ASYMPTOTIC QUANTIZATION AND APPLICATIONS TO SENSOR NETWORKS

by

Daniel Marco

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2004

Doctoral Committee:

Professor David L. Neuhoff, Chair
Professor Arthur G. Wasserman
Associate Professor Serap A. Savari
Assistant Professor Mingyan Liu
Assistant Professor Sandeep P. Sadanandarao

Logic, logic, logic. Logic is the beginning of wisdom, Valeris, not the end.

— Captain Spock, Star Trek VI: The Undiscovered Country

To my parents and my brother,
whose love is never ending.

ACKNOWLEDGEMENTS

I would like to first and foremost extend my deepest gratitude to my advisor Professor David Neuhoff for his patience and for his guidance. I thank him for investing so much time and effort, and especially for his rigor and insistence on precision. I am extremely grateful for having the privilege to work with him and learn from his expertise. I also thank Professors Wasserman, Savari, Liu and Sadanandarao (Pradhan) for serving on my doctoral committee.

I would also like to thank my friends, among them, Dimitri, Sam, Marwan, Navid, Paul, Vicki, Bob, Diane, Damian, Kostas, Doron Hai and Doron Blatt, for all the good times. Special thanks goes to Tudor for his good friendship and for introducing me to the world of ballroom dancing, to Alon and Elad for their long lasting friendships, and to Uri for always being there and for all his encouragement. Additionally, I would like to thank Ayis, Eric, John and Kevin Buell for all the fun we had playing chess and trash talking while doing so. Thanks also goes to Marc and Kevin Holt for laying a hand on top of Mount Fuji. Further, I would like to thank Whit Gray for his kindness.

I also deeply thank my dance instructors Stephen and Susan McFerran, without whom this experience would not have been nearly as enjoyable. In particular, I would like to thank Stephen for being the funniest person I know and making me laugh like no other. Also, I thank my dance partners Nina and Lillian for their patience and for all the good dancing.

I must, of course, thank Panchero's for keeping me well fed with their delicious burritos, and Potbelly's for serving the best milkshakes in town. Thanks need also go to the anonymous juggler from Boston, who introduced me to juggling, and to Dave, Fred, Marc, Bill, Ben and Ajit, from the Ann Arbor juggling club, for showing me a whole bunch of tricks.

To my dear brother, Talmon, I extend the warmest thank you for his love and support, for his generosity, and for teaching me the ins and outs of the airline industry and making my visits home possible. Finally, no words can express my gratitude to my parents, who brought me up, given me endless love, and supported me in all my endeavors.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

ASYMPTOTIC QUANTIZATION AND APPLICATIONS TO SENSOR NETWORKS

by

Daniel Marco

Chair: David L. Neuhoff

This dissertation considers three asymptotic scalar quantization problems, the last of which is applied to sensor networks.

First, the widely used additive noise model for high resolution uniform scalar quantizers is considered. Although this model is frequently used, its validity has never been rigorously demonstrated. This dissertation does so by finding conditions on the input density under which this model is valid (e.g. continuity). In addition, alternate models are provided for cases that this model is invalid.

Secondly, the operational rate-distortion function of scalar quantization in the low resolution domain of low rate and high distortion is examined. Little is known about it except that it equals zero when distortion equals the source variance. It is shown that for stationary memoryless Gaussian sources, it approaches zero with the same slope as Shannon's rate-distortion function, thus implying that scalar quantization is asymptotically, as distortion tends to source variance, optimal – a fact not previously known.

Next, sampling and identically scalar quantizing a stationary random process over a finite interval is considered. The question is if N samples are taken in the interval, what happens to their joint entropy as N goes to infinity? This is not obvious, since it is the product of the N^{th} order entropy times N, where the former tends to zero. It is shown that this product tends to infinity under a very mild condition. The rate at which this happens is upper bounded in the case of uniform quantizers and a Gaussian process, by deriving an asymptotic formula for the conditional entropy of one quantized sample conditioned on another.

Finally, field-gathering sensor networks, whose sensors use identical scalar quantizers, are examined. Their purpose is to transport quantized snapshots of a field to a collector, where the field is reconstructed. The question is with what frequency can such snapshots be transported, subject to a fidelity constraint. It is shown, using the joint entropy result, that as sensor density increases to infinity, the frequency goes to zero. This implies that beyond some optimal density, sensors should be suppressed. Furthermore, using the conditional entropy formula, an upper bound is found for the rate at which frequency goes to zero in the case of uniform quantizers and a Gaussian field.

CHAPTER I

Introduction

This dissertation is composed of six chapters, the core of which are the four middle chapters, each of which has its own introduction and its own list of references. Necessary notation is also introduced independently in each of these chapters. Chapters II – IV are self contained manuscripts, of which the first two have been submitted for publication. These three chapters are concerned with various types of asymptotic quantization problems and Chapter V applies the results of Chapter IV to sensor networks. The last chapter, Chapter VI, summarizes the contributions of this dissertation and considers possible future avenues of research. The present chapter, which is the first, has its own list of references as well and provides a general introduction to communications, source coding, quantization and sensor networks, and lists the main contributions of this dissertation. Those readers who are well familiar with these subjects, and are in particularly familiar with quantization theory, may skip the first three sections of this introduction and go directly to the last section, where the contributions are listed.

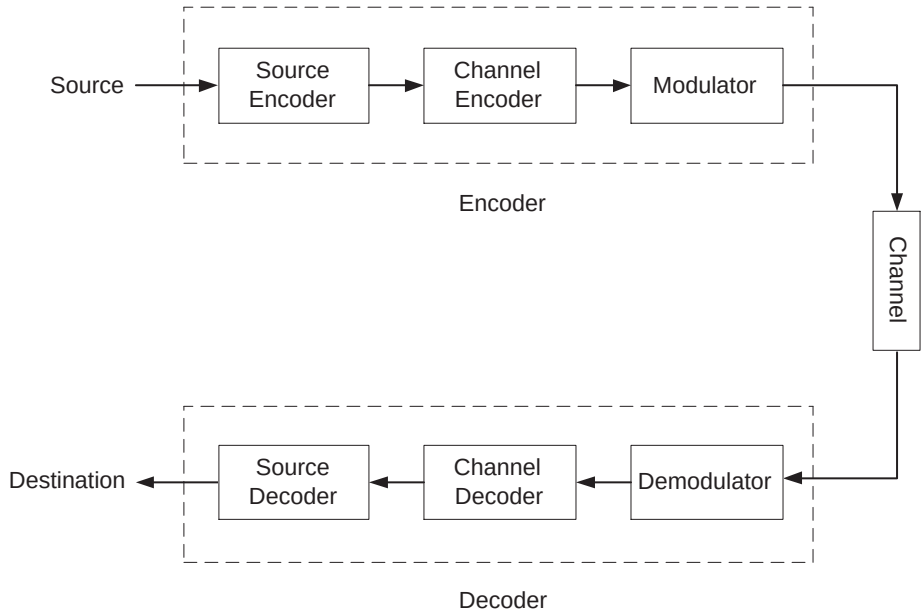

Figure 1.1: A point-to-point communication system.

1.1 Communication Systems and Coding

The purpose of a point-to-point communication system is to reliably and efficiently transfer information (e.g. radio talk show, winning lottery numbers, the weather report) from one location, referred to as the *source*, to another location, referred to as the *destination*, over some noisy channel (e.g. the atmosphere, telephone line, optic fiber). Figure 1.1 depicts such a standard communication system.

As can be seen from the figure, the information to be sent is first encoded, then transmitted over the channel, and finally decoded at the destination. The encoding procedure is composed of three stages: *source encoding*, *channel encoding* and *modulating*. The goal of the first is to compress the information into bits, the goal of the second is to make the information robust to channel corruption, and the goal of the third is to translate the channel encoded stream into a stream that is consistent with the channel (e.g. turn bits into voltage). The decoding procedure performs the

reverse operations in reverse order. Specifically, the received stream is first demodulated, so as to obtain an input stream that conforms to the same alphabet used by the channel encoder, then the channel decoder operates on the demodulated stream, and finally source decoding is performed to obtain a reconstruction of the original information.

The information that is generated at the source location is normally modeled as a random process, either continuous-time or discrete-time, and referred to simply as the *source*. Sources are often modeled as stationary and/or memoryless. The former means that the source statistics do not depend on time, and the latter means that the source value at a given time does not depend on past values. For sources with memory, the length of the memory is the length of time, or number of past samples, if the source is continuous-time or discrete-time, respectively, which effect the present source value. Some sources have infinite memory.

Similar to sources, channels are also modeled in a probabilistic manner. The most commonly used channel model is a stationary memoryless channel, which is described using a conditional probability density or mass function, depending on whether the channel is discrete or continuous, respectively. Figure 1.2 illustrates one of the simplest and most common channels, known as the binary symmetric channel, which is a stationary and memoryless channel, where the input and output take values 0 and 1, and $p(1|0) = p(0|1) = e$, $p(0|0) = p(1|1) = 1 - e$, are the conditional probabilities of the channel (also called transition probabilities). In this case, e is called the crossover probability.

We observe that the goals of the source and channel encoders are contradictory in nature. The first aims at reducing redundancy from the raw information, so as to produce the shortest description possible (data compression), while the second

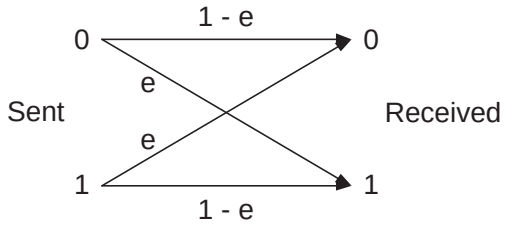

Figure 1.2: Binary symmetric channel with crossover probability e.

aims at adding redundancy to the raw information so as to protect it from channel corruption (error protection).

In 1948 Shannon showed [1] that for ordinary sources and channels, source coding and channel coding can be performed separately and sequentially, while maintaining optimality. Namely, the performance of a separate and sequential scheme does not degrade relative to that of a joint source-channel coding scheme (which is, clearly, at least as good as the former scheme). This is known as the *separation theorem*. Thus, Figure 1.1, ordinarily, depicts an optimal communication system.

A large portion of information theory is dedicated towards obtaining good source coding and channel coding performances. In both cases, there are theoretical limits to performance. Shannon's famous channel coding theorem [1] associates with certain channels a quantity called *channel capacity*, denoted as C, which is the largest rate at which communication is possible with arbitrarily small probability of error, where rate is measured by the number of bits transmitted per channel use. Large rates are desirable. The capacity of a channel depends on its probabilistic behavior. For example, consider a stationary discrete memoryless channel, whose input X has distribution $p(x)$, and whose output Y is obtained via the conditional probability distribution $p(y|x)$. Its capacity [2] (p. 184) is given by

$$C \;=\; \max_{p(x)} I(X;Y) \;,$$

where $I(X;Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$ is the mutual information between X and Y [2] (p. 18), $(p(x,y), p(y)$ are obtained from $p(x), p(y|x))$, and the maximum is taken over all possible input distributions $p(x)$.

Similarly, source coding has theoretical limits as well. Specifically, when the source is discrete-time, discrete-valued, stationary and memoryless (i.e. its samples are independently and identically distributed – i.i.d.), we identify a quantity called *entropy*, denoted as H, which is the least rate at which the output of the source can be described with no error, i.e. *losslessly* (for a detailed discussion on entropy see Chapters 2 and 4 in [2]). Here, rate is measured by the number of bits per source symbol, and small rates are desirable. The entropy H of such a source, X, is given by,

$$H(X) \;=\; \sum_{x} p(x) \log_2 \frac{1}{p(x)} \;,$$

where $p(x)$ is the probability of the source symbol x. Note that this, in fact, is the entropy of a random variable that has a probability mass function p.

In general, for stationary sources we define the N^{th} order entropy, denoted H_N, to be the joint entropy of N source samples divided by N. Namely,

$$\begin{aligned} H_N(X) \;&=\; \frac{1}{N} H(X_1, X_2, \ldots, X_N) \\ &=\; \frac{1}{N} \sum_{x_1, x_2, \ldots, x_N} p(x_1, x_2, \ldots, x_N) \log_2 \frac{1}{p(x_1, x_2, \ldots, x_N)} \;, \end{aligned} \tag{1.1}$$

where $p(x_1, x_2, \ldots, x_N)$ is the probability of the N-tuple of source symbols $(x_1, x_2, \ldots, x_N)$. Finally, (1.1) can be extended in a natural way by considering infinite order entropy, which is called entropy-rate and denoted H_∞. It is given by

$$H_\infty(X) \;=\; \lim_{N \to \infty} H_N(X) \;. \tag{1.2}$$

Let us further define for two discrete random variables X and Y with joint distribu-

tion $p(x, y)$, the *conditional entropy* of Y given X to be

$$H(Y \mid X) \;=\; \sum_x H(Y \mid X = x)\, p(x) \;=\; \sum_{x,y} p(x, y) \log_2 p(y|x) \;.$$

This definition extends straightforwardly to conditional entropies that condition on more than one random variable. We comment that conditioning never increases entropy. We comment further that for stationary processes the limit in (1.2) equals $\lim_{N\to\infty} H(X_N \mid X_{N-1}, X_{N-2}, \ldots, X_1)$.

The entropy of a source is a measure of its uncertainty. That is, the larger the entropy the greater the uncertainty, and hence the greater the rate required to describe the source output. It is not hard to see that continuous-valued sources cannot be described losslessly. For example, it would take an infinite number of bits to describe the number $\sqrt{2}$, when it is a possible outcome of a source that is equally likely to assume any number in the interval $(0, 2)$.

Since continuous-valued sources cannot be described losslessly, the question that naturally arises is: What is the least rate with which a continuous source can be described given a permitted level of fidelity? In fact, this question is valid for discrete-valued sources as well. This type of source coding is known as *lossy* source coding, whereas when sources are described exactly with perfect fidelity, the coding is called *lossless* source coding. To analyze lossy source coding, fidelity is quantified by a nonnegative function from the source alphabet crossed with the reconstruction alphabet (notice that these need not be the same) to $[0, \infty)$. This function is called a *distortion measure*, and is denoted by $d(x, \hat{x})$, where x represents a source symbol and $\hat{x}$ a reconstruction symbol.

The answer to the above question comes in the form of *rate-distortion theory* [3]. Specifically, for a given source, $\mathcal{R}(D)$ is the least rate of any encoding scheme that

achieves on average distortion D or less for this source. $\mathcal{R}(D)$ is called the *rate-distortion function* (also referred to as Shannon's rate-distortion function). Different sources have different rate-distortion functions. For example, the rate-distortion function for a memoryless Gaussian source with variance σ^2 and squared error distortion measure [4] (p. 477) is

$$\mathcal{R}(D) = \begin{cases} \frac{1}{2}\log_2 \frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2 \\ 0, & D > \sigma^2 \end{cases}.$$

The focus of this dissertation is on the source coding part of the communication system, hence we shall restrict our attention to that. Source coding comes in various flavors. For example, a discrete-time and discrete-valued source may be coded by encoding a fixed length block of source symbols into a fixed length block of code symbols. This is called fixed-length to fixed-length block coding (FFB). Similarly, one may use FVB, VFB or VVB coding.

The next section is concerned with lossy source coding, and specifically, with a particular kind of lossy source coding known as *quantization*, which is the main focus of this dissertation.

1.2 Quantization

A quantizer consists of four parts, as illustrated in Figure 1.3. A countable (finite or infinite) partition of the source space $\mathbb{R}^k$, $k \geq 1$ (sometimes referred to as a lossy encoder), a lossless encoder, a lossless decoder and a codebook (sometimes referred to as a lossy decoder). k is called the dimension of the quantizer and is the number of source samples that are jointly quantized. The partition and lossless encoder are sometimes referred to as the encoder of the quantizer. The lossless decoder and the codebook are often referred to as the decoder of the quantizer. The elements of the

collection of subsets of $\mathbb{R}^k$ that form the partition of $\mathbb{R}^k$, (which are disjoint and whose union equals $\mathbb{R}^k$) are called *quantization cells* and are denoted $S_1, S_2, \ldots S_M$, where M may be infinite. The codebook consists of M points in $\mathbb{R}^k$ called *codevectors*.

The input to the partition is a sequence of blocks of k source samples, each of which represents a point in $\mathbb{R}^k$. The output of the partition is a sequence of integers called *quantization indices*, which represent the quantization cells in which the corresponding blocks of source samples lie. The input to the lossless encoder is the sequence of quantization indices that are encoded (either individually or in blocks) into bits, which are the output of the lossless encoder. These bits are the input to the lossless decoder, whose output are the corresponding quantization indices. Finally, these quantization indices are the input to the codebook, which are mapped (either individually or in blocks) into codevectors that are the output of the codebook. We will not consider the case that the codebook maps block of quantization indices into codevectors, but rather assume throughout that each quantization index is mapped individually into a codevector. In such a case the codevectors are called *reconstruction vectors*, the set of which is denoted $\mathcal{Y} = \{\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_M\}$.

The *quantization rule* is a mapping $q : \mathbb{R}^k \to \mathcal{Y}$, $k \geq 1$. where $\mathcal{Y}$ is the countable set of reconstruction vectors. Associated with each reconstruction vector, $\underline{y}_i$, is a quantization cell, S_i, which is its inverse image. Thus, $q(\underline{x}) = \underline{y}_i$ if and only if $\underline{x} \in S_i$. When we refer to a quantizer, sometimes we shall refer only to the quantization rule, i.e. to the partition and the reconstruction vectors, and allow for various lossless encoders and decoders, which are jointly referred to as the *lossless code*. When doing so, it will be clear from context.

Since the space $\mathbb{R}^k$ is uncountable, and since there are only a countable number of reconstruction vectors, it follows that quantization is a lossy source coding technique.

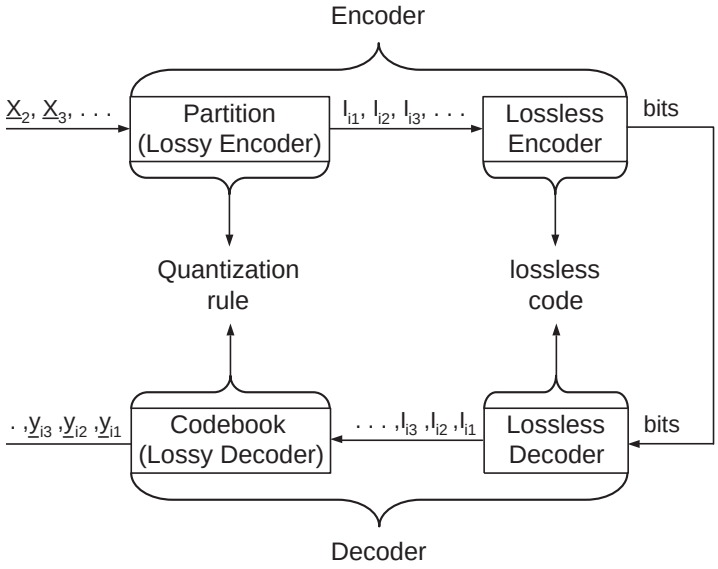

Figure 1.3: A quantizer scheme – consisting of a partition, lossless encoder, lossless decoder, and a codebook.

We thus refer to the partition part of the quantizer as a lossy encoder. Consequently, the encoder of the quantizer can be viewed as having two cascaded encoders. First, a lossy encoder that maps real-valued source inputs into quantization indices, followed by a lossless encoder that encodes quantization indices into bits.

The performance of a quantizer with quantization rule q, is measured by its rate and distortion, where rate is the average number of bits per source sample, and distortion is given by

$$D(q) \;=\; E\big[d_k(\underline{\mathrm{x}}, q(\underline{\mathrm{x}}))\big] \;=\; \sum_i \int_{S_i} d_k(\underline{\mathrm{x}}, \underline{\mathrm{y}}_i) f(\underline{\mathrm{x}})\, d\underline{\mathrm{x}} \;,$$

where $d_k(\cdot,\cdot)$ is a distortion measure, and f is the probability density function (pdf) of the source. The most commonly used distortion measure is per sample squared-error: $d_k(\underline{\mathrm{x}}, \hat{\underline{\mathrm{x}}}) = \frac{1}{k}\sum_{j=1}^{k}(x_j - \hat{x}_j)^2$. The distortion, of course, is only due to the partition and the codebook, i.e. due to the first encoder and the second decoder, which are lossy. The rate, on the other hand, depends on the design of both the lossy and lossless encoders. Low rate and low distortion are desirable.

The rate of a quantizer, as mentioned, is the average number of bits per source input, or equivalently, the average number of bits needed to encode a quantization index. Two main methods for encoding quantization indices are fixed-length and variable-length coding (also referred to as fixed-rate and variable-rate coding, respectively). The former allots a fixed number of bits to each quantization index. Clearly, such a method requires a finite number of reconstruction vectors. The latter, assigns a variable number of bits to each quantization index, thus permitting an infinite number of reconstruction vectors.

The rate of an M level quantizer with fixed-length coding is $R = \frac{1}{k}\lceil \log_2 M \rceil$, where $\lceil \log_2 M \rceil$ is the number of bits needed to represent M different outcomes. The rate of a quantizer with variable-length coding is $R = \frac{1}{k}\sum_i L_i P_i$, where L_i is the length of the string of bits representing the i^{th} quantization index, and $P_i = \int_{S_i} f(\underline{x})\, d\underline{x}$ is the probability of the i^{th} quantization cell.

When using variable-length coding, it is often the case that blocks of quantization indices are jointly encoded. If the size of the block is N, then we say that the encoder is of order N. In such a case, the rate is given by $R = \frac{1}{k}\frac{1}{N}\sum_{\underline{i}} L_{\underline{i}} P_{\underline{i}}$, where $\underline{i}$ is an N-tuple of quantization indices, $L_{\underline{i}}$ is the length of the string of bits representing this N-tuple, and $P_{\underline{i}}$ is the N-tuple's probability.

A scheme that utilizes variable-rate coding is often called quantization with entropy coding, since variable-rate encoders can be designed so as to achieve rate close to the entropy of the quantizer output, i.e. the entropy of the quantization indices (e.g. an encoder that uses a Huffman code [5]). Thus, a quantizer with N^{th} order entropy coding is a quantizer whose lossless encoder encodes blocks of N quantization indices into a number of bits that is close to the joint entropy of a block of N indices.

The *operational rate-distortion function* of a family of quantizers is defined to be the the least rate for a given distortion attainable by the given family. Thus, for example, one may choose to consider the operational rate-distortion function of quantizers of dimension 5 that use variable rate-coding and have a finite number of cells. When we describe a family of quantizers, we will specify its limitations; for example, we shall specify that the quantizers in the family use fixed-rate coding if only fixed-rate coding is considered. If both variable-rate and fixed-rate coding are considered there would be no reference to either. In comparison to the operational rate-distortion function, the Shannon rate-distortion function is the least rate for a given distortion, attainable by any source coding scheme whatsoever.

Optimizing the performance of a quantizer involves minimizing both the distortion of the lossy encoder and the rate of the lossless encoder that follows. When using quantization with entropy coding, it is often the case that the quantizer is designed subject to a constraint on its output entropy. This is known as entropy-constrained quantization.

When using fixed-rate quantization, optimizing performance reduces to minimizing quantizer distortion (since rate is fixed). In 1957 Lloyd [6] established necessary and sufficient conditions for local optimality of quantization with fixed-rate coding, i.e. so that small perturbations to the partition or reconstruction vectors would result in distortion increase. Clearly these conditions are necessary for globally optimal quantizers as well. The conditions insure that the quantizer partition is optimal for the given reconstruction vectors and vice versa. For mean-squared error distortion the first condition translates to the *nearest neighbor* condition, i.e. a source sample should be mapped to the closest reconstruction vector, and the second condition becomes the *centroid* condition, i.e. the reconstruction vector of a cell should be its

centroid, where the centroid of a cell is the expected value of the source given that it lies in the specified cell, which in turn equals the minimum mean-squared error estimator of the source sample given that it lies in the given cell. (Another way of viewing the cell centroid is as the center of gravity of the cell, where the source pdf is the weighing function). It is well-known that having reconstruction vectors at cell centroids minimizes the mean-squared error induced by the quantizer.

For the case of variable-rate quantization, i.e. entropy-constrained quantization, Lloyd's centroid condition is still necessary for the quantizer to be optimal, that is, given a partition the reconstruction vectors must be the cell centroids. The condition for the partition, however, is somewhat more complicated as shown in [7]. Specifically, the partition must satisfy for all $\underline{x} \in \mathbb{R}^k$ and $1 \leq j \leq M$

$$d_k(\underline{x}, q(\underline{x})) - \lambda \log_2 P_i \;\leq\; d_k(\underline{x}, q(\underline{y}_j)) - \lambda \log_2 P_j \;,$$

where $\lambda > 0$ may be chosen arbitrarily. The choice of λ determines the tradeoff between rate and distortion. If, however, we choose $\lambda = 0$, then this reduces to the fixed-rate case. We observe that although the above is well defined as an optimality condition, it is circular in terms of designing the partition.

When $k > 1$ the quantizer is called a *vector quantizer*. Figure 1.4 illustrates the quantization rule of a two-dimensional vector quantizer, whose cells are represented by polygons or open polygons, and the reconstruction vector associated with a given cell is denoted by a dot within the cell.

When $k = 1$, the quantizer is called a *scalar quantizer*, for which single real-valued inputs are quantized separately. Thus, the quantization rule of a scalar quantizer is a mapping $q : \mathbb{R} \to \mathcal{Y}$, where $\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}$ is the set of reconstruction levels, and is a subset of $\mathbb{R}$. The cells of such a quantizer are usually intervals, whose

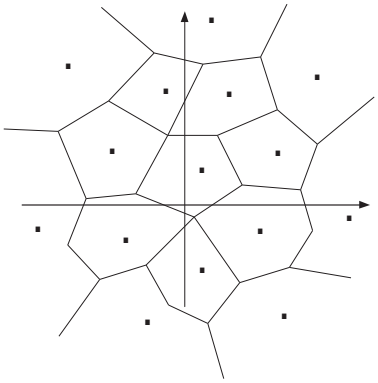

Figure 1.4: The quantization rule of a two-dimensional vector quantizer.

endpoints are called *thresholds*. We sometimes denote the left threshold of the i^{th} cell by t_i, and its right threshold by t_{i+1}. The reconstruction levels normally lie inside the cells that map to them, and can be chosen in various ways. The two most common choices for reconstruction levels are cell midpoints or cell centroids. While, as mentioned, the latter minimizes mean-squared error, the former is the simplest and most common choice of reconstruction levels.

A scalar quantizer that has infinitely many cells, all of which are of equal size (i.e. length), is called an *infinite-level uniform threshold scalar quantizer*. The quantization rule of such a quantizer is characterized by three parameters: The size of its cells, which is often called its *step size* and denoted by Δ, the set of reconstruction levels, and a number between zero and one, which we call *offset*. The offset is the fractional position of the origin within its quantization cell. Thus, the left threshold of the i^{th} cell is given by $t_i = (i - \theta)\Delta$, where θ denotes the offset. Figure 1.5 illustrates the quantization rule of an infinite-level uniform threshold scalar quantizer with step size Δ, offset zero, and reconstruction levels at cell midpoints.

A well-known result, first shown by Bennett in 1948 [8], is that the mean-squared error of a uniform scalar quantizer with small step size Δ, relative to the source vari-

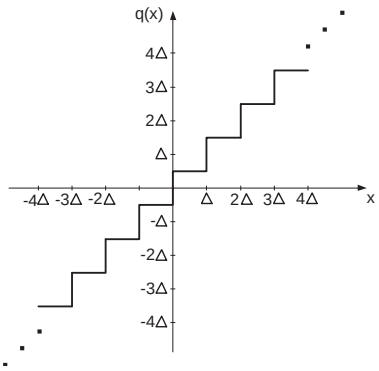

Figure 1.5: The quantization rule of an infinite-level uniform threshold scalar quantizer with step size Δ, offset zero, and reconstruction levels at cell midpoints.

ance, is approximately $\Delta^2/12$. When cells are small relative to the source variance, as is the case when the approximate formula for distortion holds, the quantizer is called *high resolution*. (Note that high resolution quantizers need not be scalar only; they may be vector quantizers as well.) Such quantizers play an important role and we shall consider them later in more detail.

Since scalar quantization is a special case of vector quantization, the latter is at least as good as the former. The question is how much better and why. Clearly, if the source has memory, scalar quantization with fixed-rate coding cannot exploit it to improve performance. (Notice that if variable-rate coding is used memory can be exploited by jointly encoding blocks of quantization indices. We shall return to this point shortly.) However, surprisingly, even for memoryless sources, fixed-rate vector quantization outperforms fixed-rate scalar quantization.

When operating in high resolution, it is possible to quantify the performance gain of vector quantizers relative to scalar quantizers, by comparing distortions of optimal quantizers of each type at a common large rate. There are three factors that contribute to the superior performance of vector quantizers, when considering

fixed-rate coding and stationary memoryless sources, as follows from [9, 10, 11, 12], and discussed in the extensive review [13].

The first is *space-filling loss*, which represents the amount by which the normalized moment of inertia of a cube is larger than that of a sphere. The second is *oblongitis loss*, which represents the amount of loss incurred due to larger normalized moments of inertia of rectangular cells relative to cubic cells. (The more oblong a cell, the higher its oblongitis loss.) The product of these two factors is collectively referred to as *cell-shape loss*. The third factor is the *point-density loss*, which represents the loss due to suboptimal distribution of cells, namely, due to not having the proper numbers of cells in the proper places. One can trade oblongitis loss for point-density loss, but cannot minimize both simultaneously.

As an example consider an i.i.d. Gaussian source and compare the best high resolution scalar quantizer with fixed-rate coding versus the best high resolution fixed-rate vector quantizer with arbitrarily large dimension. The space-filling loss in this case is 1.53 dB, the oblongitis loss is 0.94 dB, and the point density loss is 1.88 dB. Combining these gives a total loss of 4.35 dB.

When using variable-rate quantization, i.e. quantization with entropy coding, the scalar quantizer can be uniform, thus eliminating the oblongitis loss, while the entropy coder in effect eliminates the point-density loss. Furthermore, if the lossless encoder of the scalar quantizer is of sufficiently high order, then no memory loss is incurred either, when the source has memory. Therefore, variable-rate quantization incurs only space-filling loss, which can only be reduced by increasing the dimension of the quantizer.

As far as complexity is concerned, clearly, the higher the order of the entropy coder the greater its complexity. Similarly, the higher the dimension of a vector

quantizer the greater its complexity. However, these complexities can be traded, even to the extreme – e.g. using a scalar quantizer with high order entropy coding, or using a high dimensional vector quantizer with fixed-rate coding. Note that the latter will have smaller space-filling loss, but for sufficiently small vector quantizer dimensions, this difference is not significant. In general, when dimension is large, variable-rate quantization does not significantly outperform fixed-rate quantization.

We comment that the same sort of analysis as above can be used to compare vector quantizers of different dimensions, rather than comparing a vector quantizer to a scalar quantizer.

