Functional Quantization

by

Vinith Misra

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

Bachelor of Science

and

Master of Engineering in Electrical Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2008

Author .....
Department of Electrical Engineering and Computer Science
May 9, 2008

Certified by .....
Vivek K Goyal
Associate Professor
Thesis Supervisor

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

Functional Quantization

by

Vinith Misra

Submitted to the Department of Electrical Engineering and Computer Science
on May 9, 2008, in partial fulfillment of the
requirements for the degrees of
Bachelor of Science
and
Master of Engineering in Electrical Science and Engineering

Abstract

Data is rarely obtained for its own sake; oftentimes, it is a function of the data that we care about. Traditional data compression and quantization techniques, designed to recreate or approximate the data itself, gloss over this point. Are performance gains possible if source coding accounts for the user's function? How about when the encoders cannot themselves compute the function? We introduce the notion of functional quantization and use the tools of high-resolution analysis to get to the bottom of this question.

Specifically, we consider real-valued raw data X_1^N and scalar quantization of each component X_i of this data. First, under the constraints of fixed-rate quantization and variable-rate quantization, we obtain asymptotically optimal quantizer point densities and bit allocations. Introducing the notions of *functional typicality* and *functional entropy*, we then obtain asymptotically optimal block quantization schemes for each component. Next, we address the issue of non-monotonic functions by developing a model for high-resolution non-regular quantization. When these results are applied to several examples we observe striking improvements in performance.

Finally, we answer three questions by means of the functional quantization framework: (1) Is there any benefit to allowing encoders to communicate with one another? (2) If transform coding is to be performed, how does a functional distortion measure influence the optimal transform? (3) What is the rate loss associated with a suboptimal quantizer design? In the process, we demonstrate how functional quantization can be a useful and intuitive alternative to more general information-theoretic techniques.

Thesis Supervisor: Vivek K Goyal
Title: Associate Professor

Acknowledgments

- Mom/Dad, for (mostly) tolerating the sparsity of my home-calling patterns, and for shielding me from the fallout. Despite my many statements to the contrary, you are both very cool.
- My crazy delicious advisor, Vivek Goyal, for humoring the ideas and thoughts of a blundering undergraduate, for demonstrating that academia doesn't need to be crusty, and for not killing me about the missed deadlines (my bad!). Most of this work would have died in its sleep were it not for his suggestions and encouragement. With all due respect to the wonderful advisers of the world, I couldn't have picked one better.
- Lav Varshney for being a (good?) role model for an even younger researcher. I've always been troubled by the conflict between knowledge and creativity; he's shown me that one can excel at both. On a more tangible level, his contributions are peppered throughout this thesis, and several of the results would have been impossible without his help. I'm going to miss working with him.
- The members of the STIR group, for a research environment where questions never felt stupid, and where distractions only occasionally felt like distractions.
- Mike Maddox for moral support. The light at the end is a lot easier to see with two sets of eyes.
- Patrick Mangat for his never-ending supply of humor. I cry myself to sleep sometimes when I think about the time we've wasted.
- Michael Ouellette for not only stomaching my presence on his stage, but encouraging it. He seriously deserves a medal.
- Spritely Roche and Dave Hagymas for giving me a chance to play. It's meant a lot.
- And to the muse of self-improvement: I've a ways to go pal, but I'm getting there.

If a man comes to the door of poetry untouched by the madness of the Muses, believing that technique alone will make him a good poet, he and his sane compositions never reach perfection, but are utterly eclipsed by the performances of the inspired madman. - PLATO

Contents

List of Figures

Campus Preview Weekend

Modularity and abstraction are amongst the most fundamental principles of electrical engineering. Without them, complex systems would be both unimaginable and unrealizable; for instance, it is difficult to understand the workings of a computer purely from device physics. Nonetheless, it is frequently profitable to break the boundaries of abstraction: an engineer might improve performance by considering the inner workings of system A and system B together. We focus on a particular example of this — data compression followed by computation.

Consider a system that digitizes an analog voltage waveform. Somewhere towards the front end of this system will most likely be an analog-to-digital converter (ADC) sampling and quantizing the input data. Rarely are end users interested in seeing this voltage data itself; oftentimes they won't even know what a "volt" is. Instead, some computation will be performed on this data, with the results of this computation going to the end user. It is worth noting that most neurological signals are extraordinarily low-rate and low-precision; Gabor for instance makes reference to "the 20 bits per second which, the psychologists assure us, the human eye is capable of taking in" [1]. As such, a human is most likely interested in only a small fraction of the information contained within captured data [2].

The principles of abstraction and modularity require that the ADC block be designed to produce a "good" digital approximation to a continuous voltage. The word "good" is taken as shorthand for "as close as possible to the original voltage waveform," typically in terms of a generic signal distortion measure such as the mean squared error (MSE). The computation block takes the output of the ADC and produces the one or two bits of actual interest to the user.

All is not well, however. More optimistically speaking, there is considerable room for improvement in this picture. The digitization process has been designed to minimize distortion to the voltage waveform — but the end user could care less about it! A far better design philosophy would cater

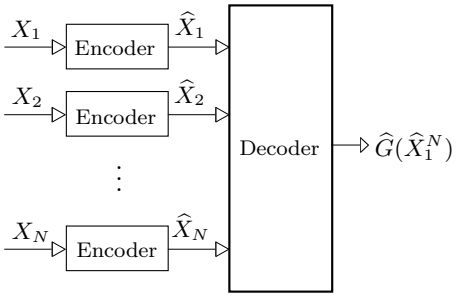

Figure 0-1: Functional source coding. Note the disjoint encoders and the computation of a function at the decoder

towards "digitizing the data to minimize distortion of the function seen by the end user." In other words, one might perform analog to digital conversion *taking into account the computation to be performed*.

As another example, consider a distributed sensor network. Each sensor collects some data and communicates it to a common fusion center after performing some compression on it. Just as with the ADC example, this data is usually not the end product of the system: there will generally be some computation performed on it at the fusion center. So how can the sensor network adjust its source coding to take this computation into account?

These problem statements are typical of a much broader class of source coding questions, represented abstractly in Fig. 0-1. N variables, $X_1^N = \{X_1, X_2, \ldots, X_N\}$, with some joint distribution are separately observed, compressed, and communicated to a common decoder. The decoder then performs some computation on the compressed representations of each source, represented as a function $G(\widehat{X}_1^N)$. In the case of the ADC, each X_i would be a sample of the voltage waveform, and the process of compression would be the act of quantization to a discrete set of possible values. Similarly with the sensor networks, each sensor would separately quantize and compress its observations before sending them to the decoder.

Both of these manifestations of the problem have a distributed flavor to them. The sensors in a sensor network are physically separated, and virtually all ADC architectures require the samples to be quantized separately or — in the case of sigma-delta — with limited collaboration. This has two consequences. First, the problem immediately becomes nontrivial: one cannot merely compute the function of the uncompressed sources, $G(X_1^N)$, and perform compression directly on G. Instead, one must devise some scheme to compress each source variable X_i into a compressed representation $\widehat{X}_i$ so as to minimize the error between the "ideal" value of $G(X_1^N)$ and the approximation $\widehat{G}(\widehat{X}_1^N)$. For

instance, suppose that two automated surveillance video cameras are monitoring a store. A decision on whether or not to alert the police must be made from a third location. How should each camera best compress its video? Neither camera can decide for itself and communicate the decision, so the optimal strategy is far from obvious.

Secondly, we make no assumptions about the independence of the samples X_1^N. This immediately brings up connections to multiterminal information theory: how can one exploit the correlations between the source variables to reduce the redundancy between each of their compressed representations? One must consider this, additionally, within the context of the computation to be performed: does "redundancy" mean the same thing when we are not interested in representing variables for their own sake?

In this thesis, we approach this problem and its related threads from the perspective of quantization theory; the results of this are grouped under the heading of *functional quantization*. There are numerous advantages to this approach. Quantization can be seen as something of a middleman between information theory and the much more physically grounded world of circuit design and implementation. While it can be used to address the physical rounding process performed in most ADCs, it can also venture into issues of fundamental limits and asymptotic behavior. Perhaps most importantly, it gives us access to powerful analytical tools that can yield quantitative results — oftentimes in places that more abstract techniques fall short.

Our goals in exploring functional quantization are the following:

1. Develop a framework for analyzing functional quantization.

2. Identify the fundamental limits on performance.

3. Design optimal or near-optimal quantization techniques that can approach these limits

4. Attack related problems with these same techniques.

In Chapter 1 we discuss some of the theoretical tools used in the thesis. We start with a review of basic source coding concepts, including discrete and differential entropy and Slepian-Wolf coding. We then consider the quantization problem in both its scalar and vector forms. The high-resolution approximation — incredibly important to our approach — is discussed as a way to reduce the complexity of the quantization problem. Finally, transform coding is briefly touched on for its relevance to our development of "functional transform coding" in Chapter 4.

Next, in Chapter 2 we obtain the results at the heart of functional quantization. A brief review of related work places this effort in its many appropriate contexts: as we show, the functional quantization problem has connections to problems as diverse as perceptual coding and multiterminal information theory. Although our eventual interest is in the multidimensional scenario (i.e., a source vector X_1^N for $N > 1$), we start our development by considering the more easily digested single-dimensional scenario. While this situation is relatively straightforward, it provides useful insights for the higher dimensional problems.

Our approach with developing the theory for multidimensional problems is to consider increasingly unconstrained situations, from fixed-rate quantization to Slepian-Wolf variable-rate quantization. Finally, we consider an even broader scenario that cannot be captured by our previous techniques: variable-rate vector quantization at each encoder. To attack this problem, we develop the notion of functional typicality — much in the vain of the well-known asymptotic equipartition theorem. We note how this technique may be used as an alternative route to several of our previous derivations.

In Chapter 3 we apply the results of this theoretical exploration to several functions of statistical interest (for instance, the decoder might be interested in obtaining the midrange of its samples). We observe a striking gap in performance between ordinary techniques and functional techniques; for variable-rate quantization, this gap is found to grow exponentially in the number of source variables. It is found that similar behavior can be observed for an entire class of functions satisfying the properties of *selectivity* and *symmetry*.

Our results up to this point have concerned functions $G(X_1^N)$ that are monotonic in each of their arguments. We find that this restriction is overly strict, and generalize to the set of functions that are smooth, bounded, and not necessarily monotonic. In the process, we consider the problem of high-resolution *non-regular* quantization, and develop a way of describing these quantizers.

In Chapter 4 we use the techniques of the previous chapters to explore new situations. Functional transform coding is first considered as a computationally tractable alternative to unconstrained functional vector quantization. We obtain the optimal transformation under the constraint of uniform quantization and note similarities to the Karhunen-Loeve transform of traditional transform coding.

Next, we explore the possibility of encoders that can communicate with one another. Wildly different behavior is observed for fixed-rate and variable-rate quantization. For the former, any bits going from encoder to encoder are better spent going to the decoder; encoder collaboration is hence

discouraged. For the latter, the potential benefits are unbounded. We give a simple example that demonstrates this dichotomy. In the process of these derivations, we make use of a picture that sees the function's sensitivity profile ($g_i^2(x)$ is introduced in Chapter 2) as a vector in a Hilbert space whose "length" indicates the distortion.

Finally, we consider the situation where the optimal quantizer is not used due to an inaccurate source/function model or inaccuracy on the part of the system designer. In other words: if the compression system thinks the source has a probability distribution that it doesn't, or that the function in question is different from what it is, how sensitive is the system's performance to this mistake? We quantify the impact of such errors in the form of a rate loss.

Chapter 1

Freshman Year: A Review of Source Coding and Quantization

Several techniques from basic information theory and quantization are used prominently in our development of functional quantization. This chapter is meant as a brief review of these topics. For a more detailed take on source coding, refer to [3], or for quantization [4] [5].

We will start be reviewing the fundamental concepts of entropy and its role in lossless compression — wherein source data can be recreated perfectly. From here, we discuss the lossy regime, where the source data can at best be imperfectly approximated from its compressed representation, along with basic rate-distortion theory.

We then change focus to scalar quantization, where a continuous random value is approximated from a finite set of real numbers. Through several examples, we come to realize the complexity of analytically describing the quantization process. The techniques of high-resolution approximation, which convert quantizer design from a discrete to a continuous problem, are described as a means to reduce this complexity.

These problems are even more pronounced for vector quantization, which we go to next. Here, multiple random variables are together approximated from a finite set of real vectors. Transform coding is discussed as a means to improve tractability of analysis and implementation; this will come in particularly handy during the development of functional transform coding in Sec. 4.1.

Type of Coding	$X = 1$	$X = 2$	$X = 3$	Rate
Fixed-Length	00	01	10	2 bits/sample
Variable-Length	1	00	01	1.5 bit/sample

Table 1.1: Illustration of fixed and variable length lossless coding of a three-value source X

1.1 Lossless Compression and Entropy Coding

Most sources of data are repetitive; artwork, for instance, frequently contains areas of relatively similar colors. Central to lossless data compression is this notion of "redundancy," or excess information. By removing this redundant information, one may reduce the amount of information that needs to be stored or communicated. The beauty of information theory is its ability to tie an elegant quantitative framework to this abstract notion.

To illustrate the notion of redundant information more concretely, suppose we have a random variable taking values in some finite alphabet — suppose, for instance, that X is 1 with probability 1/2, 2 with probability 1/4, and 3 with probability 1/4. Now suppose we wished to store the value of this variable. How many bits would it take? The obvious approach is to assign 2 bits to represent the three values in some arbitrary manner; perhaps the binary system itself would suffice (see Table 1.1). This would result in an average of 2 bits per sample of X, but can one do better?

We made two implicit constraints when formulating this generic answer: that each codeword must be the same length, and that only one sample of X may be coded at a time. Relaxing the first of these requirements, we note that a receiver can decode variable-length codewords provided *no codeword is the prefix of another*. In line with this, suppose we use the assignment rule given by the last line of Table 1.1. The average number of bits per sample is then only

$$L = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = 1.5$$

which happens to be the best we can do.

As the alphabet size and the number of samples we code together grow, obtaining the optimal compression scheme is increasingly difficult. Nonetheless, Shannon noted a remarkable fundamental limit on the performance of this optimal compression scheme: the *entropy* of the source variable X,

defined as the "eerily self-referential" [3] expression

$$H(X) = \mathbf{E}\left[-\log_2 p_X(X)\right]$$

The forward part of his theorem states that one may code a random variable with average codeword length — referred to as the *rate* — arbitrarily close to $H(X)$ with arbitrarily small probability of error. The converse states that no coding scheme, no matter how complex its implementation, can achieve a rate-per-sample below the entropy.

It can be seen that for our random variable X with dyadic PMF, $H(X)$ is precisely 1.5 bits. In general, however, one must code many samples together to come within arbitrary precision of the entropy. Constructing the optimal codeword assignments quickly becomes a complicated task: for block coding 10 samples of a binary variable together, the source alphabet is of size 1024. Several techniques have emerged to address the very gritty task of entropy coding; we list a few of the more prominent:

1. Huffman coding [6] is a greedy algorithm that produces the optimal (but non-unique) codeword assignments for a given finite alphabet. It can be difficult to deal with large blocklengths, however.

2. Arithmetic coding [3] does not necessarily produce optiml assignments for any given block length, but through a simpler architecture allows one to work with large blocklengths.

3. Lempel-Ziv-Welch (LZW) [7] is a universal compression algorithm that does not require knowledge of the source distribution.

The literature on lossless source coding is incredibly rich; we recommend interested parties to look there for further information.

1.1.1 Slepian-Wolf Coding

The situation we have just considered can be considered a form of coding for point-to-point communications. That is, user A encodes a source in order to communicate it to user B. In general, however, we can imagine many users on both ends, connected together in some sort of network. A network situation that is of particular interest to us is the distributed source coding scenario, depicted in Fig. 1-1.

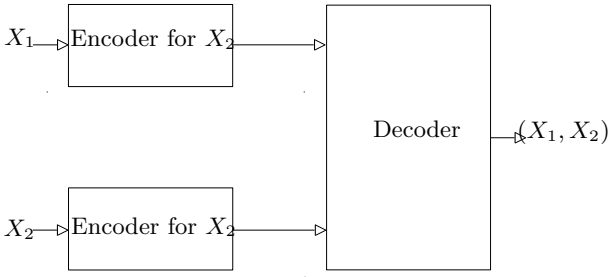

Figure 1-1: The Slepian-Wolf scenario. Two correlated variables, X_1 and X_2, are separately encoded and jointly decoded.

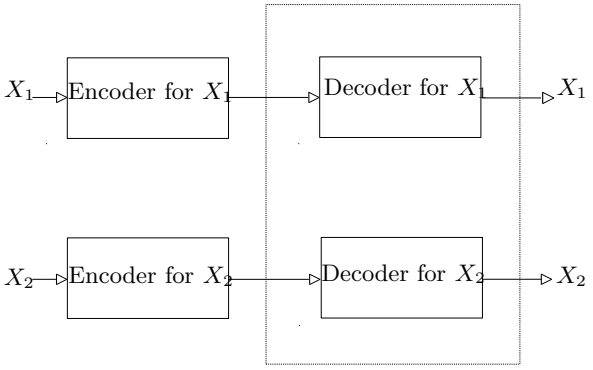

Figure 1-2: An inner bound on the Slepian-Wolf problem; one may always ignore correlations and code/decode X_1 and X_2 separately.

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	1/2	1/4
$X_2 = 1$	0	1/4

Table 1.2: An example of nontrivially correlated sources.

Two potentially correlated sources, X_1 and X_2, are separately encoded into codewords Y_1 and Y_2, at rates R_1 and R_2 respectively. These are then forwarded to a joint decoder that attempts to recover X_1 and X_2 from Y_1 and Y_2. Let us note that this is always possible if the communication is forced to be disjoint, as in Fig. 1-2. That is, one may simply losslessly encode/decode X_1 and X_2 separately. As we saw in the previous section, the rate limitation for this is given by the achievable lower bounds: $R_i \geq H(X_i)$ for each i.

However, we can do better if the sources are correlated. Consider, for instance, the joint probability mass function (pmf) depicted in table 1.2. The sum of the marginal entropies — the rate limitation under the stricter constraints of Fig. 1-2 — is $H(X_1) + H(X_2) = 1.81$ bits. Due to correlations, the "joint entropy" (the entropy of the random variable (X_1, X_2)) is only 1.5 bits. As it happens, it is possible to reconstruct X_1 and X_2 from rates that sum to the latter of these quantities: only 1.5 bits.

According to the theorem of Slepian and Wolf [8], one may block code X_1 and X_2 at a sum-rate $R_1 + R_2$ arbitrarily close to $H(X_1, X_2)$ with arbitrarily low probability of error. In other words, there is no loss associated with having to separately encode X_1 and X_2! This theorem generalizes to N sources, X_1^N, in that the sum-rate lower bound $\sum_{i=1}^N R_i \geq H(X_1^N)$ is achievable. We will make use of this property in the development of functional quantization.

1.2 Lossy Compression and Rate-Distortion Theory

Lossless compression is only half the story. Consider audio compression for instance: one might record a WAV file on a microphone before converting it to MP3 format and saving a considerable amount of disk space. Listening to the MP3, however, can be jarring, depending on one's taste in music — the compression algorithm may have introduced distortion into the audio. The original WAV cannot be perfectly recreated from the MP3 data — some information has been lost during compression. Lossy compression seeks to trade off the reduced rate from lost information with the distortion it introduces.

In the notation of entropy, we may easily define the notion of "lost" information as the difference $H(X) - H(\widehat{X})$, where $\widehat{X}$ is the approximated version of X. This corresponds to the reduction in bits due to the approximations. But not all information is equally "relevant": it's far more important, for instance, that we retain our header information for the WAV file than the last 0.1 seconds of sound!

Shannon quantified the notion of distortion by first defining an error function. For instance, the squared-error between a real value x and its approximated value $\widehat{x}$ is $|x - \widehat{x}|^2$. The expected value of the error function, d, is defined as the distortion:

$$D = \mathbf{E}\left[d(X, \widehat{X})\right]$$

where X is the random variable we care about, and $\widehat{X}$ is its compressed representation.

The rate-distortion function associated with a specified source and error function summarizes the "best possible" behavior for a lossy compression scheme. $R(D_0)$ gives the lowest possible rate at which the distortion is less than or equal to D_0. Analogously to lossless compression, one may approach $R(D_0)$ performance arbitrarily closely, but one may not simultaneously achieve a rate below R and a distortion below D_0.

The generality of this construction is both its strength and its weakness. In some cases, the rate distortion function is precisely known. For instance, Fig. 1-3 depicts the $R(D)$ performance for a memoryless Gaussian source. In most cases, it is not.

Note that the choice of error function is critical in defining both the $R(D)$ function and the implementations that can approach it. The squared-error metric is often used for real-valued sources due to its analytical tractability. One may interpret functional compression as attempting to exploit the tractability of MSE while expanding the number of applicabile scenarios.

1.3 Scalar Quantization

The rate-distortion function for a source (and distortion function) tells us how well we may approximate the source, but it does not instruct us on how to perform this approximation. Quantization provides a more literal framework for this, by explicitly dictating the lossy mapping that is to be used.

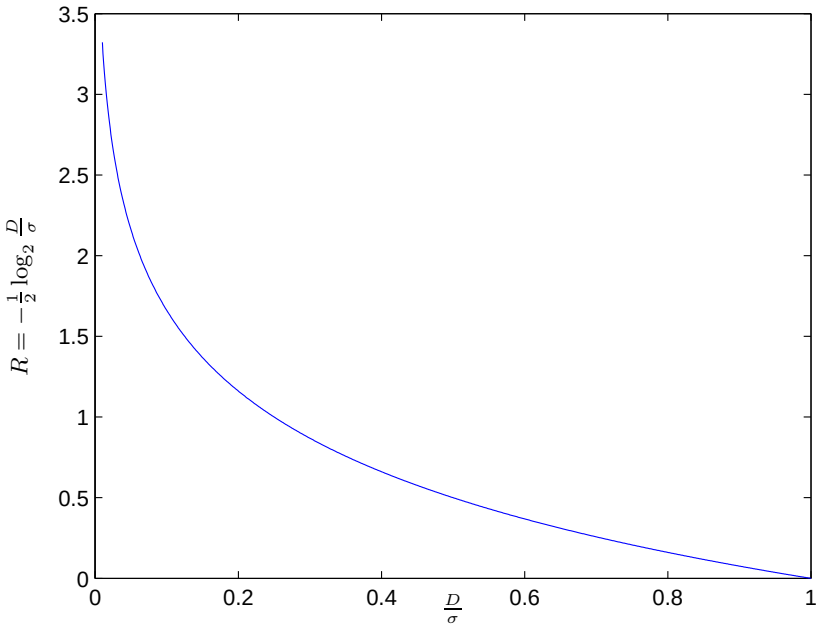

Figure 1-3: The rate-distortion function for a memoryless Gaussian source of variance σ^2.

Scalar quantization is best defined in the context of a real-valued source, X, with a probability density function (pdf) $f_X(x)$. In general, one cannot encode X exactly with any finite number of bits — some sort of "rounding" to a finite number of levels is necessary. This rounding — the process of quantization — will introduce some distortion. The distortion, in turn, will be related to both the number and placement of levels.

Quantization (as we will consider it) involves two components. First, a finite number K of reconstruction points must be specified. Second, a mapping $Q(X)$ from the source alphabet to the reconstruction points must be defined. Suppose, for instance, that X is a uniform source over $[0, 1]$, and we wish to quantize it into a discete variable $\widehat{X}$ with K levels, so as to minimize the "mean-squared error", $\mathbf{E}\left[(X - \widehat{X})^2\right]$. How should the K reconstruction points, and the corresponding cells $Q^{-1}(\widehat{X})$ be placed?

The symmetry of the problem encourages us to space the K levels uniformly over the range $[0, 1]$. Having decided on this placement, the mapping from X to $\widehat{X}$ can be chosen to minimize the distortion. This amounts to rounding each value of X to its nearest quantization level. The quantizer we have defined in this manner is summarized by two pieces of information: the placement

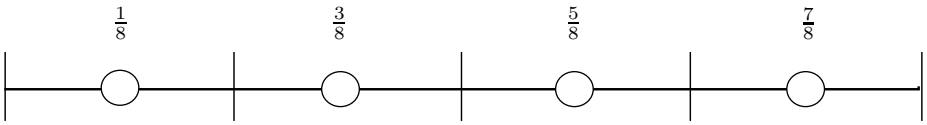

Figure 1-4: A simple 4-level quantizer for a uniform $[0, 1]$ source.

of the levels, and the "cells" in X that are rounded to each level. Fig. 1-4 illustrates both the cells and the levels for our example.

We make an important observation at this point: for the squared-error distortion metric, every quantization cell created by this rounding operation is connected (intervals in the one-dimensional case). This quality is known as *regularity*, and quantizers that obey it are *regular quantizers*. It can be seen that optimal quantizers for the squared-error metric are regular. Note that this does not necessarily extend to other distortion measures.

The performance of this quantizer mapping, $Q(X) : X \rightarrow \widehat{X}$, is described by the distortion it introduces:

$$\begin{aligned} D &= \mathbf{E}\left[(X - \widehat{X})^2\right] \\ &= \sum_{\widehat{X}} p(Q(X) = \widehat{X}) \mathbf{E}\left[(X - \widehat{X}) \mid Q(X) = \widehat{X}\right] \\ &= \sum_{\widehat{X}} p(Q(X) = \widehat{X}) \frac{1}{12K^2} \\ &= \frac{1}{12K^2} \end{aligned}$$

This is one step from taking the form of a distortion-rate function: we need only to establish a connection between the number of levels, K, and the quantizer's rate, R. This relationship is heavily dependent on the way the quantizer chooses to encode its finite-alphabet output $\widehat{X} = Q(X)$. We will consider two scenarios: fixed-rate (codebook-constrained) coding, where all codewords are of identical length, and variable-rate (entropy-constrained) coding, where the techniques of lossy compression are applied to $\widehat{X}$.

Fixed-Rate Coding. If all codewords are of identical length, the rate is set by the codebook's size. In the case of a scalar quantizer with K possible quantization points, this rate is simply the logarithm of K, $R = \log_2 K$.

Variable-Rate Coding. Things are slightly more complicated for variable rate coding. Shannon demonstrated that we may code a finite-alphabet random variable at a rate arbitrarily close to its entropy with arbitrarily small probability of error. Ignoring issues of implementation and the nuances of this statement, it more or less tells us that $\widehat{X}$ may be encoded at average rate $R = H(\widehat{X})$.

The problem is that the rate does not depend solely on K any longer; the placement of the levels plays a large part in determining the entropy of $\widehat{X}$. For instance, if I were to quantize a uniform $[0, 1]$ source with 8 levels uniformly across $[1/2, 1]$ and one level at $1/4$, the resulting entropy would be $H(\widehat{X}) = 2.5$ bits; noticeably lower than the $\log_2 9 \approx 3.17$ of a uniform quantizer with the same number of levels.

We now describe a powerful analytical tool that allows us to gracefully explore questions involving quantization. In the process, complex situations such as variable-rate coding are shown to have relatively simple interpretations.

1.4 The High-Rate Approximation for Scalar Quantization

The difficulty in analyzing quantization is symptomatic of a broader difficulty in science and engineering: the analysis of systems with mixed discrete and continuous components. For instance, the modeling of biological ion channels can be incredibly difficult if one attempts to consider the movement and behavior of each charged particle passing through the channel [9]. Given that the detailed shape of the channel and its interaction with each particle plays a critical role in regulating passage, one might consider this computational barrier a show stopper. What scientists have found, however, is that the charged particles may be approximated as a fluid with a *continuous* charge density, and the channel as a cylinder with a certain charge profile. While this approximation is incredibly rough, it yields trends and quantitative behavior that is surprisingly in line with observations [10].

A similar approximation is common in the analysis of scalar quantization. As the number of quantization levels grows, one may decouple the *design* of the quantizer from its *resolution* by means of a quantization "point density." Instead of speaking of the placement of K discrete levels, one deals with a normalized quantization *point density function*, $\lambda(x)$. We define λ in the following manner:

Let Δ be positive, K be the number of quantization points, and λ be a point density. In the limit of large K, $K\lambda(x)\Delta$ approximates the number of quantization points within the interval $[x - \frac{\Delta}{2}, x + \frac{\Delta}{2}]$. The spacing of these points will be roughly uniform.

The function λ allows us to describe a regular quantizer with a continuous function, instead of with a set of discrete levels. In order for it to be a useful construction, however, several approximations prove necessary:

1. The conditional probability density within any quantization interval, $f_{X|Q(X)}(x \mid Q(X) = \widehat{x})$, is roughly uniform. This is a reasonable assumption if the source distribution is smooth and the rate is high. See Fig. 1-5 for an illustration.

2. The quantization point density, $\lambda(x)$, is similarly approximated as constant within any quantization cell.

3. Neighboring intervals are similarly spaced. This is a reasonable assumption if the quantization point density, $\lambda(x)$, is smooth and the rate is high.

