocklength and a reproduction codebook constrained to be the possible outputs of a nonlinear filter or a finite-state quantizer or vector quantizer of smaller dimension. Both structures produce long codewords with a trellis structure, i.e., successive reproduction symbols label the branches of a trellis and the encoder is just a minimum-distortion trellis search algorithm such as the Viterbi algorithm.

Trellis-Coded Quantization: Trellis-coded quantization, both scalar and vector, improves upon traditional trellis-encoded systems by labeling the trellis branches with entire subcodebooks (or "subsets") rather than with individual reproduction levels [345], [344], [166], [167], [522], [343], [478], [514]. The primary gain resulting is a reduction in encoder complexity for a given level of performance. As the original trellis encoding systems were motivated by convolutional channel codes with Viterbi decoders, trellis-coded quantization was motivated by Ungerboeck's enormously successful coded-modulation approach to channel coding for narrowband channels [505], [506].

Recent combinations of TCQ to coding wavelet coefficients [478] have yielded excellent performance in image coding applications, winning the JPEG 2000 contest of 1997 and thereby a position as a serious contender for the new standard.

Gaussian Quantizers: Shannon [465] showed that a Gaussian i.i.d. source had the worst rate-distortion function of any i.i.d. source with the same variance, thereby showing that the Gaussian source was an extremum in a source coding sense. It was long assumed and eventually proved by Sakrison in 1975 [456] that this provided a robust approach to quantization in the sense there exist vector quantizers designed for the i.i.d. Gaussian source with a given average distortion which will provide no worse distortion when applied to any i.i.d. source with the same variance. This provided an approach to *robust* vector quantization, having a code that might not be optimal for the actual source, but which would perform no worse than it would on the Gaussian source for which it was designed.

Sakrison extended the extremal properties of the rate distortion functions to sources with memory [453]–[455] and Lapidoth [306] (1997) showed that a code designed for a Gaussian source would yield essentially the same performance when applied to another process with the same covariance structure.

These results are essentially Shannon theory and hence should be viewed as primarily of interest for high-dimensional quantizers.

In a different approach toward using a Gaussian quantizer on an arbitrary source, Popat and Zeger (1992) took advantage of the central limit theorem and the known structure of an optimal scalar quantizer for a Gaussian random variable to code a general process by first filtering it to produce an approximately Gaussian density, scalar-quantizing the result, and then inverse-filtering to recover the original [419].

C. Robust Quantization

The Gaussian quantizers were described as being *robust* in a minimax average sense: a vector quantizer suitably designed for a Gaussian source will yield no worse average distortion for any source in the class of all sources with the same second-order properties. An alternative formulation of robust quantization is obtained if instead of dealing with average distortion, as is done in most of this paper, one places a maximum distortion requirement on quantizer design. Here a quantizer is considered to be robust if it bounds the maximum distortion for a class of sources. Morris and Vandelinde (1974) [361] developed the theory of robust quantization and provide conditions under which the uniform quantizer is optimum in this minimax sense. This can be viewed as a variation on epsilon entropy since the goal is to minimize the maximum distortion. Further results along this line may be found in [37], [275], [491]. Because these are minimax results aimed at scalar quantization, these results apply to any rate or dimension.

D. Universal Quantization

The minimax approaches provide one means of designing a fixed-rate quantizer for a source with unknown or partially known statistics: a quantizer can be designed that will perform no worse than a fixed value of distortion for all sources in some collection. An alternative approach is to be more greedy and try to design a code that yields nearly optimal performance regardless of which source within some collection is actually coded. This is the idea behind universal quantization.

Universal quantization or universal source coding had its origins in an approach to universal lossless compression developed by Rice and Plaunt [435], [436] and dubbed the "Rice machine." Their idea was to have a lossless coder that would work well for distinct sources by running multiple lossless codes in parallel and choosing the one producing the fewest bits for a period of time, sending a small amount of overhead to inform the decoder which code the encoder was using. The classic work on lossy universal source codes was Ziv's 1972 paper [577], which proved the existence of fixed-rate universal lossy codes under certain assumptions on the source statistics and the source and codebook alphabets. The multiple codebook idea was also used in 1974 [221] to extend the Shannon source coding theorem to nonergodic stationary sources by using the ergodic decomposition to interpret a nonergodic source as a universal coding problem for a family of ergodic sources. The idea is easily described and provides one means of constructing universal codes. Suppose that one has a collection of k-dimensional codebooks $\mathcal{C}_k$ with 2^{kR_k} codevectors, $k = 1, \cdots, K$, each designed for a different type of local behavior. For example, one might have different codebooks in an image coder for edges, textures, and gradients. The union codebook $\bigcup_{k=1}^{K} \mathcal{C}_k$ then contains all the codevectors in all of the codes, for a total of $\sum_{k=1}^{K} 2^{kR_k}$ codevectors. Thus for example, if all of the subcodebooks $\mathcal{C}_k$ have equal rate $R_k = R$, then the rate of the universal code is $R + k^{-1} \log K$

bits per symbol, which can be small if the dimension k is moderately large. This does not mean that it is necessary to use a large-dimensional VQ, since the VQ can be a product VQ, e.g., for an image one could have $k = 64$ by coding each square of dimension $8 \times 8 = 64$ using four applications of a VQ of dimension $4 \times 4 = 16$. If one had, say, four different codes, the resulting rate would be $R + 2/64 = R + 0.031$, which would be a small increase over the original rate if the original rate is, say, 0.25.

A universal code is in theory more complicated than an ordinary code, but in practice it can mean codes with smaller dimension might be more efficient since separate codebooks can be used for distinct short-term behavior.

Subsequently, a variety of notions of fixed-rate universal codes were considered and compared [382], and fixed-distortion codes with variable rate were developed by Mackenthun and Pursley [340] and Kieffer [277], [279].

As with the early development of block source codes, universal quantization during its early days in the 1970's was viewed as more of a method for developing the theory than as a practical code-design algorithm. The Rice machine, however, proved the practicality and importance of a simple multiple codebook scheme for handling composite sources.

These works all assumed the encoder and decoder to possess copies of the codebooks being used. Zeger, Bist, and Linder [566] considered systems where the codebooks are designed at the encoder, but must be also coded and transmitted to the decoder, as is commonly done in codebook replenishment [206].

A good review of the history of universal source coding through the early 1990's may be found in Kieffer (1993) [283].

Better performance tradeoffs can be achieved by allowing both rate and distortion to vary, and in 1996, Chou *et al.* [92] formulated the universal coding problem as an entropy-constrained vector quantization problem for a family of sources and provided existence proofs and Lloyd-style design algorithms for the collection of codebooks subject to a Lagrangian distortion measure, yielding a fixed rate-distortion slope optimization rather than fixed distortion or fixed rate. The clustering of codebooks was originally due to Chou [90] in 1991. High-resolution quantization theory was used to study rates of convergence with blocklength to the optimal performance, yielding results consistent with earlier convergence results developed by other means, e.g., Linder *et al.* [321]. The fixed-slope universal quantizer approach was further developed with other code structures and design algorithms by Yang *et al.* [558].

A different approach which more closely resembles traditional adaptive and codebook replenishment was developed by Zhang, Yang, Wei, and Liu [329], [575], [574]. Their approach, dubbed "gold washing," did not involve training, but rather created and removed codevectors according to the data received and an auxiliary random process in a way that could be tracked by a decoder without side information.

E. Dithering

Dithered quantization was introduced by Roberts [442] in 1962 as a means of randomizing the effects of uniform quantization so as to minimize visual artifacts. It was further developed for images by Limb (1969) [317] and for speech by Jayant and Rabiner (1972) [266]. Intuitively, the goal was to cause the reconstruction error to look more like signal-independent additive white noise. It turns out that for one type of dithering, this intuition is true. In a dithered quantizer, instead of quantizing an input signal X_n directly, one quantizes a signal $U_n = X_n + W_n$, where W_n is a random process, independent of the signal X_n, called a *dither* process. The dither process is usually assumed to be i.i.d.. There are two approaches to dithering. Roberts considered subtractive dithering, where the final reconstruction is formed as $\tilde{X} = q(X_n + W_n) - W_n$. An obvious problem is the need for the decoder to possess a copy of the dither signal. Nonsubtractive dithering forms the reproduction as $\hat{X} = q(X_n + W_n)$.

The principal theoretical property of nonsubtractive dithering was developed by Schuchman [461], who showed that the quantizer error

$$e_n = X_n - \tilde{X}_n = X_n - q(X_n + W_n) + W_n$$

is uniformly distributed on $(-\Delta/2, \Delta/2]$ and is independent of the original input signal X_n if and only if the quantizer does not overload and the characteristic function $M_W(ju) = E[e^{juW}]$ satisfies $M_W(j2\pi l/\Delta) = 0; l \neq 0$. Schuchman's conditions are satisfied, for example, if the dither signal has a uniform probability density function on $(-\Delta/2, \Delta/2]$. It follows from the work of Jayant and Rabiner [266] and Sripad and Snyder [477] (see also [216]) that Schuchman's condition implies that the sequence of quantization errors $\{e_n\}$ is independent. The case of uniform dither remains by far the most widely studied in the literature.

The subtractive dither result is nice mathematically because it promises a well-behaved quantization noise as well as quantization error. It is impractical in many applications, however, for two reasons. First, the receiver will usually not have a perfect analog link to the transmitter (or else the original signal could be sent in analog form) and hence a pseudorandom deterministic sequence must be used at both transmitter and receiver as proposed by Roberts. In this case, however, there will be no mathematical guarantee that the quantization error and noise have the properties which hold for genuinely random i.i.d. dither. Second, subtractive dither of a signal that indeed resembles a sample function of a memoryless random process is complicated to implement, requiring storage of the dither signal, high-precision arithmetic, and perfect synchronization. As a result, it is of interest to study the behavior of the quantization noise in a simple nonsubtractive dithered quantizer. Unlike subtractive dither, nonsubtractive dither is not capable of making the reconstruction error independent of the input signal (although claims to the contrary have been made in the literature). Proper choice of dithering function can, however, make the conditional moments of the reproduction error independent of the input signal. This can be practically important. For example, it can make the perceived quantization noise energy constant as an input signal fades from high intensity to low intensity, where otherwise it can (and does) exhibit strongly signal-dependent behavior. The properties of nonsubtractive dither

were originally developed in unpublished work by Wright [542] in 1979 and Brinton [54] in 1984, and subsequently extended and refined with a variety of proofs [513], [512], [328], [227]. For any $k = 1, 2, \cdots$ necessary and sufficient conditions on the characteristic function M_W are known which ensure that the kth moment of the quantization noise $\epsilon_n = q(X_n + W_n) - X_n$ conditional on X_n does not depend on X_n. A sufficient condition is that the dither signal consists of the sum of k independent uniformly distributed random variables on $[-\Delta/2, \Delta/2]$. Unfortunately, this conditional independence of moments comes at the expense of a loss of fidelity. For example, if $k = 2$ then the quantizer noise power (the mean-squared error) will be

$$E[\epsilon^2|X] = E[\epsilon^2] = E[W^2] + \frac{\Delta^2}{12}.$$

This means that the power in the dither signal is directly added to that of the quantizer error in order to form the overall mean-squared error.

In addition to its role in whitening quantization noise and making the noise or its moments independent of the input, dithering has played a role in proofs of "universal quantization" results in information theory. For example, Ziv [578] showed that even without high resolution theory, uniform scalar quantization combined with dithering and vector lossless coding could yield performance within 0.75 bit/symbol of the rate-distortion function. Extensions to lattice quantization and variations of this result have been developed by Zamir and Feder [565].

F. Quantization for Noisy Channels

The separation theorem of information theory [464], [180] states that nearly optimal communication of an information source over a noisy channel can be accomplished by separately quantizing or source coding the source and channel coding or error-control coding the resulting encoded source for reliable transmission over a noisy channel. Moreover, these two coding functions can be designed separately, without knowledge of each other. The result is only for point-to-point communications, however, and it is a limiting result in the sense that large blocklengths and hence large complexity must be permitted. If one wishes to perform near the Shannon limit for moderate delay or blocklengths, or in multiuser situations, it is necessary to consider joint source and channel codes, codes which jointly consider quantization and reliable communication. It may not actually be necessary to combine the source and channel codes, but simply to jointly design them. There are a variety of code structures and design methods that have been considered for this purpose, many of which involve issues of channel coding which are well beyond the focus of this paper. Here we mention only schemes which can be viewed as quantizers which are modified for use on a noisy channel and not those schemes which involve explicit channel codes. More general discussions can be found, e.g., in [122].

One approach to designing quantizers for use on noisy channels is to replace the distortion measure with respect to which a quantizer is optimized by the expected distortion over the noisy channel. This simple modification of the distortion measure allows the channel statistics to be included in an optimal quantizer design formulation. Recently, the method has been referred to as "channel-optimized quantization," where the quantization might be scalar, vector, or trellis.

This approach was introduced in 1969 by Kurtenbach and Wintz [304] for scalar quantizers. A Shannon source coding theorem for trellis encoders using this distortion measure was proved in 1981 [135] and a Lloyd-style design algorithm for such encoders provided in 1987 [19]. A Lloyd algorithm for vector quantizers using the modified distortion measure was introduced in 1984 by Kumazawa, Kasahara, and Namekawa [303] and further studied in [157], [152], and [153]. The method has also been applied to tree-structured VQ [412]. It can be combined with a maximum-likelihood detector to further improve performance and permit progressive transmission over a noisy channel [411], [523]. Simulated annealing has also been used to design such quantizers [140], [152], [354].