Finally, we conclude this section by commenting that most of the analysis of quantization theory is for high resolution (i.e. high rate). This is not surprising, since in high resolution the source density can often be approximated as constant over quantization cells, which in turn makes the analysis tractable. Low resolution results, e.g. [14], are quite sparse since such approximations are not available. A comparison of high resolution theory to rate-distortion theory shows that they are complementary in the sense that the former is valid for high rates and any dimension, while the latter is valid for large dimension and any rate. When both dimension and rate are large these two theories agree, namely, they predict the same performance.

1.3 Sensor Networks

Wireless sensor networks have recently been attracting a growing interest among researchers. With the advancements in technology, specifically, in miniaturized electronic devices, it has become possible to envision situations where a large number of small and cheap sensing devices, e.g. *sensors*, are deployed over an area of interest. Such deployment, referred to as a *sensor network*. could serve various purposes.

For example, sensor networks can be used for monitoring [15, 16], field-gathering [17, 18, 19, 20, 21, 22], target tracking and classification [23, 24, 25, 26, 27] and more. For an extensive review of a broad range of applications that sensor networks can and may in the future be used for, see [28, 29].

In this dissertation we focus on field-gathering. Specifically, consider the case where it is desirable to measure a physical phenomenon, e.g. temperature, humidity level, light intensity, radiation, etc., over some area and monitor its change over time and space. Wireless sensor networks can potentially provide the means to achieve this goal. The idea is to deploy many sensors over a region of interest, and let each measure/sample the desired phenomenon at its location, and transmit its measurement to some central location, referred to as the *collector*, where a reconstructed *snapshot* of the phenomenon, over the whole region of interest, is produced.

It is reasonable to model the measured phenomenon as a two-dimensional random field. Moreover, since real world phenomena are usually analog, it is often assumed that the random field is continuous-valued. We observe further, that in practice sensors cannot sample the field with infinite precision, instead they must use some kind of quantization, which invariably introduces distortion. Therefore, a field-gathering sensor network cannot produce an exact reproduction of the measured field, but rather an approximation thereof.

Clearly, the better the quality of the reconstructed field, i.e. the smaller its distortion relative to the original field values, the finer the quantizers that the sensors must use and/or the more sensors are needed. Consequently, the rate (i.e. average number of bits per sensor per unit time) to be transmitted by the network increases. This leads to a rate-distortion type of question, namely, what is the smallest rate for the network to produce a reconstruction of the field that is within a prescribed

distortion? This is the equivalent of the source coding problem in point-to-point communication systems. Similarly, one can ask the equivalent of the channel coding problem of point-to-point communication systems, namely, what is the capacity of a sensor network, where capacity may be measured in various ways. One measure of capacity, which is used in this work, is the *many-to-one transport capacity*, which is the average number of bits a sensor can send to the collector per unit time, not including relay bits. Another measure of capacity, which we also use, is the *total many-to-one transport capacity*, which is the total number of bits the collector can receive from the sensors per unit time. For a communication model similar to that given in the seminal work of Gupta and Kumar [30], Duarte-Melo and Liu [20] have shown that the former capacity scales as $\theta(\frac{1}{N})$, where N is the number of sensors in the network, and the latter capacity remains essentially constant with respect to the number of sensors in the network.

A separation theorem like the one showed by Shannon for point-to-point communication, has not been demonstrated for the sensor network setting. However, one may consider the sensor network equivalent problems of source coding and channel coding separately, as was done, for example, in [18], where Slepian-Wolf distributed coding [31] was used to consider the source coding problem. When doing so, one runs the risk of designing a suboptimal sensor network. There has been work done with regard to these two problems. By considering various transmission models, various capacity results have been shown, e.g. [30, 20, 32, 33, 34]. Similarly, results have been derived regarding network source coding, for example, [35, 36, 37].

When designing a field-gathering sensor network, there are tradeoffs to be considered. On the one hand, few number of bits to be transmitted is an obvious requirement (which translates to low power). On the other hand, low reconstruction

distortion is also desirable, which can be attained by means of having a denser network and finer quantization. Thus, the smaller the required distortion the more bits will be transmitted. An equivalent problem is to maximize the frequency, equivalently throughput, with which snapshots can be transported to the collector, subject to a distortion constraint, thus producing a "video" like representation of the field.

Finally, one of the major ideas behind field-gathering sensor networks is to have them be dense, so as to be able to reconstruct the sensed field at various distortion levels and obtain high robustness to sensor failures (a common assumption is that individual sensors might be unreliable). Therefore, a key question, which is the main focus of Chapter V, is how frequently can snapshots be transported to the collector, subject to a distortion constrained, as the density of the sensors in the network increases? The answer to this question depends on the behavior of the number of bits per sensor per unit time that need to be transmitted by each sensor, combined with the behavior of the many-to-one transport capacity of the network, as the density of the sensors increases. It will be shown that under certain relatively general scenarios, increasing the density of the network without bound is not a good strategy in the sense that it makes the frequency with which snapshots can be transported to the collector go to zero.

1.4 Contributions

Following is a description of the contributions of the main four chapters of this dissertation, Chapters II – V. Each of the first three of these chapters examines a different aspect of asymptotic scalar quantization. Chapter V considers applications to sensor networks.

In Chapter II high resolution infinite-level uniform scalar quantization is consid-

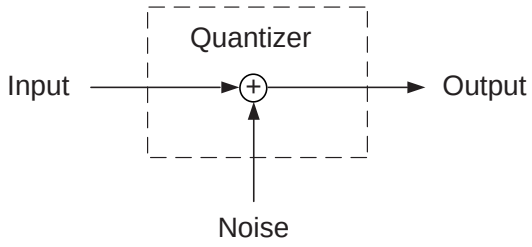

Figure 1.6: The additive noise model.

ered. A uniform scalar quantizer with small step size, large support and reconstruction levels at cell midpoints is frequently modeled as adding orthogonal noise to the quantizer input (c.f. [38] (pp. 193ff.) [39] (pp. 753ff.) and [40], to name a few). This is known as the *additive noise model*, which is illustrated in Figure 1.6 and whose validity has never been rigorously shown.

In this chapter we rigorously demonstrate the asymptotic validity of the additive noise model when the input pdf is continuous and satisfies several other mild conditions. Specifically, as step size decreases, the correlation between input and quantization error becomes asymptotically negligible relative to the mean-squared error. The model is even valid when the input density has a discontinuity (of any kind) at the origin, but discontinuities elsewhere, specifically jump discontinuities of finite height, can prevent the correlation from being negligible. Though this invalidates the additive model, an asymptotic formula for the correlation is found in terms of the step size and the heights and positions of the jump discontinuities.

For an input density with finite support, such as uniform, it is shown that the support of the uniform quantizer can be matched to that of the density in ways that make the correlation approach a variety of limits.

The derivations in this chapter are based on an analysis of the asymptotic convergence of cell centroids to cell midpoints. This convergence is fast enough that the

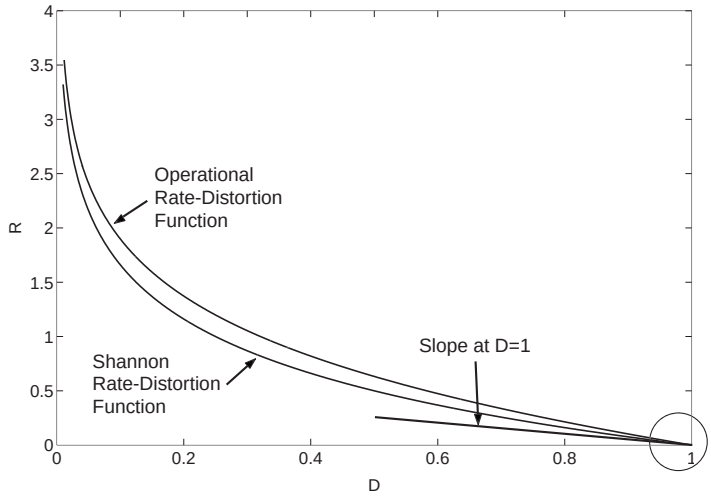

Figure 1.7: The operational rate-distortion function for scalar quantizers and the Shannon rate-distortion function for a memoryless stationary Gaussian source with variance one. The latter is described qualitatively outside the high resolution region.

centroids and midpoints induce the same asymptotic mean-squared error, but not fast enough to induce the same correlations.

In contrast to Chapter II, Chapter III analyzes asymptotic low resolution scalar quantization. Specifically, it examines the performance of scalar quantization for stationary memoryless Gaussian sources. The goal is to find the rate at which the operational rate-distortion function of scalar quantization approaches zero as distortion goes to the source variance. This is illustrated in Figure 1.7, where both the operational and Shannon rate-distortion functions are plotted (the former is plotted qualitatively). Notice, that the operational rate-distortion function is not known at low rates, and all that is known is that it equals zero when distortion equals the source variance. Thus, this work is important in that it adds additional understanding of the operational rate-distortion function of scalar quantization.

To find this rate of convergence, uniform and binary quantizers with entropy coding are analyzed. It is shown that for a stationary memoryless Gaussian source,

as distortion D, approaches its variance, the least entropy of such quantizers with mean-squared error D or less approaches zero with slope $-\frac{\log_2 e}{2\sigma^2}$. Since, as can be seen from (1.3), the Shannon rate-distortion function also approaches zero with the same slope, it follows that in low resolution scalar quantization is asymptotically optimal, i.e. it is as good as any coding technique.

We notice that since this result is demonstrated using uniform and binary quantizers, it not only shows that scalar quantization can be optimal in general, but rather it provides specific quantizers that achieve such optimality.

Motivated by the need to analyze dense sensor networks, in Chapter IV we consider a somewhat different setting than that considered in the previous two chapters. There are two main thrusts in this chapter.

The first examines the case that a stationary random process is sampled over some finite interval, and each sample is separately quantized with arbitrary, yet identical, scalar quantizers. It is shown that if the random process crosses some quantization threshold with positive probability, then the joint entropy of the quantized samples tends to infinity as the sampling interval goes to zero. This is not a trivial result since the joint entropy can be viewed as the product of the number of samples and the average number of bits needed to describe a quantized sample, where the first quantity tends to infinity, while the second tends to zero, as the sampling interval goes to zero. Thus, it is not a priori clear what this product might be. Much work has been done with regard to oversampling and quantization, e.g. [41, 42, 43, 44, 45, 46, 47, 48, 49, 50], however, the analysis in some of this work is for deterministic signals and does not address the question posed here.

Having established that the joint entropy above tends to infinity as the sampling interval goes to zero, it is of interest to find the rate at which this happens. In general,

this is a difficult problem. Instead the second result of Chapter IV provides an upper bound to the rate at which the joint entropy tends to infinity, by considering the conditional entropy of one quantized sample conditioned on a neighboring quantized sample. This, too however, is a difficult problem in its own right. For it is clear that as the correlation between neighboring samples tends to one, the entropy of the output of one quantizer conditioned on the output of a neighboring quantizer tends to zero, but it is not clear how rapidly it does so, which is what we seek. Therefore, to make this problem tractable, we consider the case that the scalar quantizers are infinite-level uniform threshold, and the random process is stationary and Gaussian such that its mean lies at a midpoint of some quantization cell. Under these assumptions, the following simple asymptotic formula for the conditional entropy is derived.

$$\lim_{\rho \to 1} \frac{H(I_2|I_1)}{-M_\lambda \sqrt{1-\rho} \log \sqrt{1-\rho}} \;=\; 1 \;,$$

where ρ is the correlation coefficient, λ is the ratio of quantization step size to source variance, and M_λ is a constant that depends on λ. The rate of convergence of conditional entropy is evident from this formula. It is intuitively apparent that the higher the resolution of the quantizers (i.e. the smaller their step size), the larger the conditional entropy should be, for a given value of ρ. This effect is quantified by the multiplicative constant M_λ. When considering asymptotically high resolution quantizers, it is shown that $M_\lambda \approx \frac{2}{\sqrt{\pi}} \frac{1}{\lambda}$, which shows that the increase in conditional entropy is of the order of $\frac{1}{\lambda}$, i.e. it is exactly inversely proportional to the ratio of step size to source variance.

Finally in Chapter V, we consider field-gathering sensor networks. Specifically, we examine their behavior as their density increases to infinity. We assume a network communication model similar to that given in [30] and use the result concerning the

many-to-one transport capacity shown in [20] for such networks. We further assume that the sensors utilize identical scalar quantizers, and that the field is stationary and crosses some quantization threshold with positive probability. Furthermore, our analysis is for one-dimensional fields, although we have no doubt that it extends to two-dimensions as well (see the Future Work Section of Chapter VI).

Under these assumptions, combining the many-to-one transport capacity result, together with the first result shown in Chapter IV, namely, that the joint entropy of the quantized samples over a finite interval tends to infinity as the sampling interval goes to zero, we obtain that as the density of the network tends to infinity, the frequency, equivalently throughput, with which snapshots of the field can be transported to the collector goes to zero. Using the second result of Chapter IV, namely, the rate at which conditional entropy of jointly Gaussian random variables tends to zero as their correlation goes to one, we upper bound the rate at which the throughput of such networks degrades to zero, for the case of a stationary Gaussian random field.

This analysis shows that under the considered scenario, making sensor networks too dense is not a good strategy, regardless of the scheme used by the lossless encoders of the quantizers. It implies further, that given a desired fidelity, there is some optimal sensor density (in terms of maximizing throughput) that should be used.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, 623–656, July, Oct. 1948.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John-Wiley & Sons Inc., New York, 1991.

[3] T. Berger, *Rate Distortion Theory*, Prentice-Hall, Englewood Cliffs, 1971.

[4] R. G. Gallager, *Information Theory and Reliable Communication*, John-Wiley & Sons Inc., New York, 1968.

[5] D. A. Huffman, "A method for the construction of minimum redundancy codes," *Proc. IRE.*, pp. 1098–1101, Sep. 1952.

[6] S. P. Lloyd, "Least squares quantization in pcm," *unpublished Bell Lab. Also IEEE Trans. Info. Theory*, vol. 28, pp. 129–137, Mar. 1982.

[7] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 31–42, Jan. 1989.

[8] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, Jul. 1948.

[9] P. L. Zador, "Development and evaluation of procedure for quantizing multivariate distributions," *Ph.D. Thesis, Stanford University*, 1963.

[10] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Info. Theory*, vol. 25, pp. 373–380, Jul. 1979.

[11] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage," *IEEE Trans. Info. Theory*, vol. 35, pp. 1020–1033, Sep. 1989.

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

[13] R.M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Info. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[14] D.F. Lyons, "Fundamental limits of low-rate transform codes," *Ph.D. Thesis, EECS Department, University of Michigan*, 1992.

[15] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless sensor networks for habitat monitoring," *ACM International Workshop on Wireless Sensor Networks and Applications (WSNA)*, pp. 88–97, Sep. 2002.

[16] H. Wang, D. Estrin, and L. Girod, "Preprocessing in a tiered sensor network for habitat monitoring," *EURASIP JASP Special Issue on Sensor Networks*, pp. 392–401, May 2003.

[17] A. Scaglione and S. D. Servetto, "On the interdependence of routing and data compression in multi-hop sensor networks," *International Conference on Mobile Computing and Networking (MobiCom), Atlanta, GA.*, pp. 140–147, Sep. 2002.

[18] D. Marco, E. Duarte-Melo, M. Liu, and D. L. Neuhoff, "On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data," *Workshop on Information Processing in Sensor Networks (IPSN), Palo Alto, CA.*, pp. 1–16, Apr. 2003.

[19] S. D. Servetto, "Sensing lena – massively distributed compression of sensor images," *IEEE International Conference on Image Processing (ICIP), Barcelona, Spain.*, Sep. 2003.

[20] E. J. Duarte-Melo and M. Liu, "Data-gathering wireless sensor networks: Organization and capacity," *Special Issue Computer Networks (Elsevier) on Wireless Sensor Networks*, vol. 43, no. 4, pp. 519–537, Nov. 2003.

[21] R. Cristescu, "Efficient decentralized communications in sensor networks," *Ph.D. Thesis, EPFL.*, Mar. 2004.

[22] D. Ganesan, R. Cristescu, and B. Beferull-Lozano, "Power-efficient sensor placement and transmission structure for data gathering under distortion constraints," *Workshop on Information Processing in Sensor Networks (IPSN), Berkeley, CA.*, pp. 142–150, Apr. 2004.

[23] H. Pasula, S. Russel, M. Ostland, and Y. Ritov, "Tracking many objects with many sensors," *Int. Joint Conf. on Artificial Intelligence (IJCAI), Stockholm, Sweden*, pp. 1160–1171, 1999.

[24] K. Chakrabarty, S. S. Iyengar, H. Qi, and E. Cho, "Grid coverage of surveillance and target location in distributed sensor networks," *IEEE Trans. Comp.*, vol. 51, no. 12, pp. 1448–1453, Dec. 2002.

[25] J. Liu, J. Reich, and F. Zhao, "Collaborative in-network processing for target tracking," *EURASIP J. Appl. Sig. Proc.*, pp. 378–391, Mar. 2003.

[26] R. R. Brooks, P. Ramanathan, and A. M. Sayeed, "Distributed target classification and tracking in sensor networks," *Proc. IEEE*, vol. 91, no. 8, pp. 1163–1171, Aug. 2003.

[27] S. Venkatesh, M. Alanyali, O. Savas, and S. Aeron, "Classification in sensor networks," *IEEE Int. Symp. Info. Theory (ISIT), Chicago, IL.*, p. 251, July 2004.

[28] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar, "Next century challenges : scalable coordination in sensor networks," *Proc. Int. Conf. on Mobile Computing and Networks (MobiCOM), Seattle, WA.*, pp. 263–270, Aug. 1999.

[29] D. Estrin and et.al., "Embedded everywhere: A research agenda for networked systems of embedded computers," *Natioal Academy Press Computer Science and Telecommunications Board (CSTB) Report*, 2001.

[30] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Info. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[31] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Info. Theory*, vol. 19, pp. 471–480, Jul. 1973.

[32] H. El Gamal, "On the scaling laws of dense wireless sensor networks," *submitted to IEEE Trans. Info. Theory*, Apr 2003.

[33] O. Arpacioglu and Z. J. Haas, "On the scalability and capacity of wireless networks with omnidirectional antennas," *Workshop on Information Processing in Sensor Networks (IPSN), Berkeley, CA.*, pp. 169–177, Apr. 2004.

[34] F. Xue, L.-L. Xie, and P. R. Kumar, "The transport capacity of wireless networks over fading channels," *IEEE Int. Symp. Info. Theory (ISIT), Chicago, IL.*, p. 370, July 2004.

[35] P. Ishwar, R. Puri, S. S. Pradhan, and K. Ramchandran, "On rate-constrained estimation in unreliable sensor networks," *Workshop on Information Processing in Sensor Networks (IPSN), Palo Alto, CA.*, pp. 178–192, Apr. 2003.

[36] M. Gastpar and M. Vetterli, "Power-bandwidth-distortion scaling laws for sensor networks," *Workshop on Information Processing in Sensor Networks (IPSN), Berkeley, CA.*, pp. 320–329, Apr. 2004.

[37] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "Scaling laws for correlated data gathering," *IEEE Int. Symp. Info. Theory (ISIT), Chicago, IL.*, p. 471, July 2004.

[38] A. V. Oppenheim, R. W. Schafer, and J.R. Buck, *Discrete-Time Signal Processing*, Prentice-Hall, Upper Saddle River, second edition, 1999.

[39] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, Prentice-Hall, Upper Saddle River, third edition, 1996.

[40] A. Gersho, "Principles of quantization," *IEEE Trans. Circuits and Systems*, vol. 25, no. 7, pp. 427–436, Jul. 1978.

[41] S. K. Tewksbury and R. W. Hallock, "Oversampled, linear predictive and noise-shaping coders of order $N > 1$," *IEEE Trans. Cir. Sys.*, vol. 25, no. 7, pp. 436–447, July 1978.

[42] M. W. Hauser, "Principles of oversampling A/D conversion," *J. Audio Eng. Soc.*, vol. 39, no. 1/2, pp. 3–26, Jan./Feb. 1991.

[43] K. Benhenni and S. Cambanis, "The effect of quantization on the performance of sampling designs," *IEEE Trans. Info. Theory*, vol. 44, no. 5, pp. 1981–1992, Sep. 1998.

[44] J. Tuqan and P. P. Vaidynathan, "Oversampling PCM techniques and optimum noise shapers for quantizing a class of nonbandlimited signals," *IEEE Trans. Signal Proccessing*, vol. 47, no. 2, pp. 389–407, Feb. 1999.

[45] Z. Cvetkovic and M. Vetterli, "Error-rate characteristics of oversampled analog-to-digital conversion," *IEEE Trans. Info. Theory*, vol. 44, no. 5, pp. 1961–1964, Sep. 1998.

[46] Z. Cvetkovic and M. Vetterli, "On simple oversampled A/D conversion in $L^2(\mathbb{R})$," *IEEE Trans. Info. Theory*, vol. 47, no. 1, pp. 146–154, Jan. 2001.

[47] Z. Cvetkovic and I. Daubechies, "Single-bit oversampled A/D conversion with exponential accuracy in the bit-rate," *Data Compression Conference, DCC, Snowbird, UT*, pp. 343–352, Mar. 2000.

[48] N. T. Thao and M. Vetterli, "Reduction of the MSE in R-times oversampled

A/D conversion $O(1/R)$ to $O(1/R^2)$," *IEEE Trans. Signal Processing*, vol. 42, no. 1, pp. 200–203, Jan. 1994.

[49] N. T. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Trans. Signal Processing*, vol. 42, no. 3, pp. 519–531, Mar. 1994.

[50] N. T. Thao and M. Vetterli, "Lower bound on the mean-squared error in oversampled quantization of periodic signals using vector quantization analysis," *IEEE Trans. Info. Theory*, vol. 42, no. 2, pp. 469–479, Mar. 1996.

CHAPTER II

The Validity of the Additive Noise Model for Uniform Scalar Quantizers[1]

2.1 Introduction

In his pioneering 1948 paper, Bennett [1] argued that the quantization error of a uniform scalar quantizer with small cells, reproduction levels at the cell midpoints and large support region can be approximately modeled as being orthogonal to the quantizer input. That is, with X and Y denoting the quantizer input and output, respectively,

$$EX(Y-X) \approx 0 . \tag{2.1}$$

It follows that, as illustrated in Figure 2.1a, the quantizer output Y can be modeled as the sum of X plus orthogonal quantization error $N = Y - X$. This is the *additive noise model*. Since $EY^2 = EX^2 + D + 2EX(Y-X)$, where $D = E(Y-X)^2$ is the mean-squared error (MSE), an equivalent property is

$$EY^2 \approx EX^2 + D , \tag{2.2}$$

[1]This work was supported by NSF Grant ANI-0112801. This chapter was submitted as a paper for publication with co-author David L. Neuhoff to the IEEE Transactions on Information Theory. Portions of this work were published in the proceedings of the IEEE International Symposium on Information Theory, Yokohama, Japan, July 2003.

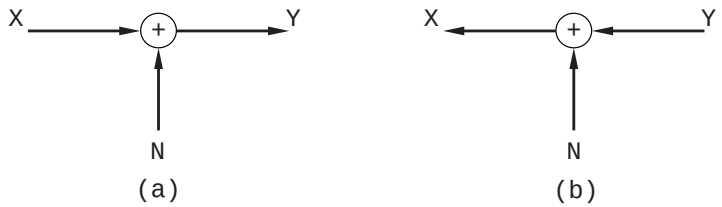

Figure 2.1: Additive models of uniform scalar quantization. (a) The levels are midpoints and the quantization error is orthogonal to the input. (b) The levels are centroids and the quantization error is orthogonal to the output.

i.e. the output power approximately equals the input power plus the MSE. Though the additive noise model is very widely used (c.f. [2] (pp. 193ff.), [3] (pp. 753ff.), [4]), its validity has never been rigorously demonstrated. The principal goal of this paper is to do this and, in addition, to discover the correlation structure when the additive noise model is not valid.

It is easy to see that the left and right-hand sides of (2.1), respectively (2.2), tend to the same values as $\Delta \to 0$. This, however, is not sufficient to validate the additive noise model. Instead, we assert that the additive noise model is *asymptotically valid* when and only when

$$EX(Y - X) \;=\; o(D) \;,$$

where $o(z)$ denotes a quantity such that $o(z)/z \to 0$ as $z \to 0$. Equivalently, it is asymptotically valid when and only when $EY^2 = EX^2 + D + o(D)$. In other words, the discrepancies in the approximations (2.1) and (2.2) must be asymptotically negligible relative to the MSE. Equivalently, using the well known approximation $D = \frac{\Delta^2}{12} + o(\Delta^2)$, where Δ denotes the width of a quantization cell, the errors must be asymptotically negligible relative to Δ^2.

With this definition in mind, our principal result, Corollary 12, shows that the additive noise model is asymptotically valid when, in addition to satisfying several

mild technical conditions, the probability density function (pdf) of X is continuous, except possibly for tending to infinity at the origin or having a finite jump discontinuity at the origin. If, on the other hand, there are finite jump discontinuities not at the origin, then Corollary 11 shows that

$$\frac{EX(Y-X)}{\Delta^2} \;=\; \frac{1}{12}\sum_{k=1}^{N} t_k e_k\big(1-6\alpha_\Delta(t_k)(1-\alpha_\Delta(t_k))\big)+o(1)\;,$$

where $t_1,\ldots,t_N$ are the positions of the jumps in the pdf, $e_1,\ldots,e_N$ are their heights, $\alpha_\Delta(t_k)$ is the fractional position of t_k within its quantization cell, and $o(1)$ denotes a quantity that approaches 0 as $\Delta\to 0$. It would be nice if the right-hand side of the above converged to some function of the t_k's and e_k's, with no dependence on the $\alpha_\Delta(t_k)$'s. In this case one could easily estimate the correlation, even in the presence of jumps. However, Theorem 13 shows that this is not possible. We conclude that when there are jumps in the pdf, the correlation structure depends intimately on the positions of such jumps within quantization cells.

To avoid overload issues, we focus on uniform quantizers with infinitely many levels, i.e. with infinite support. However, the results have significance for uniform quantizers with finitely many levels. Specifically, since the performance of a uniform quantizer with n levels approaches that of an infinite uniform quantizer as n tends to infinity, the results indicate conditions under which the additive noise model is asymptotically valid when Δ is sufficiently small and n is sufficiently large.

In deriving our results, we find it necessary to explore and exploit relations between quantization cell midpoints and centroids that yield insight into the behavior of uniform quantizers. It is well known that for a given Δ, MSE is minimized when centroids rather than midpoints are used as levels. Not surprisingly, as can be deduced from the results of [5] (p. 15), the MSE with centroids is again well approximated by

$\Delta^2/12$ when Δ is small. Thus, asymptotically, midpoints and centroids induce the same distortion. On the other hand, midpoints and centroids lead to rather different correlation structures. Specifically, with centroids, it is well known that for any Δ, the quantization error is exactly orthogonal to the quantizer output Y, rather than the quantizer input X, i.e.

$$EY(Y - X) \;=\; 0 \;, \tag{2.3}$$

or equivalently,

$$EY^2 \;=\; EX^2 - D, \tag{2.4}$$

i.e. the output power equals the input power minus the MSE. Thus, instead of the the usual additive noise model, we have the additive model illustrated in Figure 1b. This is somewhat surprising in light of the fact that the cell centroids approach the cell midpoints as Δ decreases. (This intuitive fact is shown in [6].) Clearly, there is subtle behavior here. In this paper, we strengthen previous results on the convergence of cell centroids to midpoints, and we show that this convergence happens fast enough to account for the fact that the MSE with centroids is asymptotically the same as that with midpoints. However, it is not fast enough to cause them to have the same asymptotic correlation structure. We also note that the proofs of the principal theorems are based on a measure of the difference between the values assumed by EY^2 when centroids are used vs. midpoints.

For completeness, we mention that Widrow [7], and Sripad and Snyder [8] found conditions, on Δ and the pdf, involving zeros of the characteristic function, under which the quantizer input and error are exactly orthogonal. Note, however, that these results are not asymptotic and that the conditions are rather restrictive. We also mention that Bennett's paper [1] argued that in addition to being orthogonal to the input, the quantization errors of a uniform scalar quantizer are, approximately,

white. A rigorous demonstration of this was given in [9].

The remainder of the paper is organized as follows. Section 2.2 introduces infinite uniform scalar quantizers and the framework for considering such with step size Δ decreasing to zero, as well as notation and other essential background material. Section 2.3 shows that centroids approach the midpoints rapidly enough to account for the fact that the MSE due to centroids is asymptotically the same as for midpoints. Section 2.4 discusses the additive noise model and introduces a key functional $r(f)$ measuring the closeness of cell midpoints and centroids, whose value determines the validity of the additive noise model. Section 2.5 evaluates $r(f)$ and states the main results regarding the correlation of input and quantization error and the asymptotic validity of the additive noise model. Section 2.6 discusses alternative noise models for uniform quantizers whose support is matched to that of a pdf with finite support. Section 2.7 proves the principal results. Section 2.8 offers concluding remarks. Finally, the Appendix contains proofs of certain lemmas.

2.2 Background

An infinite level uniform scalar quantizer is characterized by a *step size* $\Delta > 0$, an *offset* θ, $0 \leq \theta < 1$, and a set of *(reconstruction) levels* $\ldots < y_{-2} < y_{-1} < y_0 < y_1 < y_2 < \ldots$. The *thresholds* of such a quantizer are $\ldots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \ldots$, where $x_i = (i - \theta)\Delta$, and the i^{th} (quantization) cell is $S_i = [x_i, x_{i+1})$. Note that θ is the fractional position of the origin within its quantization cell. Given an input x, the quantizer outputs $q(x) = y_i$ when $x \in S_i$. The quantization error is $q(x) - x$, and when the input is a random variable X with pdf f, the MSE is $D = E(q(X) - X)^2 = \int_{-\infty}^{\infty}(q(x) - x)^2 f(x)\, dx$.

We focus on two choices for the levels: midpoints and centroids. In the former

case, $y_i = x_i + \Delta/2 = (i - \theta + 1/2)\Delta$. In the latter, $y_i = E[X|x_i \le X < x_{i+1}] = \frac{\int_{x_i}^{x_{i+1}} x f(x)\,dx}{\int_{x_i}^{x_{i+1}} f(x)\,dx}$, where f is the pdf of X.[2] Let $m_{\Delta,\theta}(x)$, $c_{\Delta,\theta}(x)$ and $u_{\Delta,\theta}(x)$ denote the midpoint, centroid and left threshold, respectively, of the quantization cell in which x lies. These functions are constant on quantization cells. Let $M_{\Delta,\theta}$ and $C_{\Delta,\theta}$ denote the output random variable $Y = q(X)$ when midpoints and centroids are used, respectively. In addition, let $m_{\Delta,u} = u + \Delta/2$ and $c_{\Delta,u} = E[X|u \le X < u + \Delta]$ denote, respectively, the midpoint and centroid of the interval $[u, u + \Delta)$.[3] For brevity, we usually omit the subscript θ and frequently omit the subscript Δ from $m_{\Delta,\theta}(x)$, $c_{\Delta,\theta}(x)$, etc., when they are clear from context. When we wish to emphasize dependence on the pdf, we add a superscript, as in $c^f_{\Delta,\theta}(x)$.

