SCALAR QUANTIZATION WITH ARITHMETIC CODING

by

Ashok Chhabedia Popat

S.B. Massachusetts Institute of Technology
(1986)

Submitted in Partial fulfillment
of the Requirements of the Degree of

MASTER OF SCIENCE
in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

June 1990

© Massachusetts Institute of Technology 1990
All rights reserved

Signature of Author _____
Department of Electrical Engineering and Computer Science
March 28, 1990

Certified by _____
David H. Staelin
Professor of Electrical Engineering
Thesis Supervisor

Accepted by _____
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Scalar Quantization with Arithmetic Coding

by

Ashok C. Popat

Submitted to the Department of Electrical Engineering
and Computer Science on April 6, 1990 in partial
fulfillment of the requirements for the degree
of Master of Science

ABSTRACT:

The problem of scalar quantization of certain memoryless sources with entropy coding is considered. The work is divided into two parts.

In the first part, the problem of scalar quantization is treated, under the assumption that ideal entropy coding can be performed on the output of the quantizer. Special attention is given to Gaussian, Laplacian, and gamma memoryless sources, as these are often used to model various signals that arise in speech- and image-coding systems. A procedure for finding locally optimal quantizers is provided, and results are provided for ν-power distortion measures, where $\nu = 1, 2, ..., 6$. Evidence from a large-scale random search is provided to support the hypothesis that the locally-optimal quantizers are, in most of the cases considered, globally optimal. Based on the same evidence, it is hypothesized that in some cases globally optimal quantizers are asymmetric, despite the symmetry of the source distributions and lack of alphabet constraint. Uniform quantization, which is well-known to be optimum in the limit of vanishingly small distortion, is shown to be nearly optimal for all three sources, over a range of entropies and values of ν. However, it is also shown that locally-optimal quantizers are occasionally substantially nonuniform. It is further demonstrated that the higher the order of the distortion measure, the better entropy-constrained quantization appears, relative to alphabet-constrained quantization. The first part of this thesis also discredits the widely held view that uniform quantizers are exactly optimal with respect to mean-square error in the case of Laplacian sources, and conjectures about how this mistaken idea might have arisen.

The second part of the thesis justifies the assumption of the availability of optimal entropy coding by describing a suitable, near-optimal code. In particular, an efficient arithmetic code well-suited to the quantization problem is devised. It is shown that the performance of this code is very nearly optimal for the sources of interest, over a useful range of entropies. Although the code is based in principle on previously reported work, many details not provided elsewhere are worked out and presented, and the derivation of the code presented is original. Also original is an optimal procedure for rounding-off source letter probabilities to admissible, finite-precision values.

Thesis Supervisor: David H. Staelin

Title: Professor of Electrical Engineering

ACKNOWLEDGMENTS

I would like to thank Professor David Staelin for his support, advice, teachings, and encouragement over the years. Professor Staelin cares deeply about his students and their development. Working in his group has been extremely rewarding and enjoyable.

I also thank Dr. Ronald Cheung, Professor Jae Lim, Dr. Shahid Qureshi, and Professor William Schreiber for their support, advice, and teachings.

Conversations with Professor Robert Gallager, Professor Bernd Girod, Dr. Glen Langdon, and Dr. Henrique Malvar were extremely helpful and furthered my understanding of quantization and coding. Many other people have helped me in this work directly and indirectly; it is impossible to list all of them. A partial list is: Ashraf Alkhairy, Matt Bace, Alain Briancon, Sudhir Dixit, Dan Friedman, Owen Fordham, Cheryl Johnson, Rita Kishinchandani, Dave Kuo, Charlie Miller, Peter Monta, John Payton, Marilyn Pierce, Mike Reese, Jerry Shapiro, Andy Singer, Lon Sunshine, John Tucker, John Wang, Linna Wu, and Ken Zeger.

TABLE OF CONTENTS

ABSTRACT 2

ACKNOWLEDGMENTS 3

CHAPTER 1 Introduction 7
1.1 Encoding of a Continuous-Amplitude Source 7
1.2 Gaussian, Laplacian, and Gamma Memoryless Sources 10
1.3 Scalar Quantization 10
1.4 Arithmetic Coding 13
1.5 Prior Work 14
1.6 Overview of Present Study 16

CHAPTER 2 Optimal and Uniform Entropy-Constrained Quantization . . . 18
2.1 Necessary Condition for Optimal Scalar Quantization 19
2.2 Uniform Entropy-Constrained Quantization 21
2.3 MSE-Suboptimality of Uniform Quantization for Laplacian Sources . 22
2.4 Design of Locally-Optimal Alphabet-Constrained Quantizers 25
2.5 Design of Locally-Optimal Entropy-Constrained Quantizers 28
2.6 Random Search for Optimal Entropy-Constrained Quantizers . . . 35
2.7 Calculation of Distortion-Rate Functions 41
2.8 Performance of Locally-Optimal and Uniform Quantization 43
2.9 Summary and Conclusions 47

CHAPTER 3 Arithmetic Coding Applied to Quantization 49
3.1 Notation and Basic Entropy-Coding Concepts 50
3.2 The Elias-Gallager-Shannon Sequential Arithmetic Code 53
3.3 A Fixed-Precision K-ary Arithmetic Code 59
3.4 Optimal Rounding-Off of the Source Letter Probabilities 67
3.5 A Simple Implementation 71
3.6 Performance of Arithmetic Coding Applied to Quantization 75
3.7 Alternative Coding Procedures 78
3.8 Summary and Conclusions 81