Another approach to joint source and channel coding based on a quantizer structure and not explicitly involving typical channel-coding techniques is to design a scalar or vector quantizer for the source without regard to the channel, but then code the resulting indices in a way that ensures that small (large) Hamming distance of the channel codewords corresponds to small (large) distortion between the resulting reproduction codewords, essentially forcing the topology on the channel codewords to correspond to that of the resulting reproduction codewords. The codes that do this are often called index assignments. Several specific index assignment methods were considered by Rydbeck and Sundberg [448]. DeMarca and Jayant in 1987 [121] introduced an iterative search algorithm for designing index assignments for scalar quantizers, which was extended to vector quantization by Zeger and Gersho [568], who dubbed the approach "pseudo-Gray" coding. Other index assignment algorithms include [210], [543], [287]. For binary-symmetric channels and certain special sources and quantizers, analytical results have been obtained [555], [556], [250], [501], [112], [351], [42], [232], [233], [352]. For example, it was shown by Crimmins *et al.* in 1969 [112] that the index assignment that minimizes mean-squared error for a uniform scalar quantizer used on a binary-symmetric channel is the natural binary assignment. However, this result remained relatively unknown until rederived and generalized in [351].

When source and channel codes are considered together, a key issue is the determination of the quantization rate to be used when the total of number of channel symbols per source symbol is held fixed. For example, as quantization rate is increased, the quantization noise decreases, but channel-induced noise increases because the ability of the channel code to protect the bits is reduced. Clearly, there is an optimal choice of quantization rate. Another issue is the determination of the rate at which overall distortion decreases in an optimal system as the total number of channel uses per source symbol increases. These issues have been addressed in recent papers by Zeger and Manzella [570] and Hochwald and Zeger [244], which use both exponential formulas produced by high resolution quantization theory and exponential bounds to channel coding error probability.

There are a variety of other approaches to joint source and channel coding, including the use of codes with a channel encoder structure optimized for the source or with a special decoder matched to the source, using unequal error protection to better protect more important (lower resolution) reproduction indices, jointly optimized combinations of source and channel codes, and combinations of channel-optimized quantizers with source-optimized channel codes, but we leave these to the literature as they involve a heavy dose of channel coding ideas.

G. Quantizing Noisy Sources

A parallel problem to quantizing for a noisy channel is quantizing for a noisy source. The problem can be seen as trying to compress a dirty source into a clean reproduction, or as doing estimation of the original source based on a quantized version of a noise-corrupted version. If the underlying statistics are known or can be estimated by a training sequence, then this can be treated as a quantization problem with a modified distortion measure, where now the distortion between a noise-corrupted observation $Y = y$ of an unseen original X and a reconstruction $\hat{x}$ based on the encoded and decoded y is given as the conditional expectation $E[d(X, \hat{x})|Y = y]$. The usefulness of this modified distortion for source-coding noisy sources was first seen by Dobrushin and Tsybakov (1962) [134] and was used by Fine (1965) [162] and Sakrison (1968) [452] to obtain information-theoretic bounds an quantization and source coding for noisy sources. Berger (1971) [46] explicitly used the modified distortion in his study of Shannon source coding theorems for noise-corrupted sources.

In 1970, Wolf and Ziv [537] used the modified distortion measure for a squared-error distortion to prove that the optimal quantizer for the modified distortion could be decomposed into the cascade of a minimum mean-squared error estimator followed by an optimal quantizer for the estimated original source. This result was subsequently extended to a more general class of distortion measures include the input-weighted quadratic distortion of Ephraim and Gray [145], where a generalized Lloyd algorithm for design was presented.

Related results and approaches can be found in Witsenhausen's (1980) [535] treatment of rate-distortion theory with modified (or "indirect") distortion measures, and in the Occam filters of Natarajan (1995) [370].

H. Multiple Description Quantization

A topic closely related to quantization for noisy channels is multiple description quantization. The problem is usually formulated as a source-coding or quantization problem over a network, but it is most easily described in terms of packet communications. In the simplest case, suppose that two packets of information, each of rate R, are transmitted to describe a reproduction of a single random vector X. The encoder might receive one or the other packet or the two together and wishes to provide the best reconstruction possible for the bit rate it receives. This can be viewed as a network problem with one receiver seeing only one channel, another receiver seeing the second channel, and a third reciever seeing both channels, and the goal is that each have an optimal reconstruction for the total received bitrate. Clearly, one can do no better than having each packet alone result in in a reproduction with distortion near the Shannon distortion-rate function $D(R)$ while simultaneously having the two packets together yield a reproduction with distortion near $D(2R)$, but this optimistic performance is in general not possible. This problem was first tackled in the information theory community in 1980 by Wolf, Wyner, and Ziv [536] and Ozarow [401] who developed achievable rate regions and lower bounds to performance. The results were extended by Ahlswede (1985) [6], El Gamal and Cover (1982) [139], and Zhang and Berger (1987) [573].

In 1993, Vaishampayan *et al.* used a Lloyd algorithm to actually design fixed-rate [508] and entropy-constrained [509] scalar quantizers for the multiple description problem. High-resolution quantization ideas were used to evaluate achievable performance in 1998 by Vaishampayan and Batllo [510] and Linder, Zamir, and Zeger [324]. An alternative approach to multiple-description quantization using transform coding has also been considered, e.g., in [38] and [211].

I. Other Applications

We have not treated many interesting variations and applications of quantization, several of which have been successfully analyzed or designed using the tools described here. Examples which we would have included had time, space, and patience been more plentiful include mismatch results for quantizers designed for one distribution and applied to another, quantizers designed to provide inputs to classification, detection, or estimation systems, quantizers in multiuser systems such as simple networks, quantizers implicit in finite-precision arithmetic (the modern form of roundoff error), and quantization in noise-shaping analog-to-digital and digital-to-analog converters such as $\Delta\Sigma$-modulators. Doubtless we have failed to mention a few, but this list suffices to demonstrate how rich the theoretical and applied fields of quantization have become in their half century of active development.

ACKNOWLEDGMENT

The authors gratefully acknowledge the many helpful comments, corrections, and suggestions from colleagues, students, and reviewers. Of particular assistance were A. Gersho, B. Girod, N. Kashyap, T. Linder, N. Moayeri, P. Moo, Y. Shtarkov, S. Verdú, M. Vetterli, and K. Zeger.

REFERENCES

[1] E. Abaya and G. L. Wise, "Some notes on optimal quantization," in *Proc. Int. Conf. Communications*, June 1981, vol. 2, pp. 30.7.1–30.7.5.
[2] H. Abut, *Vector Quantization* (IEEE Reprint Collection). Piscataway, NJ: IEEE Press, 1990.
[3] J. P. Adoul, C. Collin, and D. Dalle, "Block encoding and its application to data compression of PCM speech," in *Proc. Canadian Communications and EHV Conf.* (Montreal, Que., Canada, 1978), pp. 145–148.
[4] J.-P. Adoul, J.-L. Debray, and D. Dalle, "Spectral distance measure applied to the optimum design of DPCM coders with L predictors," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)* (Denver, CO, 1980), pp. 512–515.
[5] E. Agrell and T. Eriksson, "Optimization of lattices for quantization," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1814–1828, Sept. 1998. This work also appears in "Lattice-based quantization, Part I" Dept. Inform.

Theory, Chalmers Univ. Technol., Goteborg, Sweden, Rep. 17, Oct. 1996.
[6] R. Ahlswede, "The rate-distortion region for multiple descriptions without excess rate," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 721–726, Nov. 1985.
[7] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, pp. 90–93, 1974.
[8] V. R. Algazi, "Useful approximation to optimum quantization," *IEEE Trans. Commun.*, vol. COM-14, pp. 297–301, June 1966.
[9] M. R. Anderberg, *Cluster Analysis for Applications.* San Diego, CA: Academic, 1973.
[10] J. B. Anderson and J. B. Bodie, "Tree encoding of speech," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 379–387, 1975.
[11] J. B. Anderson and F. Jelinek, "A 2-cycle algorithm for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 77–92, Jan. 1973.
[12] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using vector quantization in the wavelet transform domain," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)* (Albuquerque, NM, Apr. 1990), pp. 2297–2300.
[13] M. Antonini, M. Barlaud, and P. Mathieu, "Image coding using lattice vector quantization of wavelet coefficients," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)* (Toronto, Ont., Canada, May 1991), vol. 4, pp. 2273–2276.
[14] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 205–220, Apr. 1992.
[15] R. Aravind and A. Gersho, "Low-rate image coding with finite-state vector quantization," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, (Tokyo, Japan, 1986), pp. 137–140.
[16] ——, "Image compression based on vector quantization with finite memory," *Opt. Eng.*, vol. 26, pp. 570–580, July 1987.
[17] D. S. Arnstein, "Quantization error in predictive coders," *IEEE Trans. Commun.*, vol. COM-23, pp. 423–429, Apr. 1975.
[18] E. Ayanŏglu and R. M. Gray, "The design of predictive trellis waveform coders using the generalized Lloyd algorithm," *IEEE Trans. Commun.*, vol. COM-34, pp. 1073–1080, Nov. 1986.
[19] ——, "The design of joint source and channel trellis waveform coders," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 855–865, Nov. 1987.
[20] R. L. Baker and R. M. Gray, "Image compression using nonadaptive spatial vector quantization," in *Conf. Rec. 16th Asilomar Conf. Circuits Systems and Computers* (Asilomar, CA, Nov. 1982), pp. 55–61.
[21] ——, "Differential vector quantization of achromatic imagery," in *Proc. Int. Picture Coding Symp.*, Mar. 1983, pp. 105–106.
[22] M. Balakrishnan, W. A. Pearlman, and L. Lu, "Variable-rate tree-structured vector quantizers," *IEEE Trans. Inform. Theory*, vol. 41, pp. 917–930, July 1995.
[23] A. S. Balamesh, "Block-constrained methods of fixed-rate entropy constrained quantization," Ph.D. dissertation, Univ. Michigan, Ann Arbor, Jan. 1993.
[24] A. S. Balamesh and D. L. Neuhoff, "New methods of fixed-rate entropy-coded quantization," in *Proc. 1992 Conf. Information Sciences and Systems* (Princeton, NJ, Mar. 1992), pp. 665–670.
[25] ——, Unpublished notes, 1992.
[26] ——, "Block-constrained quantization: Asymptotic analysis," *DIMACS Ser. Discr. Math. and Theoretical Comput. Sci.*, vol. 14, pp. 67–74, 1993.
[27] ——, "A new fixed-rate quantization scheme based on arithmetic coding," in *Proc. IEEE Int. Symp. Information Theory* (San Antonio, TX, Jan. 1993), p. 435.
[28] ——, "Block-constrained methods of fixed-rate entropy-coded, scalar quantization," *IEEE Trans. Inform. Theory*, submitted for publication.
[29] G. B. Ball, "Data analysis in the social sciences: What about the details?," in *Proc. Fall Joint Computing Conf.* Washington, DC: Spartan, 1965, pp. 533–559.
[30] M. Barlaud, P. Solé, T. Gaidon, M. Antonini, and P. Mathieu, "Pyramidal lattice vector quantization for multiscale image coding," *IEEE Trans. Image Processing*, vol. 3, pp. 367–381, July 1994.
[31] C. F. Barnes, "New multiple path search technique for residual vector quantizers," in *Proc. Data Compression Conf.* (Snowbird, UT, 1994), pp. 42-51.
[32] C. F. Barnes and R. L. Frost, "Vector quantizers with direct sum codebooks," *IEEE Trans. Inform. Theory*, vol. 39, pp. 565–580, Mar. 1993.
[33] C. F. Barnes, S. A. Rizvi, and N. M. Nasrabadi, "Advances in residual vector quantization: A review," *IEEE Trans. Image Processing*, vol. 5, pp. 226–262, Feb. 1996.
[34] C. W. Barnes, B. N. Tran, and S. H. Leung, "On the statistics of fixed-point roundoff error," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-3, pp. 595–606, June 1985.
[35] E. S. Barnes and N. J. A. Sloane, "The optimal lattice quantizer in three dimensions," *SIAM J. Alg. Discr. Methods*, vol. 4, pp. 30–41, Mar. 1983.
[36] P. Bartlett, T. Linder, and G. Lugosi, "The minimax distortion redundancy in empirical quantizer design," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1802–1813, Sept. 1998.
[37] W. G. Bath and V. D. Vandelinde, "Robust memoryless quantization for minimum signal distortion," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 296–306, 1982.
[38] J.-C. Batllo and V. A. Vaishampayan, "Asymptotic performance of multiple description codes," *IEEE Trans. Inform. Theory*, vol. 43, pp. 703–707, Mar. 1997.
[39] C. D. Bei and R. M. Gray, "An improvement of the minimum distortion encoding algorithm for vector quantization," *IEEE Trans. Commun.*, vol. COM-33, pp. 1132–1133, Oct. 1985.
[40] ——, "Simulation of vector trellis encoding systems," *IEEE Trans. Commun.*, vol. COM-34, pp. 214–218, Mar. 1986.
[41] P. Bello, R. Lincoln, and H. Gish, "Statistical delta modulation," *Proc. IEEE*, vol. 55, pp. 308–319, Mar. 1967.
[42] G. Ben-David and D. Malah, "On the performance of a vector quantizer under channel errors," in *Signal Proc. VI: Theories and Applications, Proc. EUSIPCO'92*, 1992, pp. 1685–1688.
[43] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, July 1948.
[44] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. Assoc. Comput. Mach.*, pp. 209–226, Sept. 1975.
[45] T. Berger, "Rate distortion theory for sources with abstract alphabet and memory," *Inform. Contr.*, vol. 13, pp. 254–273, 1968.
[46] ——, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.
[47] ——, "Optimum quantizers and permutation codes," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 759–765, Nov. 1972.
[48] ——, "Minimum entropy quantizers and permutation codes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 149–157, Mar. 1982.
[49] T. Berger, F. Jelinek, and J. K. Wolf, "Permutation codes for sources," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 160–169, Jan. 1972.
[50] V. Bhaskaran and K. Konstantinides, *Image and Video Compression Standards.* Boston, MA: Kluwer, 1995.
[51] H. S. Black, "Pulse code modulation," *Bell Lab. Rec.*, vol. 25, pp. 265–269, July 1947.
[52] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.
[53] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Belmont, CA: Wadsworth, 1984.
[54] L. K. Brinton, "Nonsubtractive dither," M.S. thesis, Elec. Eng. Dept., Univ. Utah, Salt Lake City, UT, Aug. 1984.
[55] J. D. Bruce, "On the optimum quantization of stationary signals," in *1964 IEEE Int. Conv. Rec.*, 1964, pt. 1, pp. 118–124.
[56] J. A. Bucklew, "Companding and random quantization in several dimensions," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 207–211, Mar. 1981.
[57] ——, "A note on optimal multidimensional companders," *IEEE Trans. Inform. Theory*, vol. IT-29, p. 279, Mar. 1983.
[58] ——, "Two results on the asymptotic performance of quantizers," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 341–348, Mar. 1984.
[59] ——, "A note on the absolute epsilon entropy," *IEEE Trans. Inform. Theory*, vol. 37, pp. 142–144, Jan. 1991.
[60] J. A. Bucklew and N. C. Gallagher, Jr., "A note on optimum quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 365–366, May 1979.
[61] ——, "Quantization schemes for bivariate Gaussian random variables," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 537–543, Sept. 1979.
[62] ——, "Two-dimensional quantization of bivariate circularly symmetric densities," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 667–671, Nov. 1979.
[63] ——, "Some properties of uniform step size quantizers," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 610–613, Sept. 1980.
[64] J. A. Bucklew and G. L. Wise, "Multidimensional asymptotic quantization theory with rth power distortion measures, *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 239–247, Mar. 1982.
[65] J. Buhmann and H. Kühnel, "Vector quantization with complexity costs," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1133–1145, July 1988.
[66] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-31, pp. 532–540, Apr. 1983.