In most of the results in later sections, we consider limiting characteristics of families of uniform quantizers in which the step size Δ goes to zero and the offset θ varies arbitrarily. That is, the offset θ is an arbitrary function of Δ, denoted $\theta(\Delta)$. It can be shown that if $g(\Delta, \theta)$ is a function and c is a constant such that $\lim_{\Delta\to 0} g(\Delta, \theta(\Delta)) = c$ for any function $\theta : \mathbb{R} \to [0, 1)$, then the convergence is uniform over all such functions θ.

Throughout this paper we focus on continuous input random variables X with finite first and second moments, whose pdf's are either continuous or have finite jump discontinuities, or have points at which f goes to infinity from the left or right. (f is said to have a finite jump discontinuity at t if the following limits exist, and are finite and different: $f(t^-) \stackrel{\Delta}{=} \lim_{x\nearrow t} f(x)$, $\lim_{x\searrow t} f(x) \stackrel{\Delta}{=} f(t^+)$.) Other conditions on f will be specified as needed. It should be noted that for any result in this paper that is concerned with expected values, if f_1 and f_2 are pdf's such that $f_1 = f_2$ almost

[2] When a cell $[x_i, x_{i+1})$ has zero probability, the value of $E[X|x_i \le X < x_{i+1}]$ is of no consequence. However, for concreteness, we take it to be the midpoint of the cell.

[3] When $\Pr(u \le X < u + \Delta) = 0$, we let $c_{\Delta,u} = u + \Delta/2$.

everywhere (a.e.) and f_2 satisfies the specified conditions for the result, then the result applies to f_1 as well. If a density f has finite support, then an infinite uniform quantizer has, effectively, finitely many levels.

We will occasionally introduce a symbol like f or g to represent a function that is like a pdf, but may lack the property of integrating to one. Accordingly, in all statements of results where f is required to be a pdf, we will explicitly specify such. Where there is no specification, f denotes an arbitrary function.

Finally, a function f is said to be *piecewise differentiable* if there exists a countable collection of disjoint open intervals $\{B_i\}$ such that (a) f is differentiable on each B_i, (b) $\mathbb{R} = (\cup_i B_i) \cup E$, where E is the set of interval endpoints (not including $-\infty$ and ∞), and (c) any finite interval contains at most a finite number of B_i's. We let $B \stackrel{\Delta}{=} \cup_i B_i$.

2.3 Mean-Squared Error

As mentioned earlier, when Δ is small, the MSE when using midpoints, denoted $D_{m,\Delta}$, is approximately $\Delta^2/12$. Linder and Zeger [10] showed rigorously that this holds for any pdf. The precise statement is:

$$\lim_{\Delta \to 0} \frac{D_{m,\Delta}}{\Delta^2/12} \;=\; 1 \;, \tag{2.5}$$

or equivalently, $D_{m,\Delta} = \frac{\Delta^2}{12} + o(\Delta^2)$. Although the authors did not claim such, their proof is sufficient to show that (2.5) holds for any offset function $\theta(\Delta)$.

It is quite intuitive that M and C become closer as $\Delta \to 0$. The question is how fast. The following two lemmas, whose proofs are left to the Appendix, provide some answers.

Lemma 1. *If f is continuous and positive at x, then for any offset function*

$$\lim_{\Delta \to 0} \frac{m_\Delta(x) - c_\Delta(x)}{\Delta} \;=\; 0 \;,$$

or equivalently, $c_\Delta(x) = m_\Delta(x) + o(\Delta)$.

Lemma 2. *If f is a continuous a.e. pdf, then for any offset function*

$$\lim_{\Delta \to 0} \frac{E(M_\Delta - C_\Delta)^2}{\Delta^2} \;=\; 0 \;,$$

or equivalently, $E(M_\Delta - C_\Delta)^2 = o(\Delta^2)$.

Remark: Notice that while quantities such as $m_\Delta(x), c_\Delta(x), u_\Delta(x), M_\Delta(x)$ and $C_\Delta(x)$ depend on the offset function, limit expressions, such as in these two lemmas, usually do not. Whenever appropriate, such insensitivity to the offset function will be explicitly stated in future lemmas and theorems. In their proofs, an arbitrary fixed offset function will be assumed. However, it will not appear explicitly therein. Instead, its influence on $m_\Delta(x), c_\Delta(x)$, etc. is implicit.

It is well known that centroids minimize MSE. The following theorem uses the convergence of centroids to midpoints demonstrated in Lemma 2 to show that the MSE induced by centroids, denoted $D_{c,\Delta}$, is asymptotically the same as that induced by midpoints. This result can also be deduced from the results of [5] (p. 15) without reference to the closeness of midpoints and centroids and without requiring the pdf to be continuous a.e..

Theorem 3. *If f is a continuous a.e. pdf, then for any offset function*

$$D_{c,\Delta} \;=\; D_{m,\Delta} + o(\Delta^2) \;=\; \frac{\Delta^2}{12} + o(\Delta^2) \;. \tag{2.6}$$

Proof:

$$\begin{aligned} D_{m,\Delta} &= E[(X - C_\Delta) + (C_\Delta - M_\Delta)]^2 \\ &= E(X - C_\Delta)^2 + 2E(X - C_\Delta)(C_\Delta - M_\Delta) + E(C_\Delta - M_\Delta)^2 \\ &= D_{c,\Delta} + E(C_\Delta - M_\Delta)^2 \;=\; D_{c,\Delta} + o(\Delta^2) \;, \end{aligned}$$

where the first equality is by definition of $D_{m,\Delta}$, the second is trivial, the third is by the orthogonality principle, and the last is due to Lemma 2. The last equality in (2.6) is from (2.5). □

2.4 Additive Noise Model

Our primary goal is to determine when the additive noise model is asymptotically valid for uniform scalar quantizers with infinitely many levels located at the midpoints. With M denoting the quantizer output with midpoints levels, we consider the additive noise model to be asymptotically valid when and only when for any offset function

$$EX(M - X) \;=\; o(\Delta^2) \;, \tag{2.7}$$

or equivalently,

$$EM^2 \;=\; EX^2 + D_m + o(\Delta^2) \;. \tag{2.8}$$

We focus on the latter condition. To determine when it holds, we write

$$\begin{aligned} EM^2 &= EM^2 + EX^2 - EC^2 - D_c \\ &= EX^2 - D_m + (EM^2 - EC^2) + o(\Delta^2) \;, \end{aligned}$$

where C and D_c denote the output and MSE, respectively, of a quantizer with centroid levels, and where the first equality follows from (2.4) applied to a quantizer with centroid levels and the second follows from Theorem 3, assuming the pdf is

continuous a.e.. It is now clear that the relationship between EM^2, EX^2 and D_m depends on the quantity $EM^2 - EC^2$. This motivates us to define a functional r that captures the behavior of this quantity.

Definition 4. *Given a pdf f,*

$$r_\Delta(f) \stackrel{\Delta}{=} \frac{EM^2 - EC^2}{\Delta^2/6} = \int_{-\infty}^{\infty} G_\Delta(x)\, dx \ , \tag{2.9}$$

where $G_\Delta(x) \stackrel{\Delta}{=} \frac{m_\Delta^2(x) - c_\Delta^2(x)}{\Delta^2/6} f(x)$. When the limit of $r_\Delta(f)$ exists and is the same for all offset functions,

$$r(f) \stackrel{\Delta}{=} \lim_{\Delta \to 0} r_\Delta(f) \ . \tag{2.10}$$

Using the above definition, we obtain the following lemma.

Lemma 5. *If f is a continuous a.e. pdf, then for any offset function*

$$EM^2 = EX^2 + (2r_\Delta(f) - 1)\frac{\Delta^2}{12} + o(\Delta^2) \ . \tag{2.11}$$

Proof:

$$\frac{EM^2 - EX^2}{\Delta^2/12} = \frac{EM^2 - EC^2}{\Delta^2/12} - \frac{EX^2 - EC^2}{\Delta^2/12} = 2r_\Delta(f) - \frac{D_c}{\Delta^2/12} = 2r_\Delta(f) - 1 + \frac{o(\Delta^2))}{\Delta^2} \ ,$$

where the second equality uses (2.4) and the definition of $r_\Delta(f)$, and the last uses Theorem 3. □

We now consider the ramifications of this lemma.

Corollary 6. *If the input pdf f is continuous a.e., then for any offset function*

$$EX(M - X) = \frac{\Delta^2}{12}(r_\Delta(f) - 1) + o(\Delta^2) \ , \tag{2.12}$$

and

$$EM(M - X) = \frac{\Delta^2}{12} r_\Delta(f) + o(\Delta^2) \ . \tag{2.13}$$

Proof: The first relation is

$$\begin{aligned} EX(M-X) &= \frac{1}{2}\left[EM^2 - EX^2 - D_m\right] = \frac{1}{2}\left[(2r_\Delta(f)-1)\frac{\Delta^2}{12} + o(\Delta^2) - D_m\right] \\ &= (r_\Delta(f)-1)\frac{\Delta^2}{12} + o(\Delta^2) \ , \end{aligned}$$

where the first equality is elementary, the second is from Lemma 5, and the third is from (2.5).

The second relation is:

$$\begin{aligned} EM(M-X) &= EM^2 - EX^2 - EX(M-X) \\ &= (2r_\Delta(f)-1)\frac{\Delta^2}{12} + o(\Delta^2) - (r_\Delta(f)-1)\frac{\Delta^2}{12} + o(\Delta^2) \\ &= r_\Delta(f)\frac{\Delta^2}{12} + o(\Delta^2) \ , \end{aligned}$$

where the first equality is elementary, the second comes from Lemma 5 and the first part of this proof, and third comes from (2.5). □

By comparing (2.11) to (2.8) and using (2.5), or equivalently comparing (2.12) to (2.7), we obtain the following:

Theorem 7. *If midpoints are used and the input pdf is continuous a.e., then the additive noise model is asymptotically valid if and only if $r(f)$ exists and equals one.*

2.5 Evaluating $r(f)$

In this section we give the main results of the paper, which characterize the behavior of $r(f)$, and consequently, determine the validity or invalidity of the the additive noise model. We begin with a definition. Proofs are given in Section 2.7.

Definition 8. *A pdf f is* nice *if each of the following holds:*

1. *f has finite second moment.*
2. $\lim_{x\to 0} xf(x) = 0$.
3. *There exists $\varepsilon > 0$ such that $\lim_{\substack{x\to-\infty \\ x\in B}} |x|^{2+\varepsilon} f'(x) = 0$ and $\lim_{\substack{x\to\infty \\ x\in B}} x^{2+\varepsilon} f'(x) = 0$, where f' is the derivative of f and B is the set over which f is differentiable.*
4. *f is continuous, bounded and piecewise differentiable with bounded derivative, except perhaps at a finite set of* exceptional points *$\{s_1, \ldots, s_n\}$ such that any of the following might hold:*
   - *(a) $|f'(x)| \to \infty$ as $x \to s_i$ from left and/or right*
   - *(b) f has a finite jump discontinuity at s_i*
   - *(c) $s_i = 0$ and $f(x) \to \infty$ as $x \to 0$ from left and/or right*

   and if at any s_i, $|f'|$ goes to infinity from left (right), it does so monotonically in some left (right) neighborhood of s_i.

Remark: The class of densities of the form $b|x|^\beta e^{-a|x|^\alpha}$, $\beta > -1$ and $\alpha > 0$, which includes Gaussian, Laplacian, gamma, and one-sided versions of these, such as Rayleigh and exponential, are nice pdf's.

Theorem 9. *If f is a nice pdf with no exceptional points, then $r(f) = 1$.*

Theorem 10. *If f is a nice pdf, and $T = \{t_1, \ldots, t_N\}$ is the set of exceptional points where f has discontinuities, then for any offset function,*

$$r_\Delta(f) \;=\; s_\Delta(f) + o(1)\;, \tag{2.14}$$

where

$$s_\Delta(f) \;=\; 1 + \sum_{k=1}^{N} t_k e_k \big[1 - 6\alpha_\Delta(t_k)(1 - \alpha_\Delta(t_k))\big]\;, \tag{2.15}$$

where $e_k = f(t_k^+) - f(t_k^-)$ *is the height of the discontinuity at* t_k*, where* $\alpha_\Delta(t_k) \stackrel{\Delta}{=} \frac{t_k - u_\Delta(t_k)}{\Delta}$ *is the fractional position of* t_k *within its cell, and where the summand is taken to be zero when* $t_k = 0$*, even if* $f(t_k^+)$ *and/or* $f(t_k^-)$ *are infinite.*

From the above theorem, Lemma 5 and Corollary 6, we obtain the following corollaries:

Corollary 11. *If* f *is a nice pdf, then for any offset function*

$$\begin{aligned} EX(M-X) &= \frac{\Delta^2}{12}(s_\Delta(f) - 1) + o(\Delta^2) \ , \\ EM(M-X) &= \frac{\Delta^2}{12} s_\Delta(f) + o(\Delta^2) \ , \\ EM^2 &= EX^2 + \frac{\Delta^2}{12}(2s_\Delta(f) - 1) + o(\Delta^2) \ . \end{aligned}$$

Corollary 12. *If* f *is a nice pdf with no discontinuities, except perhaps at 0, then* $s_\Delta(f) = 1$ *for all* Δ*, and consequently* $r(f) = 1$ *and the additive noise model is asymptotically valid.*

On the one hand, when f is a nice pdf with no discontinuities except possibly at the origin, Corollary 12 shows that the additive noise model is asymptotically valid, i.e. for small values of Δ. On the other hand, when there are discontinuities elsewhere, Corollary 11 permits one to determine the validity of the additive noise model for any given Δ by computing $s_\Delta(f)$. It is conceivable that $s_\Delta(f)$ converges to some value depending on the t_k's and e_k's, but not on the offset function $\theta(\Delta)$. In this case for small values of Δ, it would be sufficient to know this value, so one would not have to be concerned about the detailed calculation of $s_\Delta(f)$ for the specific values of Δ and $\theta(\Delta)$ being used. Unfortunately, the following theorem shows that this is not possible.

Theorem 13. *If f is a nice pdf, and the set of exceptional points $T = \{t_1, \ldots, t_N\}$ where there are discontinuities is not comprised of only the single point 0, then there exists an offset function $\theta(\Delta)$ such that $\lim_{\Delta\to 0} s_\Delta(f)$ does not exist. Thus, $r(f) = \lim_{\Delta\to 0} r_\Delta(f)$ does not exist, and consequently, the additive noise model is not asymptotically valid.*

The effect of jump discontinuities: In light of Corollary 12 and Theorem 13, we observe that jump discontinuities have a determining effect on the correlation between quantizer input and error and the existence of $r(f)$. To see why, consider the case that f has a single finite jump discontinuity at t, and rewrite $r_\Delta(f)$ as

$$r_\Delta(f) \;=\; \int_{-\infty}^{u_\Delta(t)} G_\Delta(x)\,dx + \int_{u_\Delta(t)}^{u_\Delta(t)+\Delta} G_\Delta(x)\,dx + \int_{u_\Delta(t)+\Delta}^{\infty} G_\Delta(x)\,dx\ .$$

The methods used in the proof of Theorem 10 can be easily used to show that the left and right terms in the above converge to finite values. Therefore, the existence of $r(f)$ is determined by whether or not the middle term, which we now rewrite in greater detail, has a limit.

$$\int_{u_\Delta(t)}^{u_\Delta(t)+\Delta} G_\Delta(x)\,dx \;=\; 6\Big[m_\Delta(t)+c_\Delta(t)\Big]\left[\frac{1}{\Delta}\int_{u_\Delta(t)}^{u_\Delta(t)+\Delta} f(x)\,dx\right]\left[\frac{m_\Delta(t)-c_\Delta(t)}{\Delta}\right]\ , \tag{2.16}$$

where we used the fact that $m_\Delta(x)$ and $c_\Delta(x)$ are constant on quantization cells.

On the one hand, if f were continuous at t, it is easy to see that the right-hand side of the above tends to zero. Specifically, the first term in brackets approaches $2t$, the second approaches $f(t)$, and either $f(t) > 0$ in which case Lemma 1 implies that the third term goes to zero, or $f(t) = 0$ in which case the second term approaches 0, while the third has magnitude no larger than $1/2$.

On the other hand, when f has a finite jump discontinuity at $t \neq 0$, as illustrated in Figure 2.2, $\frac{m_\Delta(t)-c_\Delta(t)}{\Delta}$ no longer goes to zero, necessarily, as $\Delta \to 0$, as we will

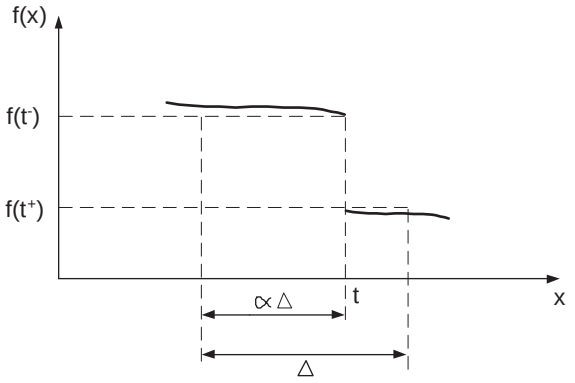

Figure 2.2: The pdf f, having a jump discontinuity at $x = t$, can be viewed as being approximately constant on the left and right parts of the cell containing t.

shortly demonstrate. In fact, it can be made to converge to different values depending on how Δ approaches zero. Furthermore, neither does $\frac{1}{\Delta}\int_{u_\Delta(t)}^{u_\Delta(t)+\Delta} f(x)\,dx$ converge. The important question is whether the product of these two terms converges. We will show that it does not. Therefore, $r(f)$ does not exist.

For example, fix $\theta(\Delta) = 0$ for all Δ and suppose Δ_n is a sequence going to zero as $n \to \infty$ in such a way that t always lies in the center of its cell; i.e. $t = u_{\Delta_n}(t) + \alpha\Delta_n$ for all n, where $\alpha = 1/2$. Assume further that f is constant in neighborhoods to the right and left of t, as will approximately be the case when Δ_n is small. Then the first term in (2.16) converges to $2t$, the second term converges to $(f(t^-)+f(t^+))/2$, and the third term can be straightforwardly shown to converge to $\frac{1}{4}\frac{f(t^-)-f(t^+)}{f(t^-)+f(t^+)}$. It follows that the right-hand side of (2.16) converges to $\frac{3}{2}t[f(t^-)-f(t^+)]$. (A careful derivation, not assuming f is constant in neighborhoods, is given later in the derivation of (2.26).) On the other hand, if $\Delta_n' \to 0$ in such a way that $t = u_{\Delta_n'}(t) + \alpha\Delta_n'$ for all n, with $\alpha \neq 1/2$, then the right-hand side of (2.16) converges to some other value. This implies that $r(f)$ does not exist.

We comment that although the contribution of the cell containing t is non-

vanishing, this fact alone is insufficient to invalidate the additive noise model, since it is conceivable that this substantial non-vanishing contribution might combine with the *sum* of vanishing contributions of all other cells (where f is continuous), which is substantial as well, so as to make $r_\Delta(f)$ converge to 1. However, as mentioned, the fact that the contribution of the cell containing t does not converge, while the sum of contributions of all other cells does converge, ultimately causes $r_\Delta(f)$ not to exist. Finally, one might imagine that if there were several jump discontinuities, then their non-converging contributions might perhaps cancel each other so that $r_\Delta(f)$ would still converge. This, however, cannot happen as shown by Theorem 13.

2.6 Uniform densities and quantizers with matched support

Consider a uniform pdf. It has discontinuities at each end of its support. Thus according to Theorem 13, the additive noise model is not asymptotically valid. While the discontinuities cause $r(f)$ not to exist, i.e. there are offset functions for which neither $s_\Delta(f)$ nor $r_\Delta(f)$ have limits, the simple fact that the midpoints are centroids (ignoring the cells containing the endpoints of the support of the pdf, for which midpoints might not equal centroids) would already lead one to suspect that $r(f)$ does not equal 1. Instead, in view of (2.3), one would more likely expect the quantization error to be approximately orthogonal to the output rather than the input, and from Corollaries 6 and 11, one would expect $r_\Delta(f) \approx s_\Delta(f) \approx 0$. Indeed, Theorem 10 shows this will be true if the endpoints of the pdf support are close to the thresholds of the quantizer, in which case the α's for the endpoints will be nearly zero or one, and the $t_k e_k$'s will sum to -1.

In view of the above discussion, for pdf's with finite support, such as uniform, it is interesting to consider the limiting characteristics of uniform quantizers in an

alternative framework. Specifically, if a pdf has support (a, b), consider the sequence of uniform quantizers such that the n^{th} quantizer partitions (a, b) into n cells of width $\Delta_n = (b - a)/n$, with thresholds exactly at a and b.[4] Such quantizers are said to have *matched support*.

Though we do not expect the usual additive noise model to be valid for quantizers with matched support, there can nevertheless be a well-defined asymptotic correlation structure, i.e. asymptotic formulas for the second moment of the output, and the correlations between input, output and quantization error. These are characterized by modified versions of $r_\Delta(f)$ and $r(f)$, namely,

$$\tilde{r}_n(f) \stackrel{\Delta}{=} \int_{-\infty}^{\infty} G_{(b-a)/n}(x)\, dx \ ,$$

and when $\tilde{r}_n(f)$ has a limit,

$$\tilde{r}(f) \stackrel{\Delta}{=} \lim_{n\to\infty} \tilde{r}_n(f) \ ,$$

which are just like $r_\Delta(f)$ and $r(f)$ except we now require that $\Delta_n = (b - a)/n$ and that there be thresholds at a and b. With $\tilde{r}(f)$ replacing $r(f)$, one may easily check that Lemma 5 and Corollary 6 apply for pdf's with finite support and uniform quantizers with matched support. Moreover, slightly modified versions Theorems of 9, 10 and Corollaries 11, 12 hold. The following is an example of what is possible.

Theorem 14. *Let f be a nice pdf with finite support (a, b) with no discontinuities except, possibly, jump discontinuities at a, b and the origin. Then for uniform quan-*

[4]Note that the offsets of the quantizers change with n.

tizers with matched support,

$$\begin{aligned} EM^2 &= EX^2 + (2\tilde{r}_n(f) - 1)D_{m,\Delta} + o(\Delta_n^2) \ , \\ EX(M-X) &= \frac{\Delta_n^2}{12}(\tilde{r}_n(f) - 1) + o(\Delta_n^2) \ , \\ EM(M-X) &= \frac{\Delta_n^2}{12}\tilde{r}_n(f) + o(\Delta_n^2) \ , \\ \tilde{r}(f) &= 1 + af(a^+) - bf(b^-) \ . \end{aligned} \tag{2.17}$$

Proof: The first three relations are derived just as in Lemma 5 and Corollary 6. The last relation follows by deriving a modified version Theorem 10, and then using the facts that $\alpha_{\Delta_n}(a) = \alpha_{\Delta_n}(b) = 0$ and that the corresponding $t_i e_i$ terms sum to $af(a^+) - bf(b^-)$. □

This theorem shows that for matched quantizers a variety of different correlations are possible, i.e. a variety of *alternative noise models* are possible. For a uniform source, $\tilde{r}(f) = 0$ and the theorem predicts the additive noise model illustrated in Figure 2.1b. However, for nonuniform pdf's, (2.17) indicates that appropriate choices of a, b, $f(a^+)$ and $f(b^-)$ can make $\tilde{r}(f)$ attain any value whatsoever, making possible a broad range of alternative noise models.

When the pdf has jump discontinuities within (a, b) that are not at the origin, $\tilde{r}$ will not exist, for reasons like those that cause r not to exist in Theorem 13. For such cases, one may develop a generalization of Theorem 10. Or one may try a more complicated matching such that all jump discontinuities occur at quantizer thresholds. This, however, is not always possible.

As a final set of options, we mention that one could also consider the family of uniform quantizers whose supports are *matched* to (a, b) in the sense of having cell midpoints at a and b, rather than boundaries at a and b. In this case, for a uniform pdf, Theorem 10 shows that the modified version of $r(f)$, again denoted $\tilde{r}(f)$, would

equal 3/2. More generally, for a uniform pdf on (a, b) and any $\alpha_1, \alpha_2 \in [0, 1)$, one could consider the family of uniform quantizers that are matched in the sense that $a = u_{\Delta_n}(a) + \alpha_1\Delta_n$ and $b = u_{\Delta_n}(b) + \alpha_2\Delta_n$. In this case, Theorem 10 implies that $\tilde{r}(f) = \frac{6}{b-a}\big(b\alpha_2(1-\alpha_2) - a\alpha_1(1-\alpha_1)\big)$. Thus, one could obtain a wide range of modified $r(f)$ values. One might even attempt to obtain $\tilde{r}(f) = 1$, so that the additive noise model would be valid.

2.7 Proofs

Proof of Theorem 9:

Let f be nice with no exceptional points. Let $\theta(\Delta)$ be an arbitrary offset function, let (a, b) be some finite interval, and let us write

$$r_\Delta(f) \;=\; \int_{-\infty}^{a} G_\Delta(x)\,dx + \int_a^b G_\Delta(x)\,dx + \int_b^{\infty} G_\Delta(x)\,dx\;. \tag{2.18}$$

The proof follows by taking the limit of the above, while using Facts 2 and 3 below.

Fact 1: For all $x \in B$, $\lim_{\Delta\to 0} G_\Delta(x)$ exists and equals $-xf'(x)$.

Fact 2: (a) $\lim_{\Delta\to 0}\int_a^b G_\Delta(x)\,dx = \int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx$.

(b) $\int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx = af(a) + \int_a^b f(x)\,dx - bf(b)$.

Fact 3: $\lim_{\Delta\to 0}\int_{-\infty}^{a} G_\Delta(x)\,dx \;=\; \int_{-\infty}^{a} f(x)\,dx - af(a)$ and $\lim_{\Delta\to 0}\int_b^{\infty} G_\Delta(x)\,dx \;=\; bf(b) + \int_b^{\infty} f(x)\,dx$.

Proof of Fact 1: $\lim_{\Delta\to 0} G_\Delta(x) = -xf'(x)$, for $x \in B$.

We will use the following lemma, proved in the Appendix, which provides a stronger statement than that of Lemma 1, under stronger conditions. A similar result was shown in [6]. However, the conditions set here are less restrictive and the statement of this lemma is more precise.

Lemma 15. *If f is positive and differentiable at x, then for any offset function*

$$c_\Delta(x) \;=\; m_\Delta(x) + \frac{\Delta^2}{12}\frac{f'(x)}{f(x)} + o(\Delta^2) \;,$$

or equivalently,

$$\lim_{\Delta\to 0} \frac{m_\Delta(x) - c_\Delta(x)}{\Delta^2} \;=\; -\frac{f'(x)}{12f(x)} \;.$$

To prove Fact 1, we begin by considering some $x \in B$, and expanding $G_\Delta(x)$:

$$G_\Delta(x) \;=\; \frac{m_\Delta^2(x) - c_\Delta^2(x)}{\Delta^2/6} f(x) \;=\; 6\left[\frac{m_\Delta(x) - c_\Delta(x)}{\Delta^2}\right][m_\Delta(x) + c_\Delta(x)]\, f(x) \;.$$

If $f(x) > 0$, then Lemma 15 shows that the first bracketed term converges to $-\frac{f'(x)}{12f(x)}$ as $\Delta \to 0$. The second bracketed term goes to $2x$ as $\Delta \to 0$. Therefore,

$$\lim_{\Delta\to 0} G_\Delta(x) \;=\; -xf'(x) \;. \tag{2.19}$$

If $f(x) = 0$, then $G_\Delta(x) = 0$ for any Δ. Also, $f'(x) = 0$ (otherwise we could move in the direction that would make f negative). Thus (2.19) holds again, which completes the proof of Fact 1.

Proof of Fact 2a: $\lim_{\Delta\to 0}\int_a^b G_\Delta(x)\,dx = \int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx$.

We will use the bounded convergence theorem [11] (p. 210) to show that the limit and integral can be swapped. Fact 1 showed that the limit of the integrand $G_\Delta(x)$ exists almost everywhere. The bounded convergence theorem also requires that $|G_\Delta(x)|$ be uniformly bounded, which we now show. Since f is nice with no exceptional points, there exists $S < \infty$ such that $|f'| \le S$ wherever f' exists. For any $x \in (a, b)$, Lemma A2 of the Appendix shows that for all sufficiently small Δ,

$$|G_\Delta(x)| \;\le\; 12(|x| + \Delta)S \;<\; 24\max\{|a|, |b|\}S \;,$$

where we used the fact that for all sufficiently small Δ, $(|x|+\Delta) < 2\max\{|a|,|b|\}$. It follows that $|G_\Delta(x)|$ is uniformly bounded for $x \in (a,b)$. Finally, since the integration is over a set of finite measure, the bounded convergence theorem[5] implies

$$\lim_{\Delta\to 0}\int_a^b G_\Delta(x)\,dx \;=\; \int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx\ .$$

Proof of Fact 2b: $\int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx = af(a) + \int_a^b f(x)\,dx - bf(b)$.

Fact 1 implies $\int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx = \int_a^b -xf'(x)\,dx$. Since f is piecewise differentiable, and B is the union of disjoint open intervals on each of which f' exists, we have that there exists some K such that $B \cap (a,b) = \cup_{i=1}^K (y_i, y_{i+1})$, where $y_1 = a$ and $y_{K+1} = b$. Applying integration by parts to each open interval (y_i, y_{i+1}), we obtain

$$\begin{aligned}
\int_a^b \lim_{\Delta\to 0} G_\Delta(x)\,dx \;&=\; \sum_{i=1}^K \int_{y_i}^{y_{i+1}} -xf'(x)\,dx \\
&=\; \sum_{i=1}^K \left(y_i f(y_i) + \int_{y_i}^{y_{i+1}} f(x)\,dx - y_{i+1}f(y_{i+1}) \right) \\
&=\; af(a) + \int_a^b f(x)\,dx - bf(b)\ . \qquad (2.20)
\end{aligned}$$

Proof of Fact 3: $\lim_{\Delta\to 0}\int_{-\infty}^a G_\Delta(x)\,dx \;=\; \int_{-\infty}^a f(x)\,dx \;-\; af(a)$ and $\lim_{\Delta\to 0}\int_b^\infty G_\Delta(x)\,dx = bf(b) + \int_b^\infty f(x)\,dx$.

We will show $\lim_{\Delta\to 0}\int_b^\infty G_\Delta(x)\,dx = bf(b) + \int_b^\infty f(x)\,dx$. The result for the other integral follows in a similar way. We decompose the integral, $\lim_{\Delta\to 0}\int_b^\infty G_\Delta(x)\,dx$ into $\lim_{\Delta\to 0}\sum_{k=0}^\infty \int_{b_k}^{b_{k+1}} G_\Delta(x)\,dx$, where $b_k \stackrel{\Delta}{=} b + k$. Our main goal is to show that the limit and sum can be swapped. To do so we shall make use of the following version of the Weierstrass M-test [11] (p. 543).