CHAPTER 4 Conclusions 83
4.1 Summary of Work 83
4.2 Suggestions for Further Research 84

APPENDIX A A Simple Optimization Procedure 87
APPENDIX B Software to calculate moments of densities 89
APPENDIX C Software to perform K-ary arithmetic coding 93

REFERENCES 97

LIST OF FIGURES

1.1 Source Coding with a Fidelity Criterion 8
1.2 Source Probability Density Functions 11
1.3 Quantizer Characteristic Staircase Function 12
2.4 Uniform Quantizer Decision Levels 22
2.5 Decision Levels of Locally-Optimal Alphabet-Constrained Quantizers . . 26
2.6 Decision Levels of Locally Optimal Quantizers 33
2.7 Random-Search Design of MMSE Quantizers for a Gaussian Source . . . 35
2.8 Random-Search Design of MMSE Quantizers for a Laplacian Source . . 36
2.9 Random-Search Design of MMSE Quantizers for a Gamma Source . . . 37
2.10 Evidence of Occasional Asymmetry in Optimal Quantizer 38
2.11 Performance of Uniform and Locally Optimal Quantization 44
2.12 Quantizer Performance Curves for a Gaussian Source 45
2.13 Quantizer Performance Curves for a Laplacian Source 46
2.14 Quantizer Performance Curves for a Gamma Source 47
3.15 Algorithms for Rounding-Off Probabilities 70
3.16 Comparison of Optimal and Naive Rounding-Off Algorithm 71
3.17 Effect on Performance of Rounding-Off Probabilities 72
3.18 Register Arrangement for an Arithmetic Coder 73
3.19 State-Transition Diagram for Carry-Over Control 75
3.20 Performance of Scalar Quantization with Arithmetic Coding 76
3.21 Inefficiency of Arithmetic Coding versus Quantizer Step Size 77

CHAPTER 1:

Introduction

Although a relatively old subject in communications engineering, quantization remains an active area of research. The reason is that quantization is a very basic operation: it must be performed in one form or another whenever a real-world analog signal is encoded, transmitted, recorded, analyzed, or enhanced, by digital means. Often, quantization plays a critical role in determining how well these processes are carried out.

In common usage, the term *quantization* is often used to refer collectively to two distinct processes — the mapping of something continuous to something discrete, and the subsequent encoding of the discrete result (usually into binary digits or *bits*). For present purposes, it is convenient to use *quantization* in the slightly more restrictive sense to refer exclusively to the former process, and to use the term *entropy-coding* to refer to the latter. In the present study the two processes are treated somewhat independently, with Chapter 2 devoted to quantization and Chapter 3 devoted to entropy-coding.

The present chapter provides background material and introduces the basic concepts of quantization and encoding.

1.1 Encoding of a Continuous-Amplitude Source

When viewed within the framework of information theory, the quantization/entropy-coding problem is a special case of the lossy source-coding problem depicted in Figure 1.1. Here, a continuous-amplitude, discrete-time source emits a sequence of real numbers which is supplied to an encoder. The encoder transforms this source sequence into a sequence of code bits. A decoder processes the code bit sequence to reconstruct, as best it can, an accurate replica of original continuous-amplitude sequence.

Invariably, some sort of restriction is placed on the number of bits that may

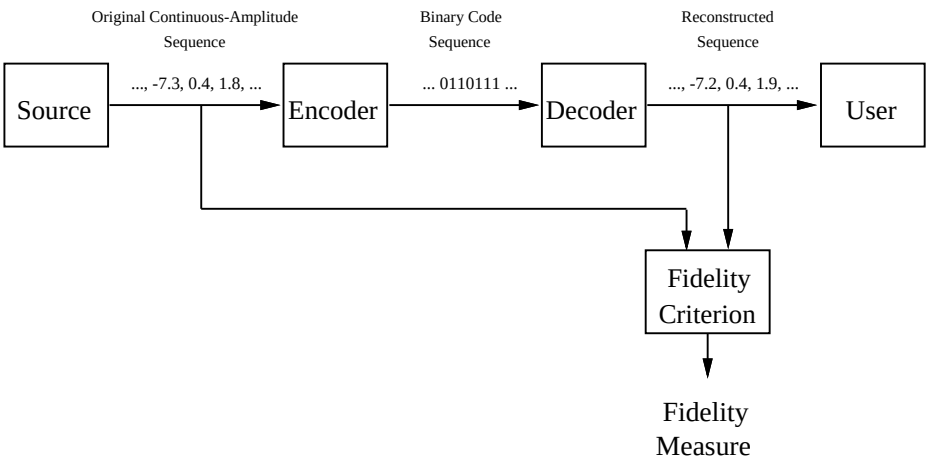

Figure 1.1: Encoding and decoding of a continuous-amplitude, discrete-time source, with measurement of the fidelity of reproduction.

be used by the encoder to represent the source sequence. The number of bits used, measured on a per-sample basis, is called the *rate* of encoding.

If the restriction limits the number of bits used per source letter to be some (perhaps fractional) constant for every letter encoded, then the encoding is said to be *fixed-rate.* Alternatively, if only the *average* number of bits used per source letter is restricted, then the encoding is termed *variable-rate.* In either case, the rate must be finite, and not every possible source sequence will be assigned a unique code sequence. This precludes exact replication of an arbitrary source sequence by the decoder, and introduces the notion of the *fidelity* of reproduction of the source sequence.