[67] A. Buzo, R. M. Gray, A. H. Gray, Jr., and J. D. Markel, "Optimal quantizations of coefficient vectors in LPC speech," in *1978 Joint Meet. Acoustical Society of America and the Acoustical Society of Japan* (Honolulu, HI, Dec. 1978).
[68] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Optimal quantizations of coefficient vectors in LPC, speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (Washington, DC, Apr. 1979), pp. 52–55.
[69] ——, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 562–574, Oct. 1980.
[70] S. Cambanis and N. Gerr, "A simple class of asymptotically optimal quantizers," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 664–676, Sept. 1983.
[71] J. C. Candy and O. J. Benjamin, "The structure of quantization noise from Sigma-Delta modulation," *IEEE Trans. Commun.*, vol. COM-29, pp. 1316–1323, Sept. 1981.
[72] J. Candy and G. Temes, Eds. *Oversampling Delta-Sigma Data Converters.* New York: IEEE Press, 1991.
[73] R. M. Capocelli and A. DeSantis, "Variations on a theme by Gallager," in *Image and Text Compression*, J. A. Storer, Ed. Boston, MA: Kluwer, 1992, pp. 181–213.
[74] J. R. Caprio, N. Westin, and J. Esposito, "Optimum quantization for minimum distortion," in *Proc. Int Telemetering Conf.*, 1978, pp. 315–323.
[75] N. Chaddha, M. Vishwanath, and P. A. Chou, "Hierarchical vector quantization of perceptually weighted block transforms," in *Proc. Compression Conf.* (Snowbird, UT). Los Alamitos, CA: IEEE Comp. Soc. Press, 1995, pp. 3–12.
[76] D. L. Chaffee, "Applications of rate distortion theory to the bandwidth compression," Ph.D. dissertation, Elec. Eng. Dept., Univ. California, Los Angeles, 1975.
[77] D. L. Chaffee and J. K. Omura, "A very low rate voice compression system," in *Abstracts of Papers IEEE Int. Symp. Information Theory*, Oct. 1974.
[78] C.-K. Chan and L.-M. Po, "A complexity reduction technique for image vector quantization," *IEEE Trans. Image Processing*, vol. 1, pp. 312–321, July 1992.
[79] W.-Y. Chan and A. Gersho, "High fidelity audio transform coding with vector quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (Albuquerque, NM, Apr. 1990), vol. 2, pp. 1109–1112.
[80] ——, "Constrained-storage vector quantization in high fidelity audio transform coding," in *Proc. IEEE Int. Conf. Acoustiics, Speech, and Signal Processing (ICASSP)* (Toronto, Ont., Canada, May 1991), pp. 3597–3600.
[81] ——, "Enhanced multistage vector quantization by joint codebook design," *IEEE Trans. Commun.*, vol. 40, pp. 1693–1697, Nov. 1992.
[82] ——, "Generalized product code vector quantization: a family of efficient techniques for signal compression," *Digital Signal Processing*, vol. 4, pp. 95–126, 1994.
[83] W.-Y. Chan, S. Gupta, and A. Gersho, "Enhanced multistage vector quantization by joint codebook design," *IEEE Trans. Commun.*, vol. 40, pp. 1693–1697, Nov. 1992.
[84] Y.-H. Chan and W. Siu, "In search of the optimal searching sequence for VQ encoding," *IEEE Trans. Commun.*, vol. 43, pp. 2891–2893, Dec. 1995.
[85] P. C. Chang and R. M. Gray, "Gradient algorithms for designing predictive vector quantizers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 679–690, Aug. 1986.
[86] P. C. Chang, J. May, and R. M. Gray, "Hierarchical vector quantizers with table-lookup encoders," in *Proc. 1985 IEEE Int. Conf. Communications*, June 1985, vol. 3, pp. 1452–1455.
[87] D. T. S. Chen, "On two or more dimensional optimum quantizers," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (Hartford, CT, 1977), pp. 640–643.
[88] D.-Y. Cheng, A. Gersho, B. Ramamurthi, and Y. Shoham, "Fast search algorithms for vector quantization and pattern matching," in *Proc. IEEE Int. Conf. Acoust.ics, Speech, and Signal Processing (ICASSP)* (San Diego, CA, Mar. 1984), pp. 911.1–911.4.
[89] D.-Y. Cheng and A. Gersho, "A fast codebook search algorithm for nearest-neighbor pattern matching," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (Tokyo, Japan, Apr. 1986), vol. 1, pp. 265–268.
[90] P. A. Chou, "Code clustering for weighted universal VQ and other applications," in *Proc. IEEE Int. Symp. Information Theory* (Budapest, Hungary, 1991), p. 253.
[91] ——, "The distortion of vector quantizers trained on n vectors decreases to the optimum as $O_p(1/n)$," in *Proc. IEEE Int. Symp. Information Theory* (Trondheim, Norway, 1994).
[92] P. A. Chou, M. Effros, and R. M. Gray, "A vector quantization approach to universal noiseless coding and quantization," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1109–1138, July 1996.
[93] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 31–42, Jan. 1989.
[94] ——, "Optimal pruning with applications to tree-structured source coding and modeling," *IEEE Trans. Inform. Theory*, vol. 35, pp. 299–315, Mar. 1989
[95] P. A. Chou and T. Lookabaugh, "Conditional entropy-constrained vector quantization of linear predictive coefficients," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1990, pp. 187–200.
[96] J. Chow and T. Berger, "Failure of successive refinement for symmetric Gaussian mixtures," *IEEE Trans. Inform. Theory*, vol. 43, pp. 350–352, Jan. 1957.
[97] T. A. C. M. Claasen and A. Jongepier, "Model for the power spectral density of quantization noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 914–917, Aug. 1981.
[98] R. J. Clarke, *Transform Coding of Images.* Orlando, FL: Academic, 1985.
[99] A. G. Clavier, P. F. Panter, and D. D. Grieg, "Distortion in a pulse count modulation system," *AIEE Trans.*, vol. 66, pp. 989–1005, 1947.
[100] ——, "PCM, distortion analysis," *Elec. Eng.*, pp. 1110–1122, Nov. 1947.
[101] D. Cohn, E. Riskin, and R. Ladner, "Theory and practice of vector quantizers trained on small training sets," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, pp. 54–65, Jan. 1994.
[102] R. R. Coifman and M. V. Wickerhauser. "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. 38, pp. 713–718, Mar. 1992.
[103] J. H. Conway and N. J. A. Sloane, "Voronoi regions of lattices, second moments of polytopes, and quantization," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 211–226, Mar. 1982.
[104] ——, "Fast quantizing and decoding algorithms for lattice quantizers and codes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 227–232, Mar. 1982.
[105] ——, "A fast encoding method for lattice codes and quantizers, *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 820–824, Nov. 1983.
[106] ——, *Sphere Packings,Lattices and Groups.* New York: Springer-Verlag, 1988.
[107] P. C. Cosman, R. M. Gray, and M. Vetterli, "Vector quantization of image subbands: A survey," *IEEE Trans. Image Processing*, vol. 5, pp. 202–225, Feb. 1996.
[108] P. C. Cosman, K. O. Perlmutter, S. M. Perlmutter, R. M. Gray, and R. A. Olshen, "Training sequence size and vector quantizer performance," in *Proc. 25th Annu. Asilomar Conf. Signals, Systems, and Computers* (Pacific Grove, CA, Nov. 1991), pp. 434–438.
[109] P. C. Cosman, S. M. Perlmutter, and K. O. Perlmutter, "Tree-structured vector quantization with significance map for wavelet image coding," in *Proc. 1995 IEEE Data Compression Conf. (DCC)*, J. A. Storer and M. Cohn, Eds. Los Alamitos, CA: IEEE Comp. Soc. Press, Mar. 1995.
[110] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Chichester, U.K.: Wiley, 1991.
[111] D. R. Cox, "Note on grouping," *J. Amer. Statist. Assoc.*, vol. 52, 543–547, 1957.
[112] T. R. Crimmins, H. M. Horwitz, C. J. Palermo, and R. V. Palermo, "Minimization of mean-squared error for data transmitted via group codes," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 72–78, Jan. 1969.
[113] R. E. Crochiere, S. M. Webber, and J. K. L. Flanagan, "Digital coding of speech in sub-bands," *Bell Syst. Tech. J.*, vol. 55, pp. 1069–1086, Oct. 1976.
[114] I. Csiszár, "Generalized entropy and quantization problems," in *Proc. 6th Prague Conf.*, 1973, pp. 159–174.
[115] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.
[116] V. Cuperman and A. Gersho, "Vector predictive coding of speech at 16 Kb/s," *IEEE Trans. Commun.*, vol. COM-33, pp. 685–696, July 1985.
[117] C. C. Cutler, "Differential quantization of communication signals," U.S. Patent 2 605 361, July 29, 1952.
[118] T. Dalenius, "The problem of optimum stratification," *Skand. Aktuarietidskrift*, vol. 33, pp. 201–213, 1950.
[119] T. Dalenius and M. Gurney, "The problem of optimum stratification II," *Skand. Aktuarietidskrift*, vol. 34, pp. 203–213, 1951.
[120] E. M. Deloraine and A. H. Reeves, "The 25th anniversary of pulse code modulation," *IEEE Spectrum*, pp. 56–64, May 1965.
[121] J. R. B. DeMarca and N. S. Jayant, "An algorithm for assigning binary indices to the codevectors of multidimensional quantizers," in *Proc.*