[5] It is easily shown that the theorem applies when the integrand is parameterized by some t converging continuously to some t_o, rather than some integer n converging to ∞.

Lemma 16. *Let $\Phi_k(\Delta)$, $k \in \mathbb{Z}$ be a sequence of functions such that $\lim_{\Delta\to 0} \Phi_k(\Delta)$ exists, $|\Phi_k(\Delta)| \le M_k$ for $0 < \Delta < \delta$, for some $\delta > 0$, and $\sum_{k=-\infty}^{\infty} M_k < \infty$. Then $\sum_{k=-\infty}^{\infty} \Phi_k(\Delta)$ exists for $0 < \Delta < \delta$, and*

$$\lim_{\Delta\to 0} \sum_{k=-\infty}^{\infty} \Phi_k(\Delta) \;=\; \sum_{k=-\infty}^{\infty} \lim_{\Delta\to 0} \Phi_k(\Delta) \;.$$

Define $\Phi_k(\Delta) \stackrel{\Delta}{=} \int_{b_k}^{b_{k+1}} G_\Delta(x)\,dx$ and write,

$$\int_b^\infty G_\Delta(x)\,dx \;=\; \sum_{k=0}^{\infty} \int_{b_k}^{b_{k+1}} G_\Delta(x)\,dx \;=\; \sum_{k=0}^{\infty} \Phi_k(\Delta) \;. \tag{2.21}$$

We would like to apply Lemma 16 to the right-hand term of (2.21). By Fact 2, $\lim_{\Delta\to 0} \Phi_k(\Delta)$ exists. We now find a sequence M_k, whose sum is finite, that dominates the sequence $|\Phi_k(\Delta)|$. We begin by bounding $|\Phi_k(\Delta)|$. Let $S < \infty$ be the uniform bound on the derivative of f. Fix δ, $0 < \delta < 1$ and consider throughout $0 < \Delta < \delta$. Recalling that f is piecewise differentiable, let $W_k \stackrel{\Delta}{=} B \cap (b_k - \delta, b_{k+1} + \delta)$ denote the subset of the interval $(b_k - \delta, b_{k+1} + \delta)$ over which f is differentiable. Let $S_k \stackrel{\Delta}{=} \sup_{x\in W_k} |f'(x)|$. Since f is a nice pdf, $\lim_{\substack{x\to\infty \\ x\in B}} x^{2+\varepsilon} f'(x) = 0$ for some $\varepsilon > 0$. Thus, there exists a nonnegative integer N, $N > 4 - b$ (i.e. $N \ge 0$ and $b_N > 4$) such that $x^{2+\varepsilon}|f'(x)| < 1$, or equivalently, $|f'(x)| < \frac{1}{x^{2+\varepsilon}}$ for all $x \in [b_N - 1, \infty) \cap B$. Therefore, when $k \ge N$, $S_k \le \frac{1}{(b_k-\delta)^{2+\varepsilon}} < \frac{1}{(b_k-1)^{2+\varepsilon}}$. Using this we obtain

$$\begin{aligned} \big|\Phi_k(\Delta)\big| \;&\le\; \int_{b_k}^{b_{k+1}} |G_\Delta(x)|\;dx \;\stackrel{(a)}{\le}\; 12S_k(|b_k| + 1 + \delta) \;\stackrel{(b)}{<}\; 12S_k(|b_k| + 2) \\ &\stackrel{(c)}{<} \begin{cases} 12S(|b| + N + 2), & 0 \le k < N \\ 24\frac{1}{(b_k-1)^{1+\varepsilon}}, & k \ge N \end{cases} \;\stackrel{\Delta}{=}\; M_k \;, \end{aligned}$$

where (a) follows from Lemma A2, (b) uses $\delta < 1$, and (c) is due to having $|b_k| < |b| + N$ for $0 \le k < N$, and $|b_k| + 2 < 2(b_k - 1)$ for $k \ge N$.

Next, we need to show that $\sum_{k=0}^{\infty} M_k < \infty$, which can be seen as follows:

$$\begin{aligned}\sum_{k=0}^{\infty} M_k &= \sum_{k=0}^{N-1} 12S(|b|+N+2) + \sum_{k=N}^{\infty} \frac{24}{(b_k-1)^{1+\varepsilon}} \\ &< 12S(|b|+N+2)N + 24\sum_{k=3}^{\infty} \frac{1}{k^{1+\varepsilon}} < \infty ,\end{aligned}$$

where the first inequality is due to $b_N - 1 > 3$. Thus, taking the limit as $\Delta \to 0$ in (2.21) we obtain

$$\begin{aligned}\lim_{\Delta\to 0}\int_b^{\infty} G_\Delta(x)\,dx &= \lim_{\Delta\to 0}\sum_{k=0}^{\infty}\Phi_k(\Delta) \overset{a}{=} \sum_{k=0}^{\infty}\lim_{\Delta\to 0}\Phi_k(\Delta) \\ &= \sum_{k=0}^{\infty}\lim_{\Delta\to 0}\int_{b_k}^{b_{k+1}} G_\Delta(x)\,dx \\ &\overset{b}{=} \sum_{k=0}^{\infty}\left(b_k f(b_k) + \int_{b_k}^{b_{k+1}} f(x)\,dx - b_{k+1}f(b_{k+1})\right) \\ &= \lim_{N\to\infty}\sum_{k=0}^{N}\left(b_k f(b_k) + \int_{b_k}^{b_{k+1}} f(x)\,dx - b_{k+1}f(b_{k+1})\right) \\ &= bf(b) + \lim_{N\to\infty}\int_b^{b_{N+1}} f(x)\,dx - \lim_{N\to\infty} b_{N+1}f(b_{N+1}) \\ &\overset{c}{=} bf(b) + \int_b^{\infty} f(x)\,dx ,\end{aligned}$$

where (a) follows from Lemma 16, (b) follows from applying Fact 2 to intervals of the form (b_k, b_{k+1}), and (c) is obtained by applying Lemma A5 of the Appendix, which shows $\lim_{x\to\infty} xf(x) = 0$, since f is a pdf with finite mean and $\lim_{x\to\infty} xf'(x) = 0$. □.

Proof of Theorem 10:

Let $\{s_1, ..., s_n\}$ be the exceptional points of f, let $\{v_0, v_1, \ldots, v_n\}$ be chosen so that $-\infty < v_0 < s_1 < v_1 < s_2 < \ldots < v_{n-1} < s_n < v_n < \infty$, and let us write

$$\int_{-\infty}^{\infty} G_\Delta(x)\,dx = \int_{-\infty}^{v_0} G_\Delta(x)\,dx + \sum_{i=1}^{n}\int_{v_{i-1}}^{v_i} G_\Delta(x)\,dx + \int_{v_n}^{\infty} G_\Delta(x)\,dx . \quad (2.22)$$

It follows from Fact 3 in the proof of Theorem 9, which applies even if f has exceptional points, that the first and third integrals on the right-hand side above converge to $\int_{-\infty}^{v_0} f(x)\,dx - v_0 f(v_0)$ and $v_n f(v_n) + \int_{v_n}^{\infty} f(x)\,dx$, respectively, as $\Delta \to 0$. We further decompose each integral in the sum term above as follows:

$$\int_{v_{i-1}}^{v_i} G_\Delta(x)\,dx \;=\; \int_{v_{i-1}}^{u_\Delta(s_i)-2\Delta} G_\Delta(x)\,dx + \int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)\,dx + \int_{u_\Delta(s_i)+3\Delta}^{v_i} G_\Delta(x)\,dx\,. \tag{2.23}$$

Since the treatment of the first and last terms above is similar, we will only consider the first. With the above decomposition in mind, the proof will derive from following two facts.

Fact 1: (a) $\lim_{\Delta\to 0} \int_{v_{i-1}}^{u_\Delta(s_i)-2\Delta} G_\Delta(x)\,dx = \int_{v_{i-1}}^{s_i} \lim_{\Delta\to 0} G_\Delta(x)\,dx$.

(b) $\int_{v_{i-1}}^{s_i} \lim_{\Delta\to 0} G_\Delta(x)\,dx = v_{i-1} f(v_{i-1}) + \int_{v_{i-1}}^{s_i} f(x)\,dx - s_i f(s_i^-)$.

Fact 2: (a) $\int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)\,dx = o(1)$, when f is continuous at s_i or when $s_i = 0$.

(b) $\int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)\,dx = 6s_i[f(s_i^-) - f(s_i^+)]\alpha_\Delta(s_i)[1-\alpha_\Delta(s_i)] + o(1)$, when f has a finite jump discontinuity at s_i and $s_i \neq 0$.

Combining (2.22), the discussion right after it, (2.23), and the above two facts, it follows that

$$\begin{aligned}
\int_{-\infty}^{\infty} G_\Delta(x)\,dx \;=\; & \int_{-\infty}^{v_0} f(x)\,dx - v_0 f(v_0) \\
& + \sum_{i=1}^{n} \left(v_{i-1} f(v_{i-1}) + \int_{v_{i-1}}^{s_i} f(x)\,dx - s_i f(s_i^-) + s_i f(s_i^+) \right. \\
& \qquad \left. + \int_{s_i}^{v_i} f(x)\,dx - v_i f(v_i) \right) \\
& + \sum_{i=1}^{n} \Big(6s_i \big[f(s_i^-) - f(s_i^+) \big] \alpha_\Delta(s_i) \big[1 - \alpha_\Delta(s_i) \big] \Big) \\
& + v_n f(v_n) + \int_{v_n}^{\infty} f(x)\,dx + o(1)\,.
\end{aligned}$$

Therefore, recalling that $T = \{t_1, \ldots, t_N\}$ is the set of exceptional points where there are discontinuities in f, we may rewrite the above as

$$r_\Delta(f) \stackrel{\Delta}{=} \int_{-\infty}^{\infty} G_\Delta(x)\,dx = 1+\sum_{k=1}^{N} t_k(f(t_k^+)-f(t_k^-))\big[1-6\alpha_\Delta(t_k)(1-\alpha_\Delta(t_k))\big]+o(1)\,,$$

which will conclude the proof of the theorem. We now prove the two facts.

Proof of Fact 1a: $\lim_{\Delta\to 0}\int_{v_{i-1}}^{u_\Delta(s_i)-2\Delta} G_\Delta(x)\,dx = \int_{v_{i-1}}^{s_i}\lim_{\Delta\to 0} G_\Delta(x)\,dx$.

To simplify notation, we write v for v_{i-1} and s instead of s_i. We begin with

$$\int_v^{u_\Delta(s)-2\Delta} G_\Delta(x)\,dx = \int_v^s G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx\,,$$

where $I_F(x)$ denotes the indicator function of the event F. Observe that for any $x \in (v,s)$, $\lim_{\Delta\to 0} G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x) = \lim_{\Delta\to 0} G_\Delta(x)$. We will use the bounded and dominated convergence theorems to show that when taking the limit of the right-hand side, the integral can be swapped with the limit.

There are four cases to consider, depending on the behavior of f and f' on (v,s):

Case (i): f is continuous and bounded, and f' is bounded wherever it exists,

Case (ii): f is continuous and bounded, and $f' \nearrow \infty$ monotonically as $x \nearrow s$ in some left neighborhood of s,

Case (iii): f is continuous and bounded, and $f' \searrow -\infty$ monotonically as $x \nearrow s$ in some left neighborhood of s,

Case (iv): $s = 0$, and on (v,s), $f(x) \to \infty$ as $x \nearrow s$, and $f'(x) \nearrow \infty$ monotonically as $x \nearrow s$ in some left neighborhood of s.

Case (i): (On (v,s), f is continuous and bounded, and f' is bounded wherever it exists.) The proof, which uses the bounded convergence theorem, is similar to that of Fact 2a in the proof of Theorem 9. Let $S < \infty$ be such that $|f'(x)| \le S$ for

all $x \in B \cap (v-\delta, s)$ for some $\delta > 0$. Then for $\Delta < \delta$ and $x \in (v, u_\Delta(s) - 2\Delta)$, Lemma A2 shows that

$$|G_\Delta(x)| \;\le\; 12(|x|+\Delta)S \;<\; 12(\max\{|v|,|s|\}+\delta)S\;.$$

Therefore, $|G_\Delta(x)| I_{(v,u_\Delta(s)-2\Delta)}(x)$ is uniformly bounded on (v,s). The bounded convergence theorem then gives $\lim_{\Delta\to 0}\int_v^s G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx = \int_v^s \lim_{\Delta\to 0} G_\Delta(x)\,dx$.

Case (ii): (On (v,s), f is continuous and bounded, and $f' \nearrow \infty$ monotonically as $x \nearrow s$ in some left neighborhood of s.) Let w be chosen so that $v < w < s$, f' exists everywhere and increases monotonically to infinity on (w,s), $f'(w) > 0$, and $|f'(x)| < f'(w)$ for $x \in B \cap (v,w)$. Then

$$\begin{aligned}\int_v^s G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx \;&=\; \int_v^w G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx \\ &\quad + \int_w^s G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx\;. \qquad (2.24)\end{aligned}$$

Since $f'(x)$ is bounded, wherever it exists on (v,w), the same argument as that in Fact 2a in the proof of Theorem 9 yields, $\lim_{\Delta\to 0}\int_v^w G_\Delta(x) I_{(v,u_\Delta(s)-2\Delta)}(x)\,dx = \int_v^w \lim_{\Delta\to 0} G_\Delta(x)\,dx$.

To justify swapping the limit and integral in the second term of (2.24), we use the dominated convergence theorem[6] [11] (p. 209), which requires us to find an integrable function $\widetilde{G}(x)$ that dominates $|G_\Delta(x)| I_{(v,u_\Delta(s)-2\Delta)}(x)$ for all $x \in (w,s)$ and all Δ. With $M \stackrel{\Delta}{=} \max\{|w|,|s|\}$, we choose

$$\widetilde{G}(x) \;=\; 24M f'\big(\frac{x+s}{2}\big),\quad w < x < s\;.$$

To show that $\widetilde{G}$ dominates $|G_\Delta| I$, we first observe that for $w < x < u_\Delta(s) - 2\Delta$, the positivity and monotonicity of f' on (w,s) implies $|f'(x)| = f'(x) \le f'(u_\Delta(x)+\Delta)$.

[6] As with the bounded convergence theorem, the integrands index parameter is allowed to approach 0 continuously.

Then Lemma A2 implies that for all sufficiently small Δ and for $w < x < u_\Delta(s) - 2\Delta$

$$\begin{aligned} |G_\Delta(x)| \;&\le\; 12(|x| + \Delta) f'(u_\Delta(x) + \Delta) \;\le\; 24M f'(u_\Delta(x) + \Delta) \\ &\le\; 24M f'(x + \Delta) \le 24M f'\big(\frac{x+s}{2}\big) \;=\; \widetilde{G}(x)\ , \end{aligned}$$

where the third inequality uses the monotonicity of f' on (w, s), and the fourth inequality derives from the fact that $x < u_\Delta(s) - 2\Delta$ implies $x < s - 2\Delta$, which in turn implies $\Delta < \frac{s-x}{2}$. It follows that $|G_\Delta(x)| I_{(v, u_\Delta(s)-2\Delta)}(x) \le \widetilde{G}(x)$ for all $x \in (w, s)$.

We now check that $\widetilde{G}(x)$ is integrable over (w, s):

$$\begin{aligned} \int_w^s \widetilde{G}(x)\, dx \;&=\; 24M \int_w^s f'\big(\frac{x+s}{2}\big)\, dx \;=\; 24M \int_{\frac{w+s}{2}}^s 2f'(y)\, dy \\ &=\; 48M\big(f(s) - f\big(\frac{w+s}{2}\big)\big) \;<\; \infty\ , \end{aligned}$$

where the inequality follows from the fact that f is bounded. Applying the dominated convergence theorem yields $\lim_{\Delta\to 0} \int_w^s G_\Delta(x) I_{(w, u_\Delta(s)-2\Delta)}\, dx = \int_w^s \lim_{\Delta\to 0} G_\Delta(x)\, dx$, which concludes Case (ii). Case (iii) is proved in the same manner.

Case (iv): ($s = 0$, and on (v, s), $f(x) \to \infty$ as $x \nearrow s$, and $f'(x) \nearrow \infty$ monotonically as $x \nearrow s$ in some left neighborhood of s.) This case is similar to Case (ii), up to a point. Let w be chosen so that $v < w < s = 0$, f' exists everywhere and increases monotonically to ∞ on $(w, 0)$, $f'(w) > 0$, and $|f'(x)| < f'(w)$ for $x \in B \cap (v, w)$. Then,

$$\begin{aligned} \int_v^0 G_\Delta(x) I_{(v, u_\Delta(0)-2\Delta)}(x)\, dx \;&=\; \int_v^w G_\Delta(x) I_{(v, u_\Delta(0)-2\Delta)}(x)\, dx \\ &\quad + \int_w^0 G_\Delta(x) I_{(v, u_\Delta(0)-2\Delta)}(x)\, dx\ . \end{aligned} \tag{2.25}$$

Since $f'(x)$ is bounded, wherever it exists on (v, w), the same argument as that in Fact 2a in the proof of Theorem 9 yields, $\lim_{\Delta\to 0} \int_v^w G_\Delta(x) I_{(v, u_\Delta(0)-2\Delta)}(x)\, dx = \int_v^w \lim_{\Delta\to 0} G_\Delta(x)\, dx$.

To justify the swapping of limit and integral in the second term of (2.25), we use the dominated convergence theorem. As the function that dominates $|G_\Delta(x)|I_{(v,u_\Delta(0)-2\Delta)}(x)$, we choose

$$\widetilde{G}(x) = 18|x|f'\big(\frac{x}{2}\big), \quad w < x < 0 \ .$$

To show that this is indeed a dominating function, recall that Lemma A2 shows that for $w < x < u_\Delta(0) - 2\Delta$

$$|G_\Delta(x)| \ \leq \ 12(|x| + |\Delta|)f'(u_\Delta(x) + \Delta) \ ,$$

where we used the positivity and monotonicity of f' on $(w, 0)$. Now if $w < x < u_\Delta(0) - 2\Delta$, then $x < -2\Delta$, or equivalently, $\Delta < -x/2$. Using these and using again the positivity and monotonicity of f' yields for $w < x < u_\Delta(0) - 2\Delta$

$$|G_\Delta(x)| \ < \ 12(|x|+|x|/2)f'(u_\Delta(x)+\Delta) \ \leq \ 18|x|f'(x+\Delta) \ \leq \ 18|x|f'\big(\frac{x}{2}\big) \ = \ \widetilde{G}(x) \ .$$

This in turn implies, $|G_\Delta(x)|I_{(v,u_\Delta(0)-2\Delta)}(x) < \widetilde{G}(x)$, for all $x \in (w, 0)$.

We now check the integrability of $\widetilde{G}$:

$$\begin{aligned}
\int_w^0 \widetilde{G}(x)\,dx \ &= \ -18\int_w^0 xf'(\frac{x}{2})\,dx \ = \ -72\int_{\frac{w}{2}}^0 yf'(y)\,dy \\
&= \ -72\lim_{x\to 0} xf(x) + 72wf(w) + \int_w^0 f(x)\,dx \\
&= \ 0 + 72wf(w) + \int_w^0 f(x)\,dx \ < \ \infty \ ,
\end{aligned}$$

where the third equality uses integration by parts, and the fourth equality derives from the definition of a nice pdf. Applying the dominated convergence theorem yields $\lim_{\Delta\to 0}\int_w^0 G_\Delta(x)I_{(w,u_\Delta(0)-2\Delta)}\,dx = \int_w^0 \lim_{\Delta\to 0} G_\Delta(x)\,dx$, which concludes Case (iv). This completes the proof of Fact 1a.

Proof of Fact 1b: $\int_{v_{i-1}}^{s_i} \lim_{\Delta\to 0} G_\Delta(x)\,dx = v_{i-1}f(v_{i-1}^+) + \int_{v_{i-1}}^{s_i} f(x)\,dx - s_i f(s_i^-)$.

This follows in a similar way to Fact 2b in the proof of Theorem 9, where v_{i-1} and s_i play the role of a and b, respectively.

Proof of Fact 2a: $\int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)\,dx = o(1)$, when f is continuous at s_i or when $s_i = 0$.

The considered integral of $G_\Delta(x)$ is over five adjacent quantization cells. We will show that the integral over each of these cells approaches 0 as $\Delta \to 0$. To simplify notation, we write s instead of s_i. There are two cases: $s = 0$ and $s \neq 0$.

If $s = 0$, then the integral over any one of the five quantization cells has the form

$$\int_{u_\Delta(0)-j\Delta}^{u_\Delta(0)-j\Delta+\Delta} G_\Delta(x)\,dx = 6\left[\frac{m_\Delta(0-j\Delta) - c_\Delta(0-j\Delta)}{\Delta}\right]\left[\frac{m_\Delta(0-j\Delta) + c_\Delta(0-j\Delta)}{\Delta}\right] \times \left[\int_{u_\Delta(0)-j\Delta}^{u_\Delta(0)-j\Delta+\Delta} f(x)\,dx\right],$$

for some $j \in \{-2,-1,0,1,2\}$. The magnitude of the first bracketed term is at most one half, the magnitude of the second bracketed term is easily seen to be no larger than 6, and the third bracketed term goes to zero as $\Delta \to 0$. Therefore, $\int_{u_\Delta(0)-j\Delta}^{u_\Delta(0)-j\Delta+\Delta} G_\Delta(x)\,dx = o(1)$, and since this holds for the integral over each of the five adjacent cells, the result follows in the case $s = 0$.

Next, if $s \neq 0$, then f is continuous at s. In this case

$$\int_{u_\Delta(s)-j\Delta}^{u_\Delta(s)-j\Delta+\Delta} G_\Delta(x)\,dx = 6\left[\frac{m_\Delta(s-j\Delta) - c_\Delta(s-j\Delta)}{\Delta}\right]\left[m_\Delta(s-j\Delta) + c_\Delta(s-j\Delta)\right] \times \left[\frac{1}{\Delta}\int_{u_\Delta(s)-j\Delta}^{u_\Delta(s)-j\Delta+\Delta} f(x)\,dx\right].$$

Suppose $f(s) = 0$. Then the magnitude of the first bracketed term is at most one half, the second bracketed term tends to $2s$ as $\Delta \to 0$, and due to the continuity of f at s the third bracketed term tends to $f(s) = 0$. Therefore, the product of the three

bracketed terms approaches zero. Now suppose $f(s) \neq 0$. By Lemma A1, which is a slightly strengthened version of Lemma 1, the first bracketed term approaches 0. The second and third terms approach $2s$ and $f(s)$, respectively, as before. Therefore, again the product of the three bracketed terms approaches zero. This completes the proof of Fact 2a.

Proof of Fact 2b: $\int_{u_\Delta(s_i)-2\Delta}^{u_\Delta(s_i)+3\Delta} G_\Delta(x)\,dx = 6s_i[f(s_i^-)-f(s_i^+)]\alpha_\Delta(s_i)[1-\alpha_\Delta(s_i)]+o(1)$, when f has a finite jump discontinuity at s_i and $s_i \neq 0$.

As before, to simplify notation, we write s instead of s_i. We shall also write α_Δ instead of $\alpha_\Delta(s)$. First observe that having a discontinuity at s has no effect on the integral over the four non middle cells. Thus, the integral over these cells tends to zero as $\Delta \to 0$ as shown in Fact 2a. It remains to consider the integral over the middle cell, which contains the discontinuity at s. Specifically, it needs to be shown that

$$\int_{u_\Delta(s)}^{u_\Delta(s)+\Delta} G_\Delta(x)\,dx \;=\; 6s[f(s^-) - f(s^+)]\alpha_\Delta(1-\alpha_\Delta) + o(1)\ . \tag{2.26}$$

We decompose the integral above as follows:

$$\int_{u_\Delta(s)}^{u_\Delta(s)+\Delta} G_\Delta(x)\,dx \;=\; \left[\frac{m_\Delta(s) - c_\Delta(s)}{\Delta/6}\right]\left[m_\Delta(s)+c_\Delta(s)\right]\left[\frac{1}{\Delta}\int_{s-\Delta\alpha_\Delta}^{s+\Delta(1-\alpha_\Delta)} f(x)\,dx\right]\ . \tag{2.27}$$

The three terms in (2.27) converge as follows:

$$\frac{1}{\Delta}\int_{s-\Delta\alpha_\Delta}^{s+\Delta(1-\alpha_\Delta)} f(x)\,dx \;=\; \alpha_\Delta f(s^-) + (1-\alpha_\Delta)f(s^+) + o(1)\ , \tag{2.28}$$

$$m_\Delta(s) + c_\Delta(s) \;=\; 2s + o(1)\ , \tag{2.29}$$

$$\frac{m_\Delta(s) - c_\Delta(s)}{\Delta/6} \;=\; 6\frac{\frac{\alpha_\Delta(1-\alpha_\Delta)}{2}[f(s^-) - f(s^+)]}{\alpha_\Delta f(s^-) + (1-\alpha_\Delta)f(s^+)} + o(1)\ , \tag{2.30}$$

where (2.28) is due to the continuity of f on $(s-\Delta\alpha_\Delta, s)$ and $(s, s+\Delta(1-\alpha_\Delta))$. Equation (2.30) can be obtained by observing that $m_\Delta(s) = s + \Delta(\frac{1}{2} - \alpha_\Delta)$ and by

noting that it can be shown that $c_\Delta(s) = \gamma_{L,\Delta} c_{L,\Delta}(s) + \gamma_{R,\Delta} c_{R,\Delta}(s)$, where $\gamma_{L,\Delta} = \frac{\int_{u_\Delta(s)}^{s} f(x)\,dx}{\gamma_\Delta}$, $\gamma_{R,\Delta} = \frac{\int_{s}^{u_\Delta(s)+\Delta} f(x)\,dx}{\gamma_\Delta}$, $\gamma_\Delta = \int_{u_\Delta(s)}^{u_\Delta(s)+\Delta} f(x)\,dx$ and where $c_{L,\Delta}(s)$ is the centroid of $(u_\Delta(s), s)$ and $c_{R,\Delta}(s)$ is the centroid of $(s, u_\Delta(s)+\Delta)$. Next, by Lemma 15 it follows that $c_{L,\Delta}(s) = m_{L,\Delta}(s) + O(\Delta^2)$ and $c_{R,\Delta}(s) = m_{R,\Delta}(s) + O(\Delta^2)$, where $m_{L,\Delta}(s) = s - \frac{\Delta\alpha_\Delta}{2}$ is the midpoint of $(u_\Delta(s), s)$ and $m_{R,\Delta}(s) = s + \frac{\Delta(1-\alpha_\Delta)}{2}$ is the midpoint of $(s, u_\Delta(s) + \Delta)$. Thus, we have

$$\begin{aligned}\frac{m_\Delta(s) - c_\Delta(s)}{\Delta} &= \frac{m_\Delta(s)}{\Delta} - \gamma_{L,\Delta}\frac{\big(m_{L,\Delta}(s) + O(\Delta^2)\big)}{\Delta} - \gamma_{R,\Delta}\frac{\big(m_{R,\Delta}(s) + O(\Delta^2)\big)}{\Delta} \\ &= \gamma_{L,\Delta}\frac{\big(m_\Delta(s) - m_{L,\Delta}(s) + O(\Delta^2)\big)}{\Delta} + \gamma_{R,\Delta}\frac{\big(m_\Delta - m_{R,\Delta}(s) + O(\Delta^2)\big)}{\Delta}\ .\end{aligned}$$

Using the continuity of f on the intervals $(u_\Delta(s), s)$ and $(s, u_\Delta(s) + \Delta)$, it is easily seen that $\gamma_{L,\Delta} = \frac{\alpha_\Delta f(s^-)}{\alpha_\Delta f(s^-)+(1-\alpha_\Delta)f(s^+)} + o(1)$ and $\gamma_{R,\Delta} = \frac{(1-\alpha_\Delta)f(s^+)}{\alpha_\Delta f(s^-)+(1-\alpha_\Delta)f(s^+)} + o(1)$. Plugging these into the above, together with some algebraic steps, establishes (2.30).

Finally, since $\alpha_\Delta f(s^-)+(1-\alpha_\Delta)f(s^+)$, $2s$ and $6\frac{\frac{\alpha_\Delta(1-\alpha_\Delta)}{2}[f(s^-)-f(s^+)]}{\alpha_\Delta f(s^-)+(1-\alpha_\Delta)f(s^+)}$ are bounded, substituting (2.28), (2.29), (2.30) into (2.27) yields (2.26), which completes the proof of Fact 2b and Theorem 10. □

Proof of Theorem 13:

We need to show that there exists some offset function $\theta(\Delta)$ for which $\lim_{\Delta\to 0} s_\Delta(f)$ does not exist. We begin by writing

$$\begin{aligned} s_\Delta(f) &= 1 + \sum_{k=1}^{N} t_k e_k - 6\sum_{k=1}^{N} t_k e_k \alpha_\Delta(t_k)(1-\alpha_\Delta(t_k)) \\ &= 1 + \sum_{k=1}^{N} t_k e_k - 6\langle \underline{\lambda}, \underline{\beta}_\Delta\rangle\ , \end{aligned} \tag{2.31}$$

where $e_k = f(t_k^+) - f(t_k^-)$ is the jump height at t_k, $\alpha_\Delta(t_k) = \frac{t_k - u_\Delta(t_k)}{\Delta}$ is the offset of t_k within its cell, $\underline{\lambda} = (\lambda_1, \ldots, \lambda_N)$, $\lambda_k = t_k e_k$, $\underline{\beta}_\Delta = (\beta_{1,\Delta}, \ldots, \beta_{N,\Delta})$, $\beta_{k,\Delta} = \alpha_\Delta(t_k)(1-\alpha_\Delta(t_k))$, and $\langle \underline{u}, \underline{v}\rangle = \sum_{k=1}^{N} u_k v_k$ is the usual inner product. We use underbars throughout to denote vectors.

Since the first two terms of (2.31) do not depend on Δ, it suffices to find an offset function $\theta(\Delta)$ such that $\lim_{\Delta\to 0}\langle \underline{\lambda}, \underline{\beta}_\Delta\rangle$ does not exist. To this end we will set $\Delta_\tau = \frac{1}{\tau}$, and let $\tau \to \infty$. We will then show that there exists a fixed quantity τ_δ such that for any $\tau_o > 0$, there exists $\tau_1 \geq \tau_o$, a value θ_{τ_1}, and an interval $(\tau_1, \tau_1 + 2\tau_\delta)$ such that with $\theta(\tau) = \theta_{\tau_1}$ in this interval, the above inner product varies by some nonzero and known in advance amount. This will then imply that $\langle \underline{\lambda}, \underline{\beta}(\Delta_\tau)\rangle$ does not converge, which is equivalent to $\lim_{\Delta\to 0}\langle \underline{\lambda}, \underline{\beta}_\Delta\rangle$ not existing.