Consider the problem of characterizing the performance of such a system. For a given source, two quantities are of principal interest — rate and fidelity.[1] Rate is naturally measured numerically (in number of bits per source letter). It is desirable, from the point of view of characterizing system performance, to have a numerical measure of fidelity as well. In attempting to define such a measure, it becomes apparent that for any such measure to be meaningful, it must describe fidelity for a *class* of se-

[1] It is natural to expect the two to be related — in a given system, one expects that it would always be possible to improve fidelity by increasing the amount of information transmitted.

quences having partially-known characteristics, rather than for a particular sequence. In other words, a probabilistic model of the source must be specified, and the fidelity measure must be defined in terms of that probabilistic model. The most obvious such definition is some sort of average *distortion* incurred in the encoding/decoding process.

Shannon [1][2] and Kolmogorov [3] have shown that if a distortion measure is suitably defined, then there exists a bounding function, known as the *rate-distortion function*, which specifies the minimum rate $R(D)$ required by any encoding/decoding system to represent a given source with distortion not exceeding D. Equivalently, there exists a *distortion-rate function*, which specifies the minimum possible distortion $D(R)$ attainable when a source is encoded at a rate not exceeding R bits per sample. An introduction to rate-distortion theory can be found in [4] or [5, Appendix D], while a more in-depth treatments are given in [6] and [7].

The most commonly used distortion measures are *single-letter* distortion measures. A single-letter distortion measure is an average of some nonnegative function of each source letter and the corresponding reproducing letter (rather than of the two *sequences* of letters). In the present study, consideration is restricted to a particular type of single-letter distortion measure called *difference distortion*, where the function is a power of the absolute value of the difference between the source and reproducing letter.

In general, the performance of an encoding/decoding system can approach the rate-distortion bound only if many source letters are encoded at once. *Vector quantization* is one such form of encoding [8], and *trellis coding* is another [9]. Alternatively, if successive source letters are encoded independently of one another, then the encoding process is called *scalar quantization with entropy-coding,* or often simply *quantization* (particularly in the special case of fixed-rate encoding). If the source happens to be memoryless — that is, if the letters in the source sequence are independent, identically distributed random numbers — then scalar quantization with entropy-coding is

often found to perform well, although it does not, except in trivial cases, approach the rate-distortion bound. It is assumed in this thesis that the source is in fact memoryless, so that scalar quantization with entropy-coding can be expected to perform reasonably well.

1.2 Gaussian, Laplacian, and Gamma Memoryless Sources

A discrete-time memoryless source x is characterized by a probability density function (pdf) $f_x(X)$. Three particular sources are considered here; each has unit-variance, zero-mean, and symmetric pdf. The sources are: *Gaussian* or normal, having pdf

$$f_x(X) = \frac{1}{\sqrt{2\pi}} e^{-X^2/2};$$

Laplacian or two-sided exponential, having pdf

$$f_x(X) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|X|};$$

and *gamma* or broad-tailed, having pdf

$$f_x(X) = \sqrt{\frac{\alpha}{4\pi|X|}} e^{-\alpha|X|},$$

where $\alpha = \sqrt{3}/2$. These sources, often used in modeling signals encountered in speech- and image-coding practice [5][10], can be regarded as instances of the "generalized Gaussian" density as defined in [11]. The pdf's of the three sources are shown in Figure 1.2. Note that the gamma pdf grows arbitrarily large near the origin.

1.3 Scalar Quantization

A *K-level scalar quantizer* Q is a nonlinear, noninvertible mapping of the real numbers onto a set of K numbers $\{r_1, \ldots, r_K\}$ according to the relationship

$$Q(x) = r_k \quad \text{if} \quad d_{k-1} < x \leq d_k, \qquad k = 1, \ldots, K, \tag{1.1}$$

where $d_K = -d_0 = \infty$ and $d_0 < r_1 < d_1 < r_2 < \cdots < r_k < d_K$. The numbers $d_0, \ldots, d_K$ are called *decision levels,* and the numbers $r_1, \ldots, r_K$ are called *reconstruction levels* or *representation levels.* Each partition cell $(d_{k-1}, d_k]$ [2] is called a

[2] Of course, the final partition cell (d_{K-1}, d_K) must be open on the right.

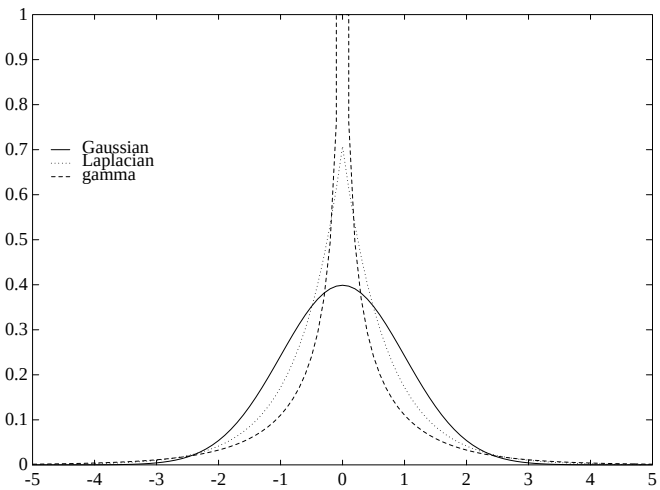

Figure 1.2: Graphs of the probability density function for a Gaussian, Laplacian, and gamma source having zero mean and unit variance.

decision region or *quantization region.* The graph of a typical quantizer function Q resembles a staircase, as shown in Figure 1.3.