IEEE Int. Conf. Communications, June 1987, pp. 1128–1132.
[122] N. Demir and K. Sayood "Joint source/channel coding for variable length codes," in *Proc. 1998 IEEE Data Compression Conf.*, J. A. Storer and M. Cohn, Eds. Los Alamitos, CA: Computer Soc. Press, Mar. 1998, pp. 139–148.
[123] C. R. Davis and M. E. Hellman, "On tree coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 373–378, July 1975.
[124] L. D. Davission, "Information rates for data compression," in *IEEE WESCON*, Session 8, Paper 1, 1968.
[125] L. D. Davisson and R. M. Gray, Eds., *Data Compression*, vol. 14, in *Benchmark Papers in Electrical Engineering and Computer Science.* Stroudsburg, PA: Dowden, Hutchinson, and Ross, 1976.
[126] L. D. Davisson, A. Leon-Garcia, and D. L. Neuhoff, "New results on coding of stationary nonergodic sources," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 137–144, Mar. 1979.
[127] L. D. Davisson and M. B. Pursley, "A direct proof of the coding theorem for discrete sources with memory," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 301–310, May 1975.
[128] F. DeJager, "Delta modulation, a method of PCM transmission using a one-unit code," *Philips Res. Repts.*, vol. 7, 1952.
[129] B. Derjavitch, E. M. Deloraine, and V. Mierlo, French Patent 932 140, Aug. 1946.
[130] R. A. DeVore, B. Jawerth, and B. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. Inform. Theory*, vol. 38, pp. 719–746, Mar. 1992.
[131] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* New York: Springer, 1996.
[132] R. J. Dick, T. Berger, and F. Jelinek, "Tree encoding of Gaussian sources," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 332–336, May 1974.
[133] E. Diday and J. C. Simon, "Clustering analysis," in *Digital Pattern Recognition*, K. S. Fu, Ed. New York: Springer-Verlag, 1976.
[134] R. L. Dobrushin and B. S. Tsybakov, "Information transmission with additional noise," *IRE Trans. Inform. Theory*, vol. IT-8, pp. S293–S304, 1962.
[135] J. G. Dunham and R. M. Gray, "Joint source and noisy channel trellis encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 516–519, July 1981.
[136] M. Ostendorf Dunham and R. M. Gray, "An algorithm for the design of labeled-transition finite-state vector quantizers," *IEEE Trans. Commun.*, vol. COM-33, pp. 83–89, Jan. 1985.
[137] J. G. Dunn, "The performance of a class of n dimensional quantizers for a Gaussian source," in *Proc. Columbia Symp. Signal Transmission Processing* (Columbia Univ., New York, 1965), pp. 76–81; reprinted in *Data Compression* (Benchmark Papers in Electrical Engineering and Computer Science, vol. 14), L. D. Davisson and R. M. Gray, Eds. Stroudsberg, PA: Dowden, Hutchinson and Ross, 1975.
[138] M. Effros, P. A. Chou, and R. M. Gray, "Variable-rate source coding theorems for stationary nonergodic sources," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1920–1925, Nov. 1994.
[139] A. E. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 851–857, Nov. 1982.
[140] A. E. El Gamal, L. A. Hemachandra, I. Shperling, and V. K. Wei, "Using simulated annealing to design good codes," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 116–123, Jan. 1987.
[141] P. Elias, "Predictive coding," Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1950.
[142] ——, "Predictive coding I, and II," *IRE Trans. Inform. Theory*, vol. IT-1, pp. 16–33, Mar. 1955.
[143] ——, "Bounds on performance of optimum quantizers," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 172–184, Mar. 1970.
[144] ——, "Bounds and asymptotes for the performance of multivariate quantizers," *Ann. Math. Statist.*, vol. 41, no. 4, pp. 1249–1259, 1970.
[145] Y. Ephraim and R. M. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive vector quantization," *IEEE Trans. Inform. Theory*, vol. 34, pp. 826–834, July 1988.
[146] W. H. Equitz, "A new vector quantization clustering algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1568–1575, Oct. 1989.
[147] W. Equitz and T. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, pp. 269–275, Mar. 1991.
[148] T. Ericson, "A result on delay-less information transmission," in *Abstracts IEEE Int. Symp. Information Theory* (Grignano, Italy, June, 1979).
[149] T. Eriksson and E. Agrell, "Lattice-based quantization, Part II," Rep. 18, Dept. Inform. Theory, Chalmers Univ. Technol., Goteborg, Sweden, Oct. 1996.
[150] A. M. Eskicioğlu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, pp. 2959–2965, Dec. 1995.
[151] M. Vedat Eyuboğlu and G. D. Forney, Jr., "Lattice and trellis quantization with lattice- and trellis-bounded codebooks-high-rate theory for memoryless sources," *IEEE Trans. Inform. Theory*, vol. 39, pp. 46–59, Jan. 1993.
[152] N. Farvardin, "A study of vector quantization for noisy channels," *IEEE Trans. Inform. Theory*, vol. 36, pp. 799–809, July 1990.
[153] ——, "On the performance and complexity of channel optimized vector quantizers," in *Speech Recognition and Coding: New Advances and Trends*, Berlin, Germany: Springer, 1995, pp. 699–704.
[154] N. Farvardin and F. Y. Lin, "Performance of entropy-constrained block transform quantizers," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1433–1439, Sept. 1991.
[155] N. Farvardin and J. W. Modestino, "Optimal quantizer performance for a class of non-Gaussian memoryless sources," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 485–497, May 1984.
[156] ——, "Rate-distortion performance of DPCM schemes," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 402–418, May 1985.
[157] N. Farvardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: An approach to combined source-channel coding," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 827–838, Nov. 1987.
[158] L. Fejes Toth, *Lagerungen in der Ebene, auf der Kugel und im Raum.* Berlin, Germany: Springer Verlag, 1953.
[159] ——, "Sur la representation d'une population infinie par un nombre fini d'elements," *Acta Math. Acad. Sci. Hung.*, vol. 10, pp. 76–81, 1959.
[160] Y. S. Feng and N. M. Nasrabadi, "Dynamic address-vector quantization of RGB color images," *Proc. Inst. Elec. Eng., Part I, Commun. Speech Vision*, vol. 138, pp. 225–231, Aug. 1991.
[161] T. L. Fine, "Properties of an optimal digital system and applications," *IEEE Trans. Inform. Theory*, vol. IT-10, pp. 287–296, Oct. 1964.
[162] ——, "Optimum mean-square quantization of a noisy input," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 293–294, Apr. 1965.
[163] ——, "The response of a particular nonlinear system with feedback to each of two random processes," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 255–264, Mar. 1968.
[164] T. R. Fischer "A pyramid vector quantizer," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 568–583, July 1986.
[165] ——, "Geometric source coding and vector quantization," *IEEE Trans. Inform. Theory*, vol. 35, pp. 137–145, July 1989.
[166] T. R. Fischer, M. W. Marcellin, and M. Wang, "Trellis-coded vector quantization," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1551–1566, Nov. 1991.
[167] T. R. Fischer and M. Wang, "Entropy-constrained trellis-coded quantization," *IEEE Trans. Inform. Theory*, vol. 38, pp. 415–426, Mar. 1992.
[168] S. Fix, "Rate distortion functions for continuous alphabet memoryless sources," Ph.D. dissertation, Univ. Michigan, Ann Arbor, 1977.
[169] J. K. Flanagan, D. R. Morrell, R. L. Frost, C.J. Read, and B. E. Nelson, "Vector quantization codebook generation using simulated annealing," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing* (Glasgow, Scotland, May 1989), pp. 1759–1762.
[170] P. Fleischer, "Sufficient conditions for achieving minimum distortion in a quantizer," in *IEEE Int. Conv. Rec.*, 1964, pp. 104–111.
[171] B. A. Flury "Principal points," *Biometrika*, vol. 77, no. 1, pp. 31–41, 1990.
[172] E. Forgey, "Cluster analysis of multivariate data: Efficiency vs. interpretability of classification," *Biometrics*, vol. 21, p. 768, 1965 (abstract).
[173] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268–278, Mar. 1973.
[174] J. Foster and R. M. Gray, "Finite-state vector quantization," in *Abstracts 1982 IEEE Int. Symp. Information Theory* (Les Arcs France, June 1982).
[175] J. Foster, R. M. Gray, and M. Ostendorf Dunham, "Finite-state vector quantization for waveform coding," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 348–359, May 1985.
[176] J. H. Friedman, F. Baskett, and L. J. Shustek, "An algorithm for finding nearest neighbors," *IEEE Trans. Comput.*, vol. C-24, pp. 1000–1006, Oct. 1975.
[177] R. L. Frost, C. F. Barnes, and F. Xu, "Design and performance of residual quantizers," in *Proc. Data Compression Conf.*, J. A. Storer and J. H. Reif, Eds. Los Alamitos, CA: IEEE Comp. Soc. Press, Apr. 1991, pp. 129–138.
[178] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 167–186, Mar. 1982.
[179] G. Gabor and Z. Gyorfi, *Recursive Source Coding.* New York: Springer-Verlag, 1986.

[180] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.
[181] ——, "Variations on a theme by Huffman," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 668–674, Nov. 1978.
[182] N. C. Gallagher, Jr., "Discrete spectral phase coding," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 622–624, Sept. 1976.
[183] ——, "Quantizing schemes for the discrete Fourier transform of a random time-series," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 156–163, Mar. 1978.
[184] N. C. Gallagher and J. A. Bucklew, "Properties of minimum mean squared error block quantizers," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 105–107, Jan. 1982.
[185] Z. Gao, F. Chen, B. Belzer, and J. Villasenor, "A comparison of the Z, E_8, and Leech lattices for image subband quantization," in *Proc. 1995 IEEE Data Compression Conf.*, J. A. Storer and M. Cohn, Eds. Los Alamitos, CA: IEEE Comp. Soc. Press, Mar. 1995, pp. 312–321.
[186] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 367–381, Sept. 1995.
[187] M. Garey, D. S. Johnson, and H. S. Witsenhausen, "The complexity of the generalized Lloyd–Max problem," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 255–266, Mar. 1982.
[188] D. P. de Garrido, L. Lu, and W. A. Pearlman, "Conditional entropy-constrained vector quantization of frame difference subband signals," in *Proc. IEEE Int. Conf. Image Processing* (Austin, TX, 1994), pt. 1 (of 3), pp. 745–749.
[189] N. L. Gerr and S. Cambanis, "Analysis of delayed delta modulation," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 496–512, July 1986.
[190] ——, "Analysis of adaptive differential PCM of a stationary Gauss–Markov input," *IEEE Trans. Inform. Theory*, vol. 35, pp. 350–359, May 1987.
[191] A. Gersho, "Stochastic stability of delta modulation," *Bell Syst. Tech. J.*, vol. 51, pp. 821–841, Apr. 1972.
[192] ——, "Principles of quantization," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 427–436, July 1978.
[193] ——, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 373–380, July 1979.
[194] ——, "Optimal nonlinear interpolative vector quantization," *IEEE Trans. Commun.*, vol. 38, pp. 1285–1287, Sept. 1990.
[195] A. Gersho and V. Cuperman, "Vector quantization: A pattern-matching technique for speech coding," *IEEE Commun. Mag.*, vol. 21, pp. 15–21, Dec. 1983.
[196] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Boston, MA: Kluwer, 1992.
[197] A. Gersho and B. Ramamurthi, "Image coding using vector quantization," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing* (Paris, France, Apr. 1982), vol. 1, pp. 428–431.
[198] J. D. Gibson, "Adaptive prediction in speech differential encoding systems," *Proc. IEEE*, vol. 68, pp. 488–525, Apr. 1980.
[199] J. D. Gibson and K. Sayood, "Lattice quantization," *Adv. Electron. Electron Phys.*, vol. 72, pp. 259–330, 1988.
[200] N. Gilchrist and C. Grewin, *Collected Papers on Digital Audio Bit-Rate Reduction.* New York: Audio Eng. Soc., 1996.
[201] B. Girod, "Rate-constrained motion estimation," in *Visual Communication and Image Processing VCIP'94, Proc. SPIE*, A. K. Katsaggelos, Ed., Sept. 1994, vol. 2308, pp. 1026–1034.
[202] B. Girod, R. M. Gray, J. Kovačević, and M. Vetterli, "Image and video coding," part of "The past, present, and future of image and multidimensional signal processing," in *Signal Proc. Mag.*, R. Chellappa, B. Girod, D. C. Munson, Jr., A. M. Telkap, and M. Vetterli, Eds., Mar. 1998, pp. 40–46.
[203] H. Gish, "Optimum quantization of random sequences," Ph.D. dissertation, Harvard Univ., Cambridge, MA, Mar. 1967.
[204] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 676–683, Sept. 1968.
[205] T. J. Goblick and J. L. Holsinger, "Analog source digitization: A comparison of theory and practice," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 323–326, Apr. 1967.
[206] M. Goldberg and H. Sun, "Image sequence coding using vector quantization," *IEEE Trans. Commun.*, vol. COM-34, pp. 703–710, July 1986.
[207] A. J. Goldstein, "Quantization noise in P.C.M.," Bell Telephone Lab. Tech. Memo., Oct. 18, 1957.
[208] R. C. Gonzales and R. C. Woods, *Digital Image Processing.* Reading, MA: Addison-Wesley, 1992.
[209] W. M. Goodall, "Telephony by pulse code modulation," *Bell Syst. Tech. J.*, vol. 26, pp. 395–409, July 1947.
[210] D. J. Goodman and T. J. Moulsley "Using simulated annealing to design transmission codes for analogue sources," *Electron. Lett.*, vol. 24, pp. 617–618, May 1988.
[211] V. K. Goyal and J. Kovačević, "Optimal multiple description transform coding of Gaussian vectors," in *Proc. Data Compression Conf.*, J. A. Storer and M. Cohn, Eds. Los Alamitos, CA: Comp. Soc. Press, Mar./Apr. 1998, pp. 388–397.
[212] R. M. Gray, "Information rates of autoregressive processes," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 516–523, Mar. 1971.
[213] ——, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 480–489, July 1973.
[214] ——, "Vector quantization," *IEEE ASSP Mag.*, vol. 1, pp. 4–29, Apr. 1984.
[215] ——, "Oversampled sigma–delta modulation," *IEEE Trans. Commun.*, vol. COM-35, pp. 481–489, Apr. 1987.
[216] ——, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1220–1244, Nov. 1990.
[217] ——, *Source Coding Theory.* Boston, MA: Kluwer, 1990.
[218] ——, *Entropy and Information Theory.* New York: Springer-Verlag, 1990.
[219] ——, "Combined compression and segmentation of images," in *Proc. 1997 Int. Workshop Mobile Multimedia Communication (MoMuC97)* (Seoul, Korea, Sept./Oct. 1997).
[220] R. M. Gray, A. Buzo, Y. Matsuyama, A. H. Gray, Jr., and J. D. Markel, "Source coding and speech compression," in *Proc. Int. Telemetering Conf.* (Los Angeles, CA, Nov. 1978), vol. XIV, pp. 871–878.
[221] R. M. Gray and L. D. Davission, "Source coding theorems without the ergodic assumption," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 502–516, July 1974.
[222] R. M. Gray and A. H. Gray, Jr., "Asymptotically optimal quantizers," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 143–144, Feb. 1977.
[223] R. M. Gray, A. H. Gray, Jr., and G. Rebolledo, "Optimal speech compression," in *Proc. 13th Asilomar Conf. Circuits Systems and Computers* (Pacific Grove, CA, 1979).
[224] R. M. Gray and E. Karnin, "Multiple local optima in vector quantizers," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 708–721, Nov. 1981.
[225] R. M. Gray and Y. Linde, "Vector quantizers and predictive quantizers for Gauss-Markov sources," *IEEE Trans. Commun.*, vol. COM-30, pp. 381–389, Feb. 1982.
[226] R. M. Gray, S. J. Park, and B. Andrews, "Tiling shapes for image vector quantization," in *Proc. 3rd Int. Conf. Advances in Commun. and Control Systems (COMCON III)* (Victoria, BC, Canada, Sept. 1991).
[227] R. M. Gray and T. G. Stockham, Jr. "Dithered quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 805–812, May 1993.
[228] R. M. Gray and A. D. Wyner, "Source coding over simple networks," *Bell Syst. Tech. J.*, vol. 53, pp. 1681–1721, Nov. 1974.
[229] L. Guan and M. Kamel, "Equal-average hyperplane partitioning method for vector quantization of image data," *Patt. Recogn. Lett.*, vol. 13, pp. 605–609, Oct. 1992.
[230] D. J. Hall and G. B. Ball, "ISODATA: A novel method of data analysis and pattern classification," Stanford Res. Inst., Menlo Park, CA, Tech. Rep., 1965.
[231] P. J. Hahn and V. J. Mathews, "Distortion-limited vector quantization," in *Proc. Data Compression Conf.—DCC'96.* Los Alamitos, CA: IEEE Comp. Soc. Press, 1996, pp. 340–348.
[232] R. Hagen and P. Hedelin, "Robust vector quantization by linear mappings of block-codes," in *Proc. IEEE Int. Symp. Information Theory* (San Antonio, TX, Jan. 1993), p. 171.
[233] ——, "Design methods for VQ by linear mappings of block codes," in *Proc. IEEE Int. Symp. Information Theory* (Trondheim, Norway, June 1994), p. 241.
[234] H. Hang and B. Haskell, "Interpolative vector quantization of color images," *IEEE Trans. Commun.*, vol. 36, pp. 465–470, 1988.
[235] H.-M. Hang and J. W. Woods, "Predictive vector quantization of images," *IEEE Trans. Commun.*, vol. COM-33, pp. 1208–1219, Nov. 1985.
[236] A. Haoui and D. G. Messerschmitt, "Predictive vector quantization," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing* (San Diego, CA, Mar. 1984), vol. 1, pp. 10.10.1–10.10.4.
[237] C. W. Harrison, "Experiments with linear prediction in television," *Bell Syst. Tech. J.*, vol. 31, pp. 764–783, July 1952.
[238] J. A. Hartigan, *Clustering Algorithms.* New York: Wiley, 1975.
[239] B. Haskell, "The computation and bounding of rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 525–531, Sept. 1969.
[240] A. Hayashi, "Differential pulse code modulation of the Wiener process," *IEEE Trans. Commun.*, vol. COM-26, pp. 881–887, June 1978.
[241] ——, "Differential pulse code modulation of stationary Gaussian inputs," *IEEE Trans. Commun.*, vol. COM-26, pp. 1137–1147, Aug. 1978.
[242] E. E. Hilbert, "Cluster compression algorithm: a joint clustering/data