We notice that if $t_k = 0$, then by definition of $s_\Delta(f)$, the discontinuity at t_k contributes nothing to $s_\Delta(f)$. Thus, without loss of generality, we assume $t_k \neq 0$ for all $k \in \{1, \ldots, N\}$. Furthermore, without loss of generality we assume that the components of $\underline{t} = (t_1, \ldots, t_N)$ are ordered by magnitude, i.e. $0 < |t_1| \leq |t_2| \leq \cdots \leq |t_{N-1}| \leq t_N$, where we also assume without loss of generality that $t_N > 0$. To simplify matters, we change slightly our notation for the α's and β's. Specifically, let $\alpha_k^\phi(\tau)$ denote the α value at t_k when $\Delta = \Delta_\tau$ and $\theta(\Delta) = \phi$. Similarly, let $\beta_k^\phi(\tau)$ denote the corresponding β value. In this notation,

$$\begin{aligned} \alpha_k^\phi(\tau) &= \frac{\big(t_k + \Delta_\tau\phi\big) \bmod \Delta_\tau}{\Delta_\tau} = \frac{\Delta_\tau[\big(t_k/\Delta_\tau + \phi\big) \bmod 1]}{\Delta_\tau} \\ &= \big(\frac{t_k}{\Delta_\tau} + \phi\big) \bmod 1 = \big(t_k\tau + \phi\big) \bmod 1 \,. \end{aligned} \tag{2.32}$$

In addition, from now on, we consider the offset to be a function of τ, namely $\theta(\tau)$, rather than a function of Δ. Our goal will be to find an offset function $\theta(\tau)$ for which we can show that $\lim_{\tau\to\infty}\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau)\rangle$ does not exist, where $\underline{\beta}^{\theta(\tau)} = (\beta_1^{\theta(\tau)}(\tau), \ldots, \beta_N^{\theta(\tau)}(\tau))$.

Before going into the details of the proof, we give an intuitive view of the meaning of α in light of (2.32), followed by an outline of the proof. We identify the unit interval with the unit circle, with 0 located at 12 o'clock. As τ goes to ∞, we view $\alpha_k^{\theta(\tau)}(\tau)$

as rotating around the unit circle — clockwise when $t_k > 0$, and counterclockwise when $t_k < 0$. If $\theta(\tau)$ is constant over an interval of τ's, then we see from (2.32) that each $\alpha_k^{\theta(\tau)}$ changes linearly with τ unless it *passes through 0*, in which case the mod 1 comes into effect — subtracting 1 from α if $t_k > 0$, and adding 1 if $t_k < 0$.

It is easy to see from (2.32) that if the offset were held constant, then no α_k would converge and consequently no β_k and no product $\lambda_k\beta_k$ would converge. The difficulty lies in showing that the inner product, which is the sum of products $\lambda_k\beta_k$, does not converge either. Essentially, one must show that nonconvergent terms in the sum cannot somehow negate each other's non convergence. We do this by choosing an offset function that is piecewise constant rather than constant. Specifically, we show that for any $\tau_o > 0$ there exists $\tau_1 \geq \tau_o$, an interval $[\tau_1, \tau_1 + 2\tau_\delta)$, and a constant offset in this interval that cause the following favorable property to hold. All α_k's, except α_N, do not pass through zero and, consequently, change linearly with τ over this interval. α_N, on the other hand, passes through zero in the middle of the interval (but nowhere else). Thus, it changes linearly over the first half of the interval and has a discontinuity as τ passes to the second half of the interval and the mod 1 comes into effect.

Using this property, the inner product $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau) \rangle$ turns out to be a parabolic function of τ, $A\tau^2 + B\tau + C$, in the first half of the interval. If A is not zero or B is bounded away from zero for large values of τ_o, then it is easily shown that the inner product must change by some nonzero amount that can be specified in advance. Otherwise, we use the discontinuity in α_N at the halfway point of the interval to lower bound the amount of change.

To keep notation short, we will assume throughout the proof that $k \in \{1, \ldots, N\}$, unless otherwise specified. We set $\mu = \min_{k \in \{1,\ldots,N-1\}} t_N - t_k$ and $\delta = \frac{\mu}{32Nt_N}$, which

remain fixed for the rest of the proof. Let W denote the set of all $\tau \geq 0$ such that

$$\left|\alpha_N^0(\tau) - \alpha_k^0(\tau)\right|_L \;\geq\; 4\delta \;, \quad \text{for all } k \neq N \;, \tag{2.33}$$

where $|a-b|_L$ denotes Lee distance, i.e., $|a-b|_L = \min\{(a-b) \bmod 1, 1-((a-b) \bmod 1)\}$. The following lemma asserts that W is unbounded.

Lemma 17. *For any $\tau_o > 0$ there exists $\tau \geq \tau_o$ such that $\tau \in W$.*

Proof: The proof is constructive. If $\tau_o \in W$, there is nothing to show. From now on, assume $\tau_o \notin W$. Consider first $\tau_1 = \tau_o + \frac{8\delta}{\mu}$. For every k such that $|\alpha_N^0(\tau) - \alpha_k^0(\tau)|_L < 4\delta$, we have

$$\begin{aligned} \left|\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)\right|_L \;&=\; \left|\alpha_N^0(\tau_o) + \frac{8\delta}{\mu}t_N - \alpha_k^0(\tau_o) - \frac{8\delta}{\mu}t_k\right|_L \\ &=\; \left|\alpha_N^0(\tau_o) - \alpha_k^0(\tau_o) + \frac{8\delta}{\mu}(t_N - t_k)\right|_L \;. \end{aligned} \tag{2.34}$$

Combining (2.34) and the facts that $\frac{8\delta}{\mu}(t_N - t_k) \geq 8\delta$ and $|\alpha_N^0(\tau) - \alpha_k^0(\tau)|_L < 4\delta$, it follows straightforwardly that $|\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)|_L > 4\delta$.

We have shown that $|\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)|_L > 4\delta$ for those k's considered above. However, now that τ has been increased from τ_o to τ_1, it is possible that other α_k's no longer satisfy (2.33). We can "fix" these by increasing τ, yet again, to $\tau_2 = \tau_1 + \frac{8\delta}{\mu}$. Since t_N has largest magnitude, the distance between α_N and previously fixed α's will only increase, and so a fixed α need not be fixed again. Thus repeating this process at most $N-1$ times will guarantee that all α_k's are fixed, i.e. (2.33) is satisfied for all $k \neq N$, provided we make one additional check.

From a geometrical point of view, the process of fixing some α_k involves sufficient advancement of α_N in clockwise direction, thus letting α_N gain sufficient distance from α_k. We observe, however, that the assertion that "repeating this process at most $N-1$ times will guarantee that all α_k's are fixed" is correct if the distance

between a previously fixed α_k and α_N cannot become small again due to having α_N getting close to α_k from the "other" direction as τ is increased. This, however, cannot happen, since in $N-1$ steps, τ increases from τ_o to $\tau_{N-1} = \tau_o + (N-1)\frac{8\delta}{\mu} = \tau_o + (N-1)\frac{8}{\mu}\frac{\mu}{32Nt_N} < \tau_o + \frac{1}{4t_N}$. Consequently, α_N advances less than $\frac{1}{4t_N}t_N = \frac{1}{4}$. Since t_N has largest magnitude, all other α_k's advance less than $\frac{1}{4}$. Therefore, $\alpha_N^0(\tau_{N-1})$ is at least $\frac{1}{2}$ away in clockwise direction from any $\alpha_k^0(\tau_{N-1})$ that was fixed using the above process. □

Now, we use Lemma 17 to show that for any τ_o there exists $\tau_1 \geq \tau_o$, an interval $[\tau_1, \tau_1 + 2\tau_\delta)$ and a choice of constant offset in this interval with the favorable property discussed in the proof outline described earlier.

Let $\tau_o > 0$ be given. It follows from Lemma 17 that there exist $\tau_1 \geq \tau_o$ for which $|\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)|_L \geq 4\delta$ for all $k \neq N$. Fix $\theta_{\tau_1} = \big(1 - \delta - \alpha_N^0(\tau_1)\big) \bmod 1$, and let $\theta(\tau) = \theta_{\tau_1}$ for $\tau_1 \leq \tau < \tau + 2\tau_\delta$, where $\tau_\delta \stackrel{\Delta}{=} \frac{\delta}{t_N} = \frac{\mu}{32Nt_N^2}$. This choice of τ_1 and θ_{τ_1} makes

$$\alpha_N^{\theta_{\tau_1}}(\tau_1) \;=\; (t_N\tau_1 + \theta_{\tau_1}) \bmod 1 \;=\; \big(\alpha_N^0(\tau_1) + 1 - \delta - \alpha_N^0(\tau_1)\big) \bmod 1 \;=\; 1 - \delta \;.$$

Moreover, since $t_N\tau_\delta = \delta$, we deduce from (2.32) and the above that $\alpha_N^{\theta_{\tau_1}}(\tau)$ increases linearly from its value $1-\delta$ at $\tau = \tau_1$, and passes through zero precisely at $\tau = \tau_1 + \tau_\delta$. And since $\delta = \min_{k\in\{1,\ldots,N-1\}} t_N - t_k/(32Nt_N) \leq 1/(16N)$, it will not pass through zero anywhere else in the interval $[\tau_1, \tau_1 + 2\tau_\delta)$. It follows that

$$\alpha_N^{\theta_{\tau_1}}(\tau_1 + s) \;=\; \begin{cases} \alpha_N^{\theta_{\tau_1}}(\tau_1) + st_N, & 0 \leq s < \tau_\delta \\ \alpha_N^{\theta_{\tau_1}}(\tau_1) + st_N - 1, & \tau_\delta \leq s < 2\tau_\delta \end{cases}, \tag{2.35}$$

Next, for any $k \neq N$, the facts that $\alpha_N^{\theta_{\tau_1}}(\tau_1) = 1 - \delta$, $|\alpha_N^{\theta_{\tau_1}}(\tau_1) - \alpha_k^{\theta_{\tau_1}}(\tau_1)|_L = |\alpha_N^0(\tau_1) - \alpha_k^0(\tau_1)|_L \geq 4\delta$, and $|t_N| \geq |t_k|$ imply that $\alpha_k^{\theta_{\tau_1}}(\tau)$ cannot pass through zero

in the interval $[\tau_1, \tau_1 + 2\tau_\delta)$. Therefore,

$$\alpha_k^{\theta_{\tau_1}}(\tau_1 + s) \;=\; \alpha_k^{\theta_{\tau_1}}(\tau_1) + st_k, \quad 0 \le s < 2\tau_\delta \;\text{ and }\; k \neq N \;, \tag{2.36}$$

Having established (2.35) and (2.36) we are now ready to express $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau)\rangle$ as a parabolic function of τ, when $\tau \in [\tau_1, \tau_1 + \tau_\delta)$. To do so, we observe that for all $s \in [0, \tau_\delta)$ and for all k,

$$\begin{aligned}
\beta_k^{\theta_{\tau_1}}(\tau_1 + s) \;&=\; (\alpha_k^{\theta_{\tau_1}}(\tau_1) + st_k)(1 - \alpha_k^{\theta_{\tau_1}}(\tau_1) - st_k) \\
&=\; \alpha_k^{\theta_{\tau_1}}(\tau_1)\big(1 - \alpha_k^{\theta_{\tau_1}}(\tau_1)\big) + t_k\big(1 - 2\alpha_k^{\theta_{\tau_1}}(\tau_1)\big)s - t_k^2 s^2 \\
&=\; \beta_k^{\theta_{\tau_1}}(\tau_1) + t_k\big(1 - 2\alpha_k^{\theta_{\tau_1}}(\tau_1)\big)s - t_k^2 s^2 \;. \qquad (2.37)
\end{aligned}$$

Using the above we evaluate $\langle \underline{\lambda}, \underline{\beta}^{\theta_{\tau_1}}(\tau_1 + s)\rangle$ for $s \in [0, \tau_\delta)$ as follows,

$$\begin{aligned}
\langle \underline{\lambda}, \underline{\beta}^{\theta_{\tau_1}}(\tau_1 + s)\rangle \;&=\; \sum_{k=1}^{N} \lambda_k \beta_k^{\theta_{\tau_1}}(\tau_1) + \left[\sum_{k=1}^{N} \lambda_k t_k\big(1 - 2\alpha_k^{\theta_{\tau_1}}(\tau_1)\big)\right] s + \left[-\sum_{k=1}^{N} \lambda_k t_k^2\right] s^2 \\
&=\; As^2 + B_{\tau_1} s + C_{\tau_1} \;, \qquad (2.38)
\end{aligned}$$

where $A \triangleq -\sum_{k=1}^{N} \lambda_k t_k^2$, $B_{\tau_1} \triangleq \sum_{k=1}^{N} \lambda_k t_k\big(1 - 2\alpha_k^{\theta_{\tau_1}}(\tau_1)\big)$ and $C_{\tau_1} \triangleq \sum_{k=1}^{N} \lambda_k \beta_k^{\theta_{\tau_1}}(\tau_1)$.

We proceed by showing how to lower bound the amount of change in the inner product. To keep notation short, we set $\mathcal{Y}_\tau(s) \triangleq \langle \underline{\lambda}, \underline{\beta}^{\theta_\tau}(\tau + s)\rangle$. If $A \neq 0$, then (2.38) shows that $\mathcal{Y}_{\tau_1}(s)$ is a parabolic function. Therefore,

$$\big|\mathcal{Y}_{\tau_1}(\tfrac{\tau_\delta}{4}) - \mathcal{Y}_{\tau_1}(0) \;\ge\; |A|\big(\tfrac{\tau_\delta}{4}\big)^2 \quad \text{and/or} \quad \big|\mathcal{Y}_{\tau_1}(\tfrac{\tau_\delta}{2}) - \mathcal{Y}_{\tau_1}(\tfrac{\tau_\delta}{4})\big| \;\ge\; |A|\big(\tfrac{\tau_\delta}{4}\big)^2 \;, \tag{2.39}$$

which derives from the fact that if $y(x) = ax^2 + bx + c$ and $a \neq 0$, then for any $t \in \mathbb{R}$, $|y(t) - y(0)| \ge |a|t^2$ and/or $|y(2t) - y(t)| \ge |a|t^2$. This fact can be seen as follows. $y(0) = c$, $y(t) = at^2 + bt + c$ and $y(2t) = 4at^2 + 2bt + c$. Thus, $|y(t) - y(0)| = |at^2 + bt|$ and if $|at^2 + bt| \ge |a|t^2$, then the fact is shown. Otherwise, $|y(2t) - y(t)| = |3at^2 + bt| = |(at^2 + bt) + 2at^2| > |2at^2 - at^2| = |a|t^2$, which shows the fact.

Thus, for the case $A \neq 0$, we have established a lower bound to the change of the inner product $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau) \rangle$ as τ ranges over the interval $[\tau_o, \infty)$.

Suppose next that A=0. Then (2.38) reduces to

$$\mathcal{Y}_{\tau_1}(s) \;=\; B_{\tau_1} s + C_{\tau_1}, \qquad \text{when} \;\; s \in [0, \tau_\delta) \; . \tag{2.40}$$

If $\lim_{\substack{\tau \to \infty \\ \tau \in W}} B_\tau \neq 0$ (in particular the limit need not exist), then for any $\tau_o > 0$, there exists $\tau_1 \in W$ such that $\tau_1 \geq \tau_o$ and $|B_{\tau_1}| > u \stackrel{\Delta}{=} \frac{1}{2} \limsup_{\substack{\tau \to \infty \\ \tau \in W}} |B_\tau| > 0$. Consequently,

$$\big|\mathcal{Y}_{\tau_1}(\frac{\tau_\delta}{2}) - \mathcal{Y}_{\tau_1}(0)\big| \;=\; |B_{\tau_1}| \frac{\tau_\delta}{2} \;>\; u \frac{\tau_\delta}{2} \;=\; u \frac{\delta}{2t_N} \; . \tag{2.41}$$

This establishes a lower bound to the change of the inner product $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau) \rangle$ for the case $A = 0$ and $\lim_{\substack{\tau \to \infty \\ \tau \in W}} B_\tau \neq 0$.

It remains to consider the case that $A = 0$ and $\lim_{\substack{\tau \to \infty \\ \tau \in W}} B_\tau = 0$. For any $\tau_o > 0$, there exists $\tau_1 \in W$ such that $\tau_1 \geq \tau_o$ and $|B_{\tau_1}| < \frac{|\lambda_N| \delta}{\tau_\delta}$. Using (2.35) we obtain that for $\tau_\delta \leq s < 2\tau_\delta$,

$$\begin{aligned} \beta_N^{\theta_{\tau_1}}(\tau_1 + s) \;&=\; (\alpha_N^{\theta_{\tau_1}}(\tau_1) + s t_N - 1)(2 - \alpha_N^{\theta_{\tau_1}}(\tau_1) - s t_N) \\ &=\; \beta_N^{\theta_{\tau_1}}(\tau_1) + t_N\big(1 - 2\alpha_N^{\theta_{\tau_1}}(\tau_1)\big)s - t_N^2 s^2 - 2\big[1 - \alpha_N^{\theta_{\tau_1}}(\tau_1) - s t_N\big] \; . \end{aligned} \tag{2.42}$$

Using (2.42) and observing that the expression in (2.37) holds for $s \in [\tau_\delta, 2\tau_\delta)$ and $1 \leq k \leq N-1$, it follows via a derivation similar to that of (2.40), that for all $s \in [\tau_\delta, 2\tau_\delta)$,

$$\mathcal{Y}_{\tau_1}(s) \;=\; \big(B_{\tau_1} s + C_{\tau_1}\big) + \big(- 2\lambda_N [1 - \alpha_N^{\theta_{\tau_1}}(\tau_1)] + 2\lambda_N t_N s\big) \; . \tag{2.43}$$

Consequently,

$$\big|\mathcal{Y}_{\tau_1}(\frac{7\tau_\delta}{4}) - \mathcal{Y}_{\tau_1}(\frac{5\tau_\delta}{4})\big| \;=\; \big|B_{\tau_1} \frac{\tau_\delta}{2} + \lambda_N t_N \tau_\delta\big| \;>\; \frac{|\lambda_N| \delta}{2} \; , \tag{2.44}$$

where the inequality follows from having $|B_{\tau_1}| < \frac{|\lambda_N|\delta}{\tau_\delta}$. This establishes a lower bound to the change of the inner product $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau) \rangle$ for the case $A = 0$ and $\lim_{\substack{\tau \to \infty \\ \tau \in W}} B_\tau = 0$.

Combining (2.39), (2.41) and (2.44) we have shown that for any $\tau_o > 0$ there exist $\tau_b > \tau_a \geq \tau_o$ such that

$$\left| \langle \underline{\lambda}, \underline{\beta}(\tau_b) \rangle - \langle \underline{\lambda}, \underline{\beta}(\tau_a) \rangle \right| \; > \; |A| \left(\frac{\delta}{4t_N} \right)^2 + \frac{\delta}{2|t_N|} V_1 + \frac{|\lambda_N|\delta}{2} V_2 \; ,$$

where V_1 is a quantity that equals 0 if $A \neq 0$, or if $A = 0$ and $\lim_{\substack{\tau \to \infty \\ \tau \in W}} B_\tau = 0$. Otherwise, $V_1 = u$. And V_2 is a quantity that equals 0 if $A \neq 0$, or if $A = 0$ and $\lim_{\substack{\tau \to \infty \\ \tau \in W}} B_\tau \neq 0$. Otherwise, $V_2 = 1$. This shows that $\langle \underline{\lambda}, \underline{\beta}^{\theta(\tau)}(\tau) \rangle$ does not converge and concludes the proof of the theorem. □

Remark: There is a simple way to prove Theorem 13 for almost all vectors $\underline{t} \in \mathbb{R}^N$. Specifically, for the cases that $\underline{t}$ is rationally independent (i.e. for $\underline{t}$'s such that for all nonzero $\underline{h} \in \mathbb{Z}^N$, $\langle \underline{t}, \underline{h} \rangle \notin \mathbb{Z}$), which is almost all of $\mathbb{R}^N$. The following is a brief sketch. Fix $\theta(\Delta) = 0$ and set $\Delta_n = \frac{1}{n}$ for $n \in \mathbb{Z}^+$. From (2.32) we have that $\alpha_{k,n} = t_k n \bmod 1$. Since $\underline{t}$ is rationally independent, it follows via a theorem of Kronecker [12] (c.f. [13] (p. 158), which also cites [14]) that the sequence $\{\alpha_{1,n}, \ldots, \alpha_{N,n}\}_{n=1}^{\infty}$ is dense in $[0,1)^N$ and so the sequence $\{\beta_{1,n}, \ldots, \beta_{N,n}\}_{n=1}^{\infty}$ is dense in $[0, \frac{1}{4}]^N$. Therefore, there exists a subsequence n_l of n, such that β_{k,n_l} is arbitrarily small for all $k \in \{1, \ldots, N-1\}$ and β_{N,n_l} is dense in $[0, \frac{1}{4}]$. Finally, $\langle \underline{\lambda}, \underline{\beta}_{n_l} \rangle = \sum_{k=1}^{N} \lambda_k \beta_{k,n_l} \approx \lambda_N \beta_{N,n_l}$. Since β_{N,n_l} is dense in $[0, \frac{1}{4}]$, it follows that $\langle \underline{\lambda}, \underline{\beta}_{n_l} \rangle$ does not converge (in fact it has an uncountable number of limit points) and consequently $\langle \underline{\lambda}, \underline{\beta}_n \rangle$ does not converge.

2.8 Conclusions

Corollary 12 rigorously establishes that the widely used additive noise model for uniform scalar quantization is, as one would hope, valid in an asymptotic sense whenever the input pdf is continuous and also satisfies certain other benign conditions.

Specifically, the correlation between input and quantization error is asymptotically negligible relative to the MSE, or equivalently, to the square of the quantizer level spacing Δ. The model is even valid when there is a discontinuity at the origin. On the other hand, Theorem 13 shows that discontinuities elsewhere can cause the correlation between the input and quantization error to no longer be negligible relative to the MSE. In such cases, the additive noise model is not asymptotically valid. Nevertheless, Theorem 10 permits one to estimate the correlation when Δ is small, in terms of the heights of the discontinuities and their fractional positions within quantization cells.

The derivation of these results is based on an analysis of the asymptotic convergence of cell centroids to cell midpoints, as expressed in the functional r. This convergence is shown to be fast enough to account for the fact that the distortion induced by midpoints is asymptotically the same as that induced by centroids. But it is not fast enough to cause the correlation induced by midpoints to be similar to that induced by centroids.

For a pdf with finite support, such as a uniform pdf, we have also shown that it is possible to design the uniform quantizer to be matched to the support in such a way that the correlation has an asymptotic limit. Depending on the pdf and the manner of matching, a wide variety of correlations may be possible.

Finally, it is interesting to consider that any discontinuous pdf can be well approximated by a continuous pdf. For example, suppose a pdf with jump discontinuities is approximated by a continuous pdf that replaces each jump with a ramp of width δ. When $\Delta << \delta$, $r(f) \approx 1$ and the additive noise model is valid. On the other hand, when $\Delta \geq \delta$, the value of $r(f)$ can be quite far from 1, and consequently, the correlation need not be small relative to the MSE.

Appendix

Lemma 1:

We will prove the following slightly stronger version of Lemma 1.

Lemma A1. *If f is continuous and positive at x, then for any offset function and any integer j*

$$\lim_{\Delta\to 0} \frac{m_\Delta(x-j\Delta) - c_\Delta(x-j\Delta)}{\Delta} \;=\; 0\ .$$

Proof: Suppose f is continuous and positive at x. Recall that $u_\Delta(x)$ denotes the left boundary of the cell containing x. It follows from the definitions of $m_\Delta(x)$ and $c_\Delta(x)$ that

$$\frac{m_\Delta(x-j\Delta) - c_\Delta(x-j\Delta)}{\Delta} \;=\; \frac{\frac{1}{\Delta^2}\int_{u_\Delta(x-j\Delta)}^{u_\Delta(x-j\Delta)+\Delta} (u_\Delta(x-j\Delta)+\frac{\Delta}{2}-t)f(t)\,dt}{\frac{1}{\Delta}\int_{u_\Delta(x-j\Delta)}^{u_\Delta(x-j\Delta)+\Delta} f(t)\,dt}\ . \quad \text{(A1)}$$

Since f is continuous at x, the denominator of the above converges to $f(x)$. Now consider the numerator:

$$\begin{aligned}
&\frac{1}{\Delta^2}\int_{u_\Delta(x-j\Delta)}^{u_\Delta(x-j\Delta)+\Delta} (u_\Delta(x-j\Delta)+\frac{\Delta}{2}-t)f(t)\,dt \\
&\quad= \frac{1}{\Delta^2}\int_0^{\frac{\Delta}{2}} (\frac{\Delta}{2}-y)f(x+(u_\Delta(x-j\Delta)-x+y))\,dy \\
&\qquad+ \frac{1}{\Delta^2}\int_{\frac{\Delta}{2}}^{\Delta} (\frac{\Delta}{2}-y)f(x+(u_\Delta(x-j\Delta)-x+y))\,dy\ .
\end{aligned}$$

Let $\varepsilon > 0$ be given. Since f is continuous at x, there exists $\delta > 0$ such that $|f(x+t)-f(x)| < \varepsilon$ for all $t \in (-\delta,\delta)$. Therefore, when $\Delta < \frac{\delta}{j+2}$ and $0 < y < \Delta$, we have $|u_\Delta(x-j\Delta)-x+y| < (j+2)\Delta < \delta$, which in turn implies $|f(x+(u_\Delta(x-j\Delta)-x+y))-f(x)| < \varepsilon$. Using this in the right hand side of the above, we have that for all sufficiently small Δ,

$$\frac{1}{\Delta^2}\int_0^{\frac{\Delta}{2}} (\frac{\Delta}{2}-y)f(x+(u_\Delta(x-j\Delta)-x+y))\,dy \;<\; \frac{f(x)+\varepsilon}{8}\ ,$$

and

$$\frac{1}{\Delta^2}\int_{\frac{\Delta}{2}}^{\Delta}(\frac{\Delta}{2}-y)f(x+(u_\Delta(x-j\Delta)-x+y))\,dy \;<\; \frac{\varepsilon-f(x)}{8}\;.$$

Thus, $\frac{1}{\Delta^2}\int_0^{\Delta}(\frac{\Delta}{2}-y)f(u_\Delta(x-j\Delta)+y)\,dy < \frac{\varepsilon}{4}$ for all sufficiently small Δ. In much the same way, it can be shown that $\frac{1}{\Delta^2}\int_0^{\Delta}(\frac{\Delta}{2}-y)f(u_\Delta(x-j\Delta)+y)\,dy > -\frac{\varepsilon}{4}$. Since ε is arbitrary, we obtain that $\lim_{\Delta\to 0}\frac{1}{\Delta^2}\int_0^{\Delta}(\frac{\Delta}{2}-y)f(u_\Delta(x-j\Delta)+y)\,dy = 0$, i.e. the numerator of (A1) converges to zero. Since the denominator converges to $f(x)\neq 0$, we conclude that $\lim_{\Delta\to 0}\frac{m_\Delta(x-j\Delta)-c_\Delta(x-j\Delta)}{\Delta} = 0$, which completes the proof of the lemma. □

Proof of Lemma 2:

Let f be a continuous a.e. pdf. Let us define $W_\Delta(x) \stackrel{\Delta}{=} \frac{[c_\Delta(x)-m_\Delta(x)]^2}{\Delta^2}$. We may then write,

$$\lim_{\Delta\to 0}\frac{E(C_\Delta-M_\Delta)^2}{\Delta^2} \;=\; \lim_{\Delta\to 0}\int_{-\infty}^{\infty}W_\Delta(x)f(x)\,dx\;. \tag{A2}$$

To show that the limit above is zero, we will swap the limit and the integral using the bounded convergence theorem. We may view the integration as being with respect to the measure $\mu(E) \stackrel{\Delta}{=} \int_E f(x)\,dx$ [11] (p. 214), and then the integration is over a set of finite measure. Furthermore, since $|c_\Delta(x)-m_\Delta(x)| \le \frac{\Delta}{2}$ for all x and Δ, it follows that $0 \le W_\Delta(x) \le \frac{1}{4}$ for all x and Δ. Hence, $W_\Delta(x)$ is uniformly bounded for all x and Δ. Therefore, using the bounded convergence theorem,

$$\lim_{\Delta\to 0}\int_{-\infty}^{\infty}W_\Delta(x)f(x)\,dx \;=\; \int_{-\infty}^{\infty}\lim_{\Delta\to 0}W_\Delta(x)f(x)\,dx \;=\; \int_S\lim_{\Delta\to 0}W_\Delta(x)f(x)\,dx \;=\; 0\;,$$

where S denotes the set over which f is continuous and positive, and where the last equality follows from Lemma 1. □

Proof of Lemma 15:

Suppose f is positive and differentiable at x. As in the proof of Lemma A1,with $j = 0$,

$$\frac{m_\Delta(x) - c_\Delta(x)}{\Delta^2} \;=\; \frac{\frac{1}{\Delta^3}\int_0^\Delta (\frac{\Delta}{2} - y) f(u_\Delta(x) + y)\, dy}{\frac{1}{\Delta}\int_0^\Delta f(u_\Delta(x) + y)\, dy}\ , \tag{A3}$$

where the limit as $\Delta \to 0$ of the denominator of (A3) equals $f(x)$, since f is continuous at x.