The ν-power difference distortion $D^{(\nu)}$ incurred by scalar quantization can be expressed as the sum of the distortions for each of the decision regions, i.e.,

$$\begin{aligned} D^{(\nu)} &= \int_{-\infty}^{\infty} |Q(X) - X|^{\nu} f_x(X) dX \\ &= \sum_k \int_{d_{k-1}}^{d_k} |r_k - X|^{\nu} f_x(X) dX. \end{aligned} \tag{1.2}$$

Let F_k denote the probability that $d_{k-1} < x \leq d_k$. That is, define

$$F_k = \int_{d_{k-1}}^{d_k} f_x(X) dX. \tag{1.3}$$

The *entropy* H of $Q(x)$, which measures the minimum number of bits that must be used, on the average, to represent $Q(x)$ without error [1][7], is given by

$$H = -\sum_k F_k \log_2 F_k. \tag{1.4}$$

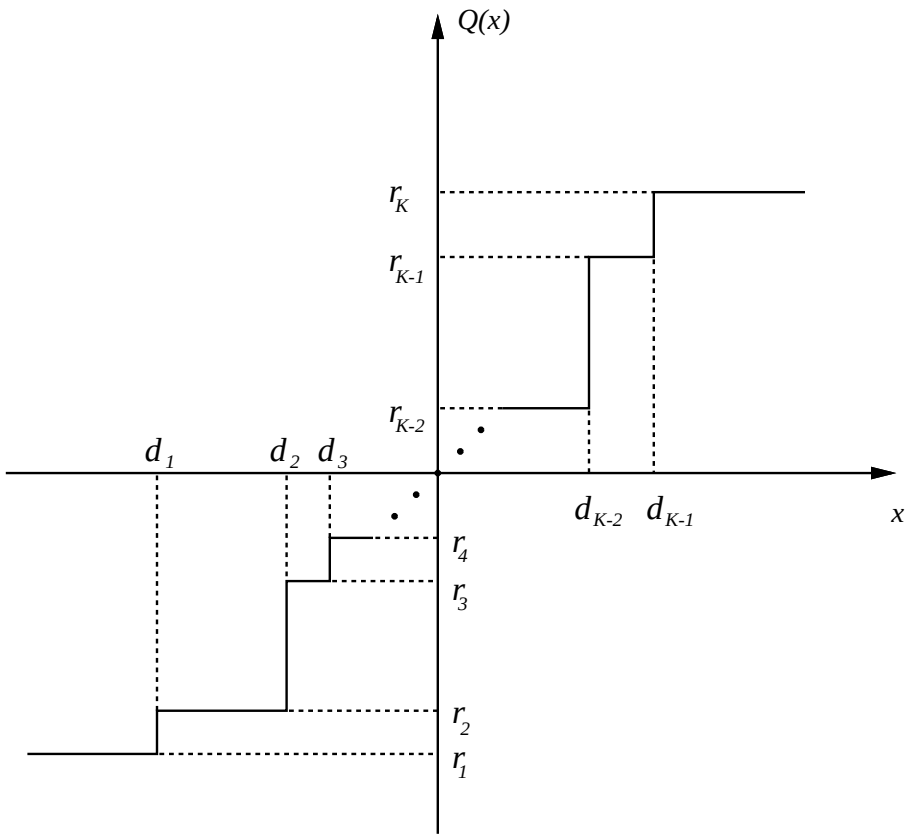

Figure 1.3: Characteristic staircase function of a K-level scalar quantizer.

It is important to note that because the set of quantization regions is discrete, the entropy can remain bounded even when the number of quantization regions K is allowed to grow arbitrarily large. To represent $Q(x)$ using a number of bits that approaches the entropy, variable-rate encoding — or entropy-coding — must be employed. Alternatively, if fixed-rate encoding is used, then the number of bits required to represent $Q(x)$ depends in a simple way on only the number of quantization regions K.

The problem of quantizer design is to select decision levels and reconstruction levels that minimize $D^{(\nu)}$ subject to a constraint on either the number of levels K or on the entropy H. A quantizer which has the least value of $D^{(\nu)}$ subject to a constraint on the number of levels is called an *optimal alphabet-constrained quantizer* or a *Lloyd-Max quantizer* (after the researchers who first described methods for designing

such quantizers [12][13]). A quantizer which has the least value of $D^{(\nu)}$ subject to a constraint on the entropy is said to be an *optimal entropy-constrained quantizer.* This is the type of quantizer of principal concern in this thesis.

In the special case of $\nu = 2$, the difference distortion measure becomes the widely-used *mean-square error* (MSE) criterion, and a quantizer which minimizes $D^{(2)}$ subject to the appropriate constraint is termed *MSE-optimal* or *minimum mean-square error* (MMSE).

A quantizer is termed *locally-optimal* if no reduction in $D^{(\nu)}$ can be obtained by making small changes in the decision and reconstruction levels while satisfying the appropriate constraint. Clearly, an optimal quantizer is also locally-optimal, but a locally-optimal quantizer is not necessarily optimal.