Armed with these assumptions, a continuous expression for the distortion in terms of $\lambda(x)$, the resolution K, and the source distribution $f_X(x)$ is possible. The MSE distortion within a single quantization cell, with reconstruction point $\widehat{x}_i$ is given by the variance of a uniform distribution, according to approximation (1). The length of this cell is given by $\Delta(\widehat{x}_i) = (\lambda(\widehat{x}_i)K)^{-1}$; therefore the distortion within the cell is

$$\frac{1}{12}\Delta^2(\widehat{x}_i) = \frac{1}{12}\frac{1}{K^2\lambda(\widehat{x}_i)^2}$$

The MSE over all quantization cells is the weighted sum of these distortions:

$$D = \sum_{i=1}^{K} p(\widehat{x}_i)\frac{1}{12}\frac{1}{K^2\lambda(\widehat{x}_i)^2} \tag{1.1}$$

Since $\lambda(x)$ is roughly constant within any interval, each term in the summation can be approximated by an integral:

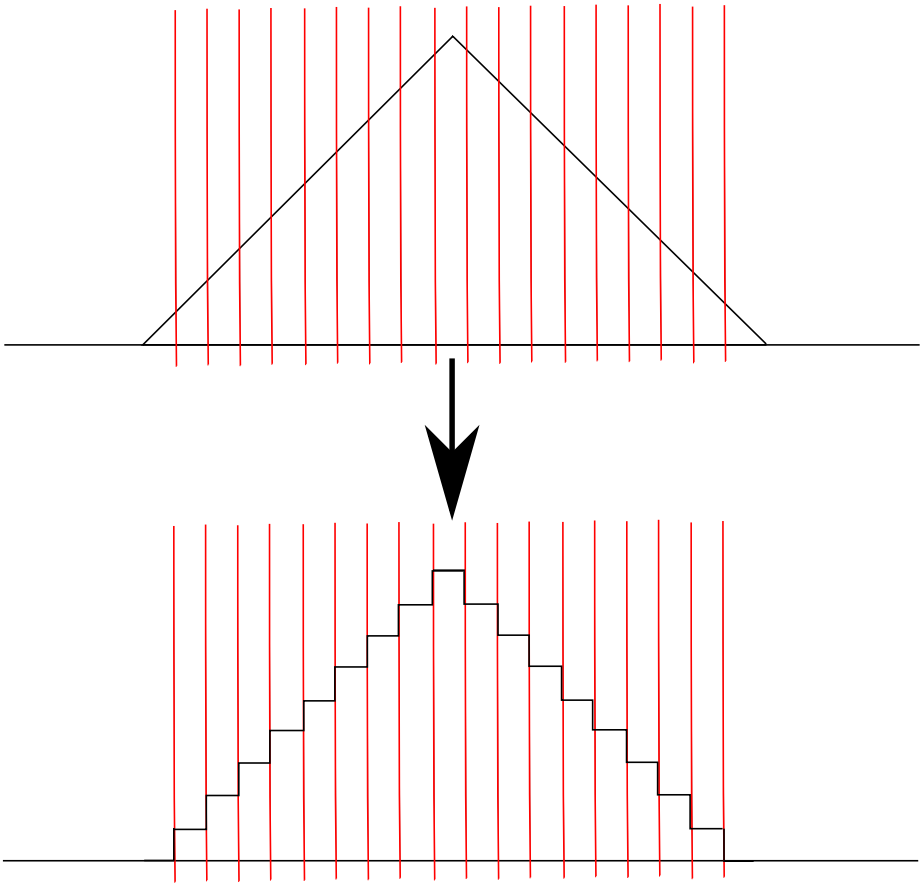

Figure 1-5: Demonstration of the high-resolution approximation. The "triangle" is the source distribution, and the vertical lines indicate quantizer cell boundaries.

$$\begin{aligned} D &\approx \sum_{i=1}^{K} \int_{x \text{s.t.} Q(x)=\widehat{x}_i} \frac{1}{12} f_X(x) \frac{1}{K^2 \lambda(x)^2} dx \\ &= \frac{1}{12K^2} \sum_{i=1}^{K} \int_{x \text{s.t.} Q(x)=\widehat{x}_i} f_X(x) \frac{1}{\lambda(x)^2} dx \\ &= \frac{1}{12K^2} \int_x f_X(x) \frac{1}{\lambda(x)^2} dx \end{aligned} \tag{1.2}$$

The design of our quantizer, represented by $\lambda(x)$, is now isolated from the resolution, K, and we have an expression that makes a good deal of qualitative sense. In fact, it is oftentimes quantitatively accurate even for relatively low rates. Our goal, however, is not merely one of modeling the system; we wish to design a quantizer to minimize this distortion. As the process of optimizing λ takes different forms for fixed- and variable-rate scalar quantization, they will be considered separately.

Fixed-Rate Quantization In this situation, $R = \log_2 K$. We therefore need to pick a point density λ so as to minimize $\mathbf{E}\left[\lambda^{-2}(X)\right]$. Application of Hölder's inequality shows that the minimizing choice is given by:

$$\lambda(x) = S f_X(x)^{1/3}$$

where S is a normalization constant and $f_X(x)$ is the probability density of the source. The resulting distortion is:

$$D = \frac{1}{12} 2^{-2R} \left[\int f_X(x)^{1/3} dx \right]^3$$

The term in brackets is the $\mathcal{L}^{1/3}$ pseudonorm of $f_X(x)$, and we denote it by $\|f_X(x)\|_{1/3}$.

Variable-Rate Quantization The construction of this optimization problem is not as trivial

as for the fixed-rate. Rather than having a simple relationship to K, the rate is given by:

$$\begin{aligned}
R &= H(\widehat{X}) && (1.3)\\
&= -\sum_{i=1}^{K} p(\widehat{x}_i)\log_2 p(\widehat{x}_i) && (1.4)\\
&\approx -\sum_{i=1}^{K} f_X(\widehat{x}_i)\Delta(\widehat{x}_i)\log_2(f_X(\widehat{x}_i)\Delta(\widehat{x}_i)) && (1.5)\\
&\approx -\int f_X(x)\log_2\left(f_X(x)\frac{1}{K\lambda(x)}\right)dx && (1.6)\\
&= \underbrace{-\int f_X(x)\log_2 f_X(x)dx}_{h(X)} + \underbrace{\int f_X(x)\log_2 K dx}_{\log_2 K} + \underbrace{\int f_X(x)\log_2\lambda(x)dx}_{\mathbf{E}[\lambda(X)]} && (1.7)\\
&= h(X) + \log_2 K + \mathbf{E}\left[\lambda(X)\right] && (1.8)
\end{aligned}$$

where Eq. 1.5 follows from the piecewise constant approximation to $f_X(x)$ and Eq. 1.6 follows from the Riemann sum approximating the integral. Inserting this relation between the rate, R, and the resolution, K, into the distortion relation 1.2 gives:

$$D = \frac{1}{12}2^{-2R+h(X)+\mathbf{E}[\log\lambda(X)]}\mathbf{E}\left[\lambda(X)^{-2}\right]$$

It can be shown by Jensen's inequality that this expression is minimized when $\lambda(x)$ is constant. That is, in the high-rate regime, the uniform quantizer is optimal. The resulting distortion is given by an aesthetically pleasing expression:

$$D = \frac{1}{12}2^{-2R+h(X)}$$

Note on Regularity Assumption: The above analysis seeks to obtain an optimal quantization profile for a given source distribution. A subtle point, however, is that the point density function can only be used to describe regular quantizers. Our solutions are therefore optimal amongst the set of regular quantizers. Since the distortion measure of concern is the squared-error, the best possible performance may be achieved by a regular quantizer, and our solutions are globally optimal. Note, however, that this ceases to be true when we consider functional quantization for arbitrary functions. The question of high-resolution non-regular quantization is settled in chapter 3.

1.5 Vector Quantization

Just as one might losslessly encode/decode several source variables at the same time via block coding, one may quantize several real-valued source variables together. This process, the quantization of a real-valued source vector, is referred to as *vector quantization* (VQ).

Formally, X_1^N is a random vector with some joint probability distribution, $f_{X_1^N}(x_1^N)$. A quantizer is a mapping from $\mathbb{R}^N$ to K reconstruction points, $\widehat{X}_1^N = Q(X_1^N)$. As before, the quantizer can be seen as a combination of two pieces of information: (1) the locations of the reconstruction points, and (2) the "cells" in $\mathbb{R}^N$ that are rounded to each of the K levels.

VQ can be seen as a generalization of, or alternative to, separately scalar quantizing each of the vector components X_i. One of its obvious advantages is to exploit correlations between the source variables. Consider as an example a two-vector source, X_1^2, where the joint pdf is given over $[-1,1]^2$ by:

$$f_{X_1^2}(x_1^2) = \begin{cases} \frac{1}{2} & \text{if} x_1 x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The marginal distributions of X_1 and X_2 are uniform over $[-1,1]$, so the optimal scalar quantizers would each be uniformly spaced over $[-1,1]$. Note, however, that half the quantization cells that are created from these scalar quantizers are never used! A vector quantizer has the option of only creating cells where $x_1 x_2 \geq 0$, and thereby using the allocated rate much more efficiently.

As it turns out, VQ has advantages over scalar quantization even when the sources are independent. Suppose, for instance, that X_1 and X_2 are independently, identically distributed with the uniform $[0,1]$ distribution. The optimal scalar quantizer, as before, is uniformly spaced over $[0,1]$ for each source. One may visualize this quantization as tiling the space $[0,1]^2$ with identical square cells. VQ, however, can reduce distortion by using hexagonally shaped cells. See Fig. 1-6 for an illustration of this.

This advantage will be referred to as the *shape gain* associated with VQ at a certain dimension. Quantitatively, the k-dimensional shape gain is the maximum constant of improvement in distortion by using a k-dimensional nonrectangular cell that (1) has normalized volume, and (2) can tile k-dimensional space. As the dimensionality grows arbitarily large, this shape gain approaches a constant that numerologists no doubt find very exciting: $\frac{1}{6}\pi e$ [4]. In this thesis, we will generally

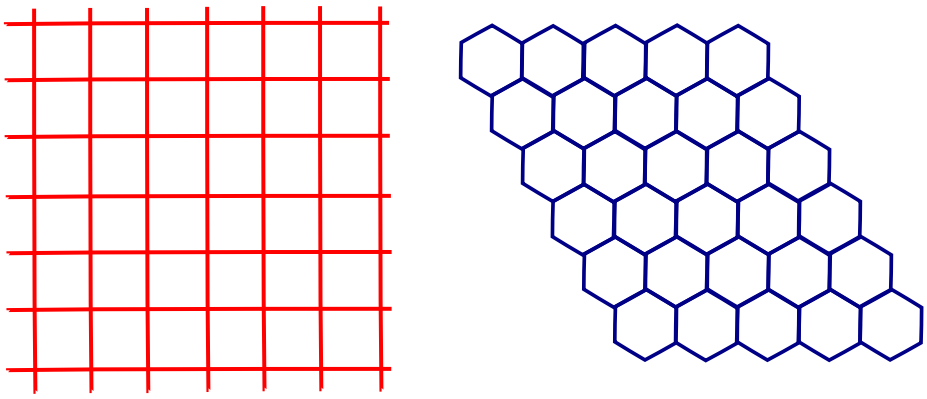

Figure 1-6: Comparison of optimal quantization cells for (left) separable scalar quantization and (right) vector quantization. The hexagonal lattice is more efficient, but more computationally intensive to implement.

tiptoe around the use of non-rectangular quantization cells; they result in a negligible boost to rate-distortion compared to the effects we will be interested in.

Just as we consider scalar quantization to come in two varieties, so does VQ. One may assign each quantization cell a constant-length codeword, and thereby generate a fixed-rate quantizer of rate $R = \frac{1}{N} \log_2 K$ bits per (scalar) source symbol. Alternatively, variable-rate entropy coding can be applied to $\widehat{X}_1^N$, and thereby generate a variable-rate quantizer of rate $R = \frac{1}{N} H(\widehat{X}_1^N)$.

Lastly, let us emphasize that while VQ has notable advantages over scalar quantization, it can prove costly to implement. In general, one must check whether the source falls into each of the quantization cells before assigning a codeword; this is an $O(2^{NR})$ operation and quickly becomes unmanageable with growing dimensionality N. Several constructions exist that seek to trade off the complexity of the VQ process with performance of the resulting quantization scheme. Of these, we will occupy ourselves primarily with the variety known as transform coding.

1.6 Transform Coding

The computational complexity of performing arbitrary vector quantization grows notoriously with increasing dimension. However, the benefits of jointly quantizing a large number of sources together is oftentimes difficult to ignore. Indeed, it can prove advantageous in many cases to constrain a vector quantizer to a more computationally tractable form — even if this constraint comes at some loss of optimality. The reduced performance at any one dimension can be more than offset by the

ability to operate at large dimensions.

Transform coding is a popular form of constrained VQ. An $N \times N$ linear transformation, U, is first applied to the source vector x_1^N to produce the transform coefficients $y_1^N = Ux_1^N$. The transform coefficients are then separately scalar quantized into the vector $\widehat{y}_1^N$. The rate of the transform code is the average of the rates for each scalar quantizer; we wish to design the transform, U, the scalar quantizers $Q_i(y_i) = \widehat{y}_i$, and the rate allocations R_i (s.t. $\sum_{i=1}^N R_i \leq R$) in order to minimize the distortion

$$D = \mathbf{E}\left[\frac{1}{N}\left|X_1^N - U^{-1}Q_1^N(UX_1^N)\right|^2\right]$$

Suppose the scalar quantization is variable-rate. As we saw in Sec. 1.4, the optimal variable-rate scalar quantizer is uniform for sufficiently high rates. The distortion for each scalar quantizer then obeys the relation

$$D_i = \frac{1}{12}2^{2h(Y_i)-2R_i} \tag{1.9}$$

For an arbitrary source, it is difficult to analyze this expression or the effect of a transformation on it. As such, we constrain our attention to the case of a jointly Gaussian source vector X_1^N. While the results from this analysis don't perfectly generalize to arbitrary source distributions, they do give useful insights.

For a jointly Gaussian source X_1^N, transformation results in another jointly Gaussian vector of coefficients Y_1^N. It can be shown [11] that Eq. 1.9 reduces to

$$D_i = \frac{\pi e}{6}\sigma_{y_i}^2 2^{-2R_i} \tag{1.10}$$

where $\sigma_{y_i}^2$ is the variance of the ith transform coefficient, y_i. Summing the contributions from the N sources, we have a total distortion of

$$D = \sum_{i=1}^{N} \frac{\pi e}{6}\sigma_{y_i}^2 2^{-2R_i} \tag{1.11}$$

This can be minimized in two steps: first the optimal distribution of rate R amongst the N encoders' rates R_i should be determined, and then the optimal transform can be chosen to minimize the resulting expression.

Optimal Rate Allocation. An application of the arithmetic/geometric mean inequality demonstrates that Eq. 1.11 is minimized when the geometric mean of the N terms equals their arithmetic mean. In other words,

$$D \geq N\frac{\pi e}{6}2^{-2R/N}\left(\prod_{i=1}^{N}\sigma_{y_i}^2\right)^{1/N} \tag{1.12}$$

Can we select individual rates R_i summing to R such that this lower bound is achieved? The rates required are $R_i = \frac{R}{N} - \frac{1}{2N}\log_2\left(\prod_{i=1}^{N}\sigma_{y_i}^2\right)$. If each of the variances $\sigma_{y_i}^2$ is nonzero, these rates are feasible for sufficiently large sum-rate R.

Optimal Transform. A linear transform, U, applied to the source vector X_1^N creates a new correlation matrix for the transform coefficients Y_1^N. Specifically,

$$\begin{aligned} K_{yy} &= \mathbf{E}\left[Y_1^N\left(Y_1^N\right)^T\right] \\ &= \mathbf{E}\left[UX_1^N\left(X_1^N\right)^T U^T\right] \\ &= U\mathbf{E}\left[X_1^N\left(X_1^N\right)^T\right]U^T \end{aligned}$$

Assuming that appropriate bit allocations will follow the transformation, our goal is to minimize the distortion given in Eq. 1.12. This is equivalent to minimizing the product of diagonal elements of K_{yy}, a quantity we refer to as the *multiplicative trace*. The Hadamard inequality demonstrates that the minimizing transformation places the matrix K_{xx} into its eigenbasis — in other words, the optimal transformation U decorrelates the source. This transformation is known as the Karhunen-Loeve Transform (KLT),

Shortcomings. As noted, the KLT was only derived as optimal for the case of a jointly gaussian source vector. In general, placing a source into its decorrelated basis reduces redundancy between the coefficients, but other factors may oppose this. For instance, if fixed-rate quantization is being performed and the source's support is not spherically symmetrical, a non-KLT basis may allow for more efficient tiling of the support with a fixed number of quantization cells.

Chapter 2

Sophomore Year: Functional Quantization for Monotonic Functions

In this chapter, we will develop the functional quantization results that form the heart of this work. The techniques and mathematical picture we work with are as important as the analytical solutions to the quantization design problems of interest; we will make use of them extensively in the ensuing chapters. To aid in the development of these techniques, we restrict our attention to functions monotonic in each argument and, thereby, to the set of regular quantizers. These restrictions will eventually be relaxed in Chapter 4.

We start by discussing a few topics within the wide spectrum of related work. The generality of the functional quantization problem creates connections to topics ranging from perceptual audio coding to multiterminal source coding. Some are obviously more closely related than others; we allocate our attention accordingly.

We then begin to develop the theory by considering the relatively simple single-dimensional functional quantization scenario. Even though the analysis for one dimension is straightforward, it suggests an approach for higher dimensions.

Optimal quantizers are then obtained for several increasingly unconstrained scenarios of multi-dimensional functional quantization. First, the N-dimensional fixed-rate problem is attacked. We

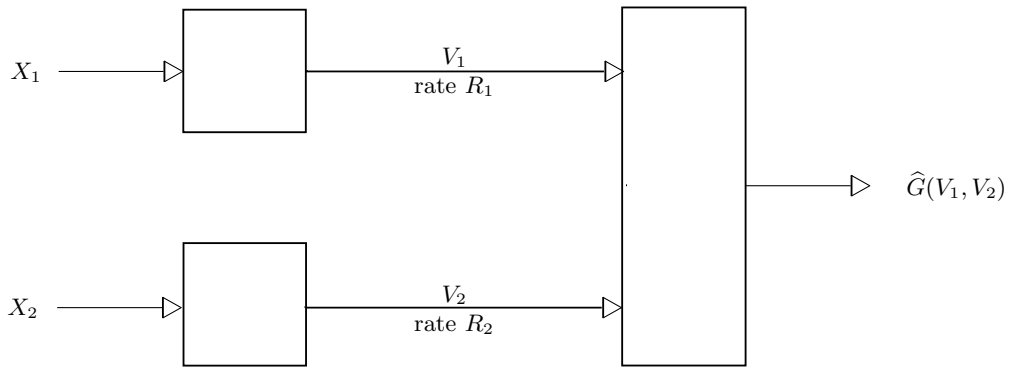

Figure 2-1: Generic functional source coding scenario shown with two variables.

build on this to solve the N-dimensional variable-rate problem, before generalizing once again to incorporate Slepian-Wolf entropy coding. The notion of "functional typicality" and "functional entropy" are then introduced, in analogy with their traditional counterparts. We demonstrate how they can provide an alternate route to many of our derivations.

2.1 Related Work

Functional quantization and, more generally, functional source coding live at the intersections of several problems: quantization, source coding, multiterminal information theory, and non-MSE distortion measures. As such, there are many topics that they relate to. We provide a brief summary of some of these connections here.

Quantization with a functional motive bears resemblance to the idea of "task-oriented quantization." There has been considerable work in this direction for classification [12], estimation [13], and detection [14]. Additionally the use of a function at the decoder can be seen as inducing a non-MSE distortion measure on the source data. In this sense, a similarity can be seen to perceptual source coding [15], where a non-MSE distortion reflects human sensitivity to audio or video.

The problem depicted in Fig. 2-1 is of central interest to us. Various special cases of it have been previously considered from different perspectives. In general, X_1 and X_2 are random variables with some joint distribution, and G is a function of the two.

- If G is the identity function, we have a general distributed source coding problem that is well-known in the lossless setting [8] and recently solved in the quadratic Gaussian case [16]. In this situation, the correlation of X_1 and X_2 is of primary interest.

- If $G(X_1, X_2) = X_1$ and R_2 is unconstrained, then X_2 can be viewed as receiver side information available at the decoder. The trade-off between R_1 and distortion to X_1 is given by the Wyner-Ziv rate-distortion function [17, 18].

- The Wyner-Ziv scenario has been examined at high resolution by Rebollo-Monedero et al. [19]. It has been shown that providing the receiver side information to the encoder yields no improvement in performance.

- For general G and R_2 unconstrained, the problem has been studied by Feng et al. [20]. Under suitable constraints on the distortion metric, one may also view X_2 as receiver side information that determines the distortion measure on X_1, drawing a connection to [21].

- Let $Y = G(X_1, X_2)$. Then Y may be interpreted as a *remote source* that is observed only through X_1 and X_2 and we have the remote source multiterminal source coding problem [22].

- Rather than having a single function G, one may consider a set of functions $\{G_i\}_{i \in I}$ and define $D_G = \mathbf{E}\left[d(G_\alpha(X_1^N), G_\alpha(\widehat{X}_1^N)\right]$, where α is a random variable taking values in index set I. In this setting, fixed- and variable-rate quantization to minimize MSE was studied by Bucklew [23]. Note that if the function were known deterministically to the encoder, one would be better off simply computing the function and encoding it directly.

A couple pieces of work are related in results, even though they make use of very different techniques. We explore these in slightly more depth below.

2.1.1 Discrete Functional Compression

One may consider the scenario of finite-alphabet sources and lossless functional compression. The Wyner-Ziv version of this problem (side information at the decoder) was shown by Orlitsky and Roche [24] to reduce to the entropy of a "characteristic graph." Later, Doshi et al. [25] generalized these results to the case of distributed sources, and demonstrated the applicability of graph-coloring to the problem. We will illustrate both the characteristic graph concept and the graph coloring approach with a simple example.

Let X_1 take values over $\{0, 1\}$ uniformly, and let X_2 take values over $\{0, 1, 2, 3\}$ uniformly. Suppose the function of interest is the modulo-2 sum of X_1 and X_2; that is, $G(X_1, X_2) = (X_1 + X_2)\%2$. If X_1 and X_2 must be separately compressed, as in Fig. 2-1, how low of a sum-rate is

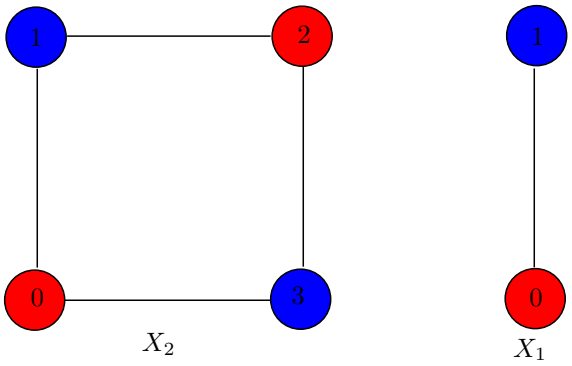

Figure 2-2: Characteristic graphs for 1- and 2-bit random variables, when the decoder is interested in the modulo-2 sum of the two.

possible while still losslessly calculating $G(X_1, X_2)$ at the decoder? To answer this question, one constructs a characteristic graph for each of the sources.

Each node of a characteristic graph corresponds to a letter in the source alphabet $\mathcal{X}_i$. An edge is drawn between two nodes a and b in the characteristic graph of X_1 if the following holds:

Condition for an edge: If there exists a symbol $y \in \mathcal{X}_2$ such that $p(X_2 = y, X_1 = a) > 0$, $p(X_2 = y, X_1 = b) > 0$, and $G(a, y) \neq G(b, y)$, we draw an edge between a and b.

For the situation we drew out, the characteristic graph for X_1 is complete, while the graph for X_2 is missing the edges $(0, 2)$ and $(1, 3)$ since those points are indistinguishible from G's perspective. According to Doshi, we can do no better than coloring each graph and having each encoder communicate the colors to the decoder. Fig. 2-2 demonstrates a possible 2-color coloring for these graphs that results in a sum-rate of 2 bits. Compare this to the ordinary compression of Slepian-Wolf — it would take 3 bits to perfectly recreate the sources according to their result.

Note that if a function can sometimes distinguish between all the values of a source, the characteristic graph is complete and functional compression yields no advantage over ordinary compression. Most functions of interest fall into this category. A similar notion of "distinguishability" shows up in our work with non-monotonic functions in Chapter 3.

2.1.2 Functional Source Coding

In their 2004 paper, Feng et al. [20] consider the problem of functional source coding in the Wyner-Ziv context; a source X_1 is coded and communicated to the receiver, which estimates a function $G(X_1, Y)$ of X_1 and side information Y. They obtain several results, which we summarize below.

1. A functional rate-distortion expression similar in form to the Wyner-Ziv equation is derived, making no assumptions about the type of distortion measure or the type of function involved.
2. The *rate loss* from using ordinary source coding instead of Wyner-Ziv coding is shown to be arbitrarily large.
3. Under the constraint of MSE distortion measure, the rate loss between providing and not providing the side information Y to the encoder is shown to be arbitrarily large when the function is not separable.
4. When the function is separable, the loss from providing the side information is at most half a bit.
5. When the function is smooth (in the same sense that we deal with it), the rate loss from side information not being present is bounded in terms of the maximum magnitude of the derivative of the function.
6. The influence of noise on the problem is also considered through the use of rate loss bounds.

A point of comfort: several of these results (2,3,4,5) may be confirmed by considering appropriate special cases of our functional quantization setup.

At this point, we will begin to develop our functional quantization picture. Our approach will be to consider added complexity step-by-step; most notably, in this chapter we only consider the highly restrictive case of functions monotonic in each of their variables. This condition will be relaxed in Chapter 4. Connections with the above prior work will be noted as they appear.

2.2 Single-Dimensional Functional Quantization

Limiting ourselves to a single dimension feels very much like a straitjacket: the applications are limited, and the results are rarely interesting. Nonetheless, it helps to avoid the full-blown distributed

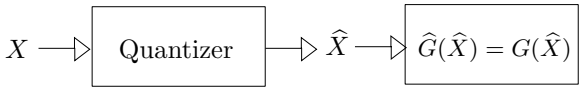

Figure 2-3: Single dimension fixed-rate functional quantization

functional quantization before we have developed the basic analytical tools to be used. The 1D scenario is perfect from this perspective: despite its simplicity, we see elements of the more complicated problems in it.

Suppose our encoder is provided a single continuous-valued source variable, $X \in [0, 1]$. Traditional source coding dictates that we are interested in recreating X itself; for instance, we may explicitly seek to minimize the mean-squared error between X and its R-bit quantized representation, $\widehat{X}$, $\mathbf{E}\left[(X - \widehat{X})^2\right]$. The functional perspective generalizes our interest from X to a function $G(X)$.

An immediately obvious approach to attacking this problem is to compute $G(X)$ first, and then quantize $G(X)$. However, while this technique is effective, it fails to generalize to the distributed N-dimensional cases where none of the encoders has sufficient information to compute the function. We instead take the approach of deriving a new distortion measure for X that reflects functional considerations. In Appendix 2.B the equivalence of these two methods is shown for one dimension.

As depicted in Fig. 2-3, our encoder performs quantization of X into $\widehat{X} = Q(X)$, and transmits $\widehat{X}$ to the receiver. The receiver then makes use of an estimator, $\widehat{G}(\widehat{X})$, to approximate the correct value of the function, $G(X)$. We quantify the receiver's performance by means of functional MSE: $\mathbf{E}\left[(G(X) - \widehat{G}(\widehat{X}))^2\right]$.

In order to analyze this problem, the function G and the source X must be restricted in several ways. For the moment, we err on the side of being too strict — Chapter 3 will loosen the requirements.

1. The probability distribution of X, $f_X(x)$ must have bounded support. For convenience, we assume a support of $[0, 1]$.

2. $f_X(x)$ should be smooth.

3. $G(x)$ must be continuous on $[0, 1]$.

4. $G(x)$ must possess bounded derivative almost everywhere.

5. $G(x)$ must be monotonic in x.

The least necessary of these requirements, #5, is also the most limiting. We include it because of the following lemma:

Lemma 2.2.1 *If $G(x)$ is monotonic, the optimal functional quantizer of X will be* regular.

Proof A *regular* quantizer is one for which every quantizer cell is connected (for $\mathbb{R}^1$ this reduces to "every quantizer cell is an interval").

We make use of the fact that the optimal functional quantizer in one dimension is induced by the optimal ordinary quantizer for the variable $Y = G(X)$. That is, one may compute the function $G(X)$, and quantize it directly. Since the optimal ordinary quantizer for a real-valued source is regular, the optimal quantizer over Y, denoted by $Q_Y(y)$, is regular.

$Q_Y(y)$ may be simulated by quantization over X with reconstruction points in X given by $G^{-1}(\hat{y}_i)$ and cells in X given by $G^{-1}(Q_Y^{-1}(\hat{y}_i))$. We know that $Q_Y^{-1}(\hat{y}_i)$ is an interval, since Q_Y must be regular. Since G is monotonic and continuous, G is a homeomorphism between $[0,1]$ and $G([0,1])$. Then G^{-1} is a continuous, well-defined mapping, and since the continuous mapping of a connected space is connected, $G^{-1}(Q_Y^{-1}(\hat{y}_i))$ is connected. This demonstrates that a regular quantizer in X will be optimal. ∎

With this restriction, we are therefore able to limit our attention to the set of regular quantizers. In Sec. 3.2, we generalize to include nonregular quantization and non-monotonic functions (in the more general N-dimensional scenario).