We consider then the numerator of (A3). We begin by expressing $f(x + z)$ using the derivative:

$$f(x + z) \;=\; f(x) + zf'(x) + z\delta_x(z)\ , \tag{A4}$$

where $\delta_x(z)$ is a quantity that goes to zero as $z \to 0$. (Note that z may be either positive or negative.) Using (A4) to evaluate the numerator of (A3), we obtain

$$\begin{aligned}
\frac{1}{\Delta^3}\int_0^\Delta (\frac{\Delta}{2} - y) f(u_\Delta(x) + y)\, dy \;&=\; \frac{1}{\Delta^3}\int_0^\Delta (\frac{\Delta}{2} - y) f(x + (u_\Delta(x) - x + y))\, dy \\
&= \frac{1}{\Delta^3}\int_0^\Delta (\frac{\Delta}{2} - y) f(x)\, dy + \frac{1}{\Delta^3}\int_0^\Delta (\frac{\Delta}{2} - y)(u_\Delta(x) - x + y) f'(x)\, dy \\
&\quad + \frac{1}{\Delta^3}\int_0^\Delta (\frac{\Delta}{2} - y)(u_\Delta(x) - x + y)[\delta_x(u_\Delta(x) - x + y)]\, dy \\
&= 0 - \frac{f'(x)}{12} + \frac{1}{\Delta^3}\int_0^\Delta (\frac{\Delta}{2} - y)(u_\Delta(x) - x + y)[\delta_x(u_\Delta(x) - x + y)]\, dy \\
&= -\frac{f'(x)}{12} + \frac{1}{\Delta^3}\int_0^\Delta (\frac{\Delta}{2} - y)(u_\Delta(x) - x)[\delta_x(u_\Delta(x) - x + y)]\, dy \\
&\quad + \frac{1}{\Delta^3}\int_0^\Delta (\frac{\Delta}{2} - y) y [\delta_x(u_\Delta(x) - x + y)]\, dy\ ,
\end{aligned} \tag{A5}$$

It remains to show that the last two integrals in (A5) converge to zero as $\Delta \to 0$. Let $\varepsilon > 0$ be given. Since $(u_\Delta(x) - x + y) \in (-\Delta, \Delta]$ when $y \in [0, \Delta]$, and since $\delta_x(z) \to 0$ as $z \to 0$, it follows that for all sufficiently small Δ, $|\delta_x(u_\Delta(x) - x + y)| < \varepsilon$ for all $y \in [0, \Delta]$. Since $-\Delta < (u_\Delta(x) - x) \le 0$, it is not hard to see that $\frac{1}{\Delta^3}\big|\int_0^\Delta (\frac{\Delta}{2} - y)(u_\Delta(x) - x)[\delta_x(u_\Delta(x) - x + y)]\, dy\big| < \frac{\varepsilon}{4}$ and that $\frac{1}{\Delta^3}\big|\int_0^\Delta (\frac{\Delta}{2} - y) y [\delta_x(u_\Delta(x) - x + y)]\, dy\big| < \frac{\varepsilon}{8}$.

Since ε is arbitrary, it follows that

$$\lim_{\Delta\to 0} \frac{1}{\Delta^3} \int_0^{\Delta} (\frac{\Delta}{2} - y) f(u_\Delta(x) + y)\, dy \;=\; -\frac{f'(x)}{12}\ .$$

By combining this with the fact that the limit of the denominator of (A3) equals $f(x)$, we obtain that $\lim_{\Delta\to 0} \frac{m_\Delta(x) - c_\Delta(x)}{\Delta^2} = -\frac{f'(x)}{12f(x)}$. □

Lemma A2. *Let f, x, $\Delta > 0$ and $S \geq 0$ be such that f is a continuous and piecewise differentiable function on $(x-\Delta, x+\Delta)$ and $|f'(x)| \leq S$ for all $x \in (x-\Delta, x+\Delta)$ where f' exists. Then for any offset function*

$$|G_\Delta(x)| \;\leq\; 12(|x| + \Delta)S\ ,$$

where $G_\Delta(x)$ is as given in Definition 4.

Proof: First note that if $S = 0$, then the lemma holds trivially, since $G_\Delta(x) = 0$ (even if $f = 0$ over the interval $(x-\Delta, x+\Delta)$, since we recall that by convention $c_\Delta(x) = 0$). Suppose then that $S > 0$. We begin by writing

$$\begin{aligned} |G_\Delta(x)| \;&=\; \frac{|m_\Delta^2(x) - c_\Delta^2(x)|}{\Delta^2/6} f(x) \;=\; 6|m_\Delta(x) + c_\Delta(x)| \frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2} f(x) \\ &\leq\; 12(|x| + \Delta) \frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2} f(x)\ . \end{aligned} \tag{A6}$$

It remains to show $\frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2} f(x) \leq S$. We do this assuming $c_\Delta(x) \leq m_\Delta(x)$. The proof is essentially the same when the reverse inequality holds. There are two cases to consider:

1. $f(u_\Delta(x)) < S\Delta$: Since f is piecewise differentiable and $|f'(y)| \leq S$ for almost all $y \in (u_\Delta(x), u_\Delta(x) + \Delta)$, we have $f(x) \leq f(u_\Delta(x)) + S\Delta < 2S\Delta$, and consequently,

$$\frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2} f(x) \;<\; \frac{\Delta/2}{\Delta^2} 2S\Delta \;=\; S\ .$$

2. $f(u_\Delta(x)) \geq S\Delta$: Define $g(y) \stackrel{\Delta}{=} f(u_\Delta(x)) - S(y - u_\Delta(x))$ for $y \in (u_\Delta(x), u_\Delta(x) + \Delta)$. From Lemma A3, which appears next, it follows that $c^g_\Delta(x) \leq c^f_\Delta(x)$. Thus

$$\frac{|m_\Delta(x) - c^f_\Delta(x)|}{\Delta^2} f(x) \;\leq\; \frac{|m_\Delta(x) - c^g_\Delta(x)|}{\Delta^2} f(x) \;.$$

$c^g_\Delta(x)$ can be simplified to

$$c^g_\Delta(x) \;=\; u_\Delta(x) + \frac{\Delta}{2} - \frac{\frac{S\Delta^2}{12}}{f(u_\Delta(x)) - \frac{S\Delta}{2}} \;.$$

It follows that

$$\begin{aligned} \frac{|m_\Delta(x) - c^g_\Delta(x)|}{\Delta^2} f(x) \;&=\; \frac{S}{12} \frac{1}{f(u_\Delta(x)) - \frac{S\Delta}{2}} f(x) \;\leq\; \frac{S}{12} \frac{f(u_\Delta(x)) + S\Delta}{f(u_\Delta(x)) - \frac{S\Delta}{2}} \\ &=\; \frac{S}{12} \frac{k+1}{k-1/2} \;\leq\; \frac{S}{12} 4 \;<\; S \;, \end{aligned} \tag{A7}$$

where $k \stackrel{\Delta}{=} \frac{f(u_\Delta(x))}{S\Delta} \geq 1$ implies $(k+1)/(k-1/2) \leq 4$. This completes the proof that $\frac{|m_\Delta(x) - c_\Delta(x)|}{\Delta^2} f(x) \leq S$ and, consequently, the proof of the lemma. □

Lemma A3. *Let f be a continuous and piecewise differentiable function on $(u, u + \Delta)$. Let also $f(u) \geq S\Delta$ and $|f'(x)| \leq S$ for all $x \in W \cap (u, u+\Delta)$ for some $S > 0$ and $\Delta > 0$, where W is the set over which f is differentiable. Let $g(x) \stackrel{\Delta}{=} f(u) - S(x - u)$. Then $c^g_{\Delta,u} \leq c^f_{\Delta,u}$.*

Proof: First note that $g(u) = f(u)$ and g decreases as fast as possible among functions that satisfy the derivative constraint. If $g = f$ on $(u, u+\Delta)$ the lemma holds trivially. Suppose then that $g \neq f$ on a subset of $(u, u + \Delta)$ with positive measure. Let $h \stackrel{\Delta}{=} f - g$, or equivalently, $f = g + h$. Observe that for $x \in W \cap (u, u + \Delta)$, the fact that $|f'(x)| \leq S$ implies $h'(x) \geq 0$ and $h(x) \geq 0$. From the definition of centroid we may write

$$c^f_u \;=\; c^g_u \frac{\int_u^{u+\Delta} g(x)\,dx}{\int_u^{u+\Delta} f(x)\,dx} + c^h_u \frac{\int_u^{u+\Delta} h(x)\,dx}{\int_u^{u+\Delta} f(x)\,dx} \;.$$

Thus, c_u^f is a weighted average of c_u^g and c_u^h. Since g is a strictly decreasing function and h is an increasing function (though not necessarily strictly increasing), it is easy to see that $c_u^g < u + \frac{\Delta}{2} < c_u^h$. It follows that $c_u^f > c_u^g$ since c_u^f is the average of c_u^g and something larger. □

Lemma A4. *Let f be a nonnegative, continuous and piecewise differentiable function such that $\lim_{\substack{x\to-\infty \\ x\in W}} f'(x) = 0$ and $\lim_{\substack{x\to\infty \\ x\in W}} f'(x) = 0$, where W denotes the set over which f is differentiable. Let also $\int_{-\infty}^{\infty} f(x)\,dx < \infty$. Then, $\lim_{x\to-\infty} f(x) = 0$ and $\lim_{x\to\infty} f(x) = 0$.*

Proof: We will show that $\lim_{x\to\infty} f(x) = 0$. The other case follows in a similar way. Let $M \stackrel{\Delta}{=} \int_{-\infty}^{\infty} f(x)\,dx < \infty$. Let $\varepsilon > 0$ be given. Set $m = \frac{\varepsilon^2}{8M}$. There exists $T_1 > 0$ such that $|f'(x)| \le m$ for all $x \in (T_1, \infty) \cap W$. Since $\int_{-\infty}^{\infty} f(x)\,dx < \infty$, it follows that there exists $T_2 > T_1$ such that $f(T_2) \le \frac{\varepsilon}{2}$. Suppose there exists $T_3 > T_2$ such that $f(T_3) \ge \varepsilon$. Then,

$$f(x) \;\ge\; \begin{cases} m(x - T_3) + \varepsilon, & T_3 - \frac{\varepsilon}{2m} \le x \le T_3 \\ 0, & T_2 \le x < T_3 - \frac{\varepsilon}{2m} \end{cases},$$

Note that $T_3 - T_2 \ge \frac{\varepsilon}{2m}$. It now follows that

$$\int_{T_2}^{T_3} f(x)\,dx \;\ge\; \int_{T_3 - \frac{\varepsilon}{2m}}^{T_3} [m(x - T_3) + \varepsilon]\,dx \;=\; \int_0^{\frac{\varepsilon}{2m}} (\varepsilon - my)\,dy \;=\; \frac{3\varepsilon^2}{8m} \;=\; 3M\ ,$$

where the last equality follows from recalling that $m = \frac{\varepsilon^2}{8M}$. The above contradicts the fact that $\int_{-\infty}^{\infty} f(x)\,dx = M$. Therefore, $f(x) < \varepsilon$ for all $x > T_2$. Since ε is arbitrary, it follows that $\lim_{x\to\infty} f(x) = 0$. □

Lemma A5. *Let f be a nonnegative, continuous and piecewise differentiable function such that $\lim_{\substack{x\to-\infty\\ x\in W}} xf'(x)=0$ and $\lim_{\substack{x\to\infty\\ x\in W}} xf'(x)=0$, where W denotes the set over which f is differentiable. Let also $\int_{-\infty}^{\infty} f(x)\,dx<\infty$ and $\int_{-\infty}^{\infty}|x|f(x)\,dx<\infty$. Then, $\lim_{x\to-\infty} xf(x)=0$ and $\lim_{x\to\infty} xf(x)=0$.*

Proof: $\lim_{\substack{x\to-\infty\\ x\in W}} xf'(x)=0$ and $\lim_{\substack{x\to\infty\\ x\in W}} xf'(x)=0$ imply that $\lim_{\substack{x\to-\infty\\ x\in W}} f'(x)=0$ and $\lim_{\substack{x\to\infty\\ x\in W}} f'(x)=0$. Thus, applying Lemma A4 to the function f, we obtain that $\lim_{x\to-\infty} f(x)=0$ and $\lim_{x\to\infty} f(x)=0$. Next, define

$$g(x) \;\stackrel{\Delta}{=}\; \begin{cases} xf(x), & x\geq 0 \\ -xf(x), & x<0 \end{cases},$$

and obtain

$$g'(x) \;=\; \begin{cases} f(x)+xf'(x), & x\in[0,\infty)\cap W \\ -f(x)-xf'(x), & x\in(-\infty,0)\cap W \end{cases}.$$

Since $\lim_{x\to-\infty} f(x)=0$, $\lim_{x\to\infty} f(x)=0$, $\lim_{\substack{x\to-\infty\\ x\in W}} xf'(x)=0$ and $\lim_{\substack{x\to\infty\\ x\in W}} xf'(x)=0$, it follows that $\lim_{\substack{x\to-\infty\\ x\in W}} g'(x)=0$ and $\lim_{\substack{x\to\infty\\ x\in W}} g'(x)=0$. Finally, since g is also nonnegative, continuous, piecewise differentiable and integrable, we may apply Lemma A4 to it, and obtain that $\lim_{x\to-\infty} g(x)=0$ and $\lim_{x\to\infty} g(x)=0$, which is equivalent to $\lim_{x\to-\infty} xf(x)=0$ and $\lim_{x\to\infty} xf(x)=0$. □

REFERENCES

[1] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, Jul. 1948.

[2] A. V. Oppenheim, R. W. Schafer, and J.R. Buck, *Discrete-Time Signal Processing*, Prentice-Hall, Upper Saddle River, second edition, 1999.

[3] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, Prentice-Hall, Upper Saddle River, third edition, 1996.

[4] A. Gersho, "Principles of quantization," *IEEE Trans. Circuits and Systems*, vol. CAS-25, no. 7, pp. 427–436, Jul. 1978.

[5] T. Linder, "On asymptotic quantization theory," *Ph.D. Thesis, Hungarian Academy of Sciences, Budapest*, 1992.

[6] R. C. Wood, "On optimum quantization," *IEEE Trans. Info. Theory*, vol. 15, no. 2, pp. 248–252, Mar. 1969.

[7] B. Widrow, "Statistical analysis of amplitude-quantized sampled-data systems," *Trans. AIEE*, pp. 555–568, Jan. 1961.

[8] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, no. 5, pp. 442–448, Oct. 1977.

[9] H. Viswanathan and R. Zamir, "On the whiteness of high-resolution quantization errors," *IEEE Trans. Info. Theory*, vol. 47, pp. 2029–2038, Jul. 2001.

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

[11] P. Billingsley, *Probability and Measure*, Wiley & Sons, NY, third edition, 1995.

[12] L. Kronecker, "Naherungsweise ganzzahilge auflosung linearer gleichungen," *Preuss. Akad. Wiss.*, pp. 1179–1193, 1271–1299, 1884. Reprint: K. Hensel, *L. Kronecker Werke*, vol. III(1), pp. 47-109, B. G. Teubner, Leipzig, 1899.

[13] K. Peterson, *Ergodic Theory*, Cambridge university Press, New York, 1983.

[14] P. L. Tchebychef, "Sur une question arithmetique," *Denkschr. Akad. Wiss. St. Petersburg No. 4*, 1866. Reprint: A. Markoff and N. Sonin, *Oeuvres de P. L. Tchebychef*, vol. I, pp. 637-684, Chelsea Publishing Company, NY, 1962.

CHAPTER III

Asymptotic Low Resolution Scalar Quantization[1]

3.1 Introduction

This paper analyzes the low resolution performance of scalar quantization with entropy coding for stationary, memoryless Gaussian sources. Let $R(D)$ denote the operational rate-distortion function of such scalar quantization for a given source with respect to mean-squared error distortion. That is, $R(D)$ equals the least output entropy of any scalar quantizer with mean-squared error D or less. While an asymptotic formula for $R(D)$ is well known in the high resolution (i.e. small distortion) region, there is no analogous formula in the low resolution (i.e. small rate) region. Specifically, the high resolution formula is (c.f. [1])

$$R(D) \;=\; h - \frac{1}{2}\log(12\,D) + o_{D\to 0}\ , \tag{3.1}$$

where h denotes the differential entropy of the source being quantized, and $o_{D\to 0}$ denotes a quantity that goes to zero as $D \to 0$. In contrast, the principal result of this paper is that for a Gaussian source with variance σ^2, in the low resolution region

$$R(D) \;=\; \frac{\log_2 e}{2}\Big(1 - \frac{D}{\sigma^2}\Big)\left[1 + o_{D\to\sigma^2}\right]\ , \tag{3.2}$$

[1]This work was supported by NSF Grant ANI-0112801, and was submitted for publication as joint work with co-author David L. Neuhoff in the IEEE Transactions on Information Theory. Portions of this work were published in the proceedings of the IEEE International Symposium on Information Theory, Chicago, Illinois, USA, July 2004.

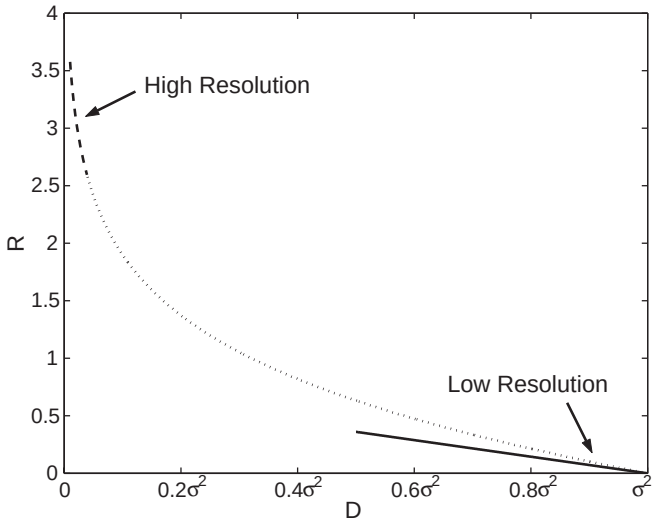

Figure 3.1: The dotted line is a qualitative representation of the operational rate-distortion curve of scalar quantization. The dashed line indicates the section of the curve that is well described by (3.1). The solid line, which shows the tangent of the curve at $D = \sigma^2$, indicates the low resolution performance given by (3.2).

where $o_{D\to\sigma^2}$ denotes a quantity that tends to zero as [2] $D \to \sigma^2$. In other words, as D increases to σ^2, the entropy approaches 0 with slope $-\frac{\log_2 e}{2\sigma^2}$. As a result, for a Gaussian source, we now have accurate approximations to the performance of optimal scalar quantization with entropy coding in both the high and low resolution regions, as illustrated in Figure 3.1.

Interestingly the Shannon rate-distortion function of a memoryless Gaussian source, $\mathcal{R}(D) = \frac{1}{2}\log_2\frac{\sigma^2}{D}$, also approaches 0 with slope $-\frac{\log_2 e}{2\sigma^2}$ as $D \to \sigma^2$. Thus, in the low resolution region, scalar quantization for such a source is asymptotically as good as any quantization technique – scalar, vector or otherwise. In contrast, in the high resolution region, $R(D)$ exceeds the Shannon rate-distortion function by $\frac{1}{2}\log_2\frac{\pi e}{6} = 0.255$ bits/sample.

We note that scalar quantization with fixed-rate coding does not attain the rate-

[2]The assumption throughout the paper is that when $D \to \sigma^2$ it does so from below.

distortion performance described in (3.2). This is because with fixed-rate coding, the smallest nonzero rate is at least 1, which implies that for any $D < \sigma^2$, the least rate of any fixed-rate scalar quantizer with mean-squared error D or less is at least 1. Consequently, the discussion throughout the paper refers to variable-rate coding, i.e. scalar quantization with entropy coding.

To derive the principal result (3.2), we focus on uniform threshold quantizers with infinitely many cells, optimal reconstruction levels, and increasingly large step sizes Δ. While it is easy to see that under ordinary conditions, $D \approx \sigma^2$ and $H \approx 0$ when Δ is large, the slope at which H approaches 0 as $D \to \sigma^2$ is not obvious. (H denotes the quantizer output entropy.) Nevertheless, we find accurate approximations to D and H from which it is straightforwardly shown that the low resolution performance of such quantizers is given by (3.2). Since this matches the Shannon rate-distortion function as $D \to \sigma^2$, no scalar quantizer could do better. Therefore, the performance of the best possible scalar quantizer is also given by (3.2). We also analyze binary quantization and show that it too has performance characterized by (3.2).

Whereas the high resolution formula (3.1) is based on the fact that the source density can be approximated as being constant on most sufficiently small cells, the low resolution formula (3.2) is based on the fact that when the cells are large, the tail of the source probability density decays sufficiently fast that only a few of the cells contribute materially to distortion and rate. This is where Gaussianity enters.

For completeness, we mention the other analytical results on low rate quantization of which we are aware. In [2, 3], the low resolution performance of fixed-rate transform codes is analyzed for Gaussian sources with memory. In this case, low rate is attained with large block lengths. In [4] and [5], upper bounds are found to the mean-squared error of dithered scalar quantization. Since these bounds apply

to all rates, they give some indication of low resolution, as well as high resolution performance.

In the remainder of the paper, Section 3.2, introduces and analyzes uniform scalar quantization. Section 3.3 does the same for binary quantizers. Section 5.5 offers concluding remarks.

3.2 Uniform Threshold Quantizers

An infinite-level uniform threshold scalar quantizer with step size Δ and offset $0 < \alpha < 1$ is a scalar quantizer with partition having cells $S_k = [(k-\alpha)\Delta, (k+1-\alpha)\Delta)$, $k \in \mathbb{Z}$, and with reconstruction levels $r_k \in S_k$, $k \in \mathbb{Z}$. Its quantization rule is $q(x) = r_k$, when $x \in S_k$. The offset α indicates the fraction of cell S_0 that lies to the left of the origin. For example, when $\alpha = 1/2$, cell S_0 is centered at the origin, whereas when $\alpha = 0$, cell S_0 begins at the origin. Let $\bar{\alpha} \stackrel{\Delta}{=} 1 - \alpha$.

We assume throughout that the source to be quantized is stationary, memoryless and Gaussian with mean zero and variance σ^2 (ordinarily we do not mention stationarity or memorylessness). The entropy of this quantizer on such a source is

$$H(\alpha, \Delta, \sigma^2) \;=\; -\sum_{k=-\infty}^{\infty} P_k \log P_k \;,$$

where P_k denotes the probability of the k^{th} cell, and all logarithms in this paper have base 2. Since the source is Gaussian

$$P_k \;=\; Q\Big((k-\alpha)\frac{\Delta}{\sigma}\Big) - Q\Big((k+1-\alpha)\frac{\Delta}{\sigma}\Big) \;,$$

where $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$. The mean-squared error of this quantizer on such a source is

$$D(\alpha, \Delta, \sigma^2) \;=\; \int_{-\infty}^{\infty} (x - q(x))^2 f(x)\, dx \;,$$

where f is the Gaussian density. Except where stated otherwise, we take the reconstruction levels to be the centroids of their respective cells; i.e., $r_k = \int_{S_k} x \frac{f(x)}{P_k}\, dx$. As is well known, this choice minimizes distortion for a given partition. The operational rate-distortion function of infinite-level uniform threshold quantization is the function

$$R_{U,\sigma^2}(D) \;=\; \inf_{0<\alpha<1,\Delta>0:D(\alpha,\Delta,\sigma^2)\le D} H(\alpha,\Delta,\sigma^2)\ ,$$

which specifies the least entropy of any such quantizer with mean-squared error D or less. Let $\mathcal{R}_{\sigma^2}(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$ denote the Shannon rate-distortion function of a Gaussian source with variance σ^2. Note that we sometimes use D to denote the distortion function $D(\alpha,\Delta,\sigma^2)$ and sometimes to denote a constant or variable representing a target distortion, as in the argument of a rate-distortion function.

It is easy to see that $H(\alpha,\Delta,\sigma^2) = H(\alpha,\Delta/\sigma,1)$, i.e. it depends only on α and $\lambda \stackrel{\Delta}{=} \Delta/\sigma$. Therefore, we will frequently use the notation $H(\alpha,\lambda)$. Similarly, $D(\alpha,\Delta,\sigma^2) = \sigma^2 D(\alpha,\Delta/\sigma,1) = \sigma^2 D(\alpha,\lambda,1)$. It follows from these remarks that $R_{U,\sigma^2}(D) = R_{U,1}\left(\frac{D}{\sigma^2}\right)$. Finally, P_k also depends only on α and λ, and for emphasis, we will sometimes denote it $P_k(\alpha,\lambda)$.

Subsections 3.2.1 and 3.2.2 find asymptotic low resolution formulas for $H(\alpha,\lambda)$ and $\sigma^2 - D(\alpha,\Delta,\sigma^2)$, respectively, and Subsection 3.2.3 uses these to find an asymptotic expression for $R_{U,\sigma^2}(D)$ as $D \to \sigma^2$. Subsection 3.2.4 relaxes the requirement that reconstruction levels be cell centroids, and provides necessary and sufficient conditions on the reconstruction levels so that the low resolution performance asymptotically remains the same as in the case of centroid levels.

We comment that in order for $H(\alpha,\lambda) \to 0$, it is necessary that $\alpha\lambda \to \infty$ and $\bar{\alpha}\lambda \to \infty$, which in turn implies that $\sigma^2 D(\alpha,\lambda,1) \to \sigma^2$. Thus we conclude that the point $(\sigma^2, 0)$ on the operational rate-distortion curve can be approached if and only

if $\alpha\lambda \to \infty$ and $\bar{\alpha}\lambda \to \infty$.

Before proceeding, we introduce notation and facts to be used later. Let $G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ denote the Gaussian density with mean zero and unit variance. Let the entropy function be defined as $\mathcal{H}(\ldots, a_{-1}, a_0, a_1, \ldots) = -\sum_{k=-\infty}^{\infty} a_k \log a_k$, where the a_k's are a finite or countably infinite set of positive numbers that need not sum to one. Let $o_{x\to x_o, y\to y_o}$ denote a quantity that approaches zero as $x \to x_o$ and $y \to y_o$, where it will be clear from context whether $x \nearrow x_o$, $x \searrow x_o$ or simply $x \to x_o$, and similarly for the variable y. Although this quantity might depend on parameters other than x and y, its convergence to zero is uniform in such parameters. To keep notation short, when $x_o = \infty$ ($y_o = \infty$), it is omitted (which will usually be the case). $o_{x\to x_o}$ is defined in an analogous fashion.

The following facts provide elementary bounds and approximations to the Q function and closely related functions.

Fact 1: $Q(x) \le \sqrt{\frac{\pi}{2}}\, G(x)$, $x \ge 0$.

Fact 2: $Q(x) < \frac{1}{x}\, G(x)$, $x > 0$.

Fact 3: $Q(x) > \frac{1}{x}(1 - \frac{1}{x^2})\, G(x)$, $x > 0$.

Fact 4: $Q(x) > \begin{cases} \frac{1}{2x}\, G(x), & x \ge \sqrt{2} \\ Q(\sqrt{2}), & x < \sqrt{2} \end{cases}$.

Fact 5: $Q(x) = \frac{1}{x}\, G(x) \left[1 + o_x\right]$, $x > 0$.

Fact 6: $Q((x+1)\lambda) = Q(x\lambda)\, o_\lambda$, $x \ge 0$; i.e. $\frac{Q((x+1)\lambda)}{Q(x\lambda)} \to 0$ as $\lambda \to \infty$, uniformly for $x \ge 0$.

Fact 7: For all sufficiently large λ, $Q\big((x+1)\lambda\big) < \frac{1}{2} Q(x\lambda)$ for all $x \ge 0$.

Fact 8: For all sufficiently large λ, $Q(x\lambda) - Q((x+1)\lambda) > \begin{cases} \frac{1}{4x\lambda}\, G(x\lambda), & x\lambda \ge \sqrt{2} \\ \frac{Q(\sqrt{2})}{2}, & 0 \le x\lambda < \sqrt{2} \end{cases}$.

Fact 9: $G^2(x) = x\,G(x)\,o_x$.

Fact 10: $C(x) \stackrel{\Delta}{=} \int_x^\infty t\,G(t)\,dt = G(x)$.

Fact 11: $V(x) \stackrel{\Delta}{=} \int_x^\infty t^2\,G(t)\,dt = x\,G(x) + Q(x) = (\text{when } x > 0)\; x\,G(x)\,[1 + o_x]$.

Fact 12: $C((x+1)\lambda) = C(x\lambda)\,o_\lambda$, $x \geq 0$; i.e. $\frac{C((x+1)\lambda)}{C(x\lambda)} \to 0$ as $\lambda \to \infty$, uniformly for $x \geq 0$.

Fact 13: $V((x+1)\lambda) = V(x\lambda)\,o_\lambda$, $x \geq 0$; i.e. $\frac{V((x+1)\lambda)}{V(x\lambda)} \to 0$ as $\lambda \to \infty$, uniformly for $x \geq 0$.

Facts 1, 2 and 3 are demonstrated in [6] (pp. 82-83). Fact 4 truncates the lower bound of Fact 3. Fact 5 follows from Facts 2 and 3. Fact 6 is derived by upper bounding $\frac{Q((x+1)\lambda)}{Q(x\lambda)}$ using Facts 1 and 4 when $x\lambda < \sqrt{2}$, and using Facts 2 and 4 when $x\lambda \geq \sqrt{2}$. Fact 7 follows from Fact 6, and Fact 8 follows from Facts 4 and 7. Fact 9 is an immediate property of exponentials. Fact 10 and the first equality of 11 derive from elementary integration. The second equality of 11 follows from Fact 5. Fact 12 follows from Fact 10 and simple manipulation of exponentials. Finally, Fact 13 is derived using Fact 11: by lower bounding $V(x\lambda)$ using Fact 4 when $x\lambda < \sqrt{2}$ and using $x\lambda < \sqrt{2}$ to upper bound $V((x+1)\lambda)$, and by upper bounding $V((x+1)\lambda)$ using Fact 2 when $x\lambda \geq \sqrt{2}$.

3.2.1 Asymptotic Entropy

We begin with several lemmas that lead to the main result of this section, a low resolution approximation for entropy.

Lemma 1. *Consider an infinite-level uniform threshold scalar quantizer with offset* $0 < \alpha < 1$ *and step size* Δ *applied to a Gaussian source with mean zero and variance* σ^2. *When* $\lambda = \frac{\Delta}{\sigma}$ *is sufficiently large,*

$$A. \quad P_{k+1}(\alpha,\lambda) \;<\; P_k(\alpha,\lambda)P_1(\alpha,\lambda) \qquad \textit{for all } \alpha \textit{ and all } k \geq 1 \;,$$

$$B. \quad P_{k-1}(\alpha,\lambda) \;<\; P_k(\alpha,\lambda)P_{-1}(\alpha,\lambda) \qquad \textit{for all } \alpha \textit{ and all } k \leq -1 \;.$$

Proof: We will show Part A; Part B follows by symmetry. To simplify notation, we omit the parameters α and λ from $P_k(\alpha,\lambda)$. Consider $k \geq 1$. First, Fact 8, with $(k-\alpha)$ playing the role of x, shows that for all sufficiently large λ, the following lower bound to P_k holds for all $k \geq 1$:

$$P_k = Q\big((k-\alpha)\lambda\big) - Q\big((k+1-\alpha)\lambda\big) > \begin{cases} \frac{1}{4}\frac{1}{(k-\alpha)\lambda}\, G((k-\alpha)\lambda), & (k-\alpha)\lambda \geq \sqrt{2} \qquad (a) \\ \frac{Q(\sqrt{2})}{2}, & (k-\alpha)\lambda < \sqrt{2} \qquad (b) \end{cases} . \tag{3.3}$$

Next, we upper bound P_{k+1} using Fact 2.

$$P_{k+1} \;=\; Q\big((k+1-\alpha)\lambda\big) - Q\big((k+2-\alpha)\lambda\big) \;<\; \frac{1}{(k+1-\alpha)\lambda}\, G((k+1-\alpha)\lambda) \;.$$

Combining the lower bound to P_k with the upper bound to P_{k+1}, we obtain

$$\frac{P_{k+1}}{P_k} \;<\; \begin{cases} 4e^{-\frac{(2(k-\alpha)+1)\lambda^2}{2}}, & (k-\alpha)\lambda \geq \sqrt{2} \qquad (a) \\ \frac{2\,G((k+1-\alpha)\lambda)}{Q(\sqrt{2})(k+1-\alpha)\lambda}, & (k-\alpha)\lambda < \sqrt{2} \qquad (b) \end{cases} . \tag{3.4}$$

It now suffices to show that for all sufficiently large λ, the above upper bound to $\frac{P_{k+1}}{P_k}$ is smaller than the lower bound to P_1 obtained from (3.3). We do so by considering two cases.