compression concept," Jet Propulsion Lab., Pasadena, CA, Publication 77-43, Dec. 1977.
[243] Y.-S. Ho and A. Gersho, "Variable-rate multi-stage vector quantization for image coding," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 1988, pp. 1156–1159.
[244] B. Hochwald and K. Zeger, "Tradeoff between source and channel coding," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1412–1424, Sept. 1997.
[245] C. H. Hsieh, P. C. Lu, and J. C. Chang, "Fast codebook generation algorithm for vector quantization of images," *Patt. Recogn. Lett.*, vol. 12, pp. 605–609, 1991.
[246] C. H. Hsieh and J. C. Chang, "Lossless compression of VQ index with search-order coding," *IEEE Trans. Image Processing*, vol. 5, pp. 1579–1582, Nov. 1996.
[247] J. Huang, "Quantization of correlated random variables," Ph.D. dissertation, School of Engi., Yale Univ., New Haven, CT, 1962.
[248] J.-Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun.*, vol. COM-11, pp. 289–296, Sept. 1963.
[249] S. H. Huang and S. H. Chen, "Fast encoding algorithm for VQ-based encoding," *Electron. Lett.*, vol. 26, pp. 1618–1619, Sept. 1990.
[250] T. S. Huang, "Optimum binary code," MIT Res. Lab. Electron., Quart. Progr. Rep. 82, pp. 223–225, July 15, 1966.
[251] D. A. Huffman, "A method for the construction of minimum redundancy codes," *Proc. IRE*, vol. 40, pp. 1098–1101, Sept. 1952.
[252] D. Hui, D. F. Lyons, and D. L. Neuhoff, "Reduced storage VQ via secondary quantization," *IEEE Trans. Image Processing*, vol. 7, pp. 477–495, Apr. 1998.
[253] D. Hui and D. L. Neuhoff, "Asymptotic analysis of optimum uniform scalar quantizers for generalized Gaussian distributions," in *Proc. 1994 IEEE Int. Symp. Information Theory* (Trondheim, Norway, June 1994), p. 461.
[254] ———, "When is overload distortion negligible in uniform scalar quantization," in *Proc. 1997 IEEE Int. Symp. Information Theory* (Ulm, Germany, July 1997), p. 517.
[255] ———, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization," *IEEE Trans. Inform. Theory*, submitted for publication.
[256] ———, "On the complexity of scalar quantization," in *Proc. 1995 IEEE Int. Symp. Information Theory* (Whistler, BC, Canada, Sept. 1995), p. 372.
[257] F. Itakura, "Maximum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.
[258] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. 6th Int. Congr. Acoustics* (Tokyo, Japan, Aug. 1968), pp. C-17–C-20.
[259] ———, "A statistical method for estimation of speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53-A, pp. 36–43, 1970.
[260] J. E. Iwersen, "Calculated quantizing noise of single-integration delta-modulation coders," *Bell Syst. Tech. J.*, vol. 48, pp. 2359–2389, Sept. 1969.
[261] A. K. Jain, *Fundamentals of Digital Image Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1989.
[262] E. Janardhanan, "Differentical PCM systems," *IEEE Trans. Commun.*, vol. COM-27, pp. 82–93, Jan. 1979.
[263] R. C. Jancey, "Multidimensional group analysis," *Austrailian J. Botany*, vol. 14, pp. 127–130, 1966.
[264] N. S. Jayant, "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," *Proc. IEEE*, vol. 62, pp. 611–632, May 1974
[265] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video.* Englewood Cliffs, NJ: Prentice-Hall, 1984.
[266] N. S. Jayant and L. R. Rabiner, "The application of dither to the quantization of speech signals," *Bell Syst. Tech. J.*, vol. 51, pp. 1293–1304, July/Aug. 1972.
[267] F. Jelinek, "Evaluation of rate distortion functions for low distortions," *Proc. IEEE* (Lett.), vol. 55, pp. 2067–2068, Nov. 1967.
[268] ———, "Tree encoding of memoryless time-discrete sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 584–590, Sept. 1969.
[269] F. Jelinek and J. B. Anderson, "Instrumentable tree encoding of information sources," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 118–119, Jan. 1971.
[270] D. G. Jeong and J. D. Gibson, "Uniform and piecewise uniform lattice vector quantization for memoryless Gaussian and Laplacian sources," *IEEE Trans. Inform. Theory*, vol. 39, pp. 786–804, May 1993.
[271] ———, "Image coding with uniform and piecewise-uniform vector quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 786–804, May. 1993.
[272] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.
[273] R. L. Joshi and P. G. Poonacha, "A new MMSE encoding algorithm for vector quantization," in *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Processing (ICASSP)* (Toronto, Ont., Canada, 1991), pp. 645–648.
[274] B.-H. Juang and A. H. Gray, Jr., "Multiple stage vector quantization for speech coding," in *Proc. Int. Conf. Acouststics, Speech, and Signal Processing (ICASSP)* (Paris, France, Apr. 1982), vol. 1, pp. 597–600.
[275] D. Kazakos, "New results on robust quantization," *IEEE Trans. Commun.*, pp. 965–974, Aug. 1983.
[276] S.-Z. Kiang, R. L. Baker, G. J. Sullivan, and C.-Y. Chiu, "Recursive optimal pruning with applications to tree structured vector quantizers," *IEEE Trans. Image Processing*, vol. 1, pp. 162–169, Apr. 1992.
[277] J. C. Kieffer, "A generalization of the Pursley–Davisson–Mackenthun universal variable-rate coding theorem," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 694–697, Nov. 1977.
[278] ———, "Block coding for an ergodic source relative to a zero-one valued fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 422–437, July 1978.
[279] ———, "A unified approach to weak universal source coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 674–682, Nov. 1978.
[280] ———, "Exponential rate of convergence for Lloyd's method I," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 205–210, Mar. 1982.
[281] ———, "Stochastic stability for feedback quantization schemes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 248–254, Mar. 1982.
[282] ———, "History of source coding," *Inform. Theory Soc. Newslett.*, vol. 43, pp. 1–5, 1993.
[283] ———, "A survey of the theory of source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1473–1490, Sept. 1993.
[284] J. C. Kieffer and J. G. Dunham, "On a type of stochastic stability for a class of encoding schemes," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 703–797, Nov. 1983.
[285] J. C. Kieffer, T. M. Jahns, and V. A. Obuljen, "New results on optimal entropy-constrained quantization," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1250–1258, Sept. 1988.
[286] T. Kim, "Side match and overlap match vector quantizers for images," *IEEE Trans. Image Processing*, vol. 1, pp. 170–185, Apr. 1992.
[287] P. Knagenhjelm and E. Agrell, "The Hadamard transform—A tool for index assignment," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1139–1151, July 1996.
[288] A. N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IEEE Trans. Inform. Theory*, vol. IT-2, pp. 102–108, Sept. 1956.
[289] H. Kodama, K. Wakasugi, and M. Kasahara, "A construction of optimum vector quantizers by simulated annealing," *Trans. Inst. Electron., Inform. Commun. Eng. B-I*, vol. J74B-I, pp. 58–65, Jan. 1991.
[290] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. Berlin Germany: Springer-Verlag, 1989.
[291] V. Koshélev, "Hierarchical coding of discrete sources," *Probl. Pered. Inform.*, vol. 16, no. 3, pp. 31–49, July–Sept. 1980.
[292] ———, "Estimation of mean error for a discrete successive-approximation scheme," *Probl. Pered. Inform.*, vol. 17, no. 3, pp. 20–33, July–Sept. 1981.
[293] T. Koski and S. Cambanis, "On the statistics of the error in predictive coding for stationary Ornstein-Uhlenbeck processes," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1029–40, May 1992.
[294] T. Koski and L.-E. Persson, "On quantizer distortion and the upper bound for exponential entropy," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1168–1172, July 1991.
[295] F. Kossentini, W. C. Chung, and M. J. T. Smith, "Subband image coding using entropy-constrained residual vector quantization," *Inform. Processing and Manag.*, vol. 30, no. 6, pp. 887–896, 1994.
[296] ———, "Conditional entropy-constrained residual VQ with application to image coding," *IEEE Trans. Image Processing*, vol. 5, pp. 311–320, Feb. 1996.
[297] F. Kossentini, M. J. T. Smith, and C. F. Barnes, "Image coding using entropy-constrained residual vector quantization" *IEEE Trans. Image Processing*, vol. 4, pp. 1349–1357, Oct. 1995.
[298] ———, "Necessary conditions for the optimality of variable-rate residual vector quantizers," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1903–1914, Nov. 1995.
[299] H. P. Kramer and M. V. Mathews, "A linear coding for transmitting a set of correlated signals," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 41–46, Sept. 1956.
[300] E. R. Kretzmer, "Statistics of television signals," *Bell Syst. Tech. J.*, vol. 31, pp. 751–763, July 1952.