The task we face is to design an estimator, $\widehat{G}(\widehat{X})$, and a K-level quantizer, $\widehat{X} = Q(X)$, so as to minimize the functional distortion $D = \mathbf{E}\left[(G(X) - \widehat{G}(\widehat{X}))^2\right]$. We attack each of these problems in turn.

Estimator, $\widehat{G}$:

App. 2.A, constrained to the one-dimensional case, demonstrates that there is no loss of optimality from selecting the estimator $\widehat{G}(\widehat{X}) = G(\widehat{X})$.

High-Resolution Distortion

Our interest lies in the high-resolution regime, where the distribution within any quantizer cell may be approximated as uniform, and the quantizer spacing may be described by a point density function.

To accommodate the function G, an additional approximation is used: G is linearized by its Taylor series coefficients within any quantizer cell. Suppose that y is the center of a quantizer cell. Then G is approximated within the cell as

$$G(x) \approx G(y) + \left.\frac{dG(x)}{dx}\right|_{x=y} (x - y)$$

Instead of repeating our work in the next sections, we make use of this approximation immediately for the N-dimensional distortion.

Theorem 2.2.2 *If N sources $X_1, \ldots, X_N$ are quantized according to point density functions $\lambda_1, \ldots, \lambda_N$ with resolutions $K_1, \ldots, K_N$, then the high-resolution distortion to a function $G(X_1, \ldots, X_N)$ is given by*

$$\boldsymbol{E}[d_G] \approx \frac{1}{12} \sum_{i=1}^{N} \boldsymbol{E}\left[g_i^2(x_i) K_i^{-2} \lambda_i^{-2}(x_i)\right]. \tag{2.1}$$

Proof of this theorem may be found in Appendix 2.C. For a single dimension — our present point of interest — the expression reduces to

$$D = \mathbf{E}\left[d_G(X, \widehat{X})\right] = \mathbf{E}\left[(G(X) - G(\widehat{X}))^2\right] = \int_0^1 f_X(x) \frac{1}{12K^2 \lambda_X(x)^2} g(x)^2 dx \tag{2.2}$$

where $d_G(x, y)$ is the functionally induced distortion measure, K is the number of quantization intervals, and we have defined $g(x) = \left|\frac{dG(x)}{dx}\right|$. Note that the quantity $g(x)^2$ summarizes the function's influence on quantizer performance.

Fixed-Rate (Codebook-Constrained) Quantization

Under a fixed-rate constraint, each quantization point, $\hat{X}_0$, is communicated with R bits — 2^R is therefore the number of intervals, K. The only remaining degree of freedom in the distortion (Eq. 2.2) is in the point density function, $\lambda_X(x)$. What choice of $\lambda_X(x)$ minimizes $D = \mathbf{E}[d_G]$?

Notice that Eq. 2.2 bears resemblance to Eq. 1.2, but with the probability density $f_X(x)$ replaced with a *surrogate* density, $f_X(x)\left|\frac{dG(x)}{dx}\right|^2$. This suggests that a similar optimization technique may be successful. Indeed, use of Holder's inequality demonstrates that the (functional) distortion is minimized when $\lambda_X(x)$ is chosen to be proportional to the cube root of the (surrogate) density. The

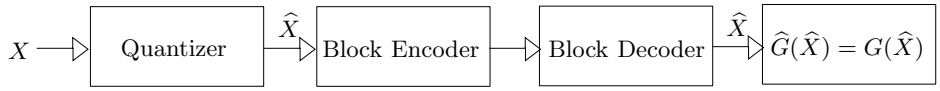

Figure 2-4: Single dimension variable rate functional quantization

resulting minimum is proportional to the $\mathcal{L}^{1/3}$ norm of the (surrogate) density. That is,

$$\lambda_X(x) \propto \left(f_X(x)g(x)^2\right)^{1/3} \tag{2.3}$$

$$\mathbf{E}\left[d_G\right] \geq \frac{1}{12}2^{-2R}\left\|f_X(x)g(x)^2\right\|_{1/3} \tag{2.4}$$

Variable-Rate (Entropy-Constrained) Quantization

Now suppose that a block entropy coder is allowed to operate on the output of the scalar quantizer, as in Fig. 2-4. While a rate constraint, $R \leq R_0$, continues to be enforced, the relationship between rate and resolution (K) is less obvious.

The lowest achievable rate of transmission is the entropy of the quantized variable, $H(\widehat{X})$. Neglecting practical considerations, we assume this rate may be precisely achieved. To consider the entropy in our optimization, we must know how it depends on the point density, $\lambda_X(x)$, and the resolution.

As demonstrated in Sec. 1.4, at high rate the discrete entropy of a quantizer output is approximated in terms of the source's differential entropy, $h(X)$. We repeat Eq. 1.8 for convenience:

$$R = H(\widehat{X}) \approx h(X) + \log_2 K + \mathbf{E}\left[\log_2 \lambda_X(x)\right] \tag{2.5}$$

An expression for distortion in terms of rate, R, and quantizer density, λ, is now available from Eqs. 2.5 and 2.4.

$$\mathbf{E}\left[d_G\right] = \frac{1}{12}2^{2h(X)-2R+2\mathbf{E}[\log_2 \lambda_X(x)]}\mathbf{E}\left[\frac{1}{\lambda_X(x)^2}g(x)^2\right] \tag{2.6}$$

In the general footsteps of Gersho [26], this may be minimized by Jensen's inequality:

$$
\begin{aligned}
D = \mathbf{E}\left[d_G\right] &= \frac{1}{12}2^{2h(X)-2R+2\mathbf{E}[\log_2 \lambda_X(x)]}\mathbf{E}\left[2^{-2\log \lambda_X(X)+2\log_2 g(X)}\right] && (2.7)\\
&\geq \frac{1}{12}2^{2h(X)-2R+2\mathbf{E}[\log_2 \lambda_X(x)]}2^{-2\mathbf{E}[\log \lambda_X(x)]+2\mathbf{E}[\log_2 g(X)]} && (2.8)\\
&= \frac{1}{12}2^{2h(X)-2R+2\mathbf{E}[\log_2 g(X)]} && (2.9)
\end{aligned}
$$

Jensen's inequality, given in general by $\mathbf{E}\left[2^Z\right] \geq 2^{\mathbf{E}[Z]}$, holds with equality when the exponent, Z, is deterministic. Therefore, the lower bound is achieved when we choose the point density appropriately:

$$\lambda_X(x)^2 \propto g(x)^2 \tag{2.10}$$

Example:

Suppose X is uniformly distributed over the interval $[0, 1]$, and that the decoder will compute $G(x) = x^2$.

The optimal ordinary quantizer — uniform for both fixed and variable rates — yields a functional distortion of

$$
\begin{aligned}
D &= \frac{1}{12}2^{-2R}\mathbf{E}\left[g(X)^2\right]\\
&= \frac{1}{9}2^{-2R} && (2.11)
\end{aligned}
$$

The optimal fixed rate functional quantizer is described by point density $\lambda_F(x) = \frac{5}{3}x^{2/3}$ and a distortion of

$$D = \frac{1}{3}\|x^2\|_{1/3} \cdot 2^{-2R} = \frac{9}{125}2^{-2R} \approx 0.0722^{-2R}$$

The optimal variable rate functional quantizer has point density $\lambda_V(x) = 2x$, proportional to dG/dX. After rearranging Eq. 2.9, the distortion is given by

$$D = \frac{2^{-2R}}{12}e^{2\mathbf{E}[\ln 2X]} = \frac{2^{-2R}}{12}e^{2(\ln 2-1)} \approx 0.045 \cdot 2^{-2R}$$

Even in a single dimension, benefits from functional quantization can be seen. Note that these

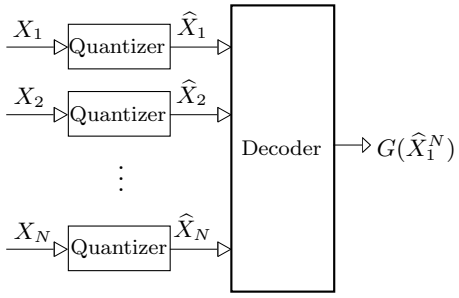

Figure 2-5: Fixed-Rate Distributed Quantization: Independent scalar quantization is performed at each source.

improvements are identical to those from computing the function prior to quantization (see App. 2.B). However, while pre-computation fails to extend to multidimensional scenarios, our techniques from this section do. We will find that the performance gap between functional and ordinary quantization can grow considerably with dimension.

2.3 N-dimensional Functional Quantization

At this point, we are prepared to attack the distributed problems at the heart of functional quantization. Our approach will be to consider increasingly open-ended scenarios, starting with fixed-rate quantization and moving towards the more general variable-rate block-quantization.

2.3.1 N-dimensional Fixed-Rate

We consider the situation depicted in Fig. 2-5. Let the source X_1^N be a random vector described by joint pdf $f_{X_1^N}(x_1^N)$; for convenience, let $f_{X_1^N}(x_1^N)$ be supported in $[0,1]^N$. Note that the components of X_1^N can be arbitrary correlated.

Let the function $G : \mathbb{R}^N \to \mathbb{R}$ be continuous, possess bounded derivative almost everywhere, and be monotonic in each of its arguments. Monotonicity is a useful property satisfied by many of the functions of interest to us; nonetheless we have found that as a requirement it can be loosened significantly (see Sec. 3.2).

Problem Statement

The ith encoder performs quantization on X_i to generate the approximation $\widehat{X}_i = Q_i(X_i)$. $\widehat{X}_i$ is chosen from a finite set of size K_i, and its index within this set is communicated at fixed rate

$R_i = \log_2 K_i$ to a centralized decoder.

The N encoders, Q_1 through Q_N, must together satisfy a sum-rate constraint. That is, $\sum R_i \leq R$ for some R. In practice, the individual rates R_i must be integer-valued, as they represent the number of bits communicated to the decoder from a single encoder. We relax this condition and allow the R_is to be (positive) real numbers for two reasons. First, real R_i may be arbitrarily closely approached by integral R_i through block coding. Second, it has been shown that optimal integral bit allocations can be obtained from optimal real bit allocations with little fuss [27].

The centralized decoder effectively receives a component-by-component quantized vector $\widehat{X}_i = Q_i(X_i)$. From this information, it must form an estimate $\widehat{G}(\widehat{X}_1^N)$ for the function $G(X_1^N)$, minimizing the functional distortion: $D = \mathbf{E}\left[d_G(X, \widehat{X})\right] = \mathbf{E}\left[|\widehat{G}(\widehat{X}_1^N) - G(X)|^2\right]$.

Choice of Estimator

Since multiple combinations of estimator/quantizer can minimize the distortion, we may restrict one or the other. App. 2.A demonstrates that optimality is still possible if the estimator is constrained as $\widehat{G} = G$.

Description of Quantizer

The cartesian product of N quantizers yields a tiling of $[0,1]^N$ by rectangular cells. In analogy with the single-dimensional situation, we make use of high-resolution approximations within each of these cells.

A1. The joint pdf, $f_{X_1^N}(x_1^N)$, is roughly uniform within any cell. A continuous $f_{X_1^N}(x_1^N)$ with derivative bounded almost everywhere will guarantee this.

A2. The function, $G(X)$, is roughly affine at any point. This permits a Taylor approximation to the function within each cell.

A3. The quantizer Q_i is described by a normalized point density function over X_i, $\lambda_i(x)$. The quantizer cell containing x_1^N is a rectangle with side lengths approximately given by $2^{-R_i}/\lambda_i(x_1^N)$.

We use these approximations to obtain the functional distortion for a specific choice of quantization densities $\lambda_i(x)$. To simplify the resulting expression, we define a quantity analogous to $g(x)^2 = \left|\frac{dG(x)}{dx}\right|^2$ from the single-dimensional case:

Definition $g_i^2(x)$, the ith *functional sensitivity* for a source X_1^N and function $G(x_1^N)$, is the expected squared partial derivative of $G(x_1^N)$ with respect to x_i:

$$g_i^2(x) = \mathbf{E}\left[\left|\frac{dG}{dX_i}\right|^2 \mid X_i = x_i\right]$$

Lemma 2.3.1 *If approximations A1, A2, and A3 hold, then functional distortion is given by*

$$D = \sum_{i=1}^{N} \frac{1}{12K_i^2} g_i^2(x_i)\lambda_i(x_i)^{-2} \tag{2.12}$$

where K_i is the number of quantization points for the ith encoder.

The proof for this may be found in App. 2.C. Because we are considering the fixed-rate situation, we may substitute the rate for the resolution: $R_i = \log_2 K_i$. This results in distortion

$$D = \sum_{i=1}^{N} \frac{2^{-2R_i}}{12} g_i^2(x_i)\lambda_i(x_i)^{-2} \tag{2.13}$$

In order to minimize this expression, we may optimize the λ_is and R_is separately.

Theorem 2.3.2 *Eq. 2.13 is minimized subject to a sum-rate constraint $\sum_{i=1}^{N} R_i \leq R$ by choice of point densities λ_i such that*

$$\lambda_i(x) \propto (f_X(x)g_i^2(x))^{1/3}$$

and choice of rate allocations R_i such that the total distortion is given by

$$\boldsymbol{E}[D] = \frac{N}{12} 2^{-2R/N} \prod_{i=1}^{N} \|f_{X_i}(x_i)g_i(x_i)^2\|_{1/3}^{1/N}.$$

Proof We choose optimal densities and rate allocations in turn.

Optimal densities, λ_i.

Eq. 2.13 is a sum of N separate expressions, each of which involves only one of the λ_is. Each of these terms is, in fact, representative of single-dimensional functional quantization with point density λ_i and squared derivative g_i^2 — the N-dimensional problem reduces to N parallel one-dimensional problems.

As such, we may separately choose each point density λ_i to minimize its corresponding term in the summation. The optimal choice is obtained from Eq. 2.3:

$$\lambda_i(x) \propto (f_X(x_i)g_i(x_i)^2)^{1/3}$$

and leads to distortion

$$\mathbf{E}[D] = \sum_{i=1}^{N} \frac{1}{12} 2^{-2R_i} \|f_{X_i}(x_i)g_i(x_i)^2\|_{1/3} \tag{2.14}$$

The only degrees of freedom remaining are in the rate allocations, R_i.

Optimal rate allocations, R_i.

Proceeding along standard lines for rate allocation problems, we first note the applicability of the arithmetic/geometric mean inequality. Applying this to Eq. 2.14, we have the following lower bound:

$$\mathbf{E}[D] \geq \frac{N}{12} \left(\prod_{i=1}^{N} 2^{-2R_i} \|f_{X_i}(x_i)g_i(x_i)^2\|_{1/3} \right)^{1/N}.$$

Recall that we constrain the rates with a sum-rate condition, $\sum_i R_i \leq R$. This may be incorporated into the lower bound:

$$\mathbf{E}[D] \geq \frac{N}{12} 2^{-2R/N} \prod_{i=1}^{N} \|f_{X_i}(x_i)g_i(x_i)^2\|_{1/3}^{1/N}. \tag{2.15}$$

If G is dependent on each of the source variables, this bound will be nonzero and achievable. If G is independent of one or more of the source variables X_i almost everywhere, then discarding these (unnecessary) components will allow a proper rate allocation amongst the remaining X_i, and Eq. 2.15 will be achievable in the adjusted source space. ∎

2.3.2 N-dimensional Variable-Rate

We now add some flexibility to the encoder and decoder by permitting the use of entropy coding, as depicted in Fig. 2-6. Specifically,

1. The ith scalar quantizer continues to scalar quantize each sample of X_i independently of other samples of X_i and independently of the happenings at other encoders. Quantization is performed at resolution K_i.

2. The ith block entropy coder losslessly encodes together M sequential outputs from the scalar quantizer, $\widehat{X}_{i1} \ldots \widehat{X}_{iM}$.

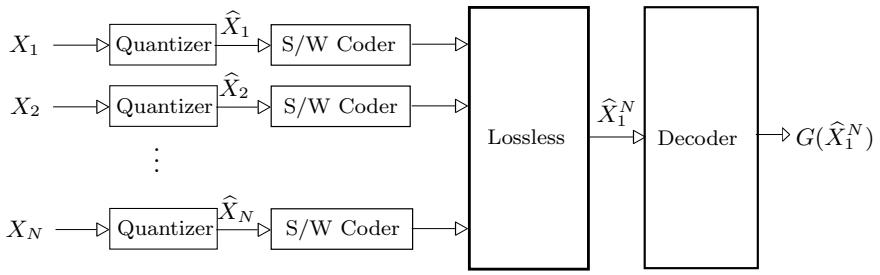

Figure 2-6: Variable Rate Quantization: Scalar quantization is now followed by block coding.

3. The average rate per symbol from the ith encoder is denoted R_i; note that $\log_2 K_i$ generally differs from R_i. We enforce a sum-rate constraint as before: $\sum_i R_i \leq R$.

4. The block decoder recreates $\widehat{X}_{i1} \ldots \widehat{X}_{iM}$ for each i.

5. Finally, the decoder computes an estimate of the values of G for each instance in the block: $\widehat{G}_j = \widehat{G}(\widehat{X}_{1j}, \ldots, \widehat{X}_{Nj})$.

Our task is to optimize the choice of estimator, lossless encoder, quantizer, and rate allocation. Each may be considered in isolation.

Estimator, $\widehat{G}$.

Assuming the lossless encoder is in fact lossless, the analysis performed in App. 2.A continues to hold for the variable-rate scenario. Therefore, the constraint $\widehat{G} = G$ is justified.

Lossless Encoder

N discrete variables, $\hat{X}_1^N$, must be separately encoded and jointly decoded — at the minimum possible sum-rate. This description fits the profile of the Slepian-Wolf problem, whose solution asserts both the achievability and optimality of a sum-rate arbitrarily close to the joint entropy, $H(\hat{X}_1^N)$.

When the source variables are independent, Slepian-Wolf coding is unnecessary. Under these circumstances, or more generally when the user desires a lower complexity entropy coding algorithm, the Slepian-Wolf decoder reduces to N disjoint encoders and decoders (Fig. 2-7). We consider the Slepian-Wolf scenario of Fig. 2-6 in more detail in the next section.

Within the framework of disjoint entropy coders, each quantized source $\widehat{X}_i$ can be encoded at rate arbitrarily close to its marginal entropy, $H(\widehat{X}_i)$. Using high resolution approximations as in

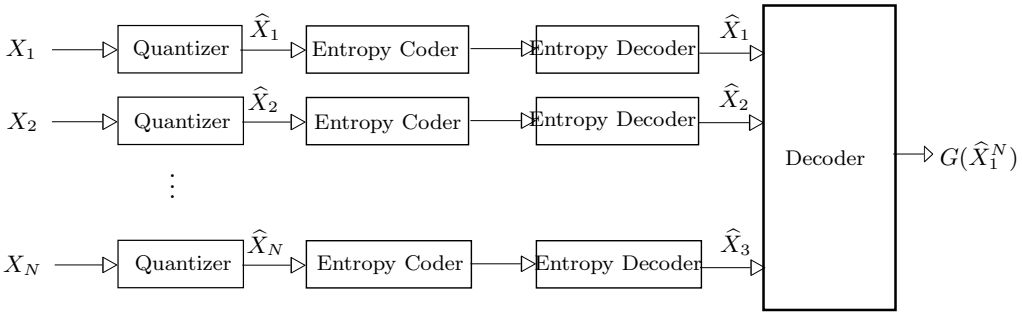

Figure 2-7: Variable Rate Quantization: The entropy coding reduces to a disjoint operation for each source component

[26], this quantity may be related to the quantizer resolution K_i, density λ_i, and source differential entropy $h(X_i)$:

$$R_i = H(\widehat{X}_i) \approx h(X_i) + \mathbf{E}\left[\log_2 \lambda_i(x_i)\right] + \log_2 K_i \tag{2.16}$$

We neglect the small error in approximation 2.16 in the derivations that follow.

Theorem 2.3.3 *The high-resolution distortion, given by Eq. 2.12, is minimized subject to a sum-rate constraint* $\sum_{i=1}^{N} H(\widehat{X}_i) \leq R$ *by choice of point density functions such that*

$$\lambda_i(x) \propto g_i(x)$$

and rate allocations such that the minimum distortion is given by

$$D = \frac{N}{12} 2^{-2R/N + \sum_{i=1}^{N} h(X_i)/N + 2\sum_{i=1}^{N} \boldsymbol{E}[\log_2 g_i(x_i)]/N}.$$

Proof As with the fixed-rate scenario, we first optimize the quantization profiles, and then perform appropriate rate allocation amongst the N quantizers. *Quantizer.*
Assuming lossless encoding, distortion is given by Eq. 2.13:

$$\begin{aligned} D &= \sum_{i=1}^{N} \frac{1}{12} g_i^2(x_i) K_i^{-2} \lambda_i(x_i)^{-2} \\ &= \sum_{i=1}^{N} \frac{1}{12} 2^{-2R_i + 2h(X_i) + 2\mathbf{E}[\log_2 \lambda_i(x_i)]} g_i^2(x_i) \lambda_i(x_i)^{-2} \end{aligned} \tag{2.17}$$

As with fixed-rate, the N-dimensional distortion reduces to N parallel one-dimensional distortions. Each λ_i may be chosen to minimize its corresponding term in the summation. According to Eq. 2.10, this minimizing choice is

$$\lambda_i^2 \propto g_i^2$$

and the corresponding distortion is

$$D = \sum_{i=1}^{N} \frac{1}{12} 2^{-2R_i + 2h(X_i) + 2\mathbf{E}[\log_2 g_i(x_i)]} \tag{2.18}$$

Optimal Rate Allocation.

Recall the sum-rate constraint: $\sum_{i=1}^{N} R_i \leq R$. The R bits may be allocated amongst the N encoders as in Sec. 2.3. The optimal allocation satisfies the arithmetic/geometric mean inequality with equality, resulting in distortion

$$D = \frac{N}{12} 2^{-2R/N + \sum_{i=1}^{N} h(X_i)/N + 2\sum_{i=1}^{N} \mathbf{E}[\log_2 g_i(x_i)]/N} \quad \blacksquare \tag{2.19}$$

2.3.3 N-dimensional Variable Rate with Slepian-Wolf Coding

Our derivations of optimal quantization profiles and distortions, summarized by Eqs. 2.15 and 2.18, have not explicitly considered correlations between the sources. Nonetheless, they are optimal within the constraints that we have placed on them. Can we loosen any of these constraints so that coding may exploit correlations? From Fig. 2-7, one can see that the answer is "yes": even though the decoding takes place at a centralized decoder, the entropy decoders for each of the sources are disjoint.

Replacing the N entropy decoders with a single block decoder, we find ourselves presented with the Slepian-Wolf situation (Fig. 1-1). The lowest achievable sum-rate is given by the joint entropy of the quantized variables, $H(\hat{X}_1^N)$; otherwise, the problem statement is identical to that for disjoint variable-rate coding.

Theorem 2.3.4 *The high-resolution distortion, given by Eq. 2.12, is minimized subject to a sum-*

rate constraint $H(\widehat{X}_1, \ldots, \widehat{X}_N) \leq R$ *by choice of point density functions such that*

$$\lambda_i(x) \propto g_i(x)$$

and rate allocations such that the minimum distortion is given by

$$D = \frac{1}{12} 2^{-2R + 2h(X_1, \ldots, X_N) + \boldsymbol{E}\left[\log_2 g_i^2\right]}.$$

Proof Recall the variable-rate distortion expression, and how it may be converted to a function of the sum-rate:

$$\begin{aligned} D &= \frac{1}{12} \sum_{i=1}^{N} \frac{1}{K_i^2} 2^{\mathbf{E}\left[\log_2 g_i^2\right]} && (2.20) \\ &\geq \frac{1}{12} \frac{1}{\prod K_i^2} 2^{\sum_{i=1}^{N} \mathbf{E}\left[\log_2 g_i^2\right]} && (2.21) \end{aligned}$$

Using a Slepian-Wolf coding scheme, the resolution quantity K_i^2 may be related to sum-rate in a manner similar to Eq. 2.16:

$$H(\widehat{X}_1^N) \approx h(X_1^N) + \sum_{i=1}^{N} \log_2 K_i + \sum_{i=1}^{N} \mathbf{E}\left[\log_2 \lambda_i\right]$$

Inserting this into the minimized distortion expression, we obtain the Slepian-Wolf achievable lower bound to the distortion-rate function.

$$D(R) \geq \frac{1}{12} 2^{-2R + 2h(X_1^N) + \mathbf{E}\left[\log_2 g_i^2(X)\right]} \quad \blacksquare$$

The gain in bits from Eq. 2.19 is, satisfyingly, the *total correlation* of the sources, $\sum_{i=1}^{N} h(X_i) - h(X_1^N)$. For the case of two sources, the total correlation is the mutual information.

Notice in the above equation the analytical separation between correlations and functional considerations. In reality, the random binning introduced by Slepian-Wolf coding allows the quantizer to be nonregular and thereby remove redundancy between sources.

Relationship to Linder et al. [28]

In [28], the authors consider the class of "locally quadratic" distortion measures for variable-rate high-resolution quantization. They define locally quadratic measures as those having the following two properties:

1. Let x be in $\mathbb{R}^N$. For y sufficiently close to x in the Euclidean metric (that is, $d(x,y) < \epsilon$), the distortion between x and y is well approximated by $\sum_{i=1}^{N} M_i(x)|x_i - y_i|^2$, where $M_i(x)$ is a positive scaling factor. In other words, the distortion is a space-varying non-isotropic scaled MSE.

2. The distance between two points is zero if and only if the points are identical. Formally $d(x,y) = 0$ implies that $x = y$.

For these distortion measures, they consider high-resolution variable-rate regular quantization, and both generalize Bucklew's results [23] (to non-functional distortion measures) and demonstrate the use of multidimensional companding functions to implement these quantizers. Of particular interest to us is the comparison they perform between joint and scalar quantization. When Slepian-Wolf coding is employed for the latter, the scenario is similar to the developments of Sec. 2.3.3.

The source of this similarity is the implicit distortion measure we work with: $d_G(x,y) = |G(x) - G(y)|^2$. When x and y are very close to one another, our Taylor approximation technique reduces this distance to a locally quadratic form:

$$|G(x) - G(y)|^2 \approx \sum_{i=1}^{N} \left| \frac{dG(x_1^N)}{dx_i} \right|^2 |x_i - y_i|^2$$

From this, we may obtain the same variable-rate S/W performance as Eq. 2.19 through the results of Linder et al.

However, there are differences both subtle and important between locally quadratic distortion measures and the functional distortion measures we consider. First and foremost: a continuous scalar function of N variables is *guaranteed* to have an *uncountable* number of point tuples (x,y) for which $G(x) = G(y)$ and therefore that $d_G(x,y) = 0$ and $x \neq y$. This loudly violates the second condition of a locally quadratic distortion measure, and the repercussions are felt most strikingly for non-monotonic functions — for whom regular quantizers are no longer necessarily optimal (discussed in Chapter 3).

This second condition is also broken by functions that are not *strictly* monotonic in each variable; one finds that without this strictness, variable-rate analysis of the centralized encoding problem is invalidated. Specifically, if the derivative vector

$$dG(X_1^N) = \left(\frac{dG(x_1^N)}{dx_1}, \ldots, \frac{dG(x_1^N)}{dx_N} \right)$$

has nonzero probability of possessing a zero component, the expected variable-rate distortion as derived by both Bucklew and Linder et al. is $D = 0$, regardless of rate. This answer is obviously nonsensical, and arrives from the null derivative having broken the high-resolution approximation. Given that even the example functions we consider in Sec. 3.1 fall into this trap, the centralized results have limited applicability to functional scenarios. We nonetheless summarize them in App. 2.D.

2.4 Block Quantization and Functional Typicality

In the previous section, we derived the design and performance of optimal variable-rate functional quantizers. Our approach there was grounded in the picture of Fig. 2-7, where an explicit quantizer is followed by an entropy coder. We allowed the entropy coder the flexibility to block code the quantized representation of our source X_i, provided it was done so losslessly.

In this section, we seek to generalize slightly further. Instead of a scalar quantizer for X_i followed by block entropy coding, we allow for block quantization; that is, an i.i.d sequence from the ith source, $(X_i)_1^M$ is vector quantized into the representation $\widehat{(X_i)_1^M}$, before being entropy coded. This is done somewhat in the spirit of [16], where the optimality of an architecture separating VQ from entropy coding was demonstrated.

Our analysis in this section takes on a slightly different flavor from before. Observing Eq. 2.9, we notice that the impact of the function G on the optimal quantizer performance is limited to the sensitivity terms $2^{\mathbf{E}\left[\log_2 g_i^2(X_i)\right]}$. These functional terms bear strong resemblance to their probabilistic counterparts: $2^{-2h(X_i)} = 2^{-\mathbf{E}\left[\log_2 f_X(X_i)^2\right]}$. This suggests that the notion of a distribution's entropy can be adjusted by the presence of a function. To explore this fact, we derive our results by using a modified notion of typicality. The relationship of this approach to the quantization point-density technique of the original derivations is completely analogous to that between typicality and codeword-length optimization in a lossless setting.

2.4.1 Shannon's Typicality

A critical step in Shannon's establishment of information theory was the development of typicality. Sequences of i.i.d samples from a discrete probability mass function are found to split into two camps: the typical and the atypical. The typical sequences dominate the sample space and, more curiously, they all have similar probabilities of occurrence.