Case 1 ($(k-\alpha)\lambda < \sqrt{2}$): In this case, $(1-\alpha)\lambda < \sqrt{2}$. Thus, by (3.3b), $P_1 > \frac{Q(\sqrt{2})}{2}$. Next, by (3.4b), $\frac{P_{k+1}}{P_k} < \frac{2\,G((k+1-\alpha)\lambda)}{Q(\sqrt{2})(k+1-\alpha)\lambda} < \frac{2\,G(\lambda)}{Q(\sqrt{2})\lambda}$, where the last inequality uses

$k+1-\alpha > 1$. Since $P_1 > \frac{Q(\sqrt{2})}{2}$ and $\frac{P_{k+1}}{P_k} < \frac{2\,G(\lambda)}{Q(\sqrt{2})\lambda} \to 0$ as $\lambda \to \infty$, we conclude that for all sufficiently large λ, $\frac{P_{k+1}}{P_k} < P_1$, for all k, α such that $(k-\alpha)\lambda < \sqrt{2}$.

Case 2 ($(k-\alpha)\lambda \geq \sqrt{2}$): We consider two subcases. First, suppose $(1-\alpha)\lambda < \sqrt{2}$. Then by (3.3b), $P_1 > \frac{Q(\sqrt{2})}{2}$. Next, by (3.4a), $\frac{P_{k+1}}{P_k} < 4e^{-\frac{(2(k-\alpha)+1)\lambda^2}{2}} < 4e^{-\frac{\lambda^2}{2}}$. We conclude that for all sufficiently large λ, $\frac{P_{k+1}}{P_k} < P_1$, for all all k, α such that $(k-\alpha)\lambda \geq \sqrt{2}$ and $(1-\alpha)\lambda < \sqrt{2}$. Next, suppose $(1-\alpha)\lambda \geq \sqrt{2}$. Then by (3.3a), $P_1 > \frac{1}{4}\frac{1}{(1-\alpha)\lambda}\,G((1-\alpha)\lambda) > \frac{1}{4}\frac{1}{\lambda}\,G(\lambda)$, using $1-\alpha < 1$. By (3.4a), $\frac{P_{k+1}}{P_k} < 4e^{-\frac{(2(k-\alpha)+1)\lambda^2}{2}} < 4e^{-\frac{\lambda^2}{2}}e^{-\sqrt{2}\lambda}$, using $(k-\alpha)\lambda \geq \sqrt{2}$. Since $e^{-\sqrt{2}\lambda} \to 0$ faster than $\frac{1}{\lambda} \to 0$, we conclude that for all sufficiently large λ, $\frac{P_{k+1}}{P_k} < P_1$, for all k, α such that $(k-\alpha)\lambda \geq \sqrt{2}$ and $(1-\alpha)\lambda \geq \sqrt{2}$. This completes the proof of Part A and the lemma. □

Lemma 2.

$$\lim_{p\to 0} \frac{\mathcal{H}(1-p+p\,o_{p\to 0})}{\mathcal{H}(p)} \;=\; 0\ .$$

Proof: We need to show that $\lim_{p\to 0} \frac{-(1-p+p\,o_{p\to 0})\ln(1-p+p\,o_{p\to 0})}{-p\ln p} \;=\; 0$. It is well-known that $\lim_{x\to 0}\frac{\ln(1-x)}{-x} = 1$, or equivalently, that $\frac{\ln(1-x)}{-x} = 1 + o_{x\to 0}$. Using this we obtain

$$\begin{aligned}
&\frac{-(1-p+p\,o_{p\to 0})\ln(1-p+p\,o_{p\to 0})}{-p\ln p} \\
&\quad= \left[\frac{-(1-p+p\,o_{p\to 0})\ln(1-p+p\,o_{p\to 0})}{(1-p+p\,o_{p\to 0})(p+p\,o_{p\to 0})}\right]\cdot\left[\frac{(1-p+p\,o_{p\to 0})(p+p\,o_{p\to 0})}{-p\ln p}\right] \\
&\quad= \big[1+o_{p\to 0}\big]\cdot\left[\big(1-p+p\,o_{p\to 0}\big)\frac{1+o_{p\to 0}}{-\ln p}\right] \longrightarrow 0 \ \text{ as } p\to 0\ ,
\end{aligned}$$

which proves the lemma. □

The next lemma shows that in the low resolution case, quantizer entropy is dominated by the cells adjacent to the center cell.

Lemma 3. *Consider an infinite-level uniform threshold scalar quantizer with offset* $0 < \alpha < 1$ *and step size* Δ *applied to a Gaussian source with mean zero and variance* σ^2*. Then*

$$\mathcal{H}(\ldots, P_{-1}(\alpha,\lambda), P_0(\alpha,\lambda), P_1(\alpha,\lambda), \ldots) \;=\; \mathcal{H}\big(P_{-1}(\alpha,\lambda), P_1(\alpha,\lambda)\big)\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] ,$$

where $\lambda = \frac{\Delta}{\sigma}$.

Proof: For brevity, we omit the parameters α and λ from $P_k(\alpha,\lambda)$. The proof is composed of two main steps. In Step 1, we show that $\mathcal{H}(\ldots, P_{-1}, P_0, P_1, \ldots)$ can be asymptotically approximated by the three middle terms; that is, $\mathcal{H}(\ldots, P_{-1}, P_0, P_1, \ldots) \;=\; \mathcal{H}\big(P_{-1}, P_0, P_1\big)\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right]$. In Step 2, it is shown that these three terms can be asymptotically approximated by only two terms; that is, $\mathcal{H}\big(P_{-1}, P_0, P_1\big) = \mathcal{H}\big(P_{-1}, P_1\big)\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right]$.

Step 1: We first show that for all sufficiently large $\alpha\lambda$ and $\bar{\alpha}\lambda$,

$$1 \;<\; \frac{\mathcal{H}(\ldots, P_{-1}, P_0, P_1, \ldots)}{\mathcal{H}(P_{-1}, P_0, P_1)} \;<\; 1 + 6P_1 + 6P_{-1} \; . \tag{3.5}$$

The left inequality is trivial. We upper bound the middle term in the following way:

$$\begin{aligned} \frac{\sum_{k=-\infty}^{\infty} -P_k \log P_k}{\sum_{k=-1}^{1} -P_k \log P_k} \;&=\; 1 + \frac{\sum_{k=-\infty}^{-2} -P_k \log P_k + \sum_{k=2}^{\infty} -P_k \log P_k}{\sum_{k=-1}^{1} -P_k \log P_k(s)} \\ &<\; 1 + \frac{\sum_{k=-\infty}^{-2} -P_k \log P_k}{-P_{-1} \log P_{-1}} + \frac{\sum_{k=2}^{\infty} -P_k \log P_k}{-P_1 \log P_1} \; . \end{aligned} \tag{3.6}$$

Consider the terms in the last summation. We claim that when $\alpha\lambda$ and $\bar{\alpha}\lambda$ are sufficiently large, $-P_k \log P_k < -P_1^k \log P_1^k$ for all $k \geq 2$. This will follow from the facts that $-p \log p$ increases monotonically when $p < \frac{1}{e}$, and that $P_k < P_1^k < \frac{1}{e}$ for all $k \geq 2$, when $\alpha\lambda$ and $\bar{\alpha}\lambda$ are sufficiently large. When $\alpha\lambda$ is large, so is λ (since $\alpha < 1$), and Lemma 1 implies $P_{k+1} < P_k P_1$ for all $k \geq 1$, which in turn implies $P_k < P_1^k < P_1$, for all $k \geq 2$. Next, Fact 1 implies that $P_1 < Q(\bar{\alpha}\lambda) < \frac{1}{e}$ when $\bar{\alpha}\lambda$

is sufficiently large. Therefore, $-P_k \log P_k < -P_1^k \log P_1^k$ for all $k \geq 2$, when $\alpha\lambda$ and $\bar{\alpha}\lambda$ are sufficiently large. Substituting this into the last summation of (3.6), we have that when $\alpha\lambda$ and $\bar{\alpha}\lambda$ are sufficiently large,

$$\frac{\sum_{k=2}^{\infty} -P_k \log P_k}{-P_1 \log P_1} < \frac{\sum_{k=2}^{\infty} -P_1^k \log P_1^k}{-P_1 \log P_1} = \sum_{k=2}^{\infty} k P_1^{k-1} = \frac{2P_1}{(1-P_1)^2} - \frac{P_1^2}{(1-P_1)^2}$$
$$< \frac{2P_1}{(1-P_1)^2} < 6P_1 ,$$

where the last inequality derives from the fact that $P_1 < \frac{1}{e}$. In much the same way it follows that when $\alpha\lambda$ and $\bar{\alpha}\lambda$ are sufficiently large, $\frac{\sum_{k=-\infty}^{-2} -P_k \log P_k}{-P_{-1} \log P_{-1}} < 6P_{-1}$. This shows (3.5).

Substituting $P_{-1} = o_{\alpha\lambda}$ and $P_1 = o_{\bar{\alpha}\lambda}$ into (3.5), we obtain

$$\mathcal{H}(\ldots, P_{-1}, P_0, P_1, \ldots) = \mathcal{H}\big(P_{-1}, P_0, P_1\big)\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] ,$$

which completes Step 1.

Step 2: We will show that

$$\mathcal{H}\big(P_{-1}, P_0, P_1\big) = \mathcal{H}\big(P_{-1}, P_1\big)\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] .$$

This is done by showing $\mathcal{H}(P_0) = \mathcal{H}(P_1, P_{-1})\, o_{\alpha\lambda,\bar{\alpha}\lambda}$. In order to show this, we find an upper bound to $\mathcal{H}(P_0)$ and show that it is asymptotically, as $\alpha\lambda \to \infty$ and $\bar{\alpha}\lambda \to \infty$, negligible compared to $\mathcal{H}(P_{-1}, P_1)$.

Define $P_t = \sum_{k=-\infty}^{-2} P_k + \sum_{k=2}^{\infty} P_k$. Using the fact that for all sufficiently large λ, $P_{k+1} < P_k P_1$ for all $k \geq 1$, we upper bound the second sum as

$$\sum_{k=2}^{\infty} P_k < \sum_{k=2}^{\infty} P_1^k = \frac{P_1^2}{1-P_1} < 2P_1^2 ,$$

where the last inequality is due to $P_1 < \frac{1}{e}$ for all sufficiently large $\bar{\alpha}\lambda$. The first sum in the definition of P_t can be upper bounded in much the same way, except that it holds

for all sufficiently large $\alpha\lambda$. Thus when $\alpha\lambda$ and $\bar{\alpha}\lambda$ are both sufficiently large, $P_t < 2(P_{-1}^2+P_1^2)$. Therefore, $P_0 > 1-P_{-1}-P_1-2(P_{-1}^2+P_1^2) > 1-P_{-1}-P_1-2(P_{-1}+P_1)^2$. Since $P_{-1}+P_1 = o_{\alpha\lambda}+o_{\bar{\alpha}\lambda}$, it follows that when $\alpha\lambda$ and $\bar{\alpha}\lambda$ are sufficiently large, $P_0 > 1-P_{-1}-P_1-2(P_{-1}+P_1)^2 > \frac{1}{e}$, which since $\mathcal{H}(p)$ decreases monotonically for $p > \frac{1}{e}$, implies that $\mathcal{H}(P_0) < \mathcal{H}\big(1-P_{-1}-P_1-2(P_{-1}+P_1)^2\big)$. Consequently,

$$\begin{aligned}\frac{\mathcal{H}(P_0)}{\mathcal{H}\big(P_{-1},P_1\big)} &< \frac{\mathcal{H}\big(1-P_{-1}-P_1-2(P_{-1}+P_1)^2\big)}{\mathcal{H}(P_{-1},P_1)} \\ &< \frac{\mathcal{H}\big(1-P_{-1}-P_1-2(P_{-1}+P_1)^2\big)}{\mathcal{H}(P_{-1}+P_1)} \\ &= \frac{\mathcal{H}(1-p-2p^2)}{\mathcal{H}(p)}\,,\end{aligned}$$

where $p \stackrel{\Delta}{=} P_{-1}+P_1$, and where the second inequality follows from the fact that for any $a,b \in \mathbb{R}^+$, $\mathcal{H}(a+b) < \mathcal{H}(a,b)$. While this is a consequence of the concavity of $\mathcal{H}$, the direct proof is quite simple. Namely, it needs to be shown that $-(a+b)\log(a+b) < -a\log a - b\log b$, which after rearranging terms, is equivalent to showing that $a\log\frac{a+b}{a} + b\log\frac{a+b}{b} > 0$. This clearly holds since the arguments of the logs are greater than one and hence the logs are positive as are a and b. We observe that as $\alpha\lambda$ and $\bar{\alpha}\lambda$ tend to infinity, p goes to zero. Therefore, by Lemma 2 it follows that $\frac{\mathcal{H}(1-p-2p^2)}{\mathcal{H}(p)} \to 0$ as $p \to 0$. This shows that $\mathcal{H}(P_0) = \mathcal{H}(P_1,P_{-1})\,o_{\alpha\lambda,\bar{\alpha}\lambda}$, which completes the proof of Step 2 and the lemma. □

Lemma 4. *Let $a(s)$ and $b(s)$ be positive functions on $\mathbb{R}$ such that $\lim_{s\to s_o}\frac{a(s)}{b(s)} = c$, $c \in \{0,1\}$, and for some $\varepsilon > 0$, $|b(s)-1| > \varepsilon$ for all s. Then*

$$\mathcal{H}(a(s)) \;=\; \mathcal{H}(b(s))\left[c + o_{s\to s_o}\right]\,.$$

Proof: It is equivalent to show $\lim_{s\to s_o} \frac{\mathcal{H}(a(s))}{\mathcal{H}(b(s))} = c$. We have the following string of equalities.

$$\begin{aligned}\frac{\mathcal{H}(a(s))}{\mathcal{H}(b(s))} &= \frac{-a(s)\log a(s)}{-b(s)\log b(s)} = \frac{a(s)}{b(s)}\frac{\log\left[\frac{a(s)}{b(s)}b(s)\right]}{\log b(s)} = \frac{a(s)}{b(s)}\left[1+\frac{\log\frac{a(s)}{b(s)}}{\log b(s)}\right] \\ &= \frac{a(s)}{b(s)} + \frac{\frac{a(s)}{b(s)}\log\frac{a(s)}{b(s)}}{\log b(s)} .\end{aligned}$$

Since $|b(s)-1| > \varepsilon$ for all s, it follows that either $\log b(s) > \log(1+\varepsilon)$ or $\log b(s) < \log(1-\varepsilon)$ for all s. Therefore, $\log b(s)$ is bounded away from zero. Combining this with the fact that $\frac{a(s)}{b(s)} \to c$ as $s \to s_o$, and that $\frac{a(s)}{b(s)}\log\frac{a(s)}{b(s)} \to c\log c$ as $s \to s_o$, the result follows. □

Lemma 5.

$$\mathcal{H}\big(Q(x)\big) = \frac{\log e}{2} x\, G(x)\left[1+o_x\right] .$$

Proof: We begin with a string of equalities, the first of which is due to Fact 5.

$$\begin{aligned}\mathcal{H}\big(Q(x)\big) &= \mathcal{H}\Big(\frac{1}{x}G(x)\left[1+o_x\right]\Big) = -\frac{1}{x}G(x)\left[1+o_x\right]\log\Big(\frac{1}{x}G(x)\left[1+o_x\right]\Big) \\ &= \frac{1}{x}G(x)\left[1+o_x\right]\left[\log\sqrt{2\pi}x + \frac{x^2}{2}\log e - \log\left[1+o_x\right]\right] . \qquad (3.7)\end{aligned}$$

Next, observing that $\frac{\log\sqrt{2\pi}x+\frac{x^2}{2}\log e-\log\left[1+o_x\right]}{\frac{x^2}{2}\log e} = 1+o_x$, it follows that $\mathcal{H}\big(Q(x)\big) = \frac{\log e}{2}x\,G(x)\left[1+o_x\right]$, which completes the proof of the lemma. □

The following theorem gives the low resolution approximation to the entropy of uniform quantization.

Theorem 6. *For an infinite-level uniform threshold scalar quantizer with offset* $0 < \alpha < 1$ *and step size* Δ *applied to a Gaussian source with mean zero and variance* σ^2*,*

$$H(\alpha,\lambda) = \frac{\log e}{2}\Big(\alpha\lambda\, G(\alpha\lambda) + \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\Big)\left[1+o_{\alpha\lambda,\bar{\alpha}\lambda}\right] , \qquad (3.8)$$

where $\lambda = \frac{\Delta}{\sigma}$ *and* $H(\alpha,\lambda) = \mathcal{H}(\ldots, P_{-1}(\alpha,\lambda), P_0(\alpha,\lambda), P_1(\alpha,\lambda), \ldots)$ *is the quantizer entropy.*

If one fixes α, this theorem shows the rate at which entropy converges to 0 as $\lambda \to \infty$. However, the convergence is not uniform in α, and this theorem shows how entropy depends on α as well as λ. In particular, it gives an accurate approximation to quantizer entropy when both $\alpha\lambda$ and $\bar{\alpha}\lambda$ are large.

Proof: For brevity, we omit the parameters α and λ from $P_k(\alpha,\lambda)$. Lemma 3 shows that

$$H(\alpha,\lambda) \;=\; \mathcal{H}(\ldots, P_{-1}, P_0, P_1, \ldots) \;=\; \mathcal{H}\big(P_{-1}, P_1\big)\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] . \tag{3.9}$$

Since $P_{-1} = Q(\alpha\lambda) - Q\big((1+\alpha)\lambda\big)$, Fact 6 implies that $\lim_{\lambda\to\infty} \frac{P_{-1}}{Q(\alpha\lambda)} = 1$. Since $|Q(\alpha\lambda) - 1| > \frac{1}{2}$ for all λ (and all α), it follows from Lemma 4 that $\mathcal{H}(P_{-1}) = \mathcal{H}(Q(\alpha\lambda))\left[1 + o_\lambda\right]$. Next, applying Lemma 5, we obtain $\mathcal{H}(Q(\alpha\lambda)) = \left(\frac{1}{2}\log e\right)\alpha\lambda\, G(\alpha\lambda)\left[1 + o_{\alpha\lambda}\right]$, where $\alpha\lambda$ plays the role of x in the lemma. Combining these yields $\mathcal{H}(P_{-1}) = \left(\frac{1}{2}\log e\right)\alpha\lambda\, G(\alpha\lambda)\left[1 + o_{\alpha\lambda}\right]$.

In a similar way, $\lim_{\lambda\to\infty} \frac{P_1}{Q(\bar{\alpha}\lambda)} = 1$, and since $|Q(\bar{\alpha}\lambda) - 1| > \frac{1}{2}$ for all λ (and all $\bar{\alpha}$) , it follows via Lemma 4 that $\mathcal{H}(P_1) = \mathcal{H}(Q(\bar{\alpha}\lambda))\left[1 + o_\lambda\right]$, and further application of Lemma 5 shows that $\mathcal{H}(P_1) = \left(\frac{1}{2}\log e\right)\bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\left[1 + o_{\bar{\alpha}\lambda}\right]$.

Combining the expressions for $\mathcal{H}(P_{-1})$ and $\mathcal{H}(P_1)$ together with (3.9) complete the proof of the theorem. □

We now comment on the cell or cells that dominate entropy when it is small. The entropy $H(\alpha,\lambda)$ will be small if and only if $P_0 \approx 1$ and $P_k \approx 0$, $k \neq 0$, which makes $-P_k \log P_k \approx 0$ for all k, and which happens if and only if $\alpha\lambda$ and $\bar{\alpha}\lambda$ are both large. Lemma 3 shows that $H(\alpha,\lambda)$ is dominated by the cells, S_{-1} and S_1, immediately adjacent to the center cell. This is not coincidental; rather it follows

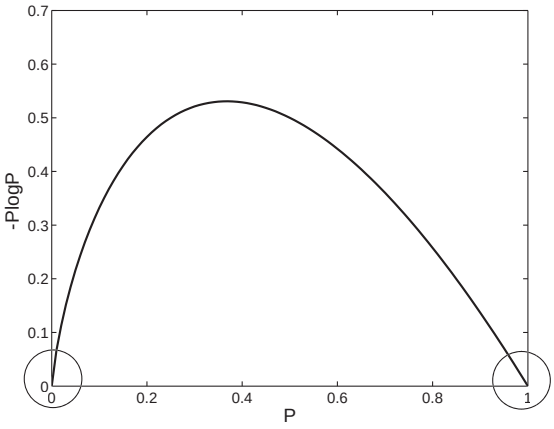

Figure 3.2: The entropy function, $-p \log p$.

from the fact, illustrated in Figure 3.2, that the entropy function, $\mathcal{H}(p) = -p \log p$, has infinite slope at $p = 0$ and finite slope at $p = 1$. Thus, when entropy is nearly zero, it is dominated by the largest of the nearly zero probabilities, which are P_{-1} and/or P_1. Indeed the two terms within the large parentheses in (3.8) correspond to $\mathcal{H}(P_{-1})$ and $\mathcal{H}(P_1)$, respectively. If $\alpha\lambda << \bar{\alpha}\lambda$, e.g. if $\alpha < \frac{1}{2}$ and λ is very large, then $P_{-1} >> P_1$, and it is cell S_{-1} and the first term within the parentheses that dominate the entropy. Conversely, if $\bar{\alpha}\lambda << \alpha\lambda$, then $P_1 >> P_{-1}$, and it is cell S_1 and the second term within the parentheses that dominate. Finally, if $\alpha\lambda \approx \bar{\alpha}\lambda$, then the two dominating cells contribute roughly the same to the entropy.

3.2.2 Asymptotic Distortion

The following theorem provides the low resolution approximation to distortion.

Theorem 7. *For an infinite-level uniform threshold scalar quantizer with offset* $0 < \alpha < 1$*, step size* Δ*, centroid reconstruction levels, and a Gaussian source with zero mean and variance* σ^2*,*

$$\frac{\sigma^2 - D(\alpha, \Delta, \sigma^2)}{\sigma^2} \;=\; \Big(\alpha\lambda\, G(\alpha\lambda) + \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\Big)\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] .$$

where $\lambda = \frac{\Delta}{\sigma}$.

Like Theorem 6, this theorem gives an accurate approximation when both $\alpha\lambda$ and $\bar{\alpha}\lambda$ are large. The proof will be structured in a way that makes evident which cell or cells dominate $\sigma^2 - D$.

Proof: For brevity we omit the arguments of $D(\alpha, \Delta, \sigma^2)$. Let $\sigma_k^2 \stackrel{\Delta}{=} \int_{S_k} x^2 f(x)\, dx = \sigma^2\big(V((k-\alpha)\lambda) - V((k+1-\alpha)\lambda)\big)$, where f is the pdf of the Gaussian source and $V(x)$ is defined in Fact 11. Let $D_k \stackrel{\Delta}{=} \int_{S_k} (x - r_k)^2 f(x)\, dx = \sigma_k^2 - r_k^2 P_k$ be the contribution of the k^{th} cell to the distortion (recalling that $r_k = \int_{S_k} x \frac{f(x)}{P_k}\, dx$). We observe that σ_k^2 and D_k both depend only on α and $\lambda = \frac{\Delta}{\sigma}$. Moreover, $\sigma^2 = \sum_k \sigma_k^2$ and $D = \sum_k D_k$. We now write

$$\sigma^2 - D = \big(\sigma^2 - D_0\big) - D_{-1} - D_1 - \sum_{k \neq -1,0,1} D_k \ . \tag{3.10}$$

We deal with these terms in reverse order. First, since $D_k \leq \sigma_k^2$,

$$\begin{aligned}
\sum_{|k|\geq 2} D_k \ &\leq \ \sum_{|k|\geq 2} \sigma_k^2 \ = \ \int_{-\infty}^{-(\alpha+1)\Delta} x^2 f(x)\, dx + \int_{(2-\alpha)\Delta}^{\infty} x^2 f(x)\, dx \\
&\stackrel{(a)}{=} \ \sigma^2 V((\alpha+1)\lambda) + \sigma^2 V((2-\alpha)\lambda) \\
&\stackrel{(b)}{=} \ \sigma^2 V(\alpha\lambda)\, o_\lambda + \sigma^2 V((1-\alpha)\lambda)\, o_\lambda \\
&\stackrel{(c)}{=} \ \sigma^2 \alpha\lambda\, G(\alpha\lambda)\, o_{\alpha\lambda} + \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\, o_{\bar{\alpha}\lambda} \ ,
\end{aligned} \tag{3.11}$$

where (a) and (c) follow from Fact 11, and (b) is obtained using Fact 13. Next,

$$\begin{aligned}
D_1 \ &= \ \sigma_1^2 - r_1^2 P_1 \\
&= \ \sigma^2 (V((1-\alpha)\lambda) - V((2-\alpha)\lambda)) \\
&\quad - \left(\frac{\sigma C((1-\alpha)\lambda) - \sigma C((2-\alpha)\lambda)}{Q((1-\alpha)\lambda) - Q((2-\alpha)\lambda)} \right)^2 \Big(Q((1-\alpha)\lambda) - Q((2-\alpha)\lambda) \Big) \\
&\stackrel{(a)}{=} \ \sigma^2 V((1-\alpha)\lambda)\big[1 + o_\lambda\big] - \frac{\sigma^2 \Big(C((1-\alpha)\lambda)\big[1 + o_\lambda\big] \Big)^2}{Q\big((1-\alpha)\lambda\big)\big[1 + o_\lambda\big]} \\
&\stackrel{(b)}{=} \ \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\,[1 + o_{\bar{\alpha}\lambda}] - \frac{\sigma^2 G^2(\bar{\alpha}\lambda)\big[1 + o_\lambda\big]}{\frac{1}{\bar{\alpha}\lambda} G(\bar{\alpha}\lambda)\big[1 + o_{\bar{\alpha}\lambda}\big]} \ = \ \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\, o_{\bar{\alpha}\lambda} \ ,
\end{aligned} \tag{3.12}$$

where (a) follows from Facts 6, 12 and 13, and (b) follows from Facts 5,10 and 11. In a similar manner, it can be shown that

$$D_{-1} \;=\; \sigma^2 \alpha\lambda\, G(\alpha\lambda)\, o_{\alpha\lambda} \;. \tag{3.13}$$

Finally,

$$\sigma^2 - D_0 \;=\; (\sigma^2 - \sigma_0^2) + (\sigma_0^2 - D_0) \;, \tag{3.14}$$

where as in (3.11) above

$$\begin{aligned} \sigma^2 - \sigma_0^2 \;&=\; \sum_{k\neq 0} \sigma_k^2 \;=\; \sigma^2 V(\alpha\lambda) + \sigma^2 V\big((1-\alpha)\lambda\big) \\ &=\; \sigma^2 \alpha\lambda\, G(\alpha\lambda)\left[1 + o_{\alpha\lambda}\right] + \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\left[1 + o_{\bar{\alpha}\lambda}\right] \;, \end{aligned} \tag{3.15}$$

where the second equality uses Fact 11, and where as in (3.12)

$$\begin{aligned} \sigma_0^2 - D_0 \;=\; r_0^2 P_0 \;&=\; \left(\frac{\sigma C(\alpha\lambda) - \sigma C((1-\alpha)\lambda)}{1 - Q(\alpha\lambda) - Q((1-\alpha)\lambda)}\right)^2 \Big(1 - Q(\alpha\lambda) - Q(\bar{\alpha}\lambda)\Big) \\ &\overset{(a)}{=}\; \frac{\sigma^2\big(G(\alpha\lambda) - G(\bar{\alpha}\lambda)\big)^2}{1 + o_{\alpha\lambda} + o_{\bar{\alpha}\lambda}} \\ &\overset{(b)}{=}\; \frac{\sigma^2 \alpha\lambda\, G(\alpha\lambda)\, o_{\alpha\lambda} + \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\, o_{\bar{\alpha}\lambda} - 2\sigma^2\, G(\alpha\lambda)\, G(\bar{\alpha}\lambda)}{1 + o_{\alpha\lambda} + o_{\bar{\alpha}\lambda}} \\ &\overset{(c)}{=}\; \frac{\sigma^2 \alpha\lambda\, G(\alpha\lambda)\, o_{\alpha\lambda} + \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\, o_{\bar{\alpha}\lambda} - \sigma^2 \alpha\lambda\, G(\alpha\lambda)\, o_{\alpha\lambda,\bar{\alpha}\lambda} - \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\, o_{\alpha\lambda,\bar{\alpha}\lambda}}{1 + o_{\alpha\lambda} + o_{\bar{\alpha}\lambda}} \\ &=\; \Big(\sigma^2 \alpha\lambda\, G(\alpha\lambda) + \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\Big)\, o_{\alpha\lambda,\bar{\alpha}\lambda} \;, \end{aligned} \tag{3.16}$$

where (a) is due to Fact 10, (b) follows from Fact 9, and (c) is obtained by observing that $G(\alpha\lambda)\, G(\bar{\alpha}\lambda) = \alpha\lambda\, G(\alpha\lambda)\, \frac{o_{\bar{\alpha}\lambda}}{\alpha\lambda} = \alpha\lambda\, G(\alpha\lambda)\, o_{\alpha\lambda,\bar{\alpha}\lambda}$, and similarly $G(\alpha\lambda)\, G(\bar{\alpha}\lambda) = \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\, o_{\alpha\lambda,\bar{\alpha}\lambda}$. Substituting (3.15) and (3.16) into (3.14) yields

$$\sigma^2 - D_0 \;=\; \Big(\sigma^2 \alpha\lambda\, G(\alpha\lambda) + \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\Big)\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] \;. \tag{3.17}$$

Substituting (3.11), (3.12), (3.13) and (3.17) into (3.10) yields

$$\sigma^2 - D \;=\; \Big(\sigma^2 \alpha\lambda\, G(\alpha\lambda) + \sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\Big)\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] \;.$$

Dividing the above by σ^2 gives the desired result. $\square$

We now consider which cell or cells make the dominating contribution to $\sigma^2 - D$, when the latter is very small. When $D \approx \sigma^2$, both $\alpha\lambda$ and $\bar{\alpha}\lambda$ are large. From (3.17), we see that $D_0 \approx \sigma^2$, and from (3.11), (3.12), and (3.13), we see that $D_k \approx 0$ for $k \neq 0$. We are interested, however, in finding the cells that dominate the rate at which distortion converges to variance. Since $D_0 \to \sigma^2$ and $D_k \to 0$, $k \neq 0$, it makes most sense to compare $\sigma^2 - D_0$ and the D_k's, $k \neq 0$. Comparing (3.17) to (3.11), (3.12), and (3.13), reveals that $\sum_{k\neq 0} D_k$ is asymptotically negligible relative to $\sigma^2 - D_0$. We conclude that when $D \approx \sigma^2$, $\sigma^2 - D_0$ is the dominant component of $\sigma^2 - D$.