[301] A. K. Krishnamurthy, S. C. Ahalt, D. E. Melton, and P. Chen, "Neural networks for vector quantization of speech and images," *IEEE J. Select. Areas Commun.*, vol. 8, pp. 1449–1457, Oct. 1990.
[302] F. Kuhlmann and J. A. Bucklew, "Piecewise uniform vector quantizers," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1259–1263, Sept. 1988.
[303] H. Kumazawa, M. Kasahara, and T. Namekawa, "A construction of vector quantizers for noisy channels," *Electron. and Eng. Japan*, vol. 67-B, pp. 39–47, 1984, translated from *Denshi Tsushin Gakkai Ronbunshi*, vol. 67-B, pp. 1–8, Jan. 1984.
[304] A. J. Kurtenbach and P. A. Wintz, *IEEE Trans. Commun. Technol.*, vol. COM-17, pp. 291–302, Apr. 1969.
[305] R. Laroia and N. Farvardin, " A structured fixed-rate vector quantizer derived from a variable-length scalar quantizer. I. Memoryless sources. II. Vector sources," *IEEE Trans. Inform. Theory*, vol. 39, pp. 851–876, May 1993.
[306] A. Lapidoth, "On the role of mismatch in rate distortion theory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 38–47, Jan. 1997.
[307] C.-H. Lee and L.-H. Chen, "Fast closest codeword search algorithm for vector quantization," *Proc. Inst. Elec. Eng.—Vis. Image Signal Processing* vol. 141, pp. 143–148, June 1994.
[308] ——, "A fast search algorithm for vector quantization using mean pyramids of codewords," *IEEE Trans. Commun.*, vol. 43, pp. 1697–1702, Feb.–Apr. 1995.
[309] D. H. Lee, "Asymptotic quantization error and cell-conditioned two-stage vector quantization," Ph.D. dissertation, Univ. Michigan, Ann Arbor, Dec. 1990.
[310] D. H. Lee and D. L. Neuhoff, "Conditionally corrected two-stage vector quantization," in *Conf. Information Sciences and Systems* (Princeton, NJ, Mar. 1990), pp. 802–806.
[311] ——, "An asymptotic analysis of two-stage vector quantization," in *1991 IEEE Int. Symp. Information Theory* (Budapest, Hungary, June 1991), p. 316.
[312] ——, "Asymptotic distribution of the errors in scalar and vector quantizers," *IEEE Trans. Inform. Theory*, vol. 42, pp. 446–460, Mar. 1996.
[313] D. H. Lee, D. L. Neuhoff, and K. K. Paliwal, "Cell-conditioned two-stage vector quantization of speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (Toronto, Ont., May 1991), vol. 4, pp. 653–656.
[314] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys.—Dokl.*, vol. 10, pp. 707–710, 1966.
[315] A. S. Lewis and G. Knowles, "Image compression using the 2-D, wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 244–250, Apr. 1992.
[316] J. Li, N. Chaddha, and R. M. Gray, "Asymptotic performance of vector quantizers with a perceptual distortion measure," in *1997 IEEE Int. Symp. Information Theory* (Ulm, Germany, June 1997); full paper submitted for publication. Preprint available online at http://www-isl.stanford.edu/gray/compression.html.
[317] J. O. Limb, "Design of dithered waveforms for quantized visual signals," *Bell Syst. Tech. J.*, vol. 48, pp. 2555–2582, Sept. 1968.
[318] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.
[319] Y. Linde and R. M. Gray, "A fake process approach to data compression," *IEEE Trans. Commun.*, vol. COM-26, pp. 840–847, June 1978.
[320] T. Linder, "On asymptotically optimal companding quantization," *Probl. Contr. Inform. Theory*, vol. 20, no. 6, pp. 465–484, 1991.
[321] T. Linder, T. Lugosi, and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1728–1740, Nov. 1994.
[322] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inform. Theory*, vol. 40, pp. 2026–2031, Nov. 1994.
[323] ——, "High-resolution source coding for nondifference distortion measures: The rate distortion function," in *Proc. 1997 IEEE Int. Symp. Information Theory* (Ulm, Germany, June 1997), p. 187. Also, submitted for publication to *IEEE Trans. Inform. Theory*.
[324] T. Linder, R. Zamir, and K. Zeger, "The multiple description rate region for high resolution source coding," in *Proc. Data Compression Conf.*, J. A. Storer and M. Cohn, Eds. Los Alamitos, CA: Comp. Soc. Press, Mar./Apr. 1998.
[325] ——, "High resolution source coding for nondifference distortion measures: multidimensional companding," *IEEE Trans. Inform. Theory*, submitted for publication.
[326] T. Linder and K. Zeger, "Asymptotic entropy-constrained performance of tessellating and universal randomized lattice quantization," *IEEE Trans. Inform. Theory*, vol. 40, pp. 575–579, Mar. 1994.
[327] Y. N. Linkov, "Evaluation of epsilon entropy of random variables for small epsilon," *Probl. Inform. Transm.*, vol. 1, pp. 12–18, 1965; translated from *Probl. Pered. Inform.*, vol. 1, pp. 18–26.
[328] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 355–75, May 1992.
[329] Q. Liu, E. Yang, and Z. Zhang, "A fixed-slope universal sequential algorithm for lossy source coding based on Gold–Washing mechanism," in *Proc. 33rd Annu. Allerton Conf. Communication, Control, and Computing* (Monticello, IL, Urbana-Champaign, IL, Oct. 1995), pp. 466–474.
[330] S. P. Lloyd, "Least squares quantization in PCM," unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical Statistics Meet., Atlantic City, NJ, Sept. 1957. Also, *IEEE Trans. Inform. Theory* (Special Issue on Quantization), vol. IT-28, pp. 129–137, Mar. 1982.
[331] ——, "Rate versus fidelity for the binary source," *Bell Syst. Tech. J.*, vol. 56, pp. 427–437, Mar. 1977.
[332] K. T. Lo and W. K. Cham, "Subcodebook searching algorithm for efficient VQ encoding of images," *Proc. Inst. Elec. Eng.–Vis. Image Signal Processing*, vol. 140, pp. 327–330, Oct. 1993.
[333] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1020–1033, Sept. 1989.
[334] A. Lowry, S. Hossain, and W. Millar, "Binary search trees for vector quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (Dallas, TX, 1987), pp. 2206–2208.
[335] G. Lugosi and A. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," *Ann. Statist.*, vol. 24, pp. 687–706, 1996.
[336] J. Łukaszewicz and H. Steinhaus, "On measuring by comparison," *Zastosowania Matematyki*, vol. 2, pp. 225–231, 1955, in Polish.
[337] S. P. Luttrell, "Self-supervised training of hierarchical vector quantizers," in *II Int. Conf. Artificial Neural Networks* (London, U.K., IEE, 1991), Conf. Publ. 349, pp. 5–9.
[338] D. F. Lyons, "Fundamental limits of low-rate transform codes," Ph.D. dissertation, Univ. Michigan, Ann Arbor, 1992.
[339] D. F. Lyons and D. L. Neuhoff, "A coding theorem for low-rate transform codes," in *Proc. IEEE Int. Symp. Information Theory* (San Antonio, TX, Jan. 1993), p. 333.
[340] K. M. Mackenthun and M. B. Pursley, "Strongly and weakly universal source coding," in *Proc. 1977 Conf. Information Science and Systems* (Baltimore, MD, The Johns Hopkins Univ., 1977), pp. 286–291.
[341] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability* 1967, vol. 1, pp. 281–296.
[342] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, pp. 1551–1588, Nov. 1985.
[343] M. W. Marcellin, "On entropy-constrained trellis-coded quantization," *IEEE Trans. Commun.*, vol. 42, pp. 14–16, Jan. 1994.
[344] M. W. Marcellin and T. R. Fischer, "Trellis coded quantization of memoryless and Gauss–Markov sources," *IEEE Trans. Commun*, vol. 38, pp. 82–93, Jan. 1990.
[345] M. W. Marcellin, T. R. Fischer, and J. D. Gibson, "Predictive trellis coded quantization of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 46–55, Jan. 1990.
[346] E. Masry and S. Cambanis, "Delta modulation of the Wiener process," *IEEE Trans. Commun.*, vol. COM-23, pp. 1297–1300 Nov. 1975.
[347] V. J. Mathews, "Vector quantization of images using the L_∞ distortion measure," in *Proc. Int. Conf. Image Processing* (Washington, DC, Oct. 1995), vol. 1, pp. 109–112.
[348] ——, "Vector quantization using the L_∞ distortion measure," *IEEE Signal Processing Lett.*, vol. 4, pp. 33–35, 1997.
[349] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.
[350] R. A. McDonald, "Signal-to-noise and idle channel performance of DPCM systems with particular application to voice signals," *Bell Syst. Tech. J.*, vol. 45, pp. 1123–1151, Sept. 1966.
[351] S. W. McLaughlin, D. L. Neuhoff, and J. K. Ashley, "Optimal binary index assignments for a class of equiprobable scalar and vector quantizers," *IEEE Trans. Inform. Theory*, vol. 41, pp. 2031–2037, Nov. 1995.
[352] A. Méhes and K. Zeger, "Binary lattice vector quantization with linear block codes and affine index assignments," *IEEE Trans. Inform. Theory*, vol. 44, pp. 79–94, Jan. 1998.
[353] J. Menez, F. Boeri, and D. J. Esteban, "Optimum quantizer algorithm for real-time block quantizing," in *Proc. 1979 IEEE Int. Conf. Acoustics, Speech, and Signal Processin*, 1979, pp. 980–984.

[354] D. Miller and K. Rose, "Combined source-channel vector quantization using deterministic annealing," *IEEE Trans. Commun.*, vol. 42, pp. 347–356, Feb.–Apr. 1994.

[355] N. Moayeri, "Some issues related to fixed-rate pruned tree-structured vector quantizers," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1523–1531, 1995.

[356] N. Moayeri and D. L. Neuhoff, "Theory of lattice-based fine-coarse vector quantization," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1072–1084, July 1991.

[357] ——, "Time-memory tradeoffs in vector quantizer codebook searching based on decision trees," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 490–506, Oct. 1994.

[358] N. Moayeri, D. L. Neuhoff, and W. E. Stark, "Fine-coarse vector quantization," *IEEE Trans. Signal Processing*, vol. 39, pp. 1503–1515, July 1991.

[359] P. W. Moo and D. L. Neuhoff, "An asymptotic analysis of fixed-rate lattice vector quantization," in *Proc. Int. Symp. Information Theory and Its Applications* (Victoria, BC, Canada, Sept. 1996), pp. 409–412.

[360] ——, "Uniform polar quantization revisited," to be published in *Proc. IEEE Int. Symp. Information Theory* (Cambridge, MA, Aug. 17–21, 1998).

[361] J. M. Morris and V. D. Vandelinde, "Robust quantization of discrete-time signals with independent samples," *IEEE Trans. Commun.*, vol. COM-22, pp. 1897–1901, 1974.

[362] K. Motoishi and T. Misumi, "On a fast vector quantization algorithm," in *Proc. VIIth Symp. Information Theory and Its Applications*, 1984, not in INSPEC.

[363] ——, "Fast vector quantization algorithm by using an adaptive searching technique," in *Abstracts IEEE Int. Symp. Information Theory* (San Diego, CA, Jan. 1990).

[364] T. Murakami, K. Asai, and E. Yamazaki, "Vector quantizer of video signals," *Electron. Lett.*, vol. 7, pp. 1005–1006, Nov. 1982.

[365] S. Na and D. L. Neuhoff, " Bennett's integral for vector quantizers," *IEEE Trans. Inform. Theory*, vol. 41, pp. 886–900, July 1995.

[366] S. Nanda and W. A. Pearlman, "Tree coding of image subbands," *IEEE Trans. Image Processing*, vol. 1, pp. 133–147, Apr. 1992.

[367] M. Naraghi-Pour and D. L. Neuhoff, "Mismatched DPCM encoding of autoregressive processes," *IEEE Trans. Inform. Theory*, vol. 36, pp. 296–304, Mar. 1990.

[368] ——, "On the continuity of the stationary state distribution of DPCM," *IEEE Trans. Inform. Theory*, vol. 36, pp. 305–311, Mar. 1990.

[369] ——, "Convergence of the projection method for an autoregressive process and a matched DPCM code," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1255–1264, Nov. 1990.

[370] B. K. Natarajan, "Filtering random noise from deterministic signals via data compression," *IEEE Trans. Signal Processing*, vol. 43, Nov. 1995.

[371] N. M. Nasrabadi and Y. Feng, "Image compression using address-vector quantization," *IEEE Trans. Commun.*, vol. 38, pp. 2166–2173 Dec. 1990.

[372] N. M. Nasrabadi and R. A. King, "Image coding using vector quantization: A review," *IEEE Trans. Commun.*, vol. 36, pp. 957–971, Aug. 1988.

[373] N. M. Nasrabadi, J. U. Roy, and C. Y. Choo, "An interframe hierarchical address-vector quantization," *IEEE Trans. Select. Areas Commun.*, vol. 10, pp. 960–967, June 1992.

[374] A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression.* New York: Plenum, 1988, 2nd ed. 1995.

[375] A. N. Netravali and J. O. Limb, "Picture coding: A review," *Proc. IEEE*, vol. 68, pp. 366–406, Mar. 1980.

[376] A. N. Netravali and R. Saigal, "Optimal quantizer design using a fixed-point algorithm," *Bell Syst. Tech. J.*, vol. 55, pp. 1423–1435, Nov. 1976.

[377] D. L. Neuhoff, "Source coding strategies: Simple quantizers vs. simple noiseless codes," in *Proc. 1986 Conf. Information Sciences and Systems*, Mar. 1986, vol. 1, pp. 267–271.

[378] ——, "Why vector quantizers outperform scalar quantizers on stationary memoryless sources," in *IEEE Int. Symp. Information Theory* (Whistler, BC, Canada, Sept. 1995), p. 438.

[379] ——, "On the asymptotic distribution of the errors in vector quantization," *IEEE Trans. Inform. Theory*, vol. 42, pp. 461–468, Mar. 1996.

[380] ——, "Polar quantization revisited," in *Proc. IEEE Int. Symp. Information Theory* (Ulm, Germany, July 1997), p. 60.

[381] D. L. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 701–713, Sept. 1982.

[382] D. L. Neuhoff, R. M. Gray, and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT–21, pp. 511–523, Sept. 1975.

[383] D. L. Neuhoff and D. H. Lee, "On the performance of tree-structured vector quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (Toronto, Ont., Canada, May 1991), vol. 4, pp. 2277–2280.

[384] D. L. Neuhoff and N. Moayeri, "Tree searched vector quantization with interblock noiseless coding," in *Proc. Conf. Information Science and Systems* (Princeton, NJ, Mar. 1988), pp. 781–783.

[385] D. J. Newman, "The hexagon theorem," Bell Lab. Tech. Memo., 1964, published in the special issue on quantization of the *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 137–139, Mar. 1982.

[386] N. B. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Trans. Commun.*, vol. COM-33, pp. 551–557, June 1985.

[387] N. B. Nill and B. H. Bouxas, "Objective image quality measure derived from digital image power spectra," *Opt. Eng.*, vol. 31, pp. 813–825, Apr. 1992.

[388] A. B. Nobel, "Vanishing distortion and shrinking cells," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1303–1305, July 1996.

[389] ——, "Recursive partitioning to reduce distortion," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1122–1133, July 1997.

[390] A. B. Nobel and R. A. Olshen, "Termination and continuity of greedy growing for tree-structured vector quantizers," *IEEE Trans. Inform. Theory*, vol. 42, pp. 191–205, Jan. 1996.

[391] P. Noll and R. Zelinski, "Bounds on quantizer performance in the low bit-rate region," *IEEE Trans. Commun.*, vol. COM-26, pp. 300–305, Feb. 1978.

[392] K. L. Oehler and R. M. Gray, "Mean-gain-shape vector quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (Minneapolis, MN, Apr. 1993), pp. 241–244.

[393] K. L. Oehler, E. A. Riskin, and R. M.Gray, "Unbalanced tree-growing algorithms for practical image compression," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (Toronto, Ont., Canada, 1991), pp. 2293–2296.

[394] B. M. Oliver, J. Pierce, and C. E. Shannon, "The philosophy of PCM," *Proc. IRE*, vol. 36, pp. 1324–1331, Nov. 1948.

[395] ——, "Efficient coding," *Bell Syst. Tech. J.*, vol. 31, pp. 724–750, July 1952.

[396] J. B. O'Neal, Jr., "A bound on signal-to-quantizing noise ratios for digital encoding systems," *Proc. IEEE*, vol. 55, pp. 287–292, Mar. 1967.

[397] ——, "Signal to quantization noise ratio for differential PCM," *IEEE Trans. Commun.*, vol. COM-19, pp. 568–569, Aug. 1971.

[398] ——, "Entropy coding in speech and television differential PCM systems," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 758–761, Nov. 1971.

[399] M. T. Orchard, "A fast nearest neighbor search algorithm," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* (Toronto, Ont., Canada, 1991), pp. 2297–2300.

[400] M. T. Orchard and C. A. Bouman, "Color quantization of images," *IEEE Trans. Signal Processing*, vol. 39, pp. 2677–2690, Dec. 1991.

[401] L. Ozarow, "On a source-coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, pp. 1909–1921, Dec. 1980.