Quantitatively, a sequence x_1^M of M samples of a discrete variable X is said to be ϵ-Shannon-typical if

$$\left| \frac{1}{M} \log_2 \mathbf{P}\left(x_1^M\right) - \mathbf{E}\left[\log_2 \mathbf{P}\left(X\right)\right] \right| < \epsilon$$

and we denote the set of such sequences by A_ϵ.

Since $\frac{1}{M} \log_2 \mathbf{P}\left(X_1^M\right)$ converges in probability to $\mathbf{E}\left[\log_2 \mathbf{P}\left(X\right)\right]$ — the entropy $H(X)$ — it can be seen that the probability of A_ϵ can be made arbitrarily close to one. One can therefore bound the probability of a typical sequence above and below as

$$(1-\epsilon)2^{-M(H(X)+\epsilon)} \leq \mathbf{P}\left(x_1^M\right) \leq 2^{-M(H(X)-\epsilon)}$$

Shannon's source coding result now follows quite blatantly: the atypical sequences may be ignored, while the (roughly equiprobable) typical sequences can each be given codewords of length $MH(X) + M\epsilon$. The beauty of this approach is its avoidance of details: by means of typicality, one has simplified the problem to one of coding a uniform distribution.

Note how this contrasts with the more direct approach of variable-length single-sample coding. If one attempts to solve for the optimal codeword lengths for each symbol of a discrete source, he or she will find that rate is minimized when each symbol is assigned a codeword with length equal to the negative logarithm of its probability. The average rate that results is then the entropy of the source — the same as by the typicality argument.

An analogous typicality construction holds for the continuous regime. If a source X is distributed over $[0, 1]$ with some distribution f_X, one may consider a sequence of M samples of X, X_1^M. Virtually all the probability will be contained within a region of volume $2^{Mh(X)}$ (in analogy to cardinality $2^{MH(X)}$) with the probability density arbitrarily close to uniform over this volume.

2.4.2 Functional Typicality: One Dimension

We seek to apply the same logic to the problem of variable rate functional VQ and observe the resulting performance. First we consider the one-dimensional problem, wherein a random variable X distributed over $[0,1]$ according to $f_X(x)$ is to be quantized to compute the function $G(x)$ (which obeys the same constraints as in our previous analysis). Unlike before, we jointly quantize/encode M samples of X: X_1^M.

Judging from its similarity to entropy in the distortion expression of Eq. 2.9, a good guess for the quantity of "functional entropy" is, in terms of the single-dimensional sensitivity profile $g(x) = \left|\frac{dG(x)}{dx}\right|$:

$$k(X,G) = \mathbf{E}\left[\log_2 g(X)\right].$$

We define functional-ϵ-typicality in the same manner as Shannon-ϵ-typicality, and we call a sequence x_1^M *ϵ-completely-typical* if it is both ϵ-functionally-typical and ϵ-Shannon-typical. The set of completely typical sequences is referred to as C_ϵ. It can be shown that the set of functionally-typical-sequences has probability arbitrarily close to 1, since

$$\frac{1}{M}\sum_{j=1}^{M}\log_2 g(x_j) \to k(X,G)$$

From this fact and the analogous statement for Shannon typicality, we may bound the probability that a sequence is ϵ-completely-typical as being greater than or equal to $1-2\epsilon$ for arbitrarily small ϵ. The question: can we somehow restrict our attention to the quantization of the completely typical sequences, and ignore the rest?

Lemma 2.4.1 *The distortion contribution from the atypical sequences can be made arbitrarily small by increasing the blocklength M.*

Proof The total distortion of our encoder can be broken into two terms: one from the typical sequences, and the other from the atypical sequences:

$$D = \mathbf{P}\left(X_1^M \text{ is atypical}\right) D_{\text{atypical}} + \mathbf{P}\left(X_1^M \text{ is typical}\right) D_{\text{typical}},$$

where D_{atypical} and D_{typical} are the conditional distortions. Because G is bounded, the distortion

between two points $d_G(x, y)$ is upper bounded by a value L. Therefore, we may bound the distortion of an atypical sequence:

$$\begin{aligned} d_G(x_1^M, \hat{x}_1^M) &= \frac{1}{M}\sum_{j=1}^{M} d_G(x_i, y_i) && (2.22)\\ &\leq \max_{x,y} d_G(x, y) && (2.23)\\ &= L && (2.24) \end{aligned}$$

We can therefore upper bound the distortion contribution from the atypical sequences as $Lp(\Omega - C_\epsilon) \leq L\epsilon$. Since this can be made arbitrarily small by raising the sequence length M, we are justified in only considering the completely-typical sequences. ∎

The completely-typical set has a roughly uniform probability distribution (consequence of Shannon-typicality), and all points have nearly identical geometrically averaged derivatives $\prod_{j=1}^{M} |g(y_j)|^{1/M} = 2^{k(X,G)}$ (consequence of functional-typicality). What does this say about the optimal quantizer and optimal performance? Note that the typical set is topologically open: it can be shown by standard topological arguments that every typical point x_1^M possesses a neighborhood of typical sequences. Because of this, high-resolution approximations may be applied. The trouble is that we are dealing with vector quantization now — not the scalar variety that the majority of this chapter has been considering. Nonetheless, using techniques borrowed from the scalar case, high-resolution optimization of the vector quantizer design is possible.

We first make an important assumption on the use of arbitrary lattices for vector quantization: any polytope quantization cell is fully determined by its shape-gain $S(M)$ and its characteristic lengths in each of the M dimenions. For both rectangular [23] and ellipsoid [29] cells, the functional distortion may be shown to take the form

$$\frac{1}{M} S(M) \sum_{j=1}^{M} g_j^2 \Delta_j^2 \tag{2.25}$$

where g_j is the function's slope in the jth direction and Δ_j is the jth characteristic length for the cell. Note that the rectangular case forms an inner bound to the distortion, while the ellipsoid case is an outer bound. It is therefore reasonable to restrict our attention to lattices obeying Eq. 2.25.

Lemma 2.4.2 *Under the high-resolution approximation, suppose K identical quantizer cells with*

shape factor $S(M)$ (equal to $1/12$ if the cells are rectangular) are to be placed in a volume $V(y)$ containing a point y_1^M. Then the minimum distortion per-cell is given by

$$D = S(M)V(y)^{2/M}2^{2k(X,G)}.$$

Proof In accordance with the high-resolution approximation, quantizer cells in close vicinity will be identical, as will the sensitivities $g^2(x_j) = \left|\left|\frac{dG(x)}{dx}\right|_{x_j}\right|^2$ in each cell. Denoting the side-lengths of these cells in the jth direction by Δ_j, one may write the expected distortion within each cell as

$$D = \frac{1}{M}S(M)\sum_{j=1}^{M} g(y_j)^2\Delta_j^2$$

Since the volume constraint may be written as $V = K\prod_{j=1}^{M}\Delta_j$, the optimization of the side-lengths Δ_j reduces to an application of the arithmetic/geometric mean inequality. The resulting minimum distortion (when each side-length Δ_j is chosen inversely proportional to its respective sensitivity $g(y_j)$) is

$$D = S(M)\left(\frac{V}{K}\right)^{2/M}\prod_{j=1}^{M} g(y_j)^2/M.$$

Since all points being quantized are functionally-typical, this may be reduced in terms of the functional entropy. We also replace V/K by the volume $V(y)$ of a cell located near point y:

$$D = S(M)V(y)^{2/M}2^{2k(X,G)} \quad \blacksquare$$

This distortion-per-cell is completely independent of the location of the quantization cell, so long as it contains typical points. Under the constraints that the sum of the volumes of all cells must equal the volume of the completely typical sequences, $2^{Mh(X)}$, it can be shown by Lagrange multipliers that the total distortion is minimized if every quantization cell has equal volume, $2^{Mh(X)}/K$. Furthermore, because the completely-typical set has uniform probability distribution, the resolution-per-sample $K^{1/M}$ may be phrased in terms of the rate-per-sample 2^R. The minimal total distortion is then given by:

$$D = S(M)2^{2h(X)-2R+2k(X,G)}.$$

This is the result we have seen before for the single-dimension case, although the 1/12 factor has been replaced by a polytope second moment, $S(M)$. As discussed in chapter 1, the shape gain is a fundamental advantage that VQ holds over scalar quantization. What we see here is that the shape gain is the *only* advantage that VQ has in the one dimensional functional case. Does this same result hold true for the distributed multi-dimensional case?

2.4.3 Functional Typicality: N Dimensions

To analyze the multidimensional scenario with typicality, we make the assumption that the optimal quantizer is regular (that is, all cells are similarly shaped and connected). As with our original derivation of multidimensional functional quantization, we note that this holds true for functions $G(x_1^N)$ that are monotonic in each variable and defer further considerations to the next chapter.

Recalling our approach from the single-dimension scenario, we start by glancing at the variable-rate distortion expression derived in Sec. 2.3.2,

$$D \propto 2^{\sum_{i=1}^{N} \mathbf{E}\left[\log_2 g_i^2(X_i)\right]}$$

In likeness to the single-dimensional $k = \mathbf{E}\left[\log_2 \left|\left|\frac{dG(x)}{dx}\right|_X\right|^2\right]$, we define the N multidimensional functional entropies

$$k_i(X_i, G) = \mathbf{E}\left[\log_2 g_i(X_i)\right]$$

We can consider the joint quantization of a length M sequence at each encoder. Notationally, let $(X_i)_j$ refer to the jth sample in the sequence seen by the ith encoder. We call a sequence $(x_i)_1^M$ completely ϵ-typical if it is both ϵ-Shannon-typical according to the distribution f_{X_i} and functionally-ϵ-typical if

$$\left| \frac{1}{M} \sum_{j=1}^{M} \log_2 g_i((x_i)_j) - k_i \right| < \epsilon.$$

Using analysis identical to the previous section, it can be shown that, asymptotically, only the completely-typical sequences contribute to the distortion. Noting that the set of completely-typical sequences is an open set (as before), we can consider the problem of high-resolution vector quantization within this set. Suppose the quantization cells in the vicinity of a point $y = (y_1^N)_1^M$ have side-length $\Delta_{i,j}$ with respect to the jth sequence sample of the ith source. Then each such cell

will have distortion

$$D = S(M) \sum_{i=1}^{N} \sum_{i=1}^{M} \Delta_{i,j}^2((x_i)_1^M) \left| \frac{dG(x_1^N)}{dx_i}(y_1^N)_j \right|^2$$

A critical detail in this expression is that the ith side-length is only a function of the sequence of M samples $(x_i)_1^M$ seen by ith quantizer. One may not adjust the side-lengths at the quantizer for X_2 based on the value of X_1. The consequence of this limitation appears when we take the expectation of this distortion over all quantization cells: as with the analysis from Sec. 2.3 we are presented with N separate one-dimensional distortion expressions, with the single-dimensional sensitivity $g^2(x)$ replaced by the multidimensional sensitivity $g_i^2(x)$.

$$D = \sum_{i=1}^{N} S(M) \sum_{j=1}^{M} \mathbf{E}\left[g_i(X_{i,j})^2 \Delta_{i,j}^2((X_i)_1^M) \right]$$

Each of these N terms may be optimized by techniques completely analagous to those of Sec. 2.4.2. The resulting total distortion, in terms of the rate-per-sample R, is then given by

$$D = NS(M)\exp\left[-2R + \frac{2}{N}\left(\sum_{i=1}^{N} h(X_i) \right) + \frac{2}{N}\left(\sum_{i=1}^{N} k_i(X_i, G) \right) \right].$$

As before, we find that the only improvement yielded by vector quantization is contained is the shape-gain term $S(M)$. If Slepian-Wolf coding is employed, the improvement in is still given by the total correlation, and we have minimum distortion

$$D = NS(M)\exp\left[-2R + \frac{2}{N}\left(h(X_1, \ldots, X_N) \right) + \frac{2}{N}\left(\sum_{i=1}^{N} k_i(X_i, G) \right) \right].$$

2.5 Notions of Optimality: How close are we to the best possible structure?

We have taken the route of increasing generality in our construction of the distributed source coding problem. First, we worked within the constraint of fixed-rate quantization, wherein each codeword is of length R_i bits. Next, lossless disjoint encoders and decoders were added for each source, and we considered the variable-rate performance of the system. Then, by allowing the entropy decoders to fuse into a single block decoder, Slepian-Wolf coding was made possible.

Slepian-Wolf coding achieves the lowest possible sum-rate $\sum_{i=1}^{N} R_i$ such that the quantized values $\widehat{X}_1^N$ are recreated at the decoder. In the spirit of functional compression we may question the constraint of perfectly recreating the quantized sources at the decoder — after all, we are interested not in $\widehat{X}_1^N$ but in $G(\widehat{X}_1^N)$. Preservation of $\widehat{X}_1^N$ is sufficient to preserve $G(\widehat{X}_1^N)$, but it may not be necessary. That is, some rate gain may be possible from seeking to represent $G(\widehat{X}_1^N)$ instead of $\widehat{X}_1^N$ itself.

The scenario we have created is identical to the discrete functional compression problem considered by Doshi et al. [25]. They demonstrate that the lowest sum-rate for representing $G(\widehat{X}_1^N)$ can be achieved by communicating the colors of a "characteristic graph" [24] for each source. When we designed the optimal variable-rate quantizer, we assumed that the entropy coding will be lossless. The use of discrete functional compression compromises this assumption and therefore our quantizer's optimality. It can be shown, however, that functional quantization followed by functional compression nearly always reduces to our Slepian-Wolf scenario (the arguments are very similar to those regarding "equivalence-free" functions in the ensuing chapter).

Even so, it is not even obvious whether the optimal approach to source coding can be found in the separation architecture we take as a starting point (quantization followed by discrete coding of some sort). We leave this question more or less unanswered.

2.A Optimal Choice of Estimator

Let X_1^N be a random vector with distribution on $[0,1]^N$, and let G be the function of interest. The goal of functional quantization is to choose (1) disjoint quantizers $(Q_{X_i})_1^N$ for each component of the source vector and (2) an estimator $\widehat{G}$ so as to minimize the distortion $\mathbf{E}\left[|G(X_1^N) - \widehat{G}(Q_1^N(X_1^N))|^2\right]$.

Suppose that a combination of quantizers and estimator, $(Q_{X_i})_1^N$ and $\widehat{G}_0$, is optimal. **We show that this performance can be matched by estimator $\widehat{G} = G$ and appropriately chosen quantizers $\widetilde{Q}_1^N$.** Specifically, $\widetilde{Q}_1^N$ will be chosen to have the same quantization intervals as Q_1^N, but (potentially) different reconstruction points.

Let $I_x = \prod_{i=1}^{N} \left(Q_{X_i}^{-1}(Q_{X_i}(x))\right)$ be the quantization cell containing the point $x \in \mathbb{R}^N$, and let $\widehat{G}'$ be an optimal estimator that is in use alongside quantizers $(Q_{X_i})_1^N$. Since I_x is regular (connected) and G is continuous, $G(I_x)$ is an interval (connected) in $\mathbb{R}$. The operation $\widehat{G}'(Q(x))$ amounts to quantizing this interval; for it to be optimal, $\widehat{G}'(Q(x)) \in G(I_x)$. Because G is continuous, we may

pick $\widetilde{Q}(I_x)$ to be a point $\hat{I_x}$ such that $G(\hat{I_x}) = \widehat{G}_1(Q(x))$.

Compared with the quantizer Q_1^N and estimator $\widehat{G}'$, $\widetilde{Q}_1^N$ and estimator $\widehat{G} = G$ have identical intervals and generate identical estimates for each interval. Therefore, there is no loss associated with limiting $\widehat{G} = G$.

2.B Equivalence of 1D Quantization Schemes

A variable X with distribution $f_X(x)$ supported on $[0, 1]$ is to be quantized at rate R. We wish to design the quantizer $Q_X(x)$ so as to minimize the distortion $\mathbf{E}[d_G] = \mathbf{E}\left[|G(X) - \widehat{G}(Q_X(X))|^2\right]$. As suggested in Sec. 2.2, a feasible strategy in this one-dimensional scenario is to calculate $G(X)$ and optimally quantize G itself. What choice of quantizer $Q_X(X)$ and estimator $\widehat{G}$ can emulate this procedure?

Estimator $\widehat{G}$:

Let $Q_G(G)$ be the optimal (regular) quantizer of G that we wish to implement via $\widehat{G}(Q_X(X))$. The most obvious choice for the estimator is $\widehat{G} = G$. Since G is monotonic and continuous, we may construct the associated quantizer $Q_X(X)$ as $Q_X(x) = G^{-1}(Q_G(G(x)))$.

The composition of G and Q_X generates the desired quantization of G. Suppose x_0 is the value to be quantized. Then $\widehat{G}(Q_X(x_0)) = G(G^{-1}(Q_G(G(x_0)))) = Q_G(G(x_0))$ — and we may constrain $\widehat{G} = G$ without problem.

Fixed-Rate (Codebook-Constrained) Quantization:

The optimal quantizer can be described by a point density function over the range of G:

$$\lambda_G(g) \propto f_G(g)^{1/3} \propto \left(f_X(G^{-1}(g))\frac{dG^{-1}(g)}{dg}\right)^{1/3} \tag{2.26}$$

where G's monotonicity justifies the second relation. λ_G induces a quantizer density on X

$$\lambda_X(x) \propto \lambda_G(g(x))\frac{dG(x)}{dx} \propto \left(f_X(x)\left(\frac{dG(x)}{dx}\right)^2\right)^{1/3} \tag{2.27}$$

The resulting distortion is that of an optimal quantizer over G. Assuming for clarity that the

range of G is $[0,1]$,

$$\begin{aligned}
\mathbf{E}[D] &= \frac{1}{12}2^{-2R}\|f_G(g)\|_{1/3} \\
&= \frac{1}{12}2^{-2R}\left\|\frac{f_X(G^{-1}(g))}{dG/dx(G^{-1}(g))}\right\|_{1/3} \\
&= \frac{1}{12}2^{-2R}\left[\int\left(\frac{f_X(x)}{dG/dx(x)}\right)^{1/3}\frac{dG}{dx}dx\right]^3 \\
&= \frac{1}{12}2^{-2R}\left\|f_X(x)\left|\frac{dG}{dx}(x)\right|^2\right\|_{1/3}
\end{aligned} \tag{2.28}$$

Eqs. 2.27 and 2.28 are identical to Eqs. 2.3 and 2.3, respectively.

Variable-Rate (Entropy-Constrained) Quantization:

A uniform quantizer over the range of G is high-rate optimal when entropy coding is employed. This induces a non-uniform quantizer on X with interval spacing

$$\Delta_X(x) \propto \Delta_G(G(x))\left|\frac{dG}{dx}\right| \tag{2.29}$$

As with the fixed-rate scenario, this quantizer will achieve the optimal distortion over G:

$$\begin{aligned}
\mathbf{E}[D] &= \frac{1}{12}2^{-2R+2h(G)} \\
&= \frac{1}{12}2^{-2R+2h(X)+2\mathbf{E}[\log_2 dG/dX]}
\end{aligned} \tag{2.30}$$

where we have made use of "derived entropy" from a coordinate transformation, as described in [30].

Once again, notice how this solution matches that of Sec. 2.2.

2.C Derivation of High-Resolution Functional Distortion

X_1^N is a random vector with distribution $f_X(x_1^N)$ supported on $[0,1]^N$, and G is a bounded function differentiable almost everywhere. Suppose each component of X_1^N, X_i, is quantized at resolution K_i using normalized quantization profile $\lambda_i(x_i)$. We obtain the high-rate functional distortion, defined as $\mathbf{E}\left[|G(X)-G(\widehat{X})|^2\right]$, in terms of λ_i, f_X, and K_i.

We start by looking at a single quantization cell, $S \subset [0,1]^N$, with centroid y_1^N. Because each

component is quantized independently, S is a rectangular region of length $\Delta_i = K^{-1}\lambda_i(y_i)^{-1}$ on the ith side. The high-rate assumptions **A1**-**A3** tell us that within S the distribution $f_X(x_1^N)$ is roughly uniform and the function G is approximately affine. Quantitatively, we replace G with its Taylor approximation around y_1^N:

$$G(x_1^N) \approx G(y_1^N) + \sum_{i=1}^{N} \frac{\partial G(x_1^N)}{\partial x_i}\bigg|_{x_i=y_i} (y_i - x_i). \tag{2.31}$$

Since $\mathbf{E}\left[G(X)|X \in S\right] = G(y_1^N)$, the midpoint y_1^N may be chosen as the reconstruction point $\widehat{X}_S$. The conditional distortion is then given by the variance of G within S:

$$\begin{aligned}
\operatorname{var}(G(X) \mid X \in S) &= \mathbf{E}\left[\left(\sum_{i=1}^{N} \frac{\partial G}{\partial X_i}\bigg|_{\widehat{X}_S} (y_i - X_i)\right)^2 \mid X_1^N \in S\right] \\
&= \sum_{i=1}^{N} \left(\frac{\partial G}{\partial X_i}\bigg|_{\widehat{X}_S}\right)^2 \mathbf{E}\left[(y_i - x_i)^2 \mid X_1^N \in S\right] \\
&= \sum_{i=1}^{N} \left(\frac{\partial G}{\partial X_i}\bigg|_{\widehat{X}_S}\right)^2 \frac{\Delta_i^2}{12} \\
&= \sum_{i=1}^{N} \left(\frac{\partial G}{\partial X_i}\bigg|_{\widehat{X}_S}\right)^2 \frac{1}{12K_i^2\lambda_i^2(\widehat{X}_S)}
\end{aligned} \tag{2.32}$$

The expectation of Eq. 2.32 across all quantizer cells S is the total distortion.

$$\begin{aligned}
\mathbf{E}\left[d_G\right] &= \sum_S \mathbf{P}\left(S\right) \sum_{i=1}^{N} \left(\frac{\partial G}{\partial X_i}\bigg|_{\widehat{X}_S}\right)^2 \frac{1}{12K_i^2\lambda_i^2(\widehat{X}_S)} \\
&= \sum_S f_X(\widehat{X}_S) \left(\prod_{i=1}^{N} \Delta_i(S)\right) \sum_{i=1}^{N} \left(\frac{\partial G}{\partial X_i}\bigg|_{\widehat{X}_S}\right)^2 \frac{1}{12K_i^2\lambda_i^2(\widehat{X}_S)} \\
&= \sum_S \left(\int_{x_1^N \in S} f_X(x_1^N) dx_1^N\right) \sum_{i=1}^{N} \left(\frac{\partial G}{\partial X_i}\bigg|_{\widehat{X}_S}\right)^2 \frac{1}{12K_i^2\lambda_i^2(\widehat{X}_S)}
\end{aligned} \tag{2.33}$$

The slope of G and the point density λ_i are both roughly constant throughout quantization cell S. Hence, we may absorb the slope and density terms into the integral:

$$\mathbf{E}\left[d_G\right] \approx \sum_S \int_{x_1^N \in S} f_X(x_1^N) \sum_{i=1}^{N} \left(\frac{dG}{dX_i}\bigg|_{x_1^N}\right)^2 \frac{1}{12K_i^2\lambda_i^2(x)} dx_1^N \tag{2.34}$$

Since the cells S cover the support of $f_X(x)$, we may remove the summation $\sum_S$ by expanding the domain of integration.

$$\begin{aligned}\mathbf{E}\left[d_G\right] &\approx \sum_{i=1}^{N} \frac{1}{12K_i^2} \int_{x_1^N} \left(\left. \frac{dG}{dX_i} \right|_{x_1^N} \right)^2 \lambda_i^{-2}(x_i) dx_1^N \\ &= \sum_{i=1}^{N} \frac{1}{12K_i^2} \mathbf{E}\left[\left(\left. \frac{dG}{dX_i} \right|_{x_1^N} \right)^2 \lambda_i^{-2}(x_i) \right] \end{aligned} \tag{2.35}$$

For convenience, we define $g_i^2(x_i)$ as the expected squared partial with respect to X_i:

$$g_i^2(x_i) = \mathbf{E}\left[\left(\frac{dG}{dX_i} \mid_{X_1^N} \right)^2 \mid X_i = x_i \right]$$

This completes the decoupling of the distortion expression into N terms for the N source components.

$$\mathbf{E}\left[d_G\right] \approx \frac{1}{12} \sum_{i=1}^{N} \mathbf{E}\left[g_i^2(x_i) K_i^{-2} \lambda_i^{-2}(x_i) \right] \tag{2.36}$$

2.D Comparison with Centralized Coding

The Slepian-Wolf theorem claims that the lossless performance of centralized encoding may be matched by distributed encoding. Does this statement hold true for functional quantization as well?

If the function, G, is precisely known to the encoders, then in the centralized case $G(X_1^N)$ may be directly quantized. All bits may be put towards the quantization of G, leading to a 2^{-2NR} rate dependence — in stark contrast to the distributed scenario's 2^{-2R}.

However, we may also compare to the scenario described by Bucklew [23], where the encoder is only aware of a distribution of possible function G_j over some index set $j \in J$. The centralized performance in this case is given by

$$D_c = L 2^{-2R + h(X_1^N) + \sum_{i=1}^{N} \mathbf{E}\left[\log_2 |dG(x_1^N)/dx_i|^2\right]}$$

where L is a constant polytope shape-gain factor. The ratio in distortion is found to be

$$D/D_c = 12L 2^{-2\sum_{i=1}^{N} \mathbf{E}[\log_2 |dG/dx_i| - \mathbf{E}[\log_2 |dG/dx_i||X_i]]} \tag{2.37}$$

We observe several points about this ratio:

1. D is greater than or equal to D_c, since M is less than or equal to $1/12$ and $\mathbf{E}\left[\log_2 |dG/dx_i|\right]$ is greater than or equal to $\mathbf{E}\left[\log_2 |dG/dx_i| \mid X_i\right]$, by concavity $\cap$ of the log function.

2. When G is almost everywhere a linear function of the N source variables, D/D_c is the polytope shape gain from vector quantization.

3. When G is an M-dimensional vector-valued function, $|dG/dx_i|^2$ is everywhere replaced by $\sum_{j=1}^{M} |dG/dx_i|^2$, and D/D_c is the polytope shape gain.

Finally, note that the quantitative result of Eq. 2.37 has been obtained in the more general context of quadratic distortion measures by Linder et al. [28].

Chapter 3

Junior Year: Scaling, Non-regular Quantization, and Non-monotonic Functions

3.1 Scaling Analysis

The use of high-resolution quantization gives us the power to derive analytical results in otherwise intractable situations. This comes in very handy when we wish to analyze the behavior of functional quantization systems as the number of sources grows arbitrarily large — one may simply leave N as an unspecified parameter.

In this section, we observe how certain classes of functions behave under distributed functional quantization. Sec. 3.1.1 considers quantization for several example functions. Striking scaling with the number of sources, N, is observed and explained, before Sec. 3.1.2 generalizes this behavior to a class of functions we call *selective* and *symmetric*.

3.1.1 The Maximum, the Median, and the Midrange

Assume the source, X_1^N, is i.i.d. uniform over $[0,1]^N$. We are interested in quantizing so as to accurately represent a function of the source, $G(X_1^N)$. Both the fixed and variable rate versions of

this problem are considered for different choices of G.

The Maximum

To illustrate the application of functional quantization, we consider a simple example: a user is interested in the maximum of N samples of X, given by $G(X_1^N) = \max(\{X_1, X_2, \ldots, X_N\})$.

For either fixed-rate or entropy-constrained quantization, the first key computation is to determine the quantity $g_i^2(x)$, the expected squared partial with respect to source variable X_i. From the symmetry of G, we may assert that g_i^2 is independent of i. For notational convenience, consider $i = 1$.

The partial derivative of $G(x_1, x_2, \ldots, x_N)$ with respect to x_1 is 1 where $x_1 \geq \max(\{x_2, x_3, \ldots, x_N\})$ and 0 otherwise. The expectation that defines $g_1^2(x)$ is thus the expectation of an indicator function of the event $X_1 \geq \max(\{X_2, X_3, \ldots, X_N\})$, so

$$\begin{aligned} g_1^2(x) &= \mathbf{P}\left(X_1 \geq \max(\{X_2, X_3, \ldots, X_N\}) \mid X_1 = x\right) && (3.1) \\ &= \prod_{i=2}^{N} \mathbf{P}\left(X_1 \geq X_i \mid X_1 = x\right) && (3.2) \\ &= x^{N-1}. && (3.3) \end{aligned}$$

We may now obtain the optimal quantization performance.