In practice, a limit must be set on the number of quantization regions K, even if variable-rate encoding is to be employed. In light of this, the term *entropy-constrained* is not strictly appropriate, and might be replaced by the more accurate term *alphabet-and-entropy-constrained.* However, if a suitable means of entropy-coding can be devised, then the alphabet constraint can be rendered inactive by allowing K to become very large. In a related vein, if the entropy-coder is significantly inefficient, then the number of bits used will be greater than the entropy. In such a situation it might make sense to optimize the quantizer and entropy-coder jointly; this problem is considered by Farvardin and Modestino [11]. However, one of the main results of this thesis is that it is in fact possible to encode $Q(x)$ in such a way that the alphabet constraint is inactive, while at the same time, the coding efficiency is sufficiently high to warrant the lone assumption of constrained entropy. A method of entropy-coding that allows this is *arithmetic coding.*

1.4 Arithmetic Coding

The usual approaches to entropy-coding involve mapping a fixed-length string of source letters to a variable-length codeword (by the method of Huffman [14]), or

mapping a variable-length string of source letters to a fixed-length codeword (by the method of Tunstall [15]), or a combination of the two (e.g., run-length coding). In the first approach, compression is achieved by minimizing the average length of the codeword used to represent each source string; in the second, it is achieved by maximizing the number of source letters represented by each codeword. Arithmetic coding uses an alternative strategy, where a long sequence of source letters is mapped to a number between zero and one, and that number is represented as a binary fraction to obtain the code bits. It is difficult to analyze arithmetic coding within the conventional framework of entropy-coding, since there is no one-to-one correspondence between subsequences of the source- and code-sequences. The technique will be explained fully in Chapter 3; here, only notable relevant characteristics of arithmetic coding are mentioned.

First, as its name implies, arithmetic coding involves performing arithmetic operations on the probabilities of sources letters as they are encoded. The probabilities are permitted to change from letter to letter, so long as the same probabilities are used in decoding a letter as were used in encoding it. This allows the code to be adaptive, which in the present application allows the rate to be adjusted "on-the-fly." Second, arithmetic coding is efficient even in the case of large-alphabet, low-entropy sources, where conventional techniques become either inefficient or prohibitively complex. Finally, arithmetic coding lends itself to easy implementation using a general-purpose microprocessor, digital-signal-processing chip, or custom VLSI hardware.

1.5 Prior Work

A vast amount of research has been done in the area of quantization, dating back to the early part of the 20^{th} century. The roots of arithmetic coding can be traced to the original work of Shannon [1], although the technique was made practical only about a decade ago [16].

For an introduction to quantization and its history, the reader is referred to Ger-

sho [17]. An extremely simple treatment of the basic concepts of quantization appears in [18, Chapter 2]. In-depth summaries of the important results of quantization research can be found in [5, Chapter 4] and [19]. In the entropy-constrained case, the most relevant prior work is by Goblick and Holsinger [20], Gish and Pierce [21], Berger [22][23], Wood [24], Noll and Zelinski [25], and Farvardin and Modestino [11][26].

Goblick and Holsinger [20] show that for a Gaussian memoryless source and some small given value of MSE, the entropy of a uniform quantizer exceeds the rate-distortion bound by only 0.25 bits/sample.

Gish and Pierce [21] establish the asymptotic optimality, in the case of low distortion, of quantizers that have uniformly spaced decision levels (hereinafter called *uniform quantizers*), under relatively mild assumptions about the pdf of the source and for a broad class of distortion measures. They further show that for a difference distortion measure $D^{(v)}$, the difference between the performance of uniform entropy-constrained quantization and the rate-distortion bound vanishes in the case of large ν, again only in the case of asymptotically low distortion. The treatment in [21] is analytical and involves some elegant approximations.

The near-optimality of the uniform quantizer is confirmed in the case of a Gaussian source and MSE criterion by Wood [24], who, like Gish and Pierce, uses an analytical approach.

A necessary condition for optimality of an entropy-constrained quantizer — which is simply the Lagrange multiplier condition — was first proposed by Berger [23]. In the alphabet-constrained case, Fleischer [27] and Trushkin [28] have obtained sufficient conditions for optimality for certain sources and distortion measures, but corresponding conditions in the entropy-constrained case have yet to be obtained.

In [22], Berger uses numerical techniques to show that in both the Gaussian and Laplacian case, uniform quantizers appear to be nearly optimal. Noll and Zelinski [25] also use a numerical technique to show that uniform quantization appears to be nearly

MSE-optimal for several memoryless sources, but in their study they unnecessarily restrict consideration to only those quantizers that have a decision level at zero, as noted in [26] and elsewhere.

Farvardin and Modestino investigate in detail the problem of designing locally-optimal quantizers for memoryless sources subject to an entropy constraint [26]. In a later paper they investigate an implementation of entropy-constrained quantization [11].

Practical arithmetic coding has a relatively short history. Although the principle of arithmetic coding was described by Shannon in 1948 [1] in proving a source coding theorem, the first practical arithmetic codes did not appear until 1976 with the work of Rissanen [29] and, as reported by Langdon [16], Pasco [30]. Since then, most of the work in arithmetic coding has been in the form of refinements to the basic technique. Most of the published work — including the recent work on the *Q-coder* [31][32] — has focused on the problem of encoding a binary source. Arithmetic coding, despite its clear advantages over better-known techniques in many applications, has been slow in gaining acceptance. Perhaps this is because it is not yet well understood by practicing engineers.

1.6 Overview of Present Study

This thesis investigates the problem of scalar quantization with entropy-coding, applied to three commonly used model memoryless sources, with respect to a ν-power difference distortion measure, where $\nu = 1, \ldots, 6$.