[402] K. K. Paliwal and V. Ramasubramanian, "Effect of ordering the codebook on the efficiency of the partial distance search algorithm for vector quantization," *IEEE Trans. Commun.*, vol. 37, pp. 538–540, May 1989.

[403] J. Pan and T. R. Fischer, "Vector quantization-lattice vector quantization of speech LPC coefficients," in *Proc. IEEE Int. Conf. Acoust.ics, Speech, and Signal Processing (ICASSP)* (Adelaide, Australia, 1994), pt. 1.

[404] ——, "Two-stage vector quantization-lattice vector quantization," *IEEE Trans. Inform. Theory*, vol. 41, pp. 155–163, Jan. 1995.

[405] P. F. Panter and W. Dite, "Quantizing distortion in pulse-count modulation with nonuniform spacing of levels," *Proc. IRE*, vol. 39, pp. 44–48, Jan. 1951.

[406] W. A. Pearlman, "Polar quantization of a complex Gaussian random variable," *IEEE Trans. Commun.*, vol. COM-27, pp. 892–899, June 1979.

[407] W. A. Pearlman and R. M. Gray, "Source coding of the discrete Fourier transform," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 683–692, Nov. 1978.

[408] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Compression Standard.* New York: Van Nostrand Reinhold, 1993.

[409] C. Pépin, J.-C. Belfiore, and J. Boutros, "Quantization of both stationary and nonstationary Gaussian sources with Voronoi constellations," in *Proc. IEEE Int. Symp. Information Theory* (Ulm, Germany, July 1997), p. 59.

[410] N. Phamdo and N. Farvardin, "Coding of speech LSP parameters using TSVQ with interblock noiseless coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (Albuquerque, NM, 1990), pp. 193–196.

[411] ——, "Optimal detection of discrete Markov sources over discrete memoryless channels—Applications to combined source-channel coding," *IEEE Trans. Inform. Theory*, vol. 40, pp. 186–193, Jan. 1994.

[412] N. Phamdo, N. Farvardin, and T. Moriya, "A unified approach to tree-structured and multistage vector quantization for noisy channels," *IEEE Trans. Inform. Theory*, vol. 39, pp. 835–850, May 1993.
[413] R. Pilc, "The transmission distortion of a source as a function of the encoding block length," *Bell Syst. Tech. J.*, vol. 47, pp. 827–885, 1968.
[414] P. Piret, "Causal sliding block encoders with feedback," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 237–240, Mar. 1979.
[415] G. Poggi, "Fast algorithm for full-search VQ encoding," *Electron. Lett.*, vol. 29, pp. 1141–1142, June 1993.
[416] ———, "Generalized-cost-measure-based address-predictive vector quantization," *IEEE Trans. Image Processing*, vol. 5, pp. 49–55, Jan. 1996.
[417] G. Poggi and R. A. Olshen, "Pruned tree-structured vector quantization of medical images with segmentation and improved prediction," *IEEE Trans. Image Processing*, vol. 4, pp. 734–742, Jan. 1995.
[418] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 199–205, Mar. 1982.
[419] K. Popat and K. Zeger, "Robust quantization of memoryless sources using dispersive FIR filters," *IEEE Trans. Commun.*, vol. 40, pp. 1670–1674, Nov. 1992.
[420] E. Posner and E. Rodemich, "Epsilon entropy and data compression," *Ann. Math. Statist.*, vol. 42, pp. 2079–2125, 1971.
[421] E. Posner, E. Rodemich, and H. Rumsey, Jr., "Epsilon entropy of stochastic processes," *Ann. Math. Statist.*, vol. 38, pp. 1000–1020, 1967.
[422] W. K. Pratt, *Image Transmission Techniques.* New York: Academic, 1979.
[423] S. W. Ra and J. K. Kim, "A fast mean-distance-ordered partial codebook search algorithm for image vector quantization," *IEEE Trans. Circuits Syst. II*, vol. 40, pp. 576–579, Sept. 1993.
[424] M. Rabbani and P. W. Jones, *Digital Image Compression Techniques*, vol. TT7 of *Tutorial Texts in Optical Engineering.* Bellingham, WA: SPIE Opt. Eng. Press, 1991.
[425] V. Ramasubramanian and K. K. Paliwal, "An optimized k-d tree algorithm for fast vector quantization of speech," in *Proc. Euro. Signal Processing Conf.* (Grenoble, France, 1988), pp. 875–878.
[426] ———, "An efficient approximation-elimination algorithm for fast nearest-neighbor search based on a spherical distance coordinate formulation," in *Proc. Euro. Signal Processing Conf.* (Barcelona, Spain, Sept. 1990).
[427] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Trans. Image Processing*, vol. 2, pp. 160–176, Apr. 1993.
[428] X. Ran and N. Farvardin, "Combined VQ-DCT coding of images using interblock noiseless coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (Albuquerque, NM, 1990), pp. 2281–2284.
[429] D. R. Rao and P. Yip, *Discrete Cosine Transform.* San Diego, CA: Academic, 1990.
[430] C. J. Read, D. M. Chabries, R. W. Christiansen, and J. K. Flanagan, "A method for computing the DFT of vector quantized data," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (Glasgow, Scotland, May 1989), pp. 1015–1018.
[431] G. Rebolledo, R. M. Gray, and J. P. Burg, "A multirate voice digitizer based upon vector quantization," *IEEE Trans. Commun.*, vol. COM-30, pp. 721–727, Apr. 1982.
[432] A. H. Reeves, French Patent 852 183, Oct. 3, 1938.
[433] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Math. Acad. Sci. Hungar.*, vol. 10, pp. 193–215, 1959.
[434] S. O. Rice, "Mathematical analysis of random noise," *Bell Syst. Tech. J.*, vol. 23, pp. 282–332, 1944, and vol. 24, pp. 46–156, 1945, reprinted in *Selected Papers on Noise and Stochastic Processes*, N. Wax and N. Wax, Eds. New York: Dover, 1954, pp. 133–294.
[435] R. F. Rice and J. R. Plaunt, "The Rice machine: Television data compression," Jet Propulsion Lab., Pasadena, CA, Tech. Rep. 900-408, Sept. 1970.
[436] ———, "Adaptive variable-length coding for efficient compression of spacecraft television data," *IEEE Trans. Commun.*, vol. COM-19, pp. 889–897, Dec. 1971.
[437] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol. 40, pp. 253–259, Jan. 1994.
[438] E. A. Riskin, "Optimal bit allocation via the generalized BFOS algorithm," *IEEE Trans. Inform. Theory*, vol. 37, pp. 400–402, Mar. 1991.
[439] E. A. Riskin and R. M. Gray, "A greedy tree growing algorithm for the design of variable rate vector quantizers," *IEEE Trans. Signal Processing*, vol. 39, pp. 2500–2507, Nov. 1991.
[440] E. A. Riskin, R. Ladner, R. Wang, and L. E. Atlas, "Index assignment for progressive transmission of full-search vector quantization," *IEEE Trans. Image Processing*, vol. 3, pp. 307–312, May 1994.
[441] S. A. Rizvi, N. M. Nasrabadi, and W. L. Cheng, "Entropy-constrained predictive residual vector quantization," *Opt. Eng.*, vol. 35, pp. 187–197, Jan. 1996.
[442] L. G. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 145–154, Feb. 1962.
[443] G. M. Roe, "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory*, vol. IT-10, pp. 384–385, Oct. 1964.
[444] K. Rose, "Mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1939–1952, Nov. 1994.
[445] K. Rose, E. Gurewitz, and G. C. Fox, "A deterministic annealing approach to clustering," *Pattern Recogn. Lett.*, vol. 11, pp. 589–594, Sept. 1990.
[446] ———, "Vector quantization by deterministic annealing," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1249–1257, July 1992.
[447] ———, "Constrained clustering as an optimization method," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 785–794, Aug. 1993.
[448] N. Rydbeck and C.-E. W. Sundberg, "Analysis of digital errors in nonlinear PCM systems," *IEEE Trans. Commun.*, vol. COM-24, pp. 59–65, Jan. 1976.
[449] M. J. Sabin and R. M. Gray, "Product code vector quantizers for speech waveform coding," in *Conf. Rec. GLOBECOM*, Dec. 1982, pp. 1087–1091.
[450] M. J. Sabin and R. M. Gray, "Product code vector quantizers for waveform and voice coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 474–488, June 1984.
[451] ———, "Global convergence and empirical consistency of the generalized Lloyd algorithm," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 148–155, Mar. 1986.
[452] D. J. Sakrison, "Source encoding in the presence of random disturbance," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 165–167, Jan. 1968.
[453] ———, "The rate distortion function of a Gaussian process with a weighted square error criterion," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 506–508, May 1968.
[454] ———, "The rate distortion function for a class of sources," *Inform. Contr.*, vol. 15, pp. 165–195, Aug. 1969.
[455] ———, "Addendum to 'The rate distortion function of a Gaussian process with a weighted-square error criterion'," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 610–611, Sept. 1969.
[456] ———, "Worst sources and robust codes for difference distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 301–309, May 1975.
[457] A. Said and W. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 6, pp. 243–50, June 1996.
[458] K. Sayood, *Introduction to Data Compression.* San Francisco, CA: Morgan Kaufmann, 1996.
[459] K. Sayood, J. D. Gibson, and M. C. Rost, "An algorithm for uniform vector quantizer design," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 805–814, Nov. 1984.
[460] T. Schmidl, P. C. Cosman, and R. M.Gray, "Unbalanced nonbinary tree-structured vector quantization," in *Proc. 27th Asilomar Conf. on Signals, Systems, and Computers* (Pacific Grove, CA, Oct./Nov. 1993), pp. 1519–1523.
[461] L. Schuchman, "Dither signals and their effects on quantization noise," *IEEE Trans. Commun.*, vol. COM-12, pp. 162–165, Dec. 1964.
[462] M. P. Schutzenberger, "On the quantization of finite dimensional messages," *Inform. Contr.*, vol. 1, pp. 153–158, 1958.
[463] T. Senoo and B. Girod, "Vector quantization for entropy coding of image subbands," *IEEE Trans. Image Processing*, vol. 1, pp. 526–532, Oct. 1992.
[464] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
[465] ———, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec., Pt. 4*, 1959, pp. 142–163.
[466] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.
[467] H. N. Shaver, "Topics in statistical quantization," Syst. Theory Lab., Stanford Electron. Lab., Stanford Univ., Stanford, CA, Tech. Rep. 7050-5, May 1965.
[468] W. F. Sheppard, "On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale," *Proc. London Math. Soc.*, vol. 24, pt. 2, pp. 353–380, 1898.
[469] P. C. Shields, D. L. Neuhoff, L. D. Davisson, and F. Ledrappier, "The distortion-rate function for nonergodic sources," *Ann. Probab.*, vol. 6, no. 1, pp. 138–143, 1978.
[470] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 36,