Optimal Fixed-Rate Quantization. According to Eq. 2.3.1, $\lambda_i(x) \propto x^{(N-1)/3}$ for optimal fixed-rate quantization. Upon correct normalization of this density (to integrate to 1), we have the exact relation

$$\lambda_i(x) = \tfrac{1}{3}(N+2)x^{(N-1)/3}.$$

See Fig. 3-1 for $\lambda_i(x)$ at different dimensionalities. Note how it reflects our intuition: a greater density of points is assigned to values more likely to affect the max. To obtain the distortion that

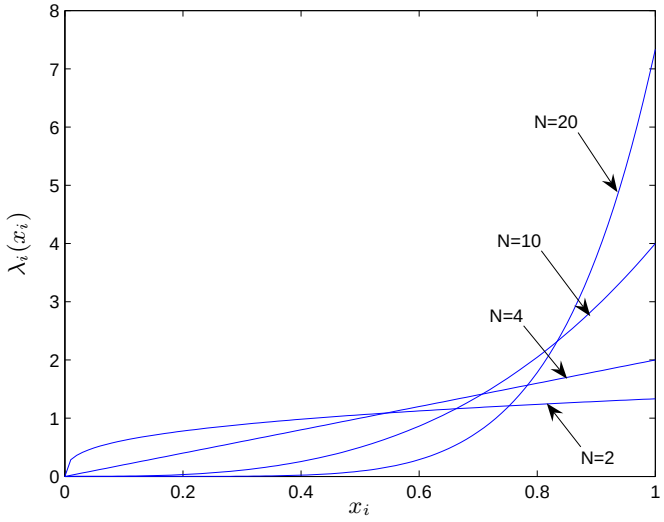

Figure 3-1: Optimal fixed-rate max quantizers for several values of N (number of sources)

results from this choice of density, we turn to Eq. 2.15. Given $g_i^2(x)$, D evaluates to

$$D = \frac{N}{12}2^{-2R/N}\left(\int_0^1 x^{(N-1)/3}\right)^3 dx \tag{3.4}$$

$$= \frac{N}{12}2^{-2R/N}\left(\frac{3}{N+2}\,x^{(N+2)/3}\Big|_0^1\right)^3 \tag{3.5}$$

$$= \frac{1}{12}2^{-2R/N}\frac{27N}{(N+2)^3} \tag{3.6}$$

$$\tag{3.7}$$

For $N = 1$, where $G(X) = X$, observe that the optimal density reduces to $\lambda = 1$ and that the resulting distortion is $\frac{1}{12}2^{-2R}$, as expected. Thus, we can see explicitly that ordinary quantization is a special case of the functional formulation.

Optimal Variable-Rate Quantization For the variable rate case, the optimal quantizer point density is proportional to $g_i = x^{(N-1)/2}$. We normalize λ_i to 1, as in the fixed-rate scenario:

$$\lambda_i(x) = \frac{N+1}{2}x^{(N-1)/2}. \tag{3.8}$$

Fig. 3-2 depicts this density for different values of the dimensionality. At a qualitative level, λ_i is not

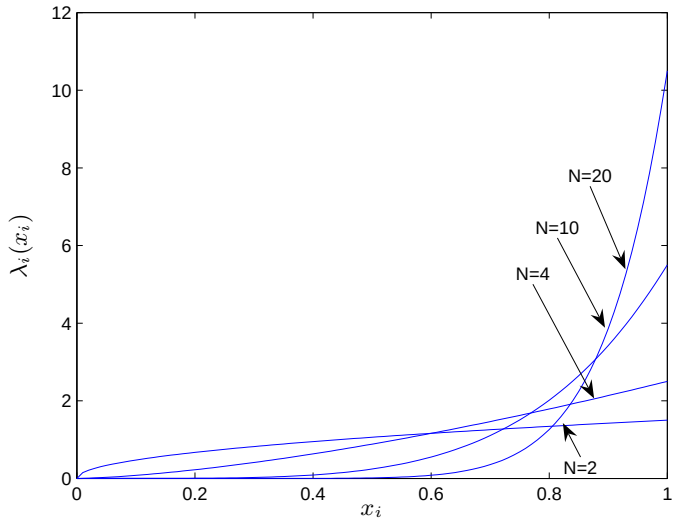

Figure 3-2: Optimal max variable-rate quantizers for several values of N.

strikingly different between the fixed and variable rate scenarios; they both demonstrate the same emphasis on larger x, and both become increasingly concentrated with increasing dimensionality N.

The distortion, obtained by Eq. 2.19, draws a much more noticeable line between fixed and variable-rate constraints (see Fig. 3-3 for an illustration of this).

$$
\begin{aligned}
D &= \frac{N}{12} 2^{-2R/N + \mathbf{E}\left[\log_2 g_i^2\right]} && (3.9)\\
&= \frac{N}{12} 2^{-2R/N + (N-1)\mathbf{E}[\log_2 x]} && (3.10)\\
&= \frac{N}{12} 2^{-2R/N - (N-1)\log_2 e} && (3.11)\\
&= \frac{N}{12} 2^{-2R/N} e^{-N+1} && (3.12)
\end{aligned}
$$

Once again, observe that both the distortion and density reduce to the appropriate form when $N = 1$.

Comparison with Ordinary Quantization We saw before (Sec. 2.2) that ordinary quantization can be far from optimal in terms of functional MSE. This can be demonstrated quantitatively by comparing functional and ordinary quantizers for the max function. How much better do functional quantizers perform?

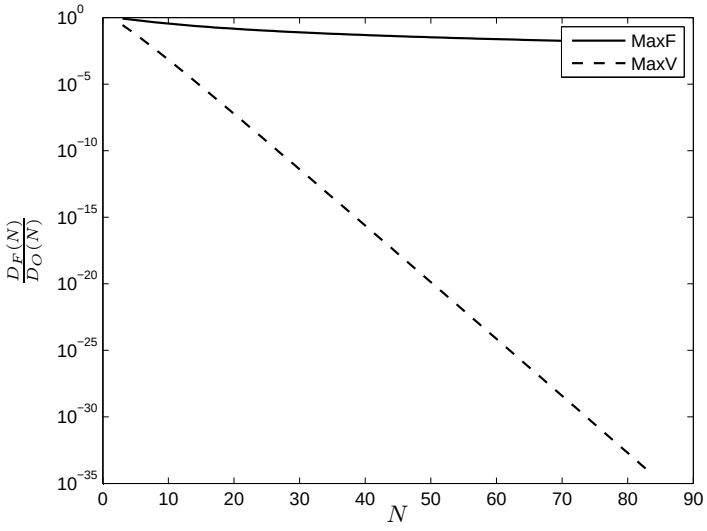

Figure 3-3: The ratio of functional to ordinary distortion for the max, as a function of dimensionality (log scale). Note that while the fixed-rate quantizer has a $1/N^2$ falloff, the distortion in the variable-rate case improves exponentially.

Ordinary quantization of a uniform $[0,1]^N$ source would result in a uniform quantizer. In the language of the high-rate approximation, $\lambda_i = 1$ for any component X_i. Eq. 2.12 then tells us the distortion is

$$D = \frac{1}{12}\sum_{i=1}^{N}\frac{1}{K_i^2}\mathbf{E}\left[g_i^2(X_i)\right] \tag{3.13}$$

By the symmetry of both the source distribution and the function with respect to the source variables, the rate must split evenly between the dimensions: $\log_2 K_i = R/N$. The distortion may then be written using the gradient operator, ∇G:

$$D = \frac{1}{12}2^{-R/N}\mathbf{E}\left[\left|\nabla G\right|^2\right] \tag{3.14}$$

To find the $\mathbf{E}\left[\left|\nabla G\right|^2\right]$ quantity for the max, we recognize the following: at almost any point $x_1^N \in [0,1]^N$ the gradient vector is 1 in the largest of its components and 0 in each of the others. Therefore, $\left|\nabla G\right|^2$ is 1 with probability 1, and the distortion is

$$D = \frac{1}{12}2^{-R/N} \tag{3.15}$$

If the number of bits per source variable, R/N, is held constant, the ordinary distortion is independent of the sample size. This lies in stark contrast to the fixed-rate functional quantizer, whose distortion falls with the square of the dimension. Even more striking is the variable-rate quantizer, whose distortion falls exponentially with the dimension (Fig. 3-3).

But where does this latter improvement come from?

Source of Improvement The max function — as an order statistic — is *selective*, in that it selects one of its inputs to be the output. The distortion of the ordinary quantizer, therefore, does not depend on the dimensionality; each component is quantized in the same manner.

However, the variance of the max itself falls with the square of the dimensionality. The fixed-rate quantizer exploits this by crowding points towards the region of interest. For a fixed resolution, increasing N will widen interval sizes for smaller values of X_i and shorten them for larger X_i. The result is an increasingly skewed probability mass function for the quantized symbols $\widehat{X}_i$. The fixed rate quantizer's output entropy, for instance, is seen to fall linearly with N:

$$\begin{aligned} H(\widehat{X}_i) &\approx h(X_i) + \log_2 K_i + 2\mathbf{E}\left[\log_2 \lambda_i\right] && (3.16)\\ &= h(X_i) + \log_2 K_i + 2\mathbf{E}\left[\log_2 N + 2 - \log_2 3 + \log_2 x^{(N-1)/3}\right] && (3.17)\\ &= h(X_i) + \log_2\left(K_i \frac{(N+2)^2}{3}\right) - \frac{N-1}{3}\log_2 e && (3.18) \end{aligned}$$

Variable rate quantization takes advantage of this fact via entropy coding. The jump from $1/N^2$ with the fixed-rate quantizer to e^{-N} with the variable-rate has little to do with the slightly different choice of λ_i, and everything to do with being able to increase the resolution K via entropy coding.

The Median

We now consider a decoder interested in computing the median of N i.i.d. uniform samples of a source, X_1^N. For simplicity, restrict N to be odd valued with $N = 2M+1$; the median is then defined as the $(M+1)$th order statistic. We first determine the performance of an ordinary (non-functional) quantizer.

Ordinary Quantization. Let x_1^N be a point in the N-dimensional space, and let x_i denote its ith coordinate. Since the median is an order statistic, $G(x_1^N) = x_i$ for some i. Therefore,

$dG/dx_j = \delta_{ij}$ within some neighborhood of x_1^N, and $|\nabla G|^2 = 1$. From this and Eq. 3.14, we have the distortion of the ordinary quantizer:

$$D_{GO}(R) = \frac{1}{12} 2^{-2R} \tag{3.19}$$

The absence of a dependence on N fits with our intuition: the median simply takes one of the source values, so the dimensionality should not affect the quantizer's accuracy.

Fixed-Rate Functional Quantization. To determine the fixed-rate functional quantizer's distortion, we must first obtain $g_i^2(x_i)$. Given a point $x_i \in [0,1]$, x_i is either itself the median of x_1^N or (differentially or locally) it has no bearing on the value of the median. Therefore, $g_i^2(x) = \mathbf{P}\left(G(X_1^N) = x \mid X_i = x\right)$. This probability may be evaluated combinatorially in terms of $M = \frac{N-1}{2}$, and the cumulative distribution function of X, $F_X(x)$. For x to be the median, M of the other sources must be greater than x, with the remaining M less than x. There are $\begin{pmatrix} 2M \\ M \end{pmatrix}$ possible selections of which sources are above or below. The probability of each of these choices is $F_X(x)^M (1 - F_X(x))^M$; the first term is the probability that M i.i.d samples will fall below x, and the second term is the probability that M will exceed x. Summing the probability of all choices leading to $G(X_1^N) = x$, we have

$$g_i^2(x) = \begin{pmatrix} 2M \\ M \end{pmatrix} F_X(x)^M (1 - F_X(x))^M.$$

For a uniform $[0,1]$ distribution, $F_X(x) = x$. This yields an optimal point density $\lambda_i \propto x^{M/3}(1-x)^{M/3}$ and total distortion

$$D_{GF}(K, M) = \frac{N}{12K^2} \left\| \begin{pmatrix} 2M \\ M \end{pmatrix} x^M (1-x)^M \right\|_{1/3} \tag{3.20}$$

The point density reflects our intuition that more quantizer intervals should be assigned to the more important middle ground — a fact that becomes increasingly true as the dimensionality is increased. Observe the increasingly concentrated point density in Fig. 3-4.

We will now demonstrate that the distortion expression of Eq. 3.20 falls with $1/N$. First, we use integration by parts and Stirling's approximation to obtain the integral $\int_0^1 x^K (1-x)^K dx$. It can be

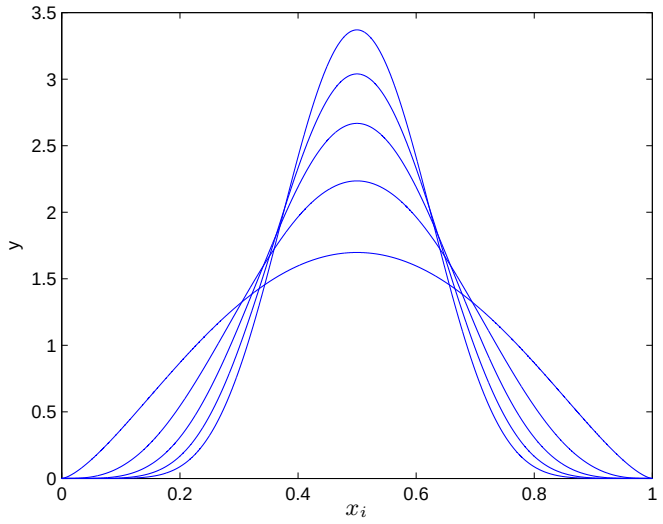

Figure 3-4: Optimal fixed-rate median quantizers for $N = 10$, 20, 30, 40, and 50 sources. Note how the quantizers become increasingly concentrated with N.

shown that iterated integration by parts reduces this integral to the form

$$\int_0^1 x^K (1-x)^K dx = \sum_{i=1}^{K+1} \frac{K!}{(K+i)!} x^{K+i} (1-x)^{K+1-i} \Bigg]_0^1 \tag{3.21}$$

$$= \frac{K!^2}{(2K+1)!} \tag{3.22}$$

Factorials are messy, so we convert them to exponentials via Stirling's formula: $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\lambda_n}$, where the error term $\frac{1}{12n+1} < \lambda_n < \frac{1}{12n}$ decays to zero. Applying this to our previous equation, we can obtain the $\mathcal{L}^{1/3}$ norm referred to in the distortion equation:

$$\int_0^1 x^K (1-x)^K dx \approx \frac{\left(\sqrt{2\pi K}\frac{K}{e}\right)^2}{\sqrt{2\pi(2K+1)}\left(\frac{2K+1}{e}\right)^{2K+1}} \tag{3.23}$$

$$= \frac{\sqrt{\pi}e}{2\sqrt{K}}2^{-2K} \tag{3.24}$$

$$\left\| \begin{pmatrix} 2M \\ M \end{pmatrix} x^M (1-x)^M \right\|_{1/3} \approx \begin{pmatrix} 2M \\ M \end{pmatrix} \left[\left(\sqrt{3\pi}\frac{e}{2}\right)\frac{2^{-2M/3}}{M^{1/2}}\right]^3 \tag{3.25}$$

$$= \begin{pmatrix} 2M \\ M \end{pmatrix} \left(\sqrt{3\pi}\frac{e}{2}\right)^3 \frac{2^{-2M}}{M^{3/2}} \tag{3.26}$$

We may apply Stirling once more to the combination term:

$$D = \frac{N}{12}2^{-2R}\begin{pmatrix} 2M \\ M \end{pmatrix}\left(\sqrt{3\pi}\frac{e}{2}\right)^3 \frac{2^{-2M}}{M^{3/2}} \tag{3.27}$$

$$= \frac{2M+1}{12}2^{-2R}\frac{2M!}{M!^2}\left(\sqrt{3\pi}\frac{e}{2}\right)^3 \frac{2^{-2M}}{M^{3/2}} \tag{3.28}$$

$$\approx \frac{2M}{12}2^{-2R}\frac{\sqrt{4\pi M}(2M/e)^{2M}}{2\pi M(M/e)^{2M}}\left(\sqrt{3\pi}\frac{e}{2}\right)^3 \frac{2^{-2M}}{M^{3/2}} \tag{3.29}$$

$$= \frac{2M}{12}2^{-2R}\frac{1}{M^2}\left(\sqrt{3\pi}\frac{e}{2}\right)^3 \tag{3.30}$$

$$\propto \frac{1}{M} \tag{3.31}$$

In contrast, since the median simply takes on one of the source values, the ordinary distortion $D_{GO}(K)$ does not even depend on M (Eq. 3.19). The ratio between these — which is independent of K — is depicted in Fig. 3-5.

Notice that the variance of the median — and with it, the distortion of a fixed-rate centralized encoder — also falls with $1/M$. In this sense, fixed-rate distributed quantization is order-optimal in M (for a system that involves no coding). This order optimality does not carry over to K, however: for centralized encoding all $N \log_2 K$ bits available can be directed towards quantizing the source that is the median. Distortion can therefore fall at $2^{-2NR} = K^{-2N}$, as opposed to the distributed $2^{-2R} = K^{-2}$.

Variable-Rate Functional Quantization. For the variable-rate scenario, the optimal quan-

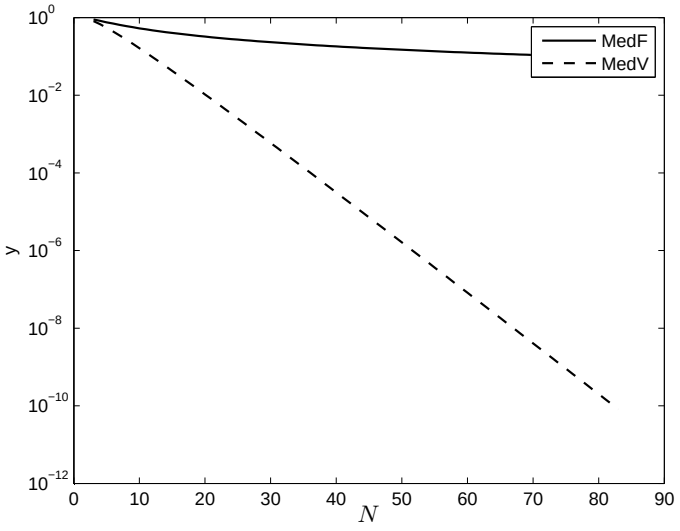

Figure 3-5: Ratio of distortion between functional and ordinary quantizers for the median. The top line is the fixed-rate performance, and the bottom is variable-rate.

tization point density is proportional to the sensitivity profile, $\lambda_i \propto g_i(x_i)$, and the distortion is given by $D = \frac{N}{12} 2^{-2R+2h(X)+\sum_{i=1}^{N} \mathbf{E}\left[\log_2 g_i^2(X_i)\right]}$. As with the fixed-rate scenario, the point density becomes increasingly peaked towards the middle of the range. How does the distortion compare? We first note, for convenience, that the minimum distortion may be rewritten in the natural base as $D = \frac{N}{12} e^{-2R+2h(X)+\sum_{i=1}^{N} \mathbf{E}\left[\ln g_i^2(X_i)\right]}$

For the uniform source distribution, we know that

$$g_i^2(x) = \left(\begin{array}{c} 2M \\ M \end{array} \right) x^M (1-x)^M$$

We can then calculate the following

$$\begin{aligned} \mathbf{E}\left[\ln g_i^2(X_i)\right] &= \int_0^1 \ln x^M dx + \int_0^1 \ln(1-x)^M dx + \ln \left(\begin{array}{c} 2M \\ M \end{array} \right) \\ &= 2M \int_0^1 \ln x dx + \sum_{i=1}^{2} M \ln i - 2 \sum_{i=1}^{M} \ln i \\ &= -2M + \sum_{i=1}^{2M} \ln i - 2 \sum_{i=1}^{M} \ln i \end{aligned}$$

The summations can be approximated by integrals; as it happens, we are making use of an intermediate step in Stirling's approximation.

$$\begin{aligned}\mathbf{E}\left[\ln g_i^2(X_i)\right] &\approx -2M + \int_0^{2M} \ln\tau d\tau - 2\int_0^{M} \ln\tau d\tau \\ &= -2M + 2M\ln 2M - 2M\ln M \\ &= -2M + 2M\ln 2\end{aligned}$$

Plugging into the distortion expression, we have that

$$\begin{aligned}D &\sim Ne^{-2M+2M\ln 2} \\ &= N\left(\frac{e}{2}\right)^{-2M}\end{aligned}$$

As with the maximum function, we see exponential reduction in distortion with N (illustrated in Fig. 3-5). One may attribute this to similar factors: while the point density did not change appreciably between fixed- and variable-rate, the use of entropy coding permitted a much higher resolution.

We also note that this geometric falloff is not restricted to situations where $f_X(x)$ is the uniform distribution. Consider the more general symmetric distribution case ($f_X(x) = f_X(1-x)$). Distortion is seen to instead take the form

$$D \sim N\exp_e\left\{2M\int_0^1 f_X(\tau)\ln F_X(\tau)d\tau + 2M\ln 2\right\}$$

which also involves a geometric falloff.

The Midrange

We now consider a scenario identical to the above, but with the maximum function replaced by the midrange (the source still being uniform i.i.d.). The midrange is defined as the average of the minimum and the maximum components of x_1^N. No parity restriction on N is necessary to obtain clear results. We start by obtaining the ordinary quantizer's distortion function, $D_{GO}(R, N)$.

For an arbitrary point $x_1^N \in [0,1]^N$, we have $G(x_1^N) = 1/2(x_{\min} + x_{\max})$, where $x_{\min}$ and

$x_{\max}$ are the minimal and maximal coordinates of x_1^N. Therefore $dG/dX_i = \frac{1}{2}\delta_{i\,\min} + \frac{1}{2}\delta_{i\,\max}$ and $|\nabla G|^2_{x_1^N} = 1/2$. Since this holds almost everywhere, $\mathbf{E}\left[|\nabla G|^2_{x_1^N}\right] = 1/2$ and the ordinary quantizer's distortion is similar to that for the median:

$$D_{GO}(R,N) = \frac{1}{24}2^{-2R} \tag{3.32}$$

To compute the analogous quantity for the functionally optimized quantizer, we first turn our attention towards the characteristic quantity $g_i^2(x)$. If x_i is not the minimum or the maximum, dG/dX_i is zero; otherwise dG/dX_i is 1/2. The latter situation occurs with probability $\mathbf{P}\left(\max(X_1^N = x) \mid X_i = x\right) + \mathbf{P}\left(\min(X_1^n = x) \mid X_i = x\right)$, since the minimal and maximal events are disjoint almost everywhere. Therefore,

$$\begin{aligned} g_i^2(x) &= \tfrac{1}{4}\left(\mathbf{P}\left(\max(X_1^n = x) \mid X_i = x\right) + \mathbf{P}\left(\min(X_1^n = x) \mid X_i = x\right)\right) \\ &= \tfrac{1}{4}(F_X(x)^{N-1} + (1 - F_X(x))^{N-1}). \end{aligned} \tag{3.33}$$

The term $F_X(x)^{N-1}$ represents the probability that all sources but X_i fall below x; likewise, $(1 - F_X(x))^{N-1}$ is the probability that x is the maximal element. The fixed-rate distortion from this expression is

$$D_{GF}(K,N) = \frac{N}{12K^2}\left\|\frac{1}{4}(F_X(x)^{N-1} + (1 - F_X(x))^{N-1})\right\|_{1/3} \tag{3.34}$$

For the uniform $[0,1]$ source, $F_X(x) = x$. For large values of N, $g_i^2(x)$ is dominated by $(1-F_X(x))^{N-1}$ when $x < 1/2$, and by $F_X(x)^{N-1}$ when x exceeds 1/2. We may then approximate the integral as

$$\begin{aligned}
D_{GF}(K,N) &\approx \frac{N}{12K^2}\frac{1}{4}\left(\int_{1/2}^{1} F_X(x)^{(N-1)/3}dx + \int_{0}^{1/2}(1-F_X(x))^{(N-1)/3}dx\right)^3 \\
&= \frac{N}{48K^2}\left(\int_{1/2}^{1} x^{(N-1)/3}dx + \int_{0}^{1/2}(1-x)^{(N-1)/3}dx\right)^3 \\
&= \frac{N}{24K^2}\left(\int_{1/2}^{1} x^{(N-1)/3}dx\right)^3 \\
&= \frac{N}{6K^2}\left(\frac{3}{N+2}(1-(1/2)^{(N+2)/3})\right)^3 \\
&\approx \frac{N}{6K^2}\left(\frac{3}{N+2}\right)^3 \\
&= 9N2^{-2R}\frac{1}{2(N+2)^3}
\end{aligned}$$

For large values of N, this follows an approximate $1/N^2$ falloff, in contrast to the constant distortion of the ordinary quantizer; see Fig. 3-7 for the exact behavior. When we examine the optimal point densities (Fig. 3-6) we find them to be increasingly concentrated towards the edges as the dimensionality increase. Since the midrange is calculated from the min and the max, this is expected.

The variance of the midrange is known to fall with $1/N^2$ [31] [32]. As with the median, this implies that a fixed-rate centralized encoder's dependence on N carries into the distributed scenario.

We may also compute the distortion-rate behavior for the variable-rate scenario.

$$\begin{aligned}
D &= \frac{N}{12}2^{-2R}e^{\mathbf{E}\left[\ln g_1^2(x_1)\right]} && (3.35)\\
&= \frac{N}{12}2^{-2R}e^{2\int_0^{1/2}\ln x^{N-1}dx} && (3.36)\\
&= \frac{N}{12}2^{-2R}e^{2(N-1)\int 0^{1/2}\ln x dx} && (3.37)\\
&= \frac{N}{12}2^{-2R}e^{(N-1)(\ln 1/2-1)}dx && (3.38)\\
&= 2^{-2R}\frac{N}{12}\left(\frac{e}{2}\right)^{1-N} && (3.39)
\end{aligned}$$

Note, once again, the differing rates of convergence. Additionally, observe in Fig. 3-6 that the optimizing point densities are not significantly different between fixed and variable rate; as before

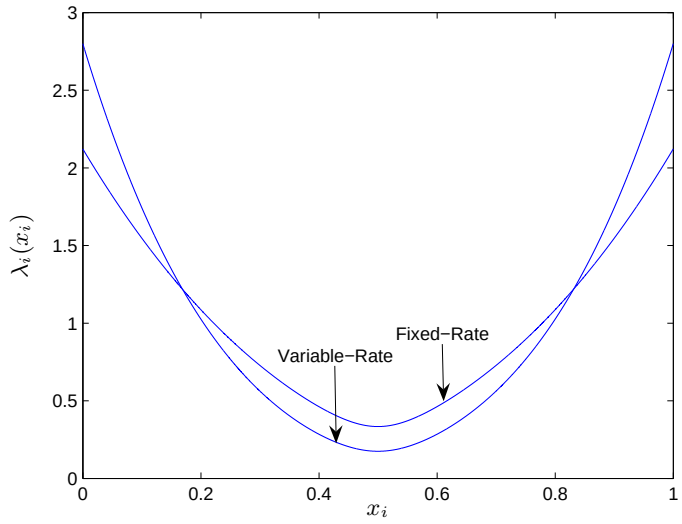

Figure 3-6: Optimal fixed and variable rate midrange quantizers for 10 disjoint sources

the improvement (drawn in Fig. 3-7) may be attributed to the entropy coding block.

3.1.2 Selective/Symmetric Functions

In the above analysis, we came to notice two curious trends in the performance of functional quantization with the median, midrange, and maximum:

1. The functional distortion of a fixed-rate quantizer fell at the same rate as the variance of the function. The median has variance $\sim \frac{1}{N}$ and fixed-rate distortion $\sim \frac{1}{N}2^{-2R}$, while the midrange/max have variance $\sim \frac{1}{N^2}$ and fixed-rate distortion $\sim \frac{1}{N^2}2^{-2R}$.

2. The functional distortion of a variable-rate quantizer fell exponentially with N for all the cases considered.

At a higher level, we observe that as the functions become increasingly deterministic, the error of the functional quantizers falls correspondingly. Can this relationship be generalized? Perhaps if we consider the set of all order-statistics, similar behavior will play out. Indeed, this turns out to be the case: for any finite linear combination of order statistics (central or extremal) for uniform i.i.d. sources, both observed trends continue to hold.

As it happens, this class of functions can be generalized even further. Let us refer to a function $G(x_1^N)$ as being *symmetric* if $G(x_1^N) = G(y_1^N)$ whenever the vector y_1^N is a permutation of x_1^N.

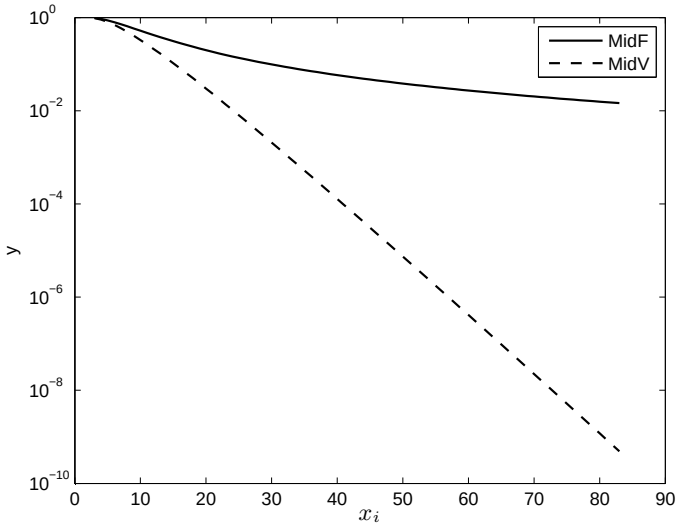

Figure 3-7: The distortion reduction from functional quantization for the midrange. Top curve is fixed-rate and bottom is variable-rate.

And let us refer to $G(x_1^N)$ as being *selective* if $G(x_1^N) = x_{I(x_1^N)}$; that is, G "selects" one of the components of x_1^N. Clearly, any order statistic is a selective function and any function of order statistics is symmetric.