Chapter 2 considers the problem of scalar quantization under the assumption that ideal entropy-coding can be performed. An iterative, numerical procedure is presented for computing a locally-optimum quantizer, and a random-search procedure is presented as a means of gaining insight into the nature of global optima. In all cases considered, it is found that uniform quantization with entropy-coding performs significantly better than alphabet-constrained quantization, with the greatest difference

being in the case of large values of the error exponent ν. It is also shown that the difference between the best locally-optimal quantizers found and uniform quantization is negligible in most cases. Chapter 2 also shows that uniform quantization is not exactly optimal in the case of a Laplacian source and MSE distortion measure, contrary to the widely held belief.

Chapter 3 develops the idea of arithmetic coding from basic principles. An arithmetic code designed for a multi-letter alphabet source is described, and simulation results are presented that show the code is almost perfectly efficient when applied to uniform quantization of Gaussian, Laplacian, and gamma sources. A procedure is also described to convert the source letter probabilities (i.e., reconstruction level probabilities) to finite-precision values in the a way that leads to the greatest possible coding efficiency.

Not addressed in the present treatment are two important implementation-related topics. One is the *buffering* which must be performed if the variable-rate code-bit stream is to be transmitted over a fixed-rate channel. The other is the suitability of the technique in the presence of transmission errors. The former topic is omitted because it is adequately treated elsewhere [11][33]. In particular, Farvardin and Modestino [11] show that little need be lost in quantizer performance by buffering, even if the size of the buffer is kept rather small. The latter topic is thought to be sufficiently deep to sustain a thesis by itself, and is left as an item for further study.

CHAPTER 2:

Optimal and Uniform Entropy-Constrained Quantization

This chapter considers the design and rate-distortion performance of scalar quantizers for Gaussian, Laplacian, and gamma sources, under the assumption that ideal entropy-coding can be performed on the quantized samples.

Discussion begins by stating necessary conditions for optimality of a quantizer subject to either an alphabet-constraint or an entropy-constraint. Uniform quantization is then considered; it is shown that such quantizers are suboptimal in the case of a Laplacian source, contrary to a widely-held belief. Next, the problem of designing optimal alphabet-constrained quantizers for ν-power difference distortion measures is discussed. In contrast to the MSE case, it is found that designing such quantizers requires overcoming a few interesting computational difficulties.

Section 2.5 presents a very simple but effective numerical procedure for obtaining a locally-optimal entropy-constrained quantizer.

The problem of demonstrating the global optimality of a particular entropy-constrained quantizer remains open. In an effort to gain insight into the general character of the best entropy-constrained quantizers, a random search is undertaken. The technique and its results are described in Section 2.6. Evidence is provided to support the hypothesis that in most cases, the locally-optimal quantizers obtained by the method of Section 2.5 are in fact globally optimum. A more surprising result suggests that in some cases, the optimal entropy-constrained quantizer can be asymmetric even when the source pdf is symmetric and the alphabet-constraint is inactive.

For the sake of completeness, a method of computing rate-distortion functions is described in Section 2.7. In Section 2.8, results are presented, and observations are made about the relative performance of optimal alphabet-constrained, locally-optimal entropy-constrained, and uniform entropy-constrained quantization.

2.1 Necessary Condition for Optimal Scalar Quantization

For a given memoryless source and difference distortion measure $D^{(\nu)}$, the problem of scalar quantizer design is to choose the decision levels and reconstruction levels in a way that minimizes $D^{(\nu)}$, subject to the appropriate constraint. In both the alphabet-constrained and entropy-constrained cases, a necessary condition for true optimality is local optimality. That is, if a quantizer is optimal, then it must not be possible to improve its performance by making small adjustments within the feasible set. In the alphabet-constrained case, this amounts to the simple requirement that the partial derivatives of $D^{(\nu)}$ with respect to the decision levels and reconstruction levels be zero [12][13]. That is, it is required that

$$\frac{\partial D^{(\nu)}}{\partial d_k} = 0; \qquad k = 1, \ldots, K-1; \tag{2.1 a}$$

$$\frac{\partial D^{(\nu)}}{\partial r_k} = 0; \qquad k = 1, \ldots, K. \tag{2.1 b}$$

In the entropy-constrained case, $D^{(\nu)}$ is to be minimized subject to the equality constraint $H = H_0$. Here, the method of Lagrange multipliers applies[1] [34][23], so that the necessary condition is

$$\frac{\partial}{\partial d_k}\left[D^{(\nu)} - \lambda(H - H_0)\right] = 0; \qquad k = 1, \ldots, K-1; \tag{2.2 a}$$

$$\frac{\partial}{\partial r_k}\left[D^{(\nu)} - \lambda(H - H_0)\right] = 0; \qquad k = 1, \ldots, K; \tag{2.2 b}$$

$$H = H_0, \tag{2.2 c}$$

for some constant λ, where K has been chosen to be sufficiently large to make the alphabet-constraint inactive.

[1] The Lagrange multiplier condition will be explained in a intuitively satisfying way in Section 2.5, where it appears as the natural termination condition of an optimization procedure.