pp. 1445–1453, Sept. 1988.
[471] V. M. Shtein, "On group transmission with frequency division of channels by the pulse-code modulation method," *Telecommun.*, pp. 169–184, 1959, a translation from *Elektrosvyaz*, no. 2, pp 43–54, 1959.
[472] D. Slepian, "A class of binary signaling alphabets," *Bell Syst. Tech. J.*, vol. 35, pp. 203–234, 1956.
[473] ———, "On delta modulation," *Bell Syst. Tech. J.*, vol. 51, pp. 2101–2136, 1972.
[474] B. Smith, "Instantaneous companding of quantized signals," *Bell Syst. Tech. J.*, vol. 36, pp. 653–709, 1957.
[475] M. R. Soleymani and S. D. Morgera, "An efficient nearest neighbor search method," *IEEE Trans. Commun.*, vol. COM-35, pp. 677–679, July 1987.
[476] ———, "A fast MMSE encoding algorithm for vector quantization," *IEEE Trans. Commun.*, vol. 37, pp. 656–659, June 1989.
[477] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 442–448, Oct. 1977.
[478] P. Sriram and M. Marcellin, "Image coding using wavelet transforms and entropy-constrained trellis-coded quantization," *IEEE Trans. Image Processing*, vol. 4, pp. 725–733, June 1995.
[479] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, pp. 63–86, Jan. 1996.
[480] H. Steinhaus, "Sur la division des corp materiels en parties," *Bull. Acad. Polon. Sci.*, C1. III, vol. IV, pp. 801–804, 1956.
[481] L. C. Stewart, R. M. Gray, and Y. Linde, "The design of trellis waveform coders," *IEEE Trans. Commun.*, vol. 30, pp. 702–710, Apr. 1982.
[482] R. W. Stroh, "Optimum and adaptive differential pulse code modulation," Ph.D. dissertation, Polytech. Inst. Brooklyn, Brooklyn, NY, 1970.
[483] P. F. Swaszek, "Uniform spherical coordinate quantization of spherically symmetric sources," *IEEE Trans. Commun.*, vol. COM-33, pp. 518–521, June 1985.
[484] P. F. Swaszek, Ed., *Quantization* (Benchmark Papers in Electrical Engineering and Computer Science), vol. 29. New York: Van Nostrand Reinhold, 1985.
[485] P. Swaszek, "Asymptotic performance of Dirichlet rotated polar quantizers," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 537–540, July 1985.
[486] ———, "A vector quantizer for the Laplace source," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1355–1365, Sept. 1991.
[487] ———, "Unrestricted multistage vector quantizers," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1169–1174, May 1992.
[488] P. F. Swaszek and T. W. Ku, "Asymptotic performance of unrestricted polar quantizers," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 330–333, Mar. 1986.
[489] P. F. Swaszek and J. B. Thomas, "Optimal circularly symmetric quantizers," *Franklin Inst. J.*, vol. 313, no. 6, pp. 373–384, 1982.
[490] ———, "Multidimensional spherical coordinates quantization," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 570–576, July 1983.
[491] ———, "Design of quantizers from histograms," *IEEE Trans. Commun.*, vol. COM-32, pp. 240–245, 1984.
[492] N. Ta, Y. Attikiouzel, and C. Crebbin, "Vector quantization of images using the competitive networks," in *Proc. 2nd Austrailian Conf. Neural Networks, ACNN'91*, 1991, pp. 258–262.
[493] S. C. Tai, C. C. Lai, and Y. C. Lin, "Two fast nearest neighbor searching algorithms for image vector quantization," *IEEE Trans. Commun.*, vol. 44, pp. 1623–1628, Dec. 1996.
[494] H. H. Tan and K. Yao, "Evaluation of rate-distortion functions for a class of independent identically distributed sources under an absolute magnitude criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 59–64, Jan. 1975.
[495] D. W. Tank and J. J. Hopfield, "Simple 'neural' optimization networks: An A/D converter, signal decision circuit, and a linear programming circuit," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 533–541, May 1986.
[496] T. Tarpey, L. Li, and B. D. Flury, "Principal points and self-consistent points of elliptical distributions," *Ann. Statist.*, vol. 23, no. 1, pp. 103–112, 1995.
[497] R. C. Titsworth, "Optimal threshold and level selection for quantizing data," JPL Space Programs Summary 37-23, vol. IV, pp. 196–200, Calif. Inst. Technol., Pasadena, CA, Oct. 1963.
[498] ———, "Asymptotic results for optimum equally spaced quantization of Gaussian data," JPL Space Programs Summary 37-29, vol. IV, pp. 242–244, Calif. Inst. Technol., Pasadena, CA, Oct. 1964.
[499] I. Tokaji and C. W. Barnes, "Roundoff error statistics for a continuous range of multiplier coefficients," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 52–59, Jan. 1987.
[500] L. Torres and J. Huhuet, "An improvement on codebook search for vector quantization" *IEEE Trans. Commun.*, vol. 42, pp. 208–210, Feb.–Apr. 1994.
[501] R. E. Totty and G. C. Clark, "Reconstruction error in waveform transmission," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 336–338, Apr. 1967.
[502] A. V. Trushkin, "Optimal bit allocation algorithm for quantizing a random vector," *Probl. Inform. Transm.*, vol. 17, no. 3, pp. 156–161, July–Sept. 1981; translated from Russian.
[503] ———, "Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 187–198, Mar. 1982.
[504] M. J. Tsai, J. D. Villasenor, and F. Chen, "Stack-run image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 519–521, Oct. 1996.
[505] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 55–67, Jan. 1982.
[506] ———, "Trellis-coded modulation with redundant signal sets, Parts I and II," *IEEE Commun. Mag.*, vol. 25, pp. 5–21, Feb. 1987.
[507] J. Vaisey and A. Gersho, "Simulated annealing and codebook design," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* (New York, Apr. 1988), pp. 1176–1179.
[508] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 821–824, May 1993.
[509] ———, "Design of entropy-constrained multiple-description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 40, pp. 245–250, Jan. 1994.
[510] V. A. Vaishampayan and J.-C. Batllo "Asymptotic analysis of multiple description quantizers," *IEEE Trans. Inform. Theory*, vol. 44, pp. 278–284, Jan. 1998.
[511] H. Van de Weg, "Quantization noise of a single integration delta modulation system with an N-digit code," *Phillips Res. Rep.*, vol. 8, pp. 568–569, Aug. 1971.
[512] J. Vanderkooy and S. P. Lipshitz, "Dither in digital audio," *J. Audio Eng. Soc.*, vol. 35, pp. 966–975, Dec. 1987.
[513] ———, "Resolution below the least significant bit in digital systems with dither," *J. Audio Eng. Soc.*, vol. 32, pp. 106–113, Nov. 1984, correction *Ibid.*, p. 889.
[514] R. J. van der Vleuten and J. H. Weber, "Construction and evaluation of Trellis-coded quantizers for memoryless sources," *IEEE Trans. Inform. Theory*, vol. 41, pp. 853–859, May 1995.
[515] M. Vetterli, "Multi-dimensional sub-band coding: Some theory and algorithms," *Signal Processing*, vol. 6, pp. 97–112, Apr. 1984.
[516] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding.* Englewood Cliffs, NJ: Prentice-Hall, 1995.
[517] E. Vidal, "An algorithm for finding nearest neighbors in (approximately) constant average time complexity," *Patt. Recogn. Lett.*, vol. 4, pp. 145–157, 1986.
[518] M. Vishwanath and P. Chou, "Efficient algorithm for hierarchical compression of video," in *Proc. Int. Conf. Image Processing* (Austin, TX, Nov. 1994). Los Alamitos, CA: IEEE Comp. Soc. Press, 1994, vol. III, pp. 275–279.
[519] A. J. Viterbi and J. K. Omura, "Trellis encoding of memoryless discrete-time sources with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 325–332, May 1974.
[520] A. G. Vitushkin, *Theory of the Transmission and Processing of Information.* New York: Pergaman, 1961. (Translation by R. Feinstein of *Otsenka Slozhnosti Zadachi Tabulirovaniya.* Moscow, USSR: Fizmatgiz., 1959.)
[521] J. C. Walrand and P. Varaiya, "Optimal causal coding-decoding problems," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 814–820, Nov. 1983.
[522] H. S. Wang and N. Moayeri, "Trellis coded vector quantization," *IEEE Trans. Commun.*, vol. 40, pp. 1273–1276, Aug. 1992.
[523] R. Wang, E. A. Riskin, and R. Ladner, "Codebook organization to enhance maximum a posteriori detection of progressive transmission of vector quantized images over noisy channels," *IEEE Trans. Image Processing*, vol. 5, pp. 37–48, Jan. 1996.
[524] J. Ward, "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Assoc.*, vol. 37, pp. 236–244, Mar. 1963.
[525] G. S. Watson, *Statistics on Spheres.* New York: Wiley, 1983.
[526] P. H. Westerink, J. Biemond, and D. E. Boekee, "An optimal bit allocation algorithm for sub-band coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1988, pp. 757–760.
[527] P. H. Westerink, D. E. Boekee, J. Biemond, and J. W. Woods, "Subband coding of images using vector quantization," *IEEE Trans. Commun.*, vol. 36, pp. 713–719, June 1988.
[528] B. Widrow, "A study of rough amplitude quantization by means of Nyquist sampling theory," *IRE Trans. Circuit Theory*, vol. CT-3, pp. 266–276, 1956.

[529] ———, "Statistical analysis of amplitude quantized sampled data systems," *Trans. AIEE, Pt. II: Appl. Ind.*, vol. 79, pp. 555–568, 1960.
[530] S. G. Wilson, "Magnitude/phase quantization of independent Gaussian variates," *IEEE Trans. Commun.*, vol. COM-28, pp. 1924–1929, Nov. 1990.
[531] S. G. Wilson and D. W. Lytle, "Trellis encoding of continuous-amplitude memoryless sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 211–226, Mar. 1982.
[532] A. P. Wilton and G. F. Carpenter, "Fast search methods for vector lookup in vector quantization," *Electron. Lett.*, vol. 28, pp. 2311–2312, Dec. 1992.
[533] P. A. Wintz, "Transform picture coding," *Proc. IEEE*, vol. 60, pp. 809–820, July 1972.
[534] H. S. Witsenhausen, "On the structure of real-time source coders," *Bell Syst. Tech. J.*, vol. 58, pp. 1437–1451, Jul./Aug. 1979
[535] ———, "Indirect rate-distortion problems," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 518–521, Sept. 1980.
[536] J. K. Wolf, A. D. Wyner, and J. Ziv, "Source coding for multiple descriptions," *Bell Syst. Tech. J.*, vol. 59, pp. 1417–1426, Oct. 1980.
[537] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 406–411, July 1970.
[538] D. Wong, B.-H. Juang, and A. H. Gray, Jr., "An 800 bit/s vector quantization LPC vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 770–779, Oct. 1982.
[539] R. C. Wood, "On optimal quantization," *IEEE Trans. Inform. Theory*, vol. IT-5, pp. 248–252, Mar. 1969.
[540] J. W. Woods, Ed., *Subband Image Coding*. Boston, MA: Kluwer, 1991.
[541] J. W. Woods and S. D. O'Neil, "Subband coding of images," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1278–1288, Oct. 1986.
[542] N. Wright, unpublished work.
[543] H.-S. Wu and J. Barba, "Index allocation in vector quantization for noisy channels," *Electron. Lett.*, vol. 29, pp. 1318–1320, July 1993.
[544] L. Wu and F. Fallside, "On the design of connectionist vector quantizers," *Comp. Speech Language*, vol. 5, pp. 207–229, 1991.
[545] ———, "Source coding and vector quantization with codebook-excited neural networks," *Comp. Speech Language*, vol. 6, pp. 43–276, 1992.
[546] X. Wu, "Globally optimum bit allocation," in *Proc. Data Compression Conf.* (Snowbird, UT, 1993), pp. 22–31.
[547] X. Wu and L. Guan, "Acceleration of the LBG algorithm," *IEEE Trans. Commun.*, vol. 42, pp. 1518–1523, Feb.–Apr. 1994.
[548] A. D. Wyner, "Communication of analog data from a Gaussian source over a noisy channel," *Bell Syst. Tech. J.*, vol. 47, pp. 801–812, May/June 1968.
[549] ———, "Recent results in the Shannon theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 2–10, Jan. 1994.
[550] A. D. Wyner and J. Ziv, "Bounds on the rate-distortion function for stationary sources with memory," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 508–513, Sept. 1971.
[551] Y. Yamada, K. Fujita, and S. Tazaki, "Vector quantization of video signals," in *Proc. Annu. Conf. IECE*, 1980, p. 1031.
[552] Y. Yamada and S. Tazaki, "Vector quantizer design for video signals," *IECE Trans.*, vol. J66-B, pp. 965–972, 1983.
[553] ———, "Recursive vector quantization for monochrome video signals," *IEICE Trans.*, vol. E74, pp. 399–405, Feb. 1991.
[554] Y. Yamada, S. Tazaki, and R. M. Gray, "Asymptotic performance of block quantizers with a difference distortion measure," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 6–14, Jan. 1980.
[555] Y. Yamaguchi and T. S. Huang, "Optimum fixed-length binary code," MIT Res. Lab. Electron., Quart. Progr. Rep. 78, pp. 231–233, July 15, 1965.
[556] ———, "Optimum binary code," MIT Res. Lab. Electron., Quart. Progr. Rep. 78, pp. 214–217, July 25, 1965.
[557] H. Yamamoto, "Source coding theory for cascade and branching communication systems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 299–308, May 1981.
[558] E. Yang, Z. Zhang, and T. Berger, "Fixed-slope universal lossy data compression," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1465–1476, Sept. 1997.
[559] K. Yao and H. H. Tan, "Some comments on the generalized Shannon lower bound for stationary finite-alphabet sources with memory," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 815–817, Nov. 1973.
[560] ———, "Absolute error rate-distortion functions for sources with constrained magnitudes," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 499–503, July 1978.
[561] P. L. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. dissertation, Stanford Univ., 1963, also Stanford Univ. Dept. Statist. Tech. Rep.
[562] ———, "Topics in the asymptotic quantization of continuous random variables," Bell Lab. Tech. Memo., 1966.
[563] ———, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 139–148, Mar. 1982, revised version of [562].
[564] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1152–1159, July 1996.
[565] ———, "Information rates of pre/post-filtered dithered quantizers," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1340–1353, Sept. 1996.
[566] K. Zeger, A. Bist, and T. Linder, "Universal source coding with codebook transmission," *IEEE Trans. Commum.*, vol. 42, pp. 336–346, Feb. 1994.
[567] K. Zeger and A. Gersho, "A stochastic relaxation algorithm for improved vector quantiser design," *Electron. Lett.*, vol. 25, pp. 896–898, July 1989.
[568] ———, "Pseudo-Gray coding," *IEEE Trans. Commun.*, vol. 38, pp. 2147–2156, May 1990.
[569] K. Zeger and M. R. Kantorovitz, "Average number of facets per cell in tree-structured vector quantizer partitions," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1053–1055, Sept. 1993.
[570] K. Zeger and V. Manzella, "Asymptotic bounds on optimal noisy channel quantization via random coding," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1926–1938, Nov. 1994.
[571] K. Zeger, J. Vaisey, and A. Gersho, "Globally optimal vector quantizer design by stochastic relaxation," *IEEE Trans. Signal Processing*, vol. 40, pp. 310–322, Feb. 1992.
[572] L. H. Zetterberg, "A comparison between delta and pulse code modulation," *Ericsson Technics*, vol. 11, no. 1, pp. 95–154, 1955.
[573] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 502–521, July 1987.
[574] Z. Zhang and V. K. Wei, "An on-line universal lossy data compression algorithm via continuous codebook refinement. I. Basic results," *IEEE Trans. Inform. Theory*, vol. 42, pp. 803–821, May 1996.
[575] Z. Zhang and E. Yang, "An on-line universal lossy data compression algorithm via continuous codebook refinement. II. Optimality for phi-mixing source models," *IEEE Trans. Inform. Theory*, vol. 42, pp. 822–836, May 1996.
[576] Z. Zhang, E.-H. Yang, and V. K. Wei, "The redundancy of source coding with a fidelity criterion—Part One: Known statistics," *IEEE Trans. Inform. Theory*, vol. 43, pp. 71–91, Jan. 1997.
[577] J. Ziv, "Coding sources with unknown statistics—Part II: Distortion relative to a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 389–394, May 1972.
[578] ———, "Universal quantization," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 344–347, May 1985.
[579] V. N. Koshelev, "Quantizataion with minimal entropy," *Probl. Pered. Inform.*, no. 14, pp. 151–156, 1993.
[580] V. F. Babkin, M. M. Lange, and Yu. M. Shtarkov, "About fixed rate lattice coding of sources with difference fidelity criterion," *Voprosi Kibernetikia, Probl. Redundancy in Inform Syst.*, vol. 34, pp. 10–30, 1977.
[581] ———, "About coding of sequence of independent continuously distributed random values after quantizing," *Voprosi Kibernetikia, Probl. Redundancy in Comp. Networks*, vol. 35, pp. 132–137, 1978.