We may then derive the sensitivity $g_i^2(x_i)$ for selective and symmetric G. For any point x_1^N, the partial derivatives are given by

$$\frac{dG(x_1^N)}{x_i} = \left\{ \begin{array}{ll} 1 & \text{if } G(x_1^N) = x_i \\ 0 & \text{otherwise} \end{array} \right\}$$

This then allows us to state that $g_i^2(x) = \mathbf{P}\left(G(X_1^N) = x \mid X_i = x\right)$. Combining the symmetry of G with Bayes' rule, we observe that:

$$\begin{aligned} \mathbf{P}\left(G(X_1^N) = x \mid X_i = x\right) &= \frac{\mathbf{P}\left(X_i = x \mid G(X_1^N) = x\right) f_G(x)}{f_X(x)} && (3.40) \\ &= \frac{1}{N}\frac{f_G(x)}{f_X(x)} && (3.41) \end{aligned}$$

Since we are restricting attention to the case of a uniform source distribution, this reduces to

$$g_i^2(x) = \frac{1}{N} f_G(x) \tag{3.42}$$

(note that $f_G(x)$ depends on N). For fixed-rate functional quantization, this yields a distortion

$$D \propto \|f_G\|_{1/3} 2^{-2R}$$

At first glance, this is both encouraging and disappointing: while we have a very simple dependence on f_G, it is not the variance that carries through but instead the $\mathcal{L}^{1/3}$ norm. At second glance, we notice some similarities between the $\mathcal{L}^{1/3}$ norm and the variance of a unimodal (single "peaked") distribution.

1. Scaling $Y = 2X$: $\|f_Y\|_{1/3} = \|2f(x/2)\|_{1/3} = 4\|f(x)\|_{1/3}$, just as $\mathrm{var}[2f(x/2)] = 4\,\mathrm{var}[f(x)]$.
2. Shifting $Y = X + \alpha$: $\|f_Y\|_{1/3} = \|f(x-\alpha)\|_{1/3} = \|f(x)\|_{1/3}$, just as $\mathrm{var}[f(x-\alpha)] = \mathrm{var}[f(x)]$.
3. Example: Uniform distribution of width Δ: $\|f(x)\|_{1/3} \propto \Delta^2$, just as $\mathrm{var}[f(x)] \propto \Delta^2$.
4. Example: Gaussian with standard deviation σ has 1/3-norm proportional to σ^2, just as $\mathrm{var}[f(x)] = \sigma^2$. Demonstration:

$$\begin{aligned}
\|f(x)\|_{1/3} &= \left[\int \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}\right)^{1/3} dx\right]^3 \\
&= \left[\int \frac{1}{\sigma^{1/3}(2\pi)^{1/6}} e^{-x^2/6\sigma^2} dx\right]^3 \\
&= \left[\sigma^{2/3}(2\pi)^{5/6}\sqrt{3} \int \frac{1}{\sqrt{3}\sigma\sqrt{2\pi}} e^{-x^2/6\sigma^2} dx\right]^3 \\
&= \sigma^2 (2\pi)^{5/2} \sqrt{3}^3 \\
&\propto \sigma^2
\end{aligned}$$

Note, however, that the $\mathcal{L}^{1/3}$ norm is significantly different from the variance in that it is invariant to permutation of the coordinate system x. A function like the one at the top of Fig. 3-8 might result in the same norm as the one at the bottom, although the two have vastly different variances. However, if we constrain our attention to the unimodal distributions, such as those of the order

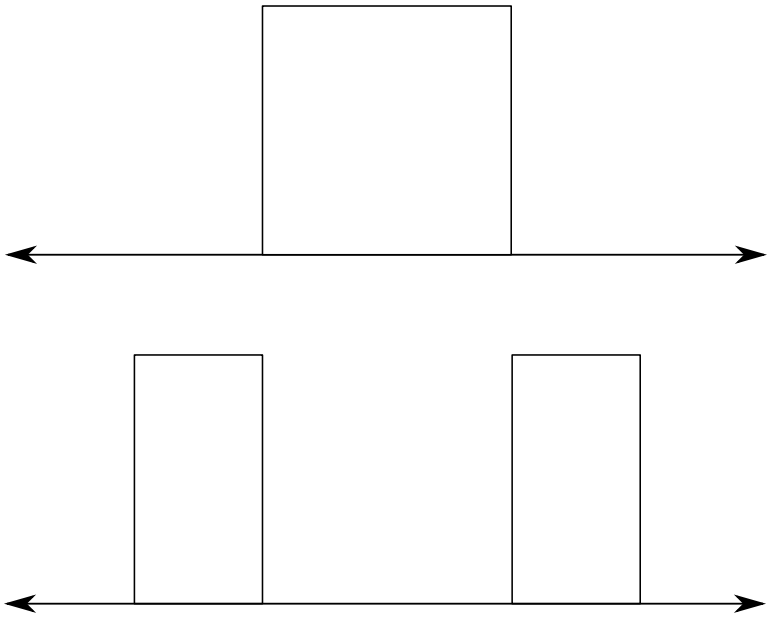

Figure 3-8: Both distributions have identical $\mathcal{L}^{1/3}$ norms, but the top distribution has smaller variance.

statistics, it is not unreasonable that the fixed-rate functional distortion — given by the $\mathcal{L}^{1/3}$ norm of the distribution — scales with N in the same manner as the variance.

The exponential falloff of the variable rate results are in some ways even more fascinating than the proportional-to-variance behavior of fixed-rate. Using Eq. 3.42, we can write the variable-rate distortion in terms of the KL-divergence between f_G and f_X:

$$
\begin{aligned}
D &= \frac{N}{12} 2^{-2R} 2^{2h(X)+\mathbf{E}\left[\log_2 g_i^2\right]} && (3.43)\\
&= \frac{N}{12} 2^{-2R} 2^{2h(X)+\mathbf{E}\left[\log_2 \frac{1}{N}+\log_2 f_G(x)\right]} && (3.44)\\
&\propto N 2^{-D(f_G \| f_X)} && (3.45)
\end{aligned}
$$

For distribution f_X that is uniform over $[0,1]$, this reduces to $D \propto N 2^{D(f_G \| f_U)}$, where f_U is the uniform distribution. In other words, the more sharply nonuniform f_G becomes, the more negative the distortion exponent grows. For the case of the order statistics, this exponent falls linearly; it is from this that we observe the exponential falloff with N.

3.2 Generalizing from the Monotonic Functions

In Chapter 2, our major results relied on a strict constraint on the functions considered: $G(X_1^N)$ must be monotonic in each of its variables. Our motivation in enforcing this requirement was to guarantee the optimality of the quantizers we designed. Since the high-resolution description only applies within the space of regular quantizers, our optimization procedure produces the optimal *regular* quantizer. When G is monotonic, the optimal quantizer can be trivially shown to be a regular; it is the restriction of G that ensures optimality amongst all quantizer designs.

To illustrate why this isn't necessarily true for non-monotonic functions, we introduce a simple one-dimensional example. Suppose X is uniformly distributed over $[0, 1]$ and that G is a simple non-monotonic function: $G(x) = \frac{1}{2} - \mid x - \frac{1}{2} \mid$. If we blindly apply the regular functional quantization machinery seen in the previous chapter to G, we will find that G is fully characterized by its squared derivative, $g^2(x) = 1$. This leads to a uniform quantizer over $[0, 1]$, for both fixed and variable rate scenarios (top of Fig. 3-9). But if the goal is to recreate G, we may save an entire bit of communication by quantizing $\mid X - \frac{1}{2} \mid$. This can be interpreted as the introduction of non-regular quantization intervals: a single codeword corresponds to the union of an interval to the left of $x = 1/2$ and one to the right (Fig. 3-9).

For a single dimension, any lack of monotonicity leads to a non-regular optimal quantizer. While this seems to affirm our earlier restriction to the monotonic functions, it turns out that a similar result does not hold for higher dimensions. Fig. 3-10 demonstrates that a function can be non-monotonic in each of its variables and still be optimally quantized by a regular quantizer. For the version of the function that is aligned with the axes (top), there is no loss from grouping the two edges of the kink in x_1 together, since the function cannot distinguish between them. When it is rotated, however, any such grouping in x_1 will introduce errors. For sufficiently high rate, these errors will outweigh the extra bits saved from the grouping.

So if monotonicity is too strict of a requirement, what is more appropriate?

In this section, we consider a much broader class of functions — those that are "equivalence-free," and demonstrate that regular quantization is asymptotically optimal for them. To do this, we first construct a model for non-regular quantization that allows for high-rate analysis. Next, this model is applied to functions that satisfy our definition for equivalence-freedom, and it is demonstrated to yield regular quantizers in the high-rate limit. Finally, we use the model to construct optimal

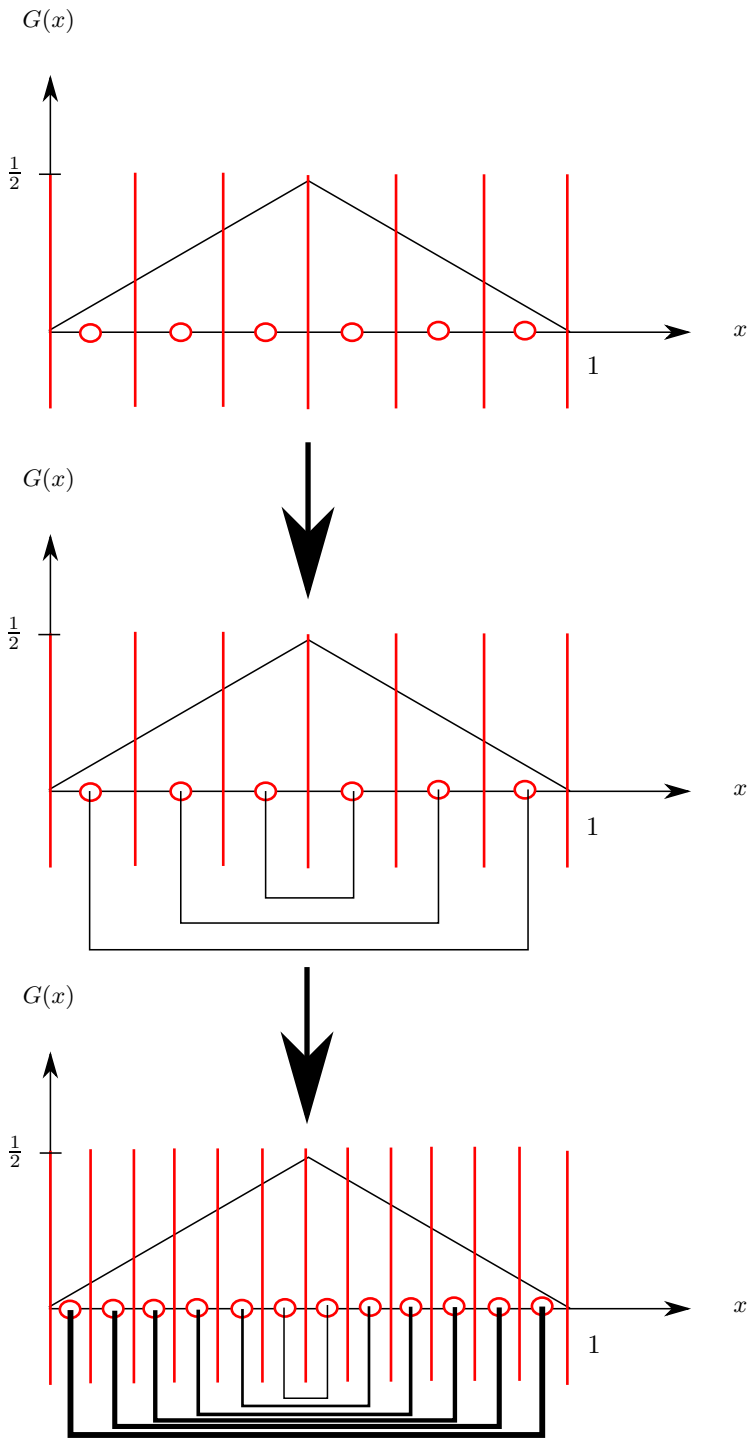

Figure 3-9: If the function G is not monotonic, non-regular quantization may be optimal. Note how the form of the binning does not change as the resolution is increased — this is a strong hint that a resolution-independent non-regular description is possible.

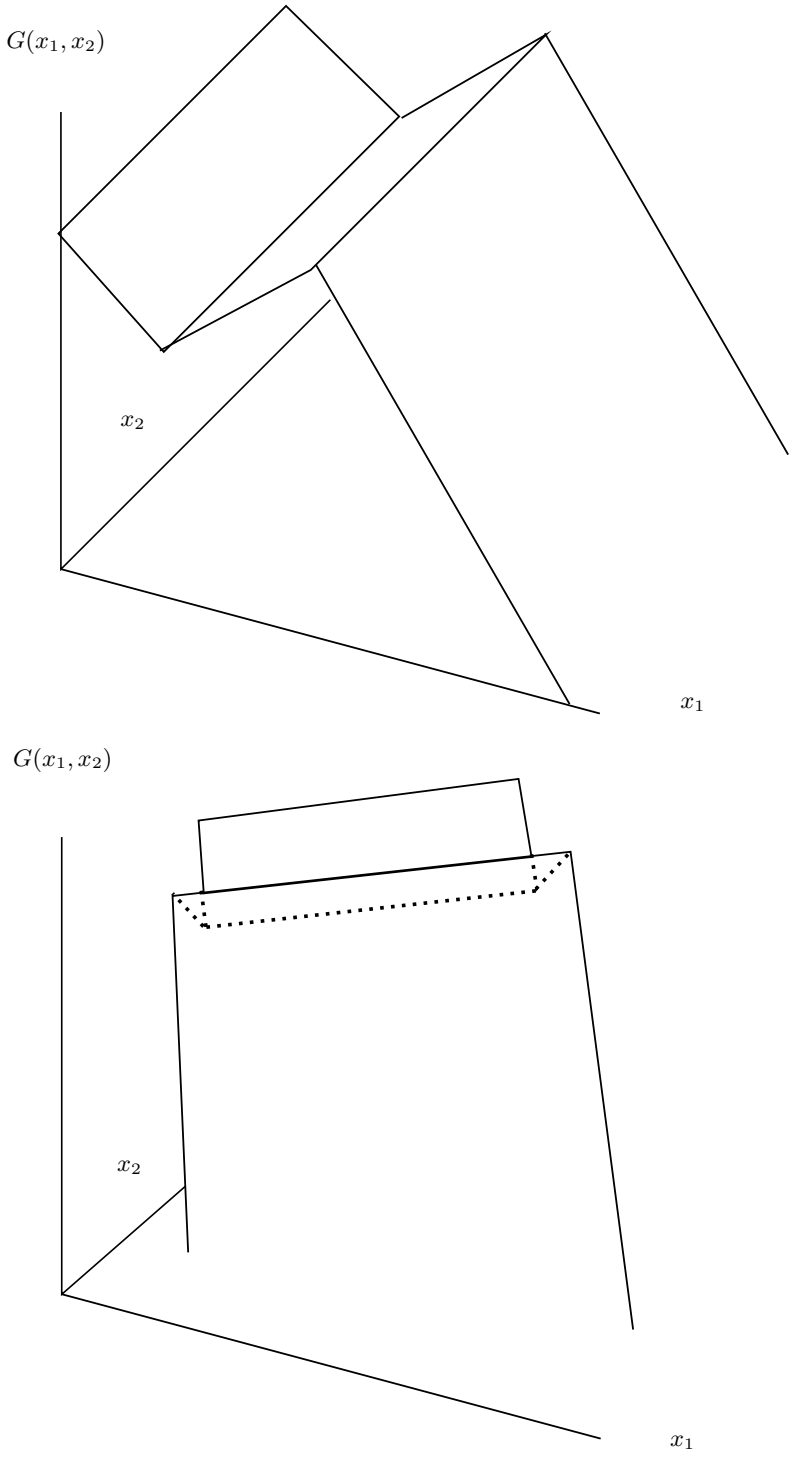

Figure 3-10: A function G of two variables is shown in both graphs. The top G (separable) is best quantized by a non-regular quantizer, while for the bottom (a rotated version of the top G) a regular quantizer is asymptotically optimal. This is due to the bottom function being "equivalence-free."

nonregular quantizers for functions that are not equivalence-free.

3.2.1 High-Rate Non-Regular Quantization

At the heart of our model is a rather simple observation: non-regular quantization cells can be seen as the union of regular quantization cells. Consider the previous example of non-regular quantization for a nonmonotonic function. Suppose the rate constraint was for a 1-bit quantizer. Then one of the quantizer cells would be, for instance, the region $[0, \frac{1}{4}] \cup [\frac{3}{4}, 1]$, which is the union of two regular cells. How can this union-of-intervals picture be incorporated into a description of a non-regular quantizer?

To approach this problem, we again turn to behavior demonstrated in the example. As the resolution is increased from 1 to 2 bits, the aforementioned cell splits into two. Each of its regular subcells is halved by the increased resolution, but the linkage between cells to the left of $x = 1/2$ and those to the right remains unaffected (Fig. 3-9).

This suggests that non-regular quantization can be seen as a two-step process. First, regular quantization is performed on the input data, producing a discrete variable $\widetilde{X}$. After this, a "binning" process is performed from $\widetilde{X}$ to the non-regularly quantized $\widehat{X}$. Each value of $\widehat{X}$ may correspond to multiple values of $\widetilde{X}$ and, therefore, to a union of regular intervals in the domain of X. Unfortunately, relying on a discrete-to-discrete mapping for a high-resolution description is at odds with the continuous-approximation nature of high-resolution analysis. In other words, we would prefer that non-regular quantizer description and design remain in the continuous realm.

Searching for inspiration, we turn to the model of compander-based quantization. This concept, closely linked to single-dimensional functional quantization, involves the application of an invertible continuous function w to X, before performing uniform quantization on $w(X)$. This process is reversed for decoding. The end result is a nonuniform quantizer implemented through appropriate selection of a companding function w. For every point density λ, there is a companding function $w(X)$ that brings about the same high-resolution quantizer as λ upon being uniformly quantized and inverted at the decoder.

Traditional companding techniques can be adapted to implement non-regular quantizers. Normally, w is restricted to be monotonic, continuous, and possess bounded derivative. We discard these conditions and replace them with the following definition.

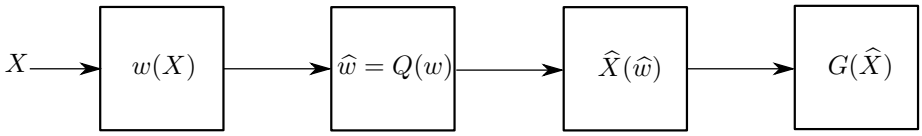

Figure 3-11: Construction for non-regular quantization. A generalized companding function $w(X)$ is applied to data prior to quantization.

Definition A function w is a *generalized compander* if it is piecewise monotonic with a finite number of pieces, continuous, and possesses bounded derivative over each piece.

The restriction of a finite number of pieces is a limitation on the types of non-regular quantizers that can be captured with this model: those for which every non-regular cell is a finite union of regular cells. It is not clear to us whether this is a necessary restriction, or if proofs of our results can be generalized to include it.

With w in place, we have the quantization structure shown in Fig. 3-11. The compander w can be seen as not only sizing the quantization intervals, as an ordinary compander would, but also binning them together to provide for non-regularity. Unlike a traditional binning function, w acts over a continuous domain of source values into a continuous range of bins. To illustrate how w may represent non-regularity, let us return to the example from before. There are many choices of w that can implement the left-right binning that we seek; we choose the most obvious for this particular case, $w(x) = 2G(x)$. One may observe that this results in points to the left of $1/2$ being grouped with points to the right. Upon performing uniform quantization of $w(X)$, the appropriate non-regular quantizer can be obtained, as demonstrated in Fig. 3-12.

The generalized compander w captures the limiting non-regularity of a quantization scheme. As the resolution is raised, the binning of the discrete values more closely resembles the continuous binning represented by w. At the same time, w is significantly more than just a binning map: its slope represents the relative size of the bins. For instance, if our function of interest over $[0, 1]$ was instead

$$G(x) = \begin{cases} \frac{4x}{3} & \text{if } x < \frac{3}{4} \\ 4 - 4x & \text{if } x \geq \frac{3}{4} \end{cases}$$

the subcells that compose each quantization cell would no longer be equally sized, as displayed in Fig. 3-13.

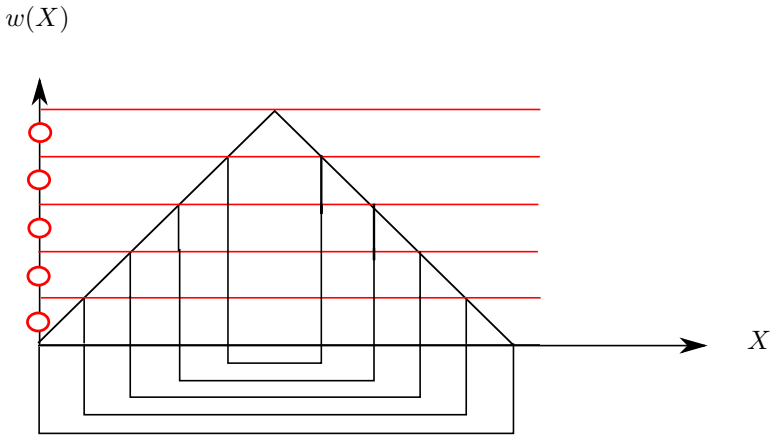

Figure 3-12: Example of non-regular quantization through a generalized companding function $w(X)$. Observe how the rate may be changed without affecting the fundamental binning structure, enforced by $w(X)$.

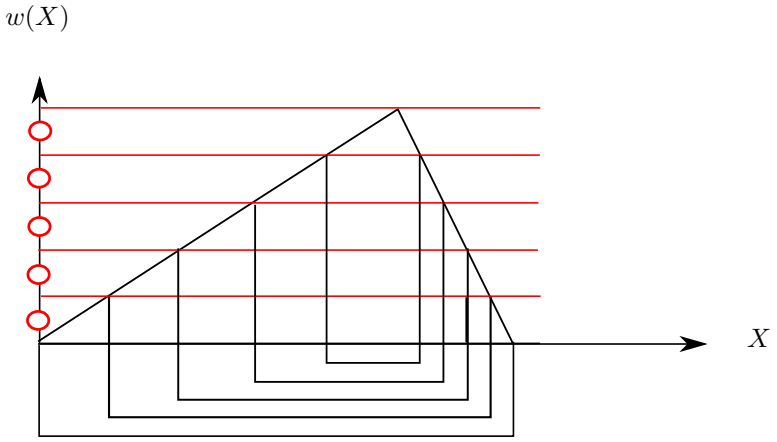

Figure 3-13: Example for a non-uniform sloped companding function $w(x)$. Notice how the relative sizes of quantization subcells are dictated by the relative slope of $w(x)$.

3.2.2 Equivalence-Free Functions

We now define a (very broad) class of functions for which regular quantization is optimal at sufficiently high resolutions. In the next subsection, the binning/companding function $w(x)$ will be used to accommodate functions outside of this class.

To start with, consider the separate quantization of the N components of $X_1^N \in [0,1]^N$ so as to minimize the mean squared error of a function $G(X_1^N)$ (i.e., the standard N-dimensional functional quantization problem). We will focus on the design of one of the quantizers — that of X_1, for example.

Let $A \subset [0,1]$ be some subset of the range of X_1. We then make the following definitions:

1. **Definition** The *conflict probability* of A is $p(A) = \mathbf{P}\left(\text{var}(G(X_1^N) \mid X_1 \in A, X_2^N) > 0\right)$, the probability that a user observing G can notice a difference between the elements of A.

2. **Definition** If $p(A)$ is zero, we say the elements of A are *functionally equivalent*, and that G possesses an *equivalence*. The former is symbolically represented as $a_i \equiv a_j$ for $a_i, a_j \in A$, and can easily be seen to justify its christening as an equivalence relation.

Our result is that non-regular quantization is asymptotically suboptimal for equivalence-free functions. Specifically, non-regular quantization will be found to introduce a nonzero lower bound on the distortion, independent of rate. We show this formally through the use of the w construction of the previous subsection.

Let $G : [0,1]^N \to [0,1]$ be equivalence-free and smooth (bounded gradient), and let X_1^N be distributed over $[0,1]^N$ according to $f_{X_1^N}(x_1^N)$. Now suppose that companding functions w_1^N are applied to X_1^N to generate the binned values Y_1^N. Y_1^N are then each quantized by quantizers $(Q_{Y_i})_1^N$ described by point density functions λ_1^N and resolutions K_i. Quantization of Y_1^N induces a quantization of X_1^N, denoted by $(Q_{X_i})_1^N$. The total distortion of the non-regular quantizers $(Q_{Yi})_1^N$ is

$$D_{TOT} = \mathbf{E}\left[\text{var}[G(X_1^N) \mid (Q_{Yi})_1^N]\right]$$

Definition We define *regularity indicator* function $I(x_i)$ in the following manner. The point $x_i \in [0,1]$ is contained within a quantization cell $Q_{X_i}^{-1}(x_i)$ that can be broken into a finite union of regular intervals $\cup_{j=1}^{M} S_j$. $I(x_i)$ returns the index, j, of the interval that x_i belongs to.

Note that $I(x_i)$ is the information gap between a regular and non-regular quantizer. We exploit this fact to reduce the dimensionality of our problem.

Lemma 3.2.1 *Let $(\widetilde{Q}_{X_i})_1^N$ be the N quantizers of X_1^N that are each potentially non-regular. Let $\widetilde{D}_{TOT}$ be the resulting distortion. $\widetilde{D}_{TOT}$ is lower bounded by the distortion of quantizers $(Q_{X_i})_1^N$ for which $(Q_{X_i})_2^N$ are regular quantizers and $\widetilde{Q}_{X_1} = Q_{X_1}$.*

Proof Suppose the regularity indicators $I(x_i)$ for components $i \in \{2, \ldots, N\}$ are communicated to the decoder in addition to the original quantization symbols $(\widetilde{Q}_{X_i}(X_i))_1^N$. If the ith encoder quantizes $x_{i>1}$ into nonregular cell $S = \cup_{j=1}^M S_j$, the indicator $I(x_i)$ tells the decoder that $x_i \in S_{I(x_i)}$. Since the subcells S_j are regular, and since this result holds for arbitrary $x_{i>1}$, the quantizers $(Q_{X_i}(X_i))_2^N$ are all regular. We bound the distortion $\widetilde{D}_{TOT}$ from the original quantizers with the distortion D from the new ones by means of the law of total variance:

$$\begin{aligned}
\widetilde{D}_{TOT} &= \mathbf{E}\left[\mathrm{var}[G(X_1^N) \mid (\widetilde{Q}_{X_i})_1^N, (I(X_i))_2^N]\right] + \mathbf{E}\left[\mathrm{var}[\mathbf{E}\left[G(X_1^N) \mid (\widetilde{Q}_{X_i})_1^N, (I(X_i))_2^N\right] \mid (\widetilde{Q}_{X_i})_1^N]\right] \\
&\geq \mathbf{E}\left[\mathrm{var}[G(X_1^N) \mid (\widetilde{Q}_{X_i})_1^N, (I(X_i))_1^N]\right] \qquad (3.46) \\
&= D \qquad (3.47)
\end{aligned}$$

We have lower bounded the total distortion of a quantizer that is potentially nonregular in each of its N dimensions by the distortion of one that is nonregular in only the first dimension. ∎

By this lemma, we may lower bound total distortion $\widetilde{D}$ by the distortion D assuming each of the quantizers $(Q_{X_i})_2^N$ is regular. No such assumption is made about Q_{X_1}.

Theorem 3.2.2 *Let G be equivalence free. Then, if $w^{-1}(w(X))$ has cardinality greater one with nonzero probability, the total distortion possesses a positive, rate-independent lower bound for rate R exceeding a constant R_0.*

Proof Consider a point $y \in w([0,1])$. Since w is uniformly quantized, y is contained within a quantization interval $Q_{Y_1}^{-1}(Q_{Y_1}(y))$, which gives rise to a potentially non-regular quantization cell for X, $w_1^{-1}(Q_{Y1}^{-1}(Q_{Y1}(y)))$. By definition, $w_1^{-1}(y)$ is a set containing finitely many points; let us enumerate these M points as $w_1^{-1}(y) = \{a_1, \ldots, a_M\}$. We note that $a_i \in w_1^{-1}(Q_{Y1}^{-1}(Q_{Y1}(y)))$ for any $i \in \{1, \ldots, M\}$. Since the points a_i are distinct, M is finite, and $w_1(x)$ has bounded derivative

within each of a finite number of pieces, for a sufficiently high rate the quantizer cell over X, $w_1^{-1}(Q_{Y1}^{-1}(Q_{Y1}(y)))$, reduces to a union of disjoint regular intervals over X — each containing one of the points a_i.