It is clear from Equations (1.3) and (1.4) that H is independent of the r_k's, so that (2.2 b) reduces immediately to (2.1 b). It is instructive to rewrite (1.2) as

$$D^{(\nu)} = \sum_k D_k^{(\nu)} \tag{2.3}$$

where

$$D_k^{(\nu)} \stackrel{\text{def}}{=} \int_{d_{k-1}}^{d_k} |r_k - X|^\nu f_x(X) dX. \tag{2.4}$$

In terms of $D_k^{(\nu)}$, (2.1 b) and (2.2 b) become

$$\frac{\partial}{\partial r_k} \int_{d_{k-1}}^{d_k} |r_k - X|^\nu f_x(X) dX = 0, \qquad k = 1, \ldots, K, \tag{2.5}$$

so that in both the alphabet-constrained and entropy-constrained cases, the optimal value of r_k is completely determined by d_{k-1} and d_k. It is assumed henceforth that the r_k's are always chosen to satisfy (2.5), so that the design problem reduces to finding the decision levels $d_1, \ldots, d_{K-1}$ that minimize $D^{(\nu)}$ subject to the appropriate constraint.

In terms of $D_k^{(\nu)}$, $\partial D^{(\nu)}/\partial d_k$ can be written

$$\frac{\partial D^{(\nu)}}{\partial d_k} = \frac{\partial D_k^{(\nu)}}{\partial d_k} + \frac{\partial D_{k+1}^{(\nu)}}{\partial d_k}.$$

Thus, in the alphabet-constrained case, a simpler statement of the necessary condition for optimality is given by rewriting (2.1 a) as

$$\frac{\partial D_k^{(\nu)}}{\partial d_k} + \frac{\partial D_{k+1}^{(\nu)}}{\partial d_k} = 0, \qquad k = 1, \ldots, K-1. \tag{2.6}$$

In the entropy-constrained case, the corresponding necessary condition is

$$\lambda = \frac{\partial D_k^{(\nu)}/\partial d_k + \partial D_{k+1}^{(\nu)}/\partial d_k}{\partial H/\partial d_k}$$

$$= \frac{\partial D_k^{(\nu)}/\partial d_k + \partial D_{k+1}^{(\nu)}/\partial d_k}{f_x(d_k)\log_2(F_{k+1}/F_k)}, \qquad k = 1, \ldots, K-1, \tag{2.7}$$

$$H = H_0,$$

for some constant λ.

In considering the entropy-constrained problem, it is occasionally instructive to regard the decision levels $d_1, \ldots, d_{K-1}$ as the components of a $(K-1)$-dimensional vector $\vec{d}$, and to regard the constraint $H = H_0$ as describing a hypersurface (in $(K-1)$-dimensional vector space) to which $\vec{d}$ is bound. The device of regarding the decision levels as a vector will be used occasionally in subsequent sections of this chapter.

It should be noted that finding the r_k's that satisfy (2.5) is generally not a simple matter, particularly when ν is odd. A related problem is the difficulty of evaluating the partial derivatives $\{\partial D_k^{(\nu)}/\partial d_k\}$ via Leibniz' formula [35] — such evaluation requires, ultimately, differentiation of the r_k's, which are not generally available in analytic form. In the present study, simple numerical techniques were used to find the optimal r_k's and to compute the partial derivatives $\{\partial D_k^{(\nu)}/\partial d_k\}$, in all cases except $\nu = 2$ (where analytical solutions are available).

2.2 Uniform Entropy-Constrained Quantization

With a constraint on entropy, uniform quantization is known to be asymptotically optimal in the case of low distortion under relatively general conditions [21], and nearly MSE-optimal at all rates for some sources [24][22][25]. A significant part of the present study is investigation of the performance of uniform entropy-constrained quantization with respect to ν-power distortion, relative to locally-optimal quantizers and the rate-distortion bound. This section introduces terminology and notation that will simplify discussion of uniform quantization.

A *uniform K-level scalar quantizer* is a scalar quantizer having decision levels $\{d_k\}$ given by

$$d_k = (k - \lfloor K/2 \rfloor)\Delta + \psi, \qquad k = 1, \ldots, K-1, \tag{2.8}$$

where Δ is called the *step size*, and ψ, which is the value of the least nonnegative decision level, is called the *offset*. The reconstruction levels of the uniform quantizer are always selected to minimize distortion, and in general are not uniformly spaced.

Here it is assumed that K is very large, so that the alphabet-constraint is inactive. The K-level uniform quantizer is thus an approximation to the uniform quantizer having an unlimited number of levels depicted in Figure 2.4. Note that Δ and ψ completely determine the characteristics of a uniform quantizer.

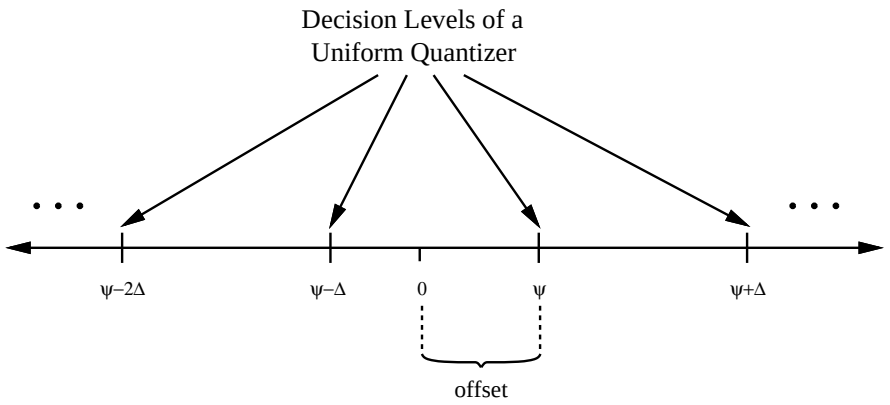

Figure 2.4: A uniform quantizer having an unlimited number of uniformly spaced decision levels can be described by two parameters: the spacing Δ between decision levels, and the value ψ of the smallest nonnegative decision level. As always, it is assumed that the reconstruction levels (not shown) are chosen to minimize quantization error and are not, in general, uniformly spaced.