Let us consider the distortion within the quantization cell containing $w_1^{-1}(y)$, and point x_2^N: $w_1^{-1}(Q_{Y1}^{-1}(Q_{Y1}(y))) \times (Q_{Xi}^{-1}(x_i))_2^N$. Note that each of the sets $Q_{Xi}^{-1}(x_i)$ is a regular interval.

$$D(y, x_2^N) = \text{var}[G(X_1^N) \mid X_1 \in Q_{Y1}^{-1}(Q_{Y1}(y)), X_{i>1} \in Q_{Xi}^{-1}(x_i)] \tag{3.48}$$

Yet again, this may be bounded by the law of total variance — this time involving the indicator function $I(X_1)$ for X_1's non-regular quantizer. To reduce notational complexity, we indicate the quantization cell $Q_{Y_1}^{-1}(Q_{Y_1}(y)) \times \prod_{i=2}^{N} Q_{X_i}^{-1}(x_i)$ by $q(y, x_2^N)$.

$$\begin{aligned} D(y, x_2^N) &= \mathbf{E}\left[\text{var}[G(X_1^N) \mid X_1^N \in q(y, x_2^N), I(X_1)]\right] \\ &\quad + \text{var}[\mathbf{E}\left[G(X_1^N) \mid X_1^N \in q(y, x_2^N), I(X_1)\right] \mid X_1^N \in q(y, x_2^N)] \\ &\geq \text{var}[\mathbf{E}\left[G(X_1^N) \mid X_1^N \in q(y, x_2^N), I(X_1)\right] \mid X_1^N \in q(y, x_2^N)] \end{aligned}$$

We now make use of G's smoothness. Specifically, we take advantage of its derivative being bounded: $\left|\frac{dG(x_1^N)}{x_i}\right| < L$ for any $i \in [1 \ldots M]$. We also introduce the notation $\Delta_i(x)$ for the width of the quantizer interval in the ith quantizer containing coordinate x.

$$
\begin{aligned}
D(y, x_2^N) &\geq \operatorname{var}[\mathbf{E}\left[G(X_1^N) \mid X_1^N \in q(y, x_2^N), I(X_1)\right] \mid X_1^N \in q(y, x_2^N)] \\
&= \operatorname{var}[\mathbf{E}\left[G(a_{I(X_1)}, x_2^N) + G(X_1^N) - G(a_{I(X_1)}, x_2^N) \mid X_1^N \in q(y, x_2^N), I(X_1)\right] \mid X_1^N \in q(y, x_2^N)] \\
&= \operatorname{var}[\mathbf{E}\left[G(a_{I(X_1)}, x_2^N) \mid X_1^N \in q(y, x_2^N), I(X_1)\right] \\
&\quad + \mathbf{E}\left[G(X_1^N) - G(a_{I(X_1)}, x_2^N) \mid X_1^N \in q(y, x_2^N), I(X_1)\right] \mid X_1^N \in q(y, x_2^N)] \qquad (3.49) \\
&\geq \operatorname{var}[G(a_{I(X_1)}, x_2^N) \mid w_1(X_1) = y] \\
&\quad - \operatorname{var}[\mathbf{E}\left[G(X_1^N) - G(a_{I(X_1)}, x_2^N) \mid X_1^N \in q(y, x_2^N), I(X_1)\right] \mid X_1^N \in q(y, x_2^N)] \qquad (3.50) \\
&\geq \operatorname{var}[G(a_{I(X_1)}, x_2^N) \mid w_1(X_1) = y] \\
&\quad - \operatorname{var}[\mathbf{E}\left[\left|G(X_1^N) - G(a_{I(X_1)}, x_2^N)\right| \mid X_1^N \in q(y, x_2^N), I(X_1)\right] \mid X_1^N \in q(y, x_2^N)] \qquad (3.51) \\
&\geq \operatorname{var}[G(a_{I(X_1)}, x_2^N) \mid w_1(X_1) = y] \\
&\quad - \operatorname{var}[\mathbf{E}\left[\sum_{i=1}^{N} \left|\left|\frac{dG}{dx_i}\right|_{X_1^N} \Delta_i(X_i)\right| \mid X_1^N \in q(y, x_2^N), I(X_1)\right] \mid X_1^N \in q(y, x_2^N)] \qquad (3.52) \\
&\geq \operatorname{var}[G(a_{I(X_1)}, x_2^N) \mid w_1(X_1) = y] - NL^2 \max_{i \in [1,\ldots,N], x \in [0,1]} \Delta_i(x)^2 \qquad (3.53)
\end{aligned}
$$

Taking the expectation of this quantity over all $y \in w_1([0,1])$ and all $x_2^N \in [0,1]^{N-1}$ yields a bound for the total distortion:

$$
D_{TOT} \geq \mathbf{E}\left[\operatorname{var}[G(a_{I(X_1)}, X_2^N) \mid w_1(X_1) = Y]\right] - NL^2 \Delta_{max}(R) \qquad (3.54)
$$

The second term in this expression decays to zero with increasing rate (the width of the largest quantizer cell), while the first — a high-rate characteristic of the system — remains constant. We demonstrate that the first is greater than zero if G is equivalence-free and w_1 is non-one-to-one over a set of nonzero probability.

Suppose the first term is, instead, zero.

$$
\begin{aligned}
0 &= \mathbf{E}\left[\operatorname{var}[G(a_{I(X_1)}, X_2^N) \mid Y]\right] \qquad (3.55) \\
&= \mathbf{P}\left(\operatorname{var}[G(a_{I(X_1)}, X_2^N) \mid Y] > 0\right) \mathbf{E}\left[\operatorname{var}[G(a_{I(X_1)}, X_2^N) \mid Y, \operatorname{var}[G(a_{I(X_1)}, X_2^N)] > 0\right] \qquad (3.56)
\end{aligned}
$$

Since G is equivalence-free and w_1 introduces non-regularity with nonzero probability,

$$\mathbf{P}\left(\text{var}[G(a_{I(X_1)}, X_2^N) \mid Y] > 0\right) > 0$$

and (clearly)

$$\mathbf{E}\left[\text{var}[G(a_{I(X_1)}, X_2^N) \mid Y, \text{var}[G(a_{I(X_1)}, X_2^N)] > 0]\right] > 0.$$

Putting these together, we may obtain the desired contradiction:

$$\begin{aligned}\mathbf{E}\left[\text{var}[G(a_{I(X_1)}, X_2^N) \mid Y]\right] \geq \mathbf{P}\left(\text{var}[G(a_{I(X_1)}, X_2^N) \mid Y] > 0\right) \mathbf{E}\left[\text{var}[G(a_{I(X_1)}, X_2^N) \mid Y]\right] \quad (3.57)\\ \geq 0\end{aligned}$$

This demonstrates that the first term is nonzero. Because $\Delta_{max}(R)$ decays to zero monotonically with rate, we may pick R_0 such that

$$NL^2\Delta_{max}(R_0) < \mathbf{E}\left[\text{var}[G(a_{I(X_1)}, X_2^N) \mid w_1(X_1) = Y]\right].$$

For $R > R_0$, the following then holds true:

$$\begin{aligned} D_{TOT} &\geq \mathbf{E}\left[\text{var}[G(a_{I(X_1)}, X_2^N) \mid w_1(X_1) = Y]\right] - NL^2\Delta_{max}(R) && (3.58)\\ &\geq \mathbf{E}\left[\text{var}[G(a_{I(X_1)}, X_2^N) \mid w_1(X_1) = Y]\right] - NL^2\Delta_{max}(R_0) && (3.59)\\ &> 0 && (3.60)\end{aligned}$$

This proves the theorem. ∎

The ramifications of this are striking: non-regular quantization introduces a nonzero lower bound to the distortion of equivalence-free functions. This is clearly suboptimal if the rate is sufficiently high; even the naive uniform quantizer possesses a 2^{-2R} dependence! Therefore, for equivalence-free functions the performance of non-regular quantization can be either improved upon or equalled by regular quantization.

A grain of salt: note that this refers to the design of *high-rate* optimal quantizers. For finite rate constraints, a non-regular quantizer may very well outperform a regular quantizer.

3.2.3 Optimal Non-Regular Functional Quantization

In the previous section we demonstrated that regular quantization is high-resolution optimal for the class of equivalence-free functions. In this section, we consider functions that possess equivalences and demonstrate that our approach to high-rate non-regular quantizer design is optimal for them in the high rate.

First of all, suppose that the function of interest, $G(X_1^N)$, is continuous, smooth (bounded gradient), and bounded. Furthermore, suppose that there exist equivalences in G from the perspective of the encoder for X_1. The analysis of the previous section, in addition to demonstrating regular quantization's optimality for equivalence-free functions, also shows that any binning that is performed over a non-equivalent set $B \in [0,1]$ will introduce a nonzero floor to the distortion. Therefore, if a non-regular quantizer is to be allowed to operate on X_1, its non-regular cells must be centered on points that are functionally equivalent. This provides a bound for the maximum possible non-regularity that will not introduce a distortion floor: for each set of functionally equivalent points $\{x_1, \ldots, x_M\}$, the quantization intervals containing any point in the set are unioned.

The only source of error with this approach is the presence of elements in the unionized quantization cells that are not equivalent to one another (they are simply near other elements that are). For instance, suppose the points $x_1 = \frac{1}{4}$ and $x_1 = \frac{3}{4}$ form the only equivalence class of cardinality greater than one. Each will be quantized within an interval of some length, $\Delta_{1/4}$ and $\Delta_{3/4}$. When these intervals are unioned, problems emerge for finite interval lengths: $\frac{1}{4} + \epsilon$ is being grouped with $[3/4 - \delta, 3/4 + \delta]$, for instance. This results in an additional distortion ΔD bounded by $\Delta D \leq 2^{-2R}$. As the resolution is raised, the quantization interval sizes become smaller and this error disappears. Therefore, we have an asymptotically optimal approach in the discrete realm: bin together the quantization cells containing each element of an equivalence class. It can be seen that this is equivalent to implementing a generalized companding function $w(X)$.

Chapter 4

Senior Year: Functional Transform Coding, Encoder Collaboration, and Uncertainty

In Chapter 2, we developed a mathematical framework in which one may analyze the performance of functional quantization systems. We then proceeded, in Chapter 3, to extend the reach of this theory to all continuous functions with derivative bounded almost everywhere. In this chapter, we use these results as a foundation from which to tackle several new scenarios and problems. Each demonstrates both the intuitiveness of the functional quantization picture and its assistance in analysis.

First, we consider functional transform coding (FTC). Transform coding, in its ordinary incarnation, has proven itself as an extremely valuable tool for practical source coding; does it continue to help us when we care about a function of the source vector? Moreover, how does the optimal functional transform code compare with the well known Karhunen-Loeve Transform?

After this, we consider the question of encoder collaboration. The picture we have been dealing with thus far does not draw any arrows between the encoders — what if they are allowed to communicate? We examine this question in the context of both fixed- and variable-rate coding, and find wildly different behavior. While collaboration can yield arbitrarily high reductions in distortion for variable-rate, for fixed-rate the bits used for encoder-encoder communication would be better spent going to the decoder and reducing distortion according to the -6dB per bit rule.

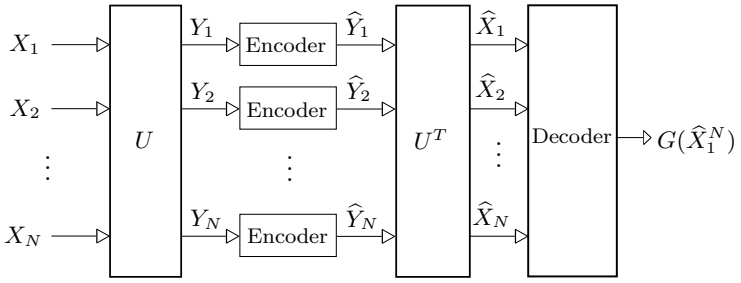

Figure 4-1: Uniform transform coding

Finally, we look to questions of uncertainty and universality. We have examined in previous chapters how to design the optimal quantizer given the source distribution — but what if the implementation was flawed? Also, a source's distribution often is not known precisely. How does this effect the design of an optimal quantizer?

4.1 Functional Transform Coding

Consider the setup common to Feng et al. [20] and Bucklew [23] (see Fig. 2-1). Both require vector quantization—a computationally expensive premise if the form of the quantizer is to be left arbitrary. By constraining the quantizer to the form of a uniform transform code, we can significantly reduce this cost while still removing redundancies between sources.

Under the transform constraint (Fig. 4-1), an N-vector of source variables, $X_1^N \in \mathbb{R}^N$ is first presented to an encoder. We constrain X_1^N so that its distribution, $f_{X_1^N}(x_1^N)$ is supported entirely within the N-sphere of radius 1. Following this,

1. An invertible linear transformation, U, is applied to X_1^N to yield vector $Y_1^N = UX_1^N$. We constrain U to be of unity determinant. This is without loss of generality, since (1) U must have a nonzero determinant from being invertible and (2) if U has a non-unity determinant c, the scaled matrix $Uc^{-1/N}$ will have unity determinant and will result in identical distortion performance as U.
2. The components of Y_1^N are uniformly scalar quantized into the vector $\widehat{Y}_1^N$. $\widehat{Y}_i$ has fixed rate R_i and resulting interval size 2^{-R_i}.
3. There is a total rate constraint: $\sum R_i \leq R$.
4. The decoder inverts the transformation $\widehat{X}_1^N = U^{-1}\widehat{Y}^N$, and computes an estimate of the

function $\widehat{G}(X_1^N) \triangleq G(\widehat{X}_1^N)$.

The goal remains unchanged from Chapter 2: U and the R_i are to be chosen to minimize the mean squared error of G, $D_G = \mathbf{E}\left[(\widehat{G}(X_1^N) - G(X_1^N))^2\right]$.

Several definitions prove helpful in obtaining the optimal transform.

Definition Define the sensitivity vector over X so that each of its components is given by $\gamma_{X_i} = \partial G(x_1^N)/\partial x_i$.

Definition Similarly, define the sensitivity vector over Y so that each of its components is given by $\gamma_{Y_i} = \partial G(U^{-1}(y_1^N))/\partial y_i$.

Note that the two vectors are related by the transform $\gamma_Y = U\gamma_X$.

Definition Define the sensitivity matrix over X_1^N, $\Gamma_X(x_1^N) = \gamma_X\gamma_X^T$ and the sensitivity matrix over Y^N, $\Gamma(y^N) = \gamma_Y\gamma_Y^T$. The (i,j)th components are more explicitly given by

$$\Gamma_X(x_1^N)_{i,j} = \frac{\partial G}{\partial x_i}\frac{\partial G}{\partial x_j}$$

and likewise for $\Gamma_Y(y_1^N)$.

Lemma 4.1.1 *The following three properties hold.*

1. Γ_X *and* Γ_Y *are real, symmetric, and positive semidefinite.*
2. $\boldsymbol{E}[\Gamma_X]$ *and* $\boldsymbol{E}[\Gamma_Y]$ *are also real, symmetric, and positive semidefinite.*
3. $\boldsymbol{E}[\Gamma_Y] = U\boldsymbol{E}[\Gamma_X]U^{-1}$.

Proof Positive semidefiniteness of Γ_X and Γ_Y follows from each matrix being the outer product of a real-valued vector with itself. Positive semidefiniteness of $\mathbf{E}[\Gamma_X]$ and $\mathbf{E}[\Gamma_Y]$ follows from their description as a sum of positively scaled positive definite matrices. The final property is the result of algebra: $\mathbf{E}[\Gamma_Y] = \mathbf{E}\left[\gamma_Y\gamma_Y^T\right] = \mathbf{E}\left[U\gamma_X\gamma_X^TU^{-1}\right] = U\mathbf{E}[\Gamma_X]U^{-1}$. ∎

Using these properties we may optimize the distortion.

Theorem 4.1.2 *The optimal transform U diagonalizes the matrix $\boldsymbol{E}[\Gamma_X]$ and results in distortion*

$$D_G = \frac{N}{12}2^{-2R/N}\det(\boldsymbol{E}[\Gamma_X])^{1/N}$$

except when $\boldsymbol{E}[\Gamma_X]$ *has any zero eigenvalues.*

Proof Uniform quantization entails high-rate point density functions over Y_1^N of the form $\lambda_i = 1$. Referring to Eq. 2.12, we may write the total distortion as

$$D_G = \frac{1}{12}\sum_i 2^{-2R_i}\mathbf{E}\left[\left|\frac{\partial G_Y(y_1^N)}{\partial y_i}\right|^2\right]$$

where $G_Y(y_1^N) = G(U^{-1}y_1^N)$.

Optimizing the rates R_i subject to the sum-rate condition,

$$D_G \geq \frac{N}{12}2^{-2R/N}\left[\prod_i \mathbf{E}\left[\left|\frac{\partial G_Y(y_1^N)}{\partial y_i}\right|^2\right]\right]^{1/N} \tag{4.1}$$

Note that this bound is achievable for finite R if and only if

$$\mathbf{E}\left[|\partial G/\partial y_i|^2\right] > 0 \text{ for all } i = 1, \ldots, N. \tag{4.2}$$

The term in brackets is the product of the diagonal elements of $\mathbf{E}[\Gamma_Y]$, referred to us as the multiplicative trace. By the Hadamard inequality, we can simultaneously minimize both the distortion and the multiplicative trace by choosing U to diagonalize $\mathbf{E}[\Gamma_X]$. This yields total distortion:

$$D_G \geq \frac{N}{12}2^{-2R/N}\det(\mathbf{E}[\Gamma_X])^{1/N}. \tag{4.3}$$

This distortion bound is achievable if $\det(\mathbf{E}[\Gamma_X]) > 0$, since in that case $\mathbf{E}[\Gamma_Y]$ has no zero diagonal elements. ∎

If $\mathbf{E}[\Gamma_Y]$ has any zero diagonal elements, $\det(\mathbf{E}[\Gamma_X]) = 0$ and condition (4.2) has been violated. To correct for this, the zero components may be discarded, for G is unaffected by them almost everywhere. We demonstrate this quantitatively.

Lemma 4.1.3 *If* $\boldsymbol{E}[\Gamma_Y]$ *has a zero diagonal element, distortion is unaffected if the corresponding component is discarded.*

Proof By our construction, the function $G(X_1^N)$ has bounded gradient. Therefore, $\Gamma_Y(y_1^N) \leq L^2$ for some constant bound L^2. Now assume that a diagonal element of $\mathbf{E}[\Gamma_Y]$, given by $\mathbf{E}[\Gamma_Y]_{jj} =$

$\mathbf{E}\left[\left|\frac{\partial G}{\partial y_j}\right|^2\right]$, is zero. Assume, for convenience, that this is the first element $j = 1$. Discarding this component forces the decoder to make the estimate

$$\widehat{G}(\widehat{y}_2^N) = \mathbf{E}\left[G(Y_1, \widehat{y}_2^N)\right]$$

from quantized components $\widehat{y}_2^N$ instead of the estimate

$$\widehat{G}(\widehat{y}_2^N) = G(\widehat{y}_1^N)$$

from quantized components $\widehat{y}_1^N$.

We can bound the effect of this change by the law of total variances.

$$\begin{aligned}
D &= \mathbf{E}\left[\mathrm{var}[G(Y_1^N) \mid \widehat{y_2^N}]\right] && (4.4)\\
&= \mathbf{E}\left[\mathrm{var}[\mathbf{E}\left[G(Y_1^N) \mid \widehat{y_2^N}, Y_1\right] \mid \widehat{y_2^N}] + \mathbf{E}\left[\mathrm{var}[G(Y_1^N) \mid \widehat{y_2^N}, Y_1] \mid \widehat{y_2^N}\right]\right] && (4.5)\\
&\leq \mathbf{E}\left[\mathrm{var}[\mathbf{E}\left[G(Y_1^N) \mid \widehat{y_2^N}, Y_1\right] \mid \widehat{y_2^N}]\right] + D_0 && (4.6)\\
&= \mathbf{E}\left[\mathrm{var}[G(Y_1, \widehat{y_2^N}) \mid \widehat{y_2^N}]\right] + D_0 && (4.7)\\
&\leq L^2 \mathbf{P}\left(|dG/dy_1| > 0\right) + D_0 && (4.8)\\
&= D_0 && (4.9)
\end{aligned}$$

Therefore, there is no cost to discarding the unused component Y_1. ∎

This process may be repeated for all the zero eigenvalues. Rederiving the optimal transform after deleting them, we notice two effects.

1. The non-zero eigenvalues are left unchanged.

2. The dimensionality has been reduced from N sources to $N - \overline{N}$, where $\overline{N}$ denotes the number of null components.

In general, use of the transform U reduces distortion by a factor $2^{-2R\overline{N}} \det(\overline{\mathbf{E}\left[\Gamma_Y\right]})/\mathbf{E}\left[\prod \gamma_{X_i}^2\right]$, where $\overline{\mathbf{E}\left[\Gamma_Y\right]}$ is the reduced dimensionality $E[\Gamma_Y]$. Unlike the separable quantization situations of the previous chapters, this distortion improvement is rate dependent when $\overline{N} > 0$.

Notice the similarity of the optimal transformation to that of the KLT for non-functional transform coding. The KLT diagonalizes the covariance of the random vector, X_1^N. U diagonalizes $\mathbf{E}[\Gamma_X]$, which may be written as the sum of the covariance of the random vector γ_X and the matrix $A_{ij} = \mathbf{E}[\partial G/\partial x_i]\,\mathbf{E}[\partial G/\partial x_j]$.

4.1.1 Example: The Selective Functions

Just as the sensitivity profile defined a function in the eyes of a functional quantizer, the sensitivity matrix defines it in the eyes of a functional transform code. For any selective function, the derivative of the function is zero with respect to all sources but one, for which the derivative is unity. Therefore, we have the following sensitivity matrix:

$$\begin{aligned} \mathbf{E}[\Gamma_X]_{ij} &= \mathbf{E}\left[\frac{dG}{dx_i}\frac{dG}{dx_j}\right] && (4.10)\\ &= \delta_{ij}\mathbf{P}\left(G(x_1^N) = x_i\right) && (4.11)\end{aligned}$$

The sensitivity matrix is already diagonalized, so the optimal transform is the identity matrix and the problem is merely one of bit allocation amongst the encoders. If G is also symmetric, the rates are split evenly amongst the sources, and functional transform coding yields no benefits.

4.1.2 Example: A Linear Combination of the Sources

Suppose we are interested in a linear function of the sources, $G(x_1^N) = A(v_1^N)^T \cdot x_1^N$, where A is a real number and v_1^N is a normalized vector. We then see that the sensitivity matrix is diagonalized if one of the basis vectors y for the transform is chosen to be v. Specifically, this will result in a single element of $\mathbf{E}[\Gamma_Y]$ being non-zero. We therefore have a distortion improvement of $2^{R(N-1)}$ by using the transform code.

4.1.3 Limitations

Our construction considers uniform quantization after transformation of a source vector that is contained within the unit sphere. This is notably different from our assumptions for distributed quantization, where the source vector was contained within $[0,1]^N$. The consequences are felt when

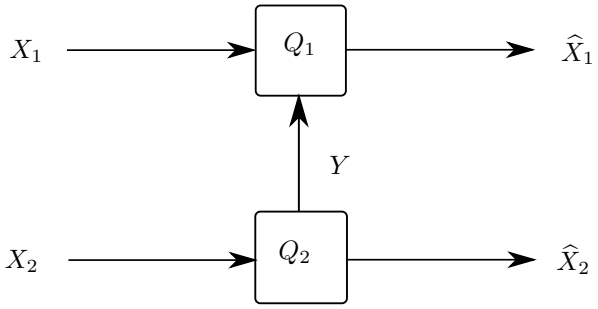

Figure 4-2: Suppose the encoder for X_2 could send a message to the encoder for X_1. Is there any benefit?

we consider sources that have a support smaller than the unit sphere.

For instance, consider a source contained within a box within the unit sphere. Geometry tells us that this box can be no larger than $[0, N^{-1/2}]^N$. On one hand, this restriction makes sense: if we were to (conceivably) rotate the box so that its diagonal was aligned with one of our axes, the uniform quantizer would have to accommodate a $[0, 1]$ support, which is the maximum it is capable of. On the other hand, when the box is in its natural $[0, N^{-1/2}]^N$ state, each uniform quantizer could scale its support down to $[0, N^{-1/2}]$ without causing overload problems. This reveals a shortcoming with our problem setup: the quantizers are not allowed to adapt their support after a transformation has taken place.

4.2 Encoder Collaboration

We now consider the scenario where one of the encoders is given the option of communication with another. For instance, X_2's encoder might choose to send a random variable of side information, Y, to X_1's encoder (Fig. 4-2). Assuming that Y must be conditionally independent of X_1 given X_2, what kind of performance gains are possible for fixed- and variable-rate coding? Our approach will be to consider the situation where Y is a single bit of information, and generalize from there.

Formally, the ith encoder may code X_i with knowledge of the binary variable Y from the jth encoder. For convenience, assume that $i = 1$ and $j = 2$ (i.e., the second encoder sends information to the first). As stated before, Y is conditionally independent of X_1 given X_2. We restrict this relationship even further: Y must be a deterministic function of $\widehat{X}_2 = Q(X_2)$. This constraint ensures that the decoder has access to Y. Note that this becomes increasingly reasonable as resolution is increased.

The best possible strategy — an upper bound for performance — is for encoder 1 to use two codebooks: one for when $Y = 0$ and one for when $Y = 1$. Since the decoder may determine the value of Y from the high-rate description of X_2, this bound is achievable. Recall that the distortion expressions for both fixed- and variable- rate coding take the form of a summation of N 1D functional distortions:

$$D = \sum_{i=1}^{N} D_i 2^{-2R_1}$$

We can immediately incorporate the effect of Y into this expression. The total distortion is merely the expected distortion over Y. Moreover, the only terms affected by Y are D_1 and R_1:

$$\begin{aligned} D_{\text{tot}} &= \mathbf{E}\left[D \mid Y\right] && (4.12) \\ &= \mathbf{E}\left[D_1(Y)2^{-2R_1(Y)}\right] + \sum_{i=2}^{N} D_i 2^{-2R_i} && (4.13) \end{aligned}$$

A brief aside is warranted on the nature of $R_1(Y)$. Under a fixed-rate constraint, $R_1(Y) = R_1$; any flexibility in its value would break the fixed-rate constraint. For variable-rate, on the other hand, the first encoder can certainly change the length of its stream based on the value of Y. However, the other encoders are not extended this privilege, as they are not privy to the value taken by Y. As demonstrated by this example, the situations are slightly different for fixed- and variable-rate coding. We therefore consider them separately from this point onward.

4.2.1 Fixed-Rate

Knowing that we are working in a fixed-rate context, Eq. 4.13 can be written more explicitly in terms of the source/side-information distribution $f_{X|Y}(X_1^N = x_1^N \mid Y = y)$. We also define the conditional sensitivity profile:

$$g_{i|Y}^2(x|y) = \mathbf{E}\left[\left|\left.\frac{dG(x_1^N)}{dx_i}\right|_{X_i}\right|^2 \mid X_i = x, Y = y\right]$$

This expression has a simple interpretation. The optimal fixed-rate encoder has two codebooks to work with: one when $Y = 0$ and one when $Y = 1$. In each scenario, it designs an optimal quantizer

based on the source distribution from its perspective, which is $f_{X|Y}$ instead of the usual f_X. The conditional sensitivity profile is merely the profile, derived from the conditional source distribution, that it uses to design the optimal quantizer in each case. Making use of $g^2_{i|Y}$ and the conditional distribution $f_{X_i|Y}$, we have:

$$
\begin{aligned}
D &= \mathbf{E}\left[\frac{1}{12}2^{-2R_1}\|f_{X_i|Y}(x|Y)g^2_{1|Y}(x|Y)\|_{1/3} \mid Y\right] + \sum_{i=2}^{N} D_i 2^{-2R_i} && (4.14)\\
&= \frac{1}{12}2^{-2R_1}\left(\mathbf{P}\left(Y=0\right)\|f_{X_i|Y}(x|0)g^2_{1|Y}(x|0)\|_{1/3} + \mathbf{P}\left(Y=1\right)\|f_{X_i|Y}(x|1)g^2_{1|Y}(x|1)\|_{1/3}\right)\\
&\quad + \sum_{i=2}^{N} D_i 2^{-2R_i} && (4.15)
\end{aligned}
$$

This equation does not seem terribly helpful until we massage out a curious relationship.

$$
\begin{aligned}
&\mathbf{P}\left(Y=0\right) f_{X_i|Y}(x|0)g^2_{1|Y}(x|0) + \mathbf{P}\left(Y=1\right) f_{X_i|Y}(x|1)g^2_{1|Y}(x|1) && (4.16)\\
&= \frac{f_{X_i}(x)}{f_{X_i}(x)}\left(\mathbf{P}\left(Y=0\right) f_{X_i|Y}(x|0)g^2_{1|Y}(x|0) + \mathbf{P}\left(Y=1\right) f_{X_i|Y}(x|1)g^2_{1|Y}(x|1)\right) && (4.17)\\
&= f_{X_i}(x)\left(\frac{\mathbf{P}\left(Y=0\right) f_{X_i|Y}(x|0)}{f_{X_i}(x)}g^2_{1|Y}(x|0) + \frac{\mathbf{P}\left(Y=0\right) f_{X_i|Y}(x|1)}{f_{X_i}(x)}g^2_{1|Y}(x|1)\right) && (4.18)\\
&= f_{X_i}(x)\left(\mathbf{P}\left(Y=0 \mid X=x\right) g^2_{1|Y}(x|0) + \mathbf{P}\left(Y=1 \mid X=x\right) g^2_{1|Y}(x|1)\right) && (4.19)\\
&= f_{X_i}(x)g^2_1(x) && (4.20)
\end{aligned}
$$

It helps to look at $f_X(\cdot)g^2_1(\cdot)$ and $\mathbf{P}\left(Y=y\right) f_{X|Y}(\cdot|y)g^2_{1|Y=y}(\cdot|y)$ as vectors in a Hilbert space. By sending side information Y, we are essentially expressing a vector $f_X(\cdot)g^2_1(\cdot)$ as the sum of two others. If we interpret the $\mathcal{L}^{1/3}$ norm as a distance measure (a mathematical stretch, but back to that in a bit), we have replaced

$$D \propto \|f_X(x)g^2_1(x)\|_{1/3}$$

with

$$D \propto \mathbf{P}(Y=0)\,\|f_{X_1|Y}(x,0)g_{1|Y}(x,0)^2\|_{1/3} + \mathbf{P}(Y=1)\,\|f_{X_1|Y}(x,1)g_{1|Y}(x,1)^2\|_{1/3}$$
$$= \|\mathbf{P}(Y=0)\,f_{X_1|Y}(x,0)g_{1|Y}(x,0)^2\|_{1/3} + \|\mathbf{P}(Y=1)\,f_{X_1|Y}(x,1)g_{1|Y}(x,1)^2\|_{1/3}$$

In other words, the length of a vector has been replaced by the sum of the lengths of two vectors that add to produce it. If the distance measure were an actual distance metric, this would be rather unfortunate, as it implies we can never reduce the distortion via side information (triangle inequality). However, the operation $\|\cdot\|_{1/3}$ qualifies as a *quasinorm*, for which gains are possible. In fact, there is a triangle-like inequality that applies to the $\mathcal{L}^{1/3}$ norm, and by relating the lengths of the components to the length of their sum it bounds the distortion improvement from usage of side information.

Saito [33] provides a quantitative relation without proof; we give a proof of this in Appendix 4.A:

$$\|x(t)+y(t)\|_{1/3} \le 4(\|x(t)\|_{1/3} + \|y(t)\|_{1/3}) \tag{4.21}$$

This can be back-substituted to yield the following:

$$\begin{aligned} D &= \frac{1}{12}2^{-2R_1}\left(\mathbf{P}(Y=0)\,\|g_{1|Y}^2(x|0)\|_{1/3} + \mathbf{P}(Y=1)\,\|g_{1|Y}^2(x|1)\|_{1/3}\right) \\ &\quad + \sum_{i=2}^{N} D_i 2^{-2R_i} \qquad (4.22) \\ &\ge \frac{1}{4}D_1 2^{-2R_1} + \sum_{i=2}^{N} D_i 2^{-2R_i} \qquad (4.23) \end{aligned}$$

Use of a single bit of side information can *at most* reduce an encoder's distortion by a factor of four. What's more, one may *guarantee* a factor-of-four reduction by using the extra bit of side communication to instead increase resolution for the encoder ($R_1 \to R_1 + 1$). This result trivially generalizes to multiple bits: side-information is generally a losing game for fixed-rate encoding. We note that the flavor of this is similar to that of Prabhakaran et al. [34] for the ordinary source coding of Gaussian sources.

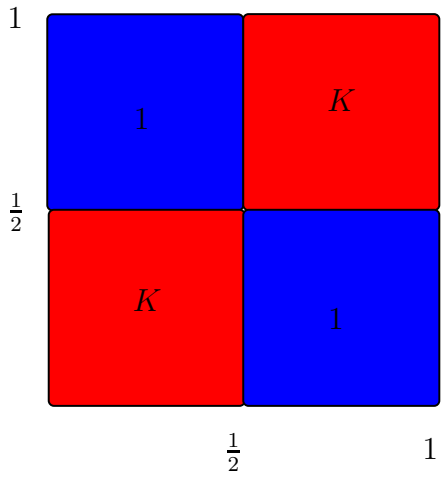

Figure 4-3: Example scenario: X_1 is the horizontal axis, and X_2 the vertical. The numbers in each quadrant are the values of the derivative of G against X_1.

4.2.2 Variable-Rate

For the variable-rate scenario, one might approach things in a similar light. The expectation of distortion may be taken over both possible values of Y, and one may write the resulting distortion in terms of the conditional sensitivities $g^2_{1|Y=y}$ defined as in the previous section. R_1 is now allowed to depend on Y, but we will ignore this degree of freedom, as it turns out to be inconsequential for our purposes.

$$D = \mathbf{E}\left[\frac{1}{12}2^{-2R_1}2^{\mathbf{E}\left[\log_2 g^2_{1|Y}(X_1|Y)|Y\right]}\right] + \sum_{i=2}^{N} D_i 2^{-2R_i} \tag{4.24}$$

$$= \frac{1}{12}2^{-2R_1}\left(\mathbf{P}\left(Y=0\right)2^{\mathbf{E}\left[\log_2 g^2_{1|Y}(X_1|0)|Y=0\right]}\|_{1/3} + \mathbf{P}\left(Y=1\right)2^{\mathbf{E}\left[\log_2 g^2_{1|Y}(X_1|1)|Y=1\right]}\right) + \sum_{i=2}^{N} D_i 2^{-2R_i} \tag{4.25}$$

Unlike the $\mathcal{L}^{1/3}$ norm, the log-norm has no bound of the form of Eq. 4.21. This implies, in theory, that arbitrary improvements are possible from a single bit of side information. We demonstrate that this is the case by means of a simple example.

Let X_1 and X_2 be uniform i.i.d. over $[0,1]^2$. Rather than explicitly defining our function, we will define its derivative profile. The function G may itself be obtained via integration. For X_2,

$\frac{dG(x_1,x_2)}{dx_2} = 1$, and the optimal quantizer is uniform. For X_1, we define its derivative as a piecewise constant function over the four quadrants of $[0,1]^2$ (Fig. 4-3). If X_1 and X_2 are both less than $1/2$ or if X_1 and X_2 are both greater than $1/2$, $\frac{dG(x_1,x_2)}{dx_1} = 1$. Otherwise (in the other two quadrants), $\frac{dG(x_1,x_2)}{dx_1} = L$, where L is some nonzero positive constant. We can derive the sensitivity profile for X_1 from this description:

$$g_1^2(x) = \frac{1+L}{2}$$

This also allows us to find the distortion of the functional quantizer without encoder communication. We only express the term from the first encoder, since the second term's contribution is unaffected by the side-information.

$$D_1 = \frac{1}{12}2^{-2R}2^{\mathbf{E}[\log(1+L)-1]} = \frac{1}{12}2^{-2R}\frac{L+1}{2}$$

Now suppose that X_2 can provide one bit of information to X_1. Functional quantization reduces to ordinary quantization for a constant sensitivity profile; the less constant the profile is, the greater the improvement from functional techniques. This suggests to us that the best piece of information allows the quantizer of X_1 to tailor itself to the nonuniformity of the joint distribution:

$$Y = \begin{cases} 0 & \text{if } X_2 > 1/2 \\ 1 & \text{otherwise} \end{cases}$$

From this, we can define the conditional sensitivity profiles for X_1:

$$g^2_{1|Y=y}(x,y) = \begin{cases} 1 & \text{if } X_1 > 1/2 \text{ and } Y = 0 \\ 1 & \text{if } X_1 \leq 1/2 \text{ and } Y = 1 \\ L & \text{otherwise} \end{cases}$$

We can now compute the distortion contribution from X_1 with Y available, D_{1Y}.

$$D_{1Y} = \frac{1}{12}2^{-2R}2^{\mathbf{E}\left[\log_2 g^2_{i|Y=0}\right]/2+\mathbf{E}\left[\log_2 g^2_{i|Y=1}\right]/2} \quad (4.26)$$

$$= \frac{1}{12}2^{-2R}2^{\frac{1}{2}\log_2 L} \quad (4.27)$$

$$= \frac{1}{12}2^{-2R}\sqrt{L} \quad (4.28)$$

This is in contrast to the performance without the side information Y. Taking the ratio between the old and the new D_1, we have

$$\frac{D_1}{D_{1Y}} = \frac{L+1}{2\sqrt{K}}$$

This performance gap grows arbitrarily large as L is increased — and in all cases it is due to a single bit of information.

4.2.3 Comparison with Ordinary (Non-Functional) Scenario

These results are strikingly different than those from ordinary source coding. Consider first the discrete scenario. X_1^N is now a vector from a discrete alphabet, and we wish to recreate X_1^N perfectly at the decoder. Can chatting between encoders assist us in reducing the total rate of communication at all? According to Slepian and Wolf, the answer is a resounding "no." Even in the case of unlimited collaboration via fused encoders, the minimum sum-rate to the decoder remains unchanged.

How about ordinary quantization? If quantization is variable-rate and Slepian-Wolf coding is employed on the quantized indices, no gains are possible from talking encoders. This result can also be seen as a consequence of Rebollo-Monedero's work [19] with high-resolution Wyner-Ziv coding, where he demonstrates that there is no gain from supplying the source encoder with the decoder side information.

4.3 Penalties for Suboptimality

For a variety of reasons, the optimal quantizer point density may not be perfectly implemented. Perhaps precision is the limiting factor, or perhaps it is complexity. One often sees piecewise constant

point density functions in practice for this latter reason. Also, one may not have perfect knowledge of the source distribution. In this section, we explore how sensitive functional quantizers are to these types of imperfections.

The worst kind of screwup is the kind that isn't recognized as a screwup. If a functional quantizer is designed for the wrong source distribution or the wrong sensitivity profile g_i^2, how much of an effect can this have on the asymptotic rate-distortion performance? This is examined for both the fixed- and variable-rate scenarios.

There are two varieties of mistakes we will consider. A *source modeling error* occurs when an incorrect source distribution, $e_X(x)$ is used to design a quantizer in place of the correct distribution, $f_X(x)$. A *functional modeling error* occurs when an incorrect functional sensitivity profile $h_i^2(x)$ is used to design a quantizer in place of the correct sensitivity profile, $g_i^2(x)$. Our approaches to analyzing the errors from these two sources proves different for fixed- and variable-rate coding.

4.3.1 Fixed-Rate Imperfect Design

The only part of a fixed-rate encoder that needs to be designed is the quantization block — the process of codeword generation requires no design decision. The quantization block is completely summarized by the quantization profile $\lambda_i(x)$. An optimal choice from the perspective of an engineer who believes the source distribution to be e_X and the sensitivity to be h_i^2 is $\lambda_i(x) \propto (e_X h_i^2)^{1/3}$. This lies in contrast to the truly optimal choice, $\lambda_i(x) \propto (f_X g_i^2)^{1/3}$. Rather than attempting to separate the effects of incorrect functional sensitivity from those of incorrect source distribution, we will consider them together as the effect of having chosen a sub-optimal λ_E.

Since the dependence on rate follows 2^{-2R} scaling for either optimal point density λ_O or erroneous point density λ_E, we can quantify the effect of suboptimal design by a ratio of distortions D_E/D_O (independent of rate) or, equivalently, an excess rate $\Delta R = R_E - R_O$ to achieve the same distortion. Clearly $\Delta R = \frac{1}{2}\log_2(D_O/D_E)$. The optimal point density $\lambda_O \propto (f_X(x)g_i^2)^{1/3}$ achieves a distortion proportional to $\|f_X(x)g_i^2(x)\|_{1/3}$, while the erroneous design λ_E achieves a distortion proportional to $\int f_X(x)g_i^2\lambda_E^{-2}(x)dx$; in other words, $\|f_X(x)g_i^2\lambda_E^{-2}\|_1$. We may compute the rate loss with this information.

$$\begin{aligned}
\Delta R &= \frac{1}{2}\left(\log_2 \|\lambda_O^3 \cdot \|(f_X(x)g_i^2)^{1/3}\|_1^3 \cdot \lambda_E^{-2}\|_1 - \log_2 \|f_X(x)g_i^2\|_{1/3}\right) && (4.29)\\
&= \frac{1}{2}\left(\log_2 \left[\|(f_X(x)g_i^2)^{1/3}\|_1^3 \cdot \|\lambda_O^3 \cdot \lambda_E^{-2}\|_1\right] - \log_2 \|f_X(x)g_i^2\|_{1/3}\right) && (4.30)\\
&= \frac{1}{2}\left(\log_2 \left[\|f_X(x)g_i^2\|_{1/3} \cdot \|\lambda_O^3 \cdot \lambda_E^{-2}\|_1\right] - \log_2 \|f_X(x)g_i^2\|_{1/3}\right) && (4.31)\\
&= \frac{1}{2}\left(\log_2 \left\|\frac{\lambda_O^3}{\lambda_E^2}\right\|_1\right) && (4.32)
\end{aligned}$$

A couple points are worth noting about the form taken by this penalty. First of all, we may be comforted in that $\Delta R = 0$ when $\lambda_E = \lambda_O$. Second, as with the KL-divergence, this metric for "distance" between designs diverges if the erroroneous density is zero somewhere that the real density is nonzero.

4.3.2 Variable-Rate Erroneous Design

Unlike with fixed-rate quantization, variable-rate quantization has two components to it: a quantizer (described by the profile $\lambda(x)$) and an entropy coder (related to the entropy of the quantized output). We consider both in turn. First, the effect of incorrect quantization will be modeled as it was for the fixed-rate scenario: by comparing the erroneous point density λ_E with the optimal one, λ_O. After that, the effect of error on the entropy coding process will be added to the rate loss.

We will first work in terms of the rate loss assuming that the entropy coding is performed properly (that is, assuming that the coding is performed with the correct source distribution in mind). Proceeding as we did for fixed-rate, we may equate the distortions and obtain the difference between the quantization-faulty rate R_E and the correct quantizer's rate R_O. Note that we expect the latter to be smaller.

$$\begin{aligned}
D_E &= D_O && (4.33)\\
2^{-2R_E+2h(X)+\mathbf{E}\left[\log_2 \lambda_E^2\right]}\mathbf{E}\left[\frac{g_i^2(X)}{\lambda_E^2(X)}\right] &= 2^{-2R_O+2h(X)+\mathbf{E}\left[\log_2 g_i^2(X)\right]} && (4.34)\\
2^{-2R_E+2h(X)+\mathbf{E}\left[\log_2 \lambda_E^2\right]+\log_2 \mathbf{E}\left[g_i^2(X)/\lambda_E^2(X)\right]} &= 2^{-2R_O+2h(X)+\mathbf{E}\left[\log_2 g_i^2(X)\right]} && (4.35)\\
2R_E - 2R_O &= \mathbf{E}\left[\log_2 \lambda_E^2\right] + \log_2 \mathbf{E}\left[\frac{g_i^2(X)}{\lambda_E^2(X)}\right] \\
&\quad -\mathbf{E}\left[\log_2 g_i^2(X)\right] && (4.36)
\end{aligned}$$

Noting that $\lambda_O^2(x) = g_i^2(x) \cdot C$, where C is a normalization constant:

$$\begin{aligned} 2R_E - 2R_O &= \mathbf{E}\left[\log_2 \lambda_E^2\right] + \log_2 \mathbf{E}\left[\frac{\lambda_O^2(X)}{\lambda_E^2(X)}\right] - \mathbf{E}\left[\log_2 \lambda_O^2(X)\right] + \log_2 C - \log_2 C \quad (4.37) \\ &= \log \mathbf{E}\left[\frac{\lambda_O^2(X)}{\lambda_E^2(X)}\right] + \mathbf{E}\left[\log \lambda_E^2(X)\right] - \mathbf{E}\left[\log_2 \lambda_O^2(X)\right] \quad (4.38) \\ R_E - R_O &= \frac{1}{2} \log \mathbf{E}\left[\frac{\lambda_O^2(X)}{\lambda_E^2(X)}\right] + D(f_X \| \lambda_E) - D(f_X \| \lambda_O) \quad (4.39) \end{aligned}$$

In the last step, we observe the emergence of the KL divergence $D(\cdot\|\cdot)$, a term seen frequently in rate loss expressions. Note, as we did in the derivation for selective/symmetric functions in Chapter 3, that the point densities λ_E and λ_O are being used as if they were probability densities.

We may add to this the effect of rate loss in the encoder. Thus far, the expression $R_E - R_O$ indicates the amount of rate that must be added for the erroneous decoder to catch up to the distortion performance of the optimal one. However, if the source distribution is not known correctly, some more rate will have to be added for it to catch up to the performance of the optimal entropy coder. We denote this second gap $R_{EE} - R_E$, and quote a well-known result [3] that the rate loss associated with designing for an incorrect probability mass function p_E instead of the correct one p_X is the divergence between the two. In our case, the PMF's of interest are those over the quantized representation $\widehat{X}$ according to the two different pdf's:

$$R_{EE} - R_E = D(p_X \| p_E)$$

We may obtain p_X and p_E from the associated probability density functions, f_X and e_X, by means of the high-rate approximation:

$$
\begin{aligned}
R_{EE} - R_E &= \sum_{\widehat{X}} p_{\widehat{X}}(\widehat{X}) \frac{p_{\widehat{X}}(\widehat{X})}{p_E(\widehat{X})} && (4.40)\\
&\approx \sum_{\widehat{X}} f_X(\widehat{X})\Delta(\widehat{X}) \frac{f_X(\widehat{X})\Delta(\widehat{X})}{e_X(\widehat{X})\Delta(\widehat{X})} && (4.41)\\
&\approx \sum_{\widehat{X}} f_X(\widehat{X}) \int_{x|Q(x)=\widehat{X}} \frac{f_X(\widehat{X})}{e_X(\widehat{X})} dx && (4.42)\\
&\approx \sum_{\widehat{X}} \int_{x|Q(x)=\widehat{X}} f_X(x) \frac{f_X(x)}{e_X(x)} dx && (4.43)\\
&= \int_0^1 f_X(x) \frac{f_X(x)}{e_X(x)} dx && (4.44)\\
&= D(f_X \| e_X) && (4.45)
\end{aligned}
$$

Once again, the KL divergence is equal to the rate loss. Adding the loss from both steps, we have the total loss from incorrect pdf e_X and incorrect quantization profile λ_E as:

$$R_{EE} - R_O = D(f_X \| e_X) + D(f_X \| \lambda_E) - D(f_X \| \lambda_O) + \frac{1}{2} \log \mathbf{E}\left[\frac{\lambda_O^2(X)}{\lambda_E^2(X)}\right]$$

4.A Proof of Quasi-triangle-inequality

Let $x(t)$ and $y(t)$ be functions from $\mathbb{R} \to \mathbb{R}$. We will demonstrate the quasi-triangle inequality:

$$\|x(t) + y(t)\|_{1/3} \leq 4(\|x(t)\|_{1/3} + \|y(t)\|_{1/3})$$

First, we prove the relation $(x+y)^3 \leq 4(x^3 + y^3)$:

$$
\begin{aligned}
(x+y)^3 - 4(x^3+y^3) &= x^3+y^3+3x^2y+3xy^2-4x^3-4y^3 && (4.46)\\
&= 3(-x^3-y^3+x^2y+xy^2) && (4.47)\\
&= 3(x^2(y-x)-y^2(y-x)) && (4.48)\\
&= 3(x^2-y^2)(y-x) && (4.49)\\
&= 3(x+y)(x-y)^2 && (4.50)\\
&\geq 0 && (4.51)
\end{aligned}
$$

We now prove the triangle equation. By the relation we have just demonstrated:

$$\left(\int x(t)^{1/3}dx\right)^3 + \left(\int y(t)^{1/3}dx\right)^3 \geq \frac{1}{4}\left(\int \left(x(t)^{1/3}+y(t)^{1/3}\right)dx\right)^3$$

Using the concavity $\cap$ of the function $t^{1/3}$, each term in the integral may be lower bounded:

$$\frac{1}{4}\left(\int \left(x(t)^{1/3}+y(t)^{1/3}\right)dx\right)^3 \geq \frac{1}{4}\left(\int (x(t)+y(t))^{1/3}dx\right)^3 = \frac{1}{4}\|x(t)+y(t)\|_{1/3}$$

This concludes the proof.

Chapter 5

Graduation

We began by considering two specific problems: quantization of data in an analog-to-digital converter, and distributed source coding in sensor networks. Both problems were given a "functional twist" when we noted that the user would likely be interested in a function of the data more than the data itself. This difference is particularly pronounced when we consider human end-users, who are physically incapable of caring about high-rate data.

From here, we generalized to the abstract functional quantization scenario represented in Fig. 0-1. With the firepower of the high-resolution quantization analysis, we attacked this problem in increasingly unconstrained forms. Eventually, the notion of functional typicality was introduced as an alternative route to our derivations. We then applied and extended these results to a wide variety of situations, some showing surprising improvements over ordinary quantization techniques. Non-monotonic functions were dealt with by means of high-resolution *non-regular* quantization, a notion and a quantitative picture that we introduced. The functional version of the transform coding problem was considered, and solved with functional quantization techniques under the constraint of uniform quantization. Encoder collaboration was also explored and found to demonstrate strikingly different behavior between variable-rate and fixed-rate coding. Finally, we considered the effect of imperfect implementations and imperfect knowledge on the performance of a functional quantization system.

The clearest message from these results is a strong endorsement for the high-resolution quantization approach to compression problems. While these results are built on continuous approximations and are therefore not as precise as those of rate-distortion theory, one can obtain meaningful solu-

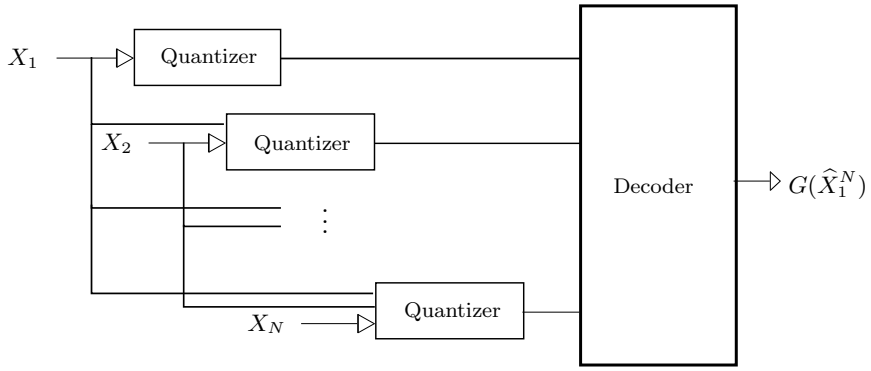

Figure 5-1: A sequential source with memory. The ith encoder knows the value of X_j for $j \leq i$.

tions to a wide variety of problems where exact approaches fail. This manifests itself in the ease with which functional quantization extends to new scenarios.

For instance, consider the problem of a sequential encoder with memory (see Fig. 5-1) — which may better represent an analog-to-digital converter than our original picture. The encoder first encodes X_1 into $\widehat{X}_1$ according to some quantization profile, and sends the latter to the decoder. From our quantization picture, we know that the best quantizer — variable- or fixed-rate — is one that works with source distribution $f_{X_1}(x)$ and functional sensitivity $g_1^2(x)$. Next, X_2 is encoded into $\widehat{X}_2$ by the same encoder. Since both the decoder and the encoder remember the value of $\widehat{X}_1 \approx X_1$, the best quantizer for X_2 works with source distribution $f_{X_2|X_1}(x \mid X_1 = x_1)$ and similarly conditioned functional sensitivity $g_2^2(x \mid X_1 = x_1)$. In general, the Nth quantizer is obtained from source distribution $f_{X_N|X_1^{N-1}}(x \mid X_1^{N-1} = x_1^{N-1})$ and functional sensitivity $g_2^2(x \mid X_1^{N-1} = x_1^{N-1})$. The problem has been solved almost trivially, due to the functional quantization framework.

We close by considering several extensions to the functional quantization theory. As demonstrated by the example above, there is no shortage of directions in which this work can take. Amongst these options, we believe the three topics below have the greatest potential for both practical and theoretical impact.

Universality. Most of the situations we have considered, with the exception of the imperfect design considerations of Sec. 4.3, assume that the probability distribution of the source is known to both the encoder and the decoder. In reality, this is rarely the case. From this comes motivation for *universal* functional quantization, where the encoder and decoder adapt to the distribution of the source as it is repeatedly sampled and quantized. It is not obvious what algorithm should be used

to refine the distribution estimate, nor is it clear what the fundamental limitations on this are. We note that it would be particularly interesting to see how functional considerations affect the design of an estimator.

Complexity Constraints. Quantization, over its history, has taken two very different directions. In one room are the theorists, chasing fundamental limits and optimal quantizers. In the other room are those interested in practical compression schemes. This thesis has fallen largely into the realm of the former. For real-world systems to take advantage of our results, the latter must be embraced as well. To this end, we suggest that the various constructions for ordinary lossy coding be investigated in a functional context.

Specific Applications. There are several potential applications of functional quantization that are interesting not just from a practical standpoint, but in the modifications to the theory that they encourage. In control theory: within a feedback loop attempting to drive the output of a system to a specific value, how should the observations be quantized? The feedback structure of this problem complicates things greatly, but it also suggests an extension of FQ to network problems where $\widehat{G}$ is itself subject to computation at a future node.

One might also consider quantization of continuous data that is to be lossy compressed according to a fixed algorithm in the discrete realm. Audio coding, for instance: how should a microphone's ADC quantize its voltage levels given that further compression will be taking place? Or compressive sampling: how should random linear projections of a vector be quantized if they will be used to recover the vector?

Bibliography

[1] D. Gabor. Guest editorial. *Information Theory, IEEE Transactions on*, 5(3):97–97, Sep 1959.

[2] Rahul Sarpeshkar. Analog versus digital: Extrapolating from electronics to neurobiology. *Neural Comput.*, 10(7):1601–1638, October 1998.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[4] Robert M. Gray and David L. Nehoff. Quantization. *IEEE Trans. Inf. Theory*, 44(6):2325–2383, October 1998.

[5] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1992.

[6] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Second Edition*. The MIT Press, September 2001.

[7] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

[8] David Slepian and Jack K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inf. Theory*, IT-19(4):471–480, July 1973.

[9] G. Truskey, F. Yuan, and D. Katz. *Transport Phenomena in Biological Systems*. Prentice Hall, 2004.

[10] A. Grodzinsky. *Fields, Forces, and Flows in Biological Systems*. Unpublished Manuscript, 2007.

[11] Vivek K Goyal. Theoretical foundations of transform coding. *IEEE Signal Process. Mag.*, 18(5):9–21, September 2001.

[12] Hua Xie and Antonio Ortega. Entropy- and complexity-constrained classified quantizer design for distributed image classification. In *IEEE Workshop on Multimedia Sig. Process*, pages 77–80, December 2002.

[13] Lavanya Vasudevan, Antonio Ortega, and Urbashi Mitra. Application-specific compression for time delay estimation in sensor networks. In *SenSys '03: Proc. 1st Int. Conf. Embedded Netw. Sensor Syst.*, pages 243–254, New York, NY, USA, 2003. ACM.

[14] S. Kassam. Optimum quantization for signal detection. *IEEE Trans. Commun.*, 25(5):479–484, May 1977.

[15] Jia Li, N. Chaddha, and R.M. Gray. Asymptotic performance of vector quantizers with a perceptual distortion measure. *Information Theory, IEEE Transactions on*, 45(4):1082–1091, May 1999.

[16] Aaron B. Wagner, Saurabha Tavildar, and Pramod Viswanath. Rate region of the quadratic Gaussian two-terminal source-coding problem. *IEEE Trans. Inf. Theory.* submitted.

[17] Aaron D. Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. Inf. Theory*, IT-22(1):1–10, January 1976.

[18] Ram Zamir. The rate loss in the Wyner-Ziv problem. *IEEE Trans. Inf. Theory*, 42(6):2073–2084, November 1996.

[19] David Rebollo-Monedero, Shantanu Rane, Anne Aaron, and Bernd Girod. High-rate quantization and transform coding with side information at the decoder. *Signal Process.*, 86(11):3160–3179, 2006.

[20] Hanying Feng, Michelle Effros, and Serap A. Savari. Functional source coding for networks with receiver side information. In *Proc. Allerton Conf. Commun. Control Comput.*, September 2004.

[21] Emin Martinian, Gregory Wornell, and Ram Zamir. Source coding with encoder side information. *IEEE Trans. Inf. Theory*, December 2004. submitted.

[22] Hirosuke Yamamoto and Kohji Itoh. Source coding theory for multiterminal communication systems with a remote source. *Trans. IECE Japan*, E63(10):700–706, October 1980.

[23] James A. Bucklew. Multidimensional digitization of data followed by a mapping. *IEEE Trans. Inf. Theory*, IT-30(1):107–110, January 1984.

[24] Alon Orlitsky and James R. Roche. Coding for computing. *IEEE Trans. Inf. Theory*, 47(3):903–917, March 2001.

[25] Vishal Doshi, Devavrat Shah, Muriel Medard, and Sidharth Jaggi. Distributed functional compression through graph coloring. In *Proc. Data Compression Conf. (DCC 2007)*, pages 93–102, Snowbird, Utah, March 2007.

[26] Allen Gersho. Asymptotically optimal block quantization. *IEEE Trans. Inf. Theory*, IT-25(4):373–380, July 1979.

[27] Benjamin Farber and Kenneth Zeger. Quantization of multiple sources using nonnegative integer bit allocation. *IEEE Trans. Inf. Theory*, 52(11):4945–4964, November 2006.

[28] T. Linder, R. Zamir, and K. Zeger. High-resolution source coding for non-difference distortion measures: multidimensional companding. *Information Theory, IEEE Transactions on*, 45(2):548–561, Mar 1999.

[29] Yoshio Yamada, Saburo Tazakia, and Robert M. Gray. Asymptotic performance of block quantizers with difference distortion measures. *IEEE Trans. Inf. Theory*, IT-26(1):6–14, January 1980.

[30] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423, 623–656, July/Oct. 1948.

[31] Harald Cramer. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1946.

[32] H. Akcay, H. Hjalmarsson, and L. Ljung. On the choice of norms in system identification. *Automatic Control, IEEE Transactions on*, 41(9):1367–1372, Sep 1996.

[33] Naoki Saito. The generalized spike process, sparsity, and statistical independence, 2001.

[34] V. Prabhakaran, K. Ramchandran, and D. Tse. On the role of interaction between sensors in the ceo problem. In *Annual Allerton Conference on Communication, Control, and Computing*.