In early investigations into uniform quantization [23][25], consideration was restricted to the case $\psi = 0$. Uniform quantizers of this type are often termed *midriser,* as their characteristic staircase functions rise at the origin. At the other extreme are uniform quantizers with $\psi = \Delta/2$, which are often termed *midtread*[5]. Clearly, for the symmetric-pdf sources of interest, a midriser uniform quantizer can never have an entropy of less than one bit per sample, since the maximum possible probability of a quantization region in that case is $1/2$.

2.3 MSE-Suboptimality of Uniform Quantization for Laplacian Sources

It has been asserted by several authors [19][25][36][37][38], all of whom cite Berger[23],

that when a Laplacian source is to be quantized subject to an entropy constraint, the uniform quantizer is exactly optimal in the MSE sense for all values of entropy. This assertion, which appears to be a misinterpretation of Berger's findings, can be seen to be erroneous by examining the data presented later in this chapter. Alternatively, it can be shown by exhibiting a single example of a quantizer that has less mean-square error than the best uniform quantizer for some particular value of H. One such example, at $H = 1$ bit/sample, is the symmetric but nonuniform quantizer having decision levels at $-\infty$, -15, -10.19, -8.113, -6.358, -4.629, -2.902, -1.1514, 1.1514, 2.902, 4.629, 6.358, 8.113, 10.19, 15, and ∞. The mean-square error for this quantizer when applied to a unit-variance Laplacian source is 0.2618 (5.82 dB signal-to-error ratio), which is a slight improvement over the best uniform quantizer at the same rate, which has decision levels at $2.2235i + 1.11175, i = 0, \pm 1, \pm 2, \ldots$, and which has mean-square error of 0.2650 (5.76 dB). It is not difficult to verify these figures with the aid of a programmable calculator. The difference in error is extremely small;[2] nevertheless, this example discredits the assertion of exact optimality of uniform quantization for Laplacian sources.

To see how the mistaken idea of exact optimality of the uniform quantizer in the case of a Laplacian source might have come about, it is instructive to examine the right-hand side of (2.7) for $k = 1, \ldots, K$. Recall from the previous section that the decision levels of a uniform quantizer can be described by two parameters, the spacing Δ between levels, and the value ψ of the smallest nonnegative decision level. Here, as in the literature cited, consideration is restricted to the case of infinitely many quantization regions. Let λ_k be defined as

$$\begin{aligned} \lambda_k &= \frac{\partial D_k^{(2)}/\partial d_k + \partial D_{k+1}^{(2)}/\partial d_k}{\partial H/\partial d_k} \\ &= \frac{(r_k - d_k)^2 - (r_{k+1} - d_k)^2}{\log_2(F_{k+1}/F_k)} \end{aligned} \tag{2.9}$$

[2] Had the difference been larger, perhaps the optimality assertion would have been discovered to be erroneous much earlier.

It can be shown after some algebraic manipulation that for those decision levels d_k that do not border the region containing zero (in the case where zero falls between decision levels),

$$\lambda_k = -\ln 2 \frac{(\Delta - \sqrt{2})e^{\sqrt{2}\Delta} + \sqrt{2} + \Delta}{\sqrt{2}(e^{\sqrt{2}\Delta} - 1)}; \tag{2.10}$$

with the important result being that λ_k turns out to be independent of k. For the d_k's that *do* border the region containing zero, λ_k exhibits a strong dependence on ψ and assumes the common value of λ_k for the other d_k's only as ψ approaches zero.[3] In this case, which Berger considers exclusively in [23], the condition (2.10) holds for all k; i.e., the Lagrange multiplier condition is satisfied. This does not imply that these quantizers are optimal — the Lagrange multiplier condition is necessary, but not sufficient, for optimality. In fact, it can be shown[4] in the Laplacian MSE case that the uniform midriser quantizers happen to be, ironically, precisely the *worst* uniform quantizers to use at any given entropy with respect to mean-square error, while the *best* uniform quantizers are those which have zero as the midpoint of a decision region (midtread). It can further be shown, by evaluating λ_k in (2.9) either analytically or numerically for all k, that these best uniform quantizers do not satisfy the necessary condition for optimality (i.e., the Lagrange multiplier condition). Thus, uniform quantizers are in fact suboptimal in the case of entropy-constrained quantization of a Laplacian source with respect to the MSE criterion.

It is important to note that Berger himself never claims in [23] that the uniform midriser quantizers he considers are optimal, and in a subsequent paper [22], he concludes that indeed they must not be optimal. It is remarkable then, that the erroneous belief of exact optimality persists within the research community.

[3] The dependence of λ_k on ψ for these innermost levels is most conveniently demonstrated numerically. Although a correct closed-form expression for λ_k for these decision levels was obtained via the *MACSYMA* symbolic algebra program, it did not appear to be very useful, aside from showing that the expression does in fact depend on ψ. The expression is not included here because it is extremely cumbersome and does not particularly lead to insight, requiring perhaps one-third of a page to display after full simplification.

[4] This is most easily shown numerically by evaluating $D^{(2)}$ as a function of ψ (adjusting Δ as needed to keep H constant).