3.2.3 Asymptotic Rate-Distortion

Directly applying Theorems 6 and 7 yields the asymptotic low resolution rate-distortion behavior of uniform threshold quantizers:

Lemma 8. *For an infinite-level uniform threshold scalar quantizer with offset $0 < \alpha < 1$, step size Δ, centroid reconstruction levels, and a Gaussian source with zero mean and variance σ^2,*

$$\frac{H(\alpha,\lambda)}{\sigma^2 - D(\alpha,\Delta,\sigma^2)} \;=\; \frac{\log e}{2\sigma^2}\left[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] ,$$

where $\lambda = \frac{\Delta}{\sigma}$.

The following is the principal result of this paper.

Theorem 9. *In the low resolution region, the operational rate-distortion function of infinite-level uniform threshold scalar quantization for a Gaussian source with variance σ^2 is*

$$R_{U,\sigma^2}(D) \;=\; \frac{\log e}{2}\Big(1 - \frac{D}{\sigma^2}\Big)\left[1 + o_{_{D\to\sigma^2}}\right] . \tag{3.18}$$

Proof: We begin by rewriting the operational rate-distortion function as

$$R_{U,\sigma^2}(D) \;=\; \inf_{0<\alpha<1} R_{U,\sigma^2,\alpha}(D) \;, \tag{3.19}$$

where $R_{U,\sigma^2,\alpha}(D) \stackrel{\Delta}{=} \inf_{\Delta>0:D(\alpha,\Delta,\sigma^2)\leq D} H(\alpha, \frac{\Delta}{\sigma})$. As a preliminary to showing (3.18), we will show $R_{U,\sigma^2,\alpha}(D)$ satisfies (3.18).

Since $R_{U,\sigma^2}(D) = R_{U,1}\big(\frac{D}{\sigma^2}\big)$, it suffices to show

$$\frac{R_{U,1}(D)}{1-D} \longrightarrow \frac{\log e}{2} \;\text{ as } D \longrightarrow 1 \;, \tag{3.20}$$

and the corresponding result for $R_{U,1,\alpha}(D)$. First,

$$\limsup_{D\to 1} \frac{R_{U,1,\alpha}(D)}{1-D} \stackrel{(a)}{=} \limsup_{\lambda\to\infty} \frac{R_{U,1,\alpha}\big(D(\alpha,\lambda,1)\big)}{1-D(\alpha,\lambda,1)} \stackrel{(b)}{\leq} \limsup_{\lambda\to\infty} \frac{H(\alpha,\lambda)}{1-D(\alpha,\lambda,1)} \stackrel{(c)}{=} \frac{\log e}{2} \;, \tag{3.21}$$

where (a) derives from the fact that $D(\alpha,\lambda,1)$ goes continuously to 1 as $\lambda\to\infty$, (b) follows from the definition of $R_{U,1,\alpha}\big(D(\alpha,\lambda,1)\big)$, and (c) is obtained from Lemma 8. Next,

$$\liminf_{D\to 1} \frac{R_{U,1,\alpha}(D)}{1-D} \stackrel{(a)}{\geq} \liminf_{\lambda\to\infty} \frac{H(\alpha,\lambda)}{1-D(\alpha,\lambda,1)} \stackrel{(b)}{=} \frac{\log e}{2} \;, \tag{3.22}$$

where (a) follows from Lemma A1 of the Appendix, and (b) is obtained from Lemma 8. It now follows from (3.21) and (3.22) that

$$\lim_{D\to 1} \frac{R_{U,1,\alpha}(D)}{1-D} \;=\; \frac{\log e}{2} \;. \tag{3.23}$$

Finally, to obtain the result of the theorem we proceed as follows

$$\limsup_{D\to 1} \frac{R_{U,1}(D)}{1-D} \stackrel{(a)}{=} \limsup_{D\to 1} \frac{\inf_\alpha R_{U,1,\alpha}(D)}{1-D} \stackrel{(b)}{\leq} \inf_\alpha \limsup_{D\to 1} \frac{R_{U,1,\alpha}(D)}{1-D} \stackrel{(c)}{=} \frac{\log e}{2} \;, \tag{3.24}$$

where (a) follows from the definition of $R_{U,1}(D)$, (b) is elementary, and (c) follows from (3.23), and

$$\liminf_{D\to 1} \frac{R_{U,1}(D)}{1-D} \stackrel{(a)}{\geq} \liminf_{D\to 1} \frac{\mathcal{R}_1(D)}{1-D} \stackrel{(b)}{=} \liminf_{D\to 1} \frac{\frac{1}{2}\log\frac{1}{D}}{1-D} \stackrel{(c)}{=} \frac{\log e}{2} \;, \tag{3.25}$$

where $\mathcal{R}_1(D)$ is the Shannon rate-distortion function of a unit variance Gaussian source, and where (a) follows from the converse rate-distortion theorem, (b) uses the well-known formula for $\mathcal{R}_1(D)$ [7] (p. 477), and (c) follows from taking the limit, for example, using L'Hospital's rule. Equation (3.20) and the theorem now follow from (3.24) and (3.25). We note that (3.22) could have been shown using Shannon's rate-distortion function as a lower bound, as was done in (3.25), instead of using Lemma A1. However, the approach taken above, demonstrates that $\lim_{D\to 1} \frac{R_{U,1,\alpha}(D)}{1-D} = \lim_{\lambda\to\infty} \frac{H(\alpha,\lambda)}{1-D(\alpha,\lambda,1)}$ without using either Gaussianity or Shannon's rate-distortion function. It requires only that the latter limit exist. □

As is easy to see, $\mathcal{R}_{\sigma^2}(D) = \frac{1}{2}\log\frac{\sigma^2}{D} = \frac{\log e}{2}\left[1-\frac{D}{\sigma^2}\right]\left[1+o_{D\to\sigma^2}\right]$. Comparing this to the theorem statement reveals that for a Gaussian source and the low resolution region, the operational rate-distortion function of infinite-level uniform threshold scalar quantization and the Shannon rate-distortion function approach 0 with the same slope as $D \to \sigma^2$. Therefore in the low resolution region, such quantizers are asymptotically as good as any quantization technique — scalar, block, or otherwise. Additionally, from (3.23) and the relation between $R_{U,\sigma^2,\alpha}(D)$ and $R_{U,1,\alpha}(D)$, one concludes that for any α, the operational rate-distortion function $R_{U,\sigma^2,\alpha}(D)$ of uniform threshold quantization with offset α also approaches the Shannon rate-distortion function, as does the operational rate-distortion function of scalar quantization of any type. Finally, we note that from the dominance results presented previously, the slope is approximately equal to $\frac{H_{-1}+H_1}{\sigma^2-D_0}$, i.e. the distortion term is dominated by the center cell and the entropy is dominated by the two adjacent cells.

3.2.4 Asymptotically Optimal Reconstruction Levels

Theorem 7 assumed the reconstruction levels were centroids, which necessarily yields the smallest possible distortion. Since distortion is a continuous function of the levels, it is natural to expect that this assumption can be relaxed somewhat, asymptotically in the low resolution region. That is, for the conclusion of the theorem to hold, it should only be necessary for the reconstruction levels to be sufficiently close to the centroids. As it turns out, there is very little sensitivity to the reconstruction levels for $k \neq -1, 0, 1$, in the sense that the requirement that $r_k \in S_k$ insures they contribute negligibly to $\sigma^2 - D$ when $\alpha\lambda$ and $\bar{\alpha}\lambda$ are large. Moreover, for $k = -1, 0, 1$, the centroids approximately equal $-\alpha\Delta$, 0 and $\bar{\alpha}\Delta$, respectively, when $\alpha\lambda$ and $\bar{\alpha}\lambda$ are large. Consequently, the next theorem shows that the result of Theorem 7 continues to hold if and only if $r_{-1} \approx -\alpha\Delta$, $r_0 \approx 0$ and $r_1 \approx \bar{\alpha}\Delta$.

Theorem 10. *For an infinite-level uniform threshold scalar quantizer with offset* $0 < \alpha < 1$*, step size* Δ*, reconstruction levels* $\{\tilde{r}_{\alpha,\Delta,k}\}$ *that are not necessarily centroids*[3]*, and a Gaussian source with zero mean and variance* σ^2*, the distortion* $\widetilde{D}(\alpha, \Delta, \sigma^2, \{\tilde{r}_{\alpha,\Delta,k}\})$ *satisfies*

$$\begin{aligned}
&\frac{\widetilde{D}(\alpha, \Delta, \sigma^2, \{\tilde{r}_{\alpha,\Delta,k}\}) - D(\alpha, \Delta, \sigma^2)}{\sigma^2} \\
&\qquad = \ \alpha\lambda\, G(\alpha\lambda)\left[\Big(\frac{\tilde{r}_{\alpha,\Delta,-1} + \alpha\Delta}{\alpha\Delta}\Big)^2\big[1 + o_{\alpha\lambda}\big] + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] \\
&\qquad\quad + \ \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\left[\Big(\frac{\tilde{r}_{\alpha,\Delta,1} - \bar{\alpha}\Delta}{\bar{\alpha}\Delta}\Big)^2\big[1 + o_{\bar{\alpha}\lambda}\big] + o_{\alpha\lambda,\bar{\alpha}\lambda}\right] \\
&\qquad\quad + \ 2\Big[G(\bar{\alpha}\lambda) - G(\alpha\lambda)\Big]\Big(\frac{\tilde{r}_{\alpha,\Delta,0}}{\sigma}\Big)\big[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\big] \\
&\qquad\quad + \ \Big(\frac{\tilde{r}_{\alpha,\Delta,0}}{\sigma}\Big)^2\big[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\big]\ , \qquad (3.26)
\end{aligned}$$

where $\lambda = \frac{\Delta}{\sigma}$ *and* $D(\alpha, \Delta, \sigma^2)$ *is the distortion induced by cell centroids.*

[3]Recall that reconstruction levels are required to lie within their respective cells.

Proof: Given an infinite-level uniform threshold quantizer with offset $0 < \alpha < 1$, step size $\Delta > 0$, and reconstruction levels $\{\tilde{r}_{\alpha,\Delta,k}\}$, denoted $\{\tilde{r}_k\}$ for short, let $\widetilde{D}_k$ represent the contribution to distortion due to the k^{th} cell. Let D_k represent the contribution to distortion due to the same cell when the cell centroid, denoted r_k, is the reconstruction level. For brevity, let $\widetilde{D}$ denote $\widetilde{D}(\alpha, \Delta, \sigma^2, \{\tilde{r}_{\alpha,\Delta,k}\})$. Also, let D represent $D(\alpha, \Delta, \sigma^2)$, the distortion induced by cell centroids. We observe that

$$\frac{\widetilde{D} - D}{\sigma^2} \;=\; \frac{\sigma^2 - D}{\sigma^2} - \frac{\sigma^2 - \widetilde{D}}{\sigma^2} \;. \tag{3.27}$$

The first term on the right hand side is known from Theorem 7. We focus on finding the second term.

Since $\widetilde{D}_k = D_k + (\tilde{r}_k - r_k)^2 P_k$ and $D_k = \sigma_k^2 - r_k^2 P_k$, we have

$$\begin{aligned}\sigma^2 - \widetilde{D} = \;& \sum_k \sigma_k^2 - \sum_k \widetilde{D}_k \;=\; \sum_k \tilde{r}_k(2r_k - \tilde{r}_k)P_k \\ =\; & \tilde{r}_0(2r_0 - \tilde{r}_0)P_0 + \tilde{r}_{-1}(2r_{-1} - \tilde{r}_{-1})P_{-1} + \tilde{r}_1(2r_1 - \tilde{r}_1)P_1 + \sum_{|k|\geq 2} \tilde{r}_k(2r_k - \tilde{r}_k)P_k \;.\end{aligned} \tag{3.28}$$

We consider these terms in reverse order. First, it easy to see that for $|k| \geq 2$, $|\tilde{r}_k| \leq 2|r_k|$ and $|2r_k - \tilde{r}_k| \leq 2|r_k|$. Therefore,

$$\begin{aligned}\left|\sum_{|k|\geq 2} \tilde{r}_k(2r_k - \tilde{r}_k)P_k\right| \;&\leq\; \sum_{|k|\geq 2} 4\, r_k^2 P_k \;\overset{(a)}{\leq}\; 4 \sum_{|k|\geq 2} \sigma_k^2 \\ &\overset{(b)}{=}\; 4\,\sigma^2 \alpha\lambda\, G(\alpha\lambda)\, o_{\alpha\lambda} + 4\,\sigma^2 \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\, o_{\bar{\alpha}\lambda} \;,\end{aligned} \tag{3.29}$$

where (a) follows from the fact that $\sigma_k^2 = D_k + r_k^2 P_k$ and (b) follows from (3.11).

Next, Facts 5 and 6 imply that $P_1 = Q((1-\alpha)\lambda) - Q((2-\alpha)\lambda) = \frac{1}{\bar{\alpha}\lambda} G(\bar{\alpha}\lambda)\left[1 + o_{\bar{\alpha}\lambda}\right]$. We also obtain an expression for r_1 as follows.

$$r_1 \;=\; \frac{1}{P_1} \int_{(1-\alpha)\lambda}^{(2-\alpha)\lambda} x f(x)\, dx \overset{(a)}{=} \; -\frac{\sigma\left[C((1-\alpha)\lambda) - C((2-\alpha)\lambda)\right]}{\frac{1}{\bar{\alpha}\lambda}\, G(\bar{\alpha}\lambda)\left[1 + o_{\bar{\alpha}\lambda}\right]} \overset{(b)}{=} \; \bar{\alpha}\Delta\left[1 + o_{\bar{\alpha}\lambda}\right] \;,$$

where (a) is due to Fact 10 and (b) follows from Facts 10 and 12. Writing $\tilde{r}_1 = \bar{\alpha}\Delta + (\tilde{r}_1 - \bar{\alpha}\Delta)$, we obtain, after some algebra,

$$\tilde{r}_1(2r_1 - \tilde{r}_1)P_1 \;=\; \sigma^2\bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\Big[1 - \Big(\frac{\tilde{r}_1 - \bar{\alpha}\Delta}{\bar{\alpha}\Delta}\Big)^2\big[1 + o_{\bar{\alpha}\lambda}\big] + o_{\bar{\alpha}\lambda}\Big]\,. \tag{3.30}$$

In a similar way we can write $\tilde{r}_{-1} = -\alpha\Delta + (\tilde{r}_{-1} + \alpha\Delta)$ and obtain

$$\tilde{r}_{-1}(2r_{-1} - \tilde{r}_{-1})P_{-1} \;=\; \sigma^2\alpha\lambda\, G(\alpha\lambda)\Big[1 - \Big(\frac{\tilde{r}_{-1} + \alpha\Delta}{\alpha\Delta}\Big)^2\big[1 + o_{\alpha\lambda}\big] + o_{\alpha\lambda}\Big]\,. \tag{3.31}$$

Finally, from Fact 1 $P_0 = 1 - Q(\alpha\lambda) - Q(\bar{\alpha}\lambda) = 1 + o_{\alpha\lambda} + o_{\bar{\alpha}\lambda}$, and from Fact 10 $r_0 = \frac{1}{P_0}\int_{-\alpha\Delta}^{(1-\alpha)\Delta} x f(x)\,dx = \sigma\big(G(\alpha\lambda) - G(\bar{\alpha}\lambda)\big)\big[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\big]$. Therefore,

$$\tilde{r}_0(2r_0 - \tilde{r}_0)P_0 \;=\; 2\sigma^2\Big[G(\alpha\lambda) - G(\bar{\alpha}\lambda)\Big]\Big(\frac{\tilde{r}_0}{\sigma}\Big)\big[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\big] - \sigma^2\Big(\frac{\tilde{r}_0}{\sigma}\Big)^2\big[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\big]\,. \tag{3.32}$$

Combining (3.28) – (3.32) gives

$$\begin{aligned}\frac{\sigma^2 - \widetilde{D}}{\sigma^2} \;=\;& \alpha\lambda\, G(\alpha\lambda)\Big[1 - \big(\tfrac{\tilde{r}_{\alpha,\Delta,-1} + \alpha\Delta}{\alpha\Delta}\big)^2\big[1 + o_{\alpha\lambda}\big] + o_{\alpha\lambda}\Big] \\ &+\; \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\Big[1 - \big(\tfrac{\tilde{r}_{\alpha,\Delta,1} - \bar{\alpha}\Delta}{\alpha\Delta}\big)^2\big[1 + o_{\bar{\alpha}\lambda}\big] + o_{\bar{\alpha}\lambda}\Big] \\ &+\; 2\Big[G(\alpha\lambda) - G(\bar{\alpha}\lambda)\Big]\Big(\frac{\tilde{r}_0}{\sigma}\Big)\big[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\big] \;-\; \Big(\frac{\tilde{r}_0}{\sigma}\Big)^2\big[1 + o_{\alpha\lambda,\bar{\alpha}\lambda}\big]\,. \end{aligned} \tag{3.33}$$

The theorem now easily follows from (3.33), (3.27) and Theorem 7. □

We observe that since $\frac{\sigma^2 - \widetilde{D}}{\sigma^2} = \frac{\sigma^2 - D}{\sigma^2} - \frac{\widetilde{D} - D}{\sigma^2}$, the distortion induced by general reconstruction levels is asymptotically the same as the distortion induced by cell centroids if and only if $\frac{\widetilde{D} - D}{\sigma^2}$ is asymptotically negligible relative to $\frac{\sigma^2 - D}{\sigma^2}$. Using Theorem 7, the next theorem shows that this happens if and only if the first two square bracketed terms in (3.26) go to 0 as $\alpha\lambda \to \infty$ and $\bar{\alpha}\lambda \to \infty$, and the last two terms in (3.26) become small relative to $\alpha\lambda G(\alpha\lambda) + \bar{\alpha}\lambda G(\bar{\alpha}\lambda)$ as $\alpha\lambda \to \infty$ and $\bar{\alpha}\lambda \to \infty$. More precisely, the following theorem provides necessary and sufficient conditions on $\tilde{r}_{-1}, \tilde{r}_0, \tilde{r}_1$ so that the distortion is asymptotically the same as that induced by centroids.

Theorem 11. *Consider a family of infinite-level uniform threshold scalar quantizers with offsets α_Δ and reconstruction levels $\{\tilde{r}_{\Delta,k}\}$ parameterized by the step size Δ, $0 < \Delta < \infty$, and with $\Delta \to \infty$ implying $\alpha_\Delta \Delta \to \infty$ and $(1-\alpha_\Delta)\Delta \to \infty$. Such reconstruction levels are asymptotically low resolution optimal for a Gaussian source with mean zero and variance σ^2 if and only if the following hold:*

$$\begin{aligned} A. \quad \tilde{r}_{\Delta,-1} &= -\alpha_\Delta \Delta \left[1 + o_\Delta + o_{(\alpha_\Delta - \frac{1}{2})\Delta \to 0}\, I_{\alpha_\Delta > \frac{1}{2}}\right], \\ B. \quad \tilde{r}_{\Delta,0} &= \sqrt{\alpha_\Delta \lambda\, G(\alpha_\Delta \lambda) + (1-\alpha_\Delta)\lambda\, G((1-\alpha_\Delta)\lambda}\; o_\Delta\,, \\ C. \quad \tilde{r}_{\Delta,1} &= (1-\alpha_\Delta)\Delta \left[1 + o_\Delta + o_{(\frac{1}{2} - \alpha_\Delta)\Delta \to 0}\, I_{\alpha_\Delta < \frac{1}{2}}\right], \end{aligned}$$

where $\lambda = \frac{\Delta}{\sigma}$, I_F denotes the indicator function of the event F, and by "asymptotically low resolution optimal" we mean $\lim_{\Delta\to\infty} \frac{\sigma^2 - \widetilde{D}(\alpha_\Delta, \Delta, \sigma^2, \{\tilde{r}_{\Delta,k}\})}{\sigma^2 - D(\alpha_\Delta, \Delta, \sigma^2)} = 1$, with $\widetilde{D}(\alpha_\Delta, \Delta, \sigma^2, \{\tilde{r}_{\Delta,k}\})$ and $D(\alpha_\Delta, \Delta, \sigma^2)$ denoting distortion assuming parametric and centroid levels, respectively.

Theorem 11 can be interpreted as follows. Suppose Δ is large. Then $\alpha\Delta >> \bar{\alpha}\Delta$ implies $\alpha > \frac{1}{2}$ and thus $I_{\alpha_\Delta < \frac{1}{2}} = 0$, from which it follows, using C above, that $\tilde{r}_{\Delta,1}$ needs to be close to the centroid of S_1. $\alpha\Delta >> \bar{\alpha}\Delta$ also implies that $I_{\alpha_\Delta > \frac{1}{2}} = 1$ and that $(\alpha_\Delta - \frac{1}{2})\Delta >> 0$, from which it follows, using A above, that there is no restriction on $\tilde{r}_{\Delta,-1}$ (except for lying in S_{-1}). Similarly, if $\alpha\Delta << \bar{\alpha}\Delta$, then $\tilde{r}_{\Delta,-1}$ needs to be close to the centroid of S_{-1} and there is no restriction on $\tilde{r}_{\Delta,1}$ (except for lying in S_1). Lastly, if $\alpha\Delta \approx \bar{\alpha}\Delta$, then both $\tilde{r}_{\Delta,-1}$ and $\tilde{r}_{\Delta,1}$ need to be close to the centroids of S_{-1} and S_1, respectively. In all cases, $\tilde{r}_{\Delta,0}$ needs to be sufficiently small, as given in B.

Proof: For brevity we omit the subscript Δ from α_Δ and from $\tilde{r}_{\Delta,k}$, and write $\widetilde{D}$ and D instead of $\widetilde{D}(\alpha_\Delta, \Delta, \sigma^2, \{\tilde{r}_{\Delta,k}\})$ and $D(\alpha_\Delta, \Delta, \sigma^2)$, respectively. We will show that the conditions given in the theorem are necessary and sufficient for $\frac{(\sigma^2 - \widetilde{D})/\sigma^2}{(\sigma^2 - D)/\sigma^2} \to 1$

as $\Delta \to \infty$, or equivalently, for $\frac{(\widetilde{D}-D)/\sigma^2}{(\sigma^2-D)/\sigma^2} \to 0$ as $\Delta \to \infty$. We note that since for the considered family of quantizers, $\Delta \to \infty$ implies $\alpha\Delta \to \infty$ and $\bar{\alpha}\Delta \to \infty$, it follows that expressions such as $o_{\alpha\Delta}$ and $o_{\bar{\alpha}\Delta}$ simply become o_Δ. We comment further that since the variance of the source σ^2 is fixed (as we are considering a family of quantizers rather than of quantizer-source pairs), letting $\Delta \to \infty$ is equivalent to letting $\lambda \to \infty$ (we will use the latter notation). Thus, expressions such as $o_{\alpha\lambda}$ and $o_{\bar{\alpha}\lambda}$ become o_λ.

From Theorem 7 we have

$$\frac{\sigma^2 - D}{\sigma^2} \;=\; \Big(\alpha\lambda\, G(\alpha\lambda) + \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\Big)\,\big[1 + o_\lambda\big]\;. \tag{3.34}$$

And from Theorem 10 we have

$$\begin{aligned}\frac{\widetilde{D} - D}{\sigma^2} \;=\;& \alpha\lambda\, G(\alpha\lambda)\bigg[\Big(\frac{\tilde{r}_{-1} + \alpha\Delta}{\alpha\Delta}\Big)^2\big[1 + o_\lambda\big] + o_\lambda\bigg] \\ &+\; \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\bigg[\Big(\frac{\tilde{r}_1 - \bar{\alpha}\Delta}{\bar{\alpha}\Delta}\Big)^2\big[1 + o_\lambda\big] + o_\lambda\bigg] \\ &+\; 2\Big[G(\bar{\alpha}\lambda) - G(\alpha\lambda)\Big]\Big(\frac{\tilde{r}_0}{\sigma}\Big)\,\big[1 + o_\lambda\big] \;+\; \Big(\frac{\tilde{r}_0}{\sigma}\Big)^2\big[1 + o_\lambda\big]\;.\end{aligned} \tag{3.35}$$

Using (3.34) and (3.35), sufficiency is easy to see. Therefore, we focus on necessity.

To show necessity, we need to show that if (3.35) is asymptotically negligible relative to (3.34), then the conditions of the theorem are met. We observe that the difficulty in showing this stems from the fact that not all terms in (3.35) need have the same sign. Specifically, the third term may be either positive or negative, while all other terms are always positive. In order to show necessity, we will show the contrapositive, namely, if the conditions given in the theorem statement are not met, then (3.35) is not asymptotically negligible relative to (3.34).

We begin by considering the condition on $\tilde{r}_0$, i.e. condition B. Suppose this condition is not satisfied. Then there exists $\varepsilon > 0$ such that for any λ_o there exists

$\lambda > \lambda_o$ for which

$$\frac{|\tilde{r}_0/\sigma|}{\sqrt{\alpha\lambda\, G(\alpha\lambda) + \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)}} \;>\; \varepsilon \;. \tag{3.36}$$

We show below that $\limsup_{\lambda\to\infty} \frac{\widetilde{D}-D}{\sigma^2-D} > 0$, which implies that $\lim_{\lambda\to\infty} \frac{(\widetilde{D}-D)/\sigma^2}{(\sigma^2-D)/\sigma^2} \neq 0$, which in turn implies that $\lim_{\lambda\to\infty} \frac{(\sigma^2-\widetilde{D})/\sigma^2}{(\sigma^2-D)/\sigma^2} \neq 1$, i.e. (3.35) is not asymptotically negligible relative to (3.34). Dropping the terms in (3.35) involving $\tilde{r}_{-1}$, $\tilde{r}_1$ and using (3.34), we have that for any λ_0 there exists $\lambda > \lambda_o$ such that

$$\begin{aligned}
\frac{\widetilde{D}-D}{\sigma^2-D} \;&>\; \frac{\left(\frac{\tilde{r}_0}{\sigma}\right)^2\big[1+o_\lambda\big] + 2\Big[G(\bar{\alpha}\lambda) - G(\alpha\lambda)\Big]\left(\frac{\tilde{r}_0}{\sigma}\right)\big[1+o_\lambda\big]}{\big(\alpha\lambda\, G(\alpha\lambda) + \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)\big)\big[1+o_\lambda\big]} \\
&>\; \frac{\left(\frac{\tilde{r}_0}{\sigma}\right)^2\Big[1 - \frac{2\left(G(\alpha\lambda)+G(\bar{\alpha}\lambda)\right)}{|\tilde{r}_0/\sigma|}\Big]}{\alpha\lambda\, G(\alpha\lambda) + \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)}\big[1+o_\lambda\big] \\
&\overset{(a)}{>}\; \varepsilon^2\Big[1 - \frac{2\big(G(\alpha\lambda)+G(\bar{\alpha}\lambda)\big)}{\varepsilon\sqrt{\alpha\lambda\, G(\alpha\lambda) + \bar{\alpha}\lambda\, G(\bar{\alpha}\lambda)}}\Big]\big[1+o_\lambda\big] \\
&>\; \varepsilon^2\Big[1 - \frac{2}{\varepsilon}\sqrt{G(\alpha\lambda)+G(\bar{\alpha}\lambda)}\Big]\big[1+o_\lambda\big] \;>\; \frac{\varepsilon^2}{2} \;>\; 0\;,
\end{aligned} \tag{3.37}$$

where (a) follows from (3.36). Therefore, $\limsup_{\lambda\to\infty} \frac{\widetilde{D}-D}{\sigma^2-D} > 0$, which proves the contrapositive, and consequently, shows the necessity of condition B.

Finally, given that condition B is satisfied, it is easy to see that conditions A and C on $\tilde{r}_{-1}$ and $\tilde{r}_1$, respectively, are necessary as well. This concludes the proof of the theorem. □

We observe that the result concerning entropy, i.e. Theorem 6, is not affected by the choice of reconstruction levels. Therefore, we obtain the following corollary, which is a direct consequence of Theorems 6, 7, 9 and 11.

Corollary 12. *Consider a family of infinite-level uniform threshold scalar quantizers, as in Theorem 11, with offsets α_Δ and reconstruction levels $\{\tilde{r}_{\Delta,k}\}$ parameterized by the step size Δ, $0 < \Delta < \infty$, and with $\Delta \to \infty$ implying $\alpha_\Delta\Delta \to \infty$ and $(1-\alpha_\Delta)\Delta \to \infty$. For this family, the rate of the quantizers goes to 0 as $D \to \sigma^2$ with*

the same slope as that of the operational rate-distortion function of infinite-level uniform threshold quantization if and only if the reconstruction levels satisfy conditions A, B and C of Theorem 11.

3.3 Binary Quantizers

A binary (two-level) scalar quantizer is characterized by three numbers: a threshold t and two reconstruction levels $r_1 < t$ and $r_2 \geq t$. Let $S_1(t) = (-\infty, t)$ and $S_2(t) = [t, \infty)$ be the two quantization cells, and let the quantization rule be $q(x) = r_k$ when $x \in S_k$, $k = 1, 2$.

As in the previous section, the source considered is stationary, memoryless Gaussian with mean zero and variance σ^2, and the reconstruction levels r_1 and r_2 are taken to be the cell centroids, unless otherwise specified. We let P_k or $P_k(t, \sigma^2)$ denote the probability of the source value lying in S_k, $k = 1, 2$.

Let $H(t, \sigma^2) = \mathcal{H}(P_1(t, \sigma^2), P_2(t, \sigma^2))$ denote the entropy of the quantizer output with threshold t for the Gaussian source. Let $D(t, \sigma^2) \stackrel{\Delta}{=} \int_{-\infty}^{\infty}(x - q(x))^2 f(x)\, dx$ denote the mean-squared error distortion of this quantizer on this source. The operational rate-distortion function of binary quantization for this source is $R_{B,\sigma^2}(D) = \inf_{t:D(t,\sigma^2)\leq D} H(t, \sigma^2)$, which specifies the least entropy of any such quantizer with mean-squared error D or less.

It is easy to see that $P(t, \sigma^2) = P(\frac{t}{\sigma}, 1)$, $H(t, \sigma^2) = H(\frac{t}{\sigma}, 1)$, $D(t, \sigma^2) = \sigma^2 D(\frac{t}{\sigma}, 1)$, and $R_{B,\sigma^2}(D) = R_{B,1}(\frac{D}{\sigma^2})$. Hence it is convenient to parameterize P_k and H by $\lambda = \frac{t}{\sigma}$, i.e. $P_k(\lambda)$ and $H(\lambda)$. Due to the symmetry of the Gaussian density, it suffices to restrict attention to $t \geq 0$.

As before, we will find asymptotic low resolution approximations to entropy and distortion, and then combine these to determine the asymptotic low resolution ex-

pression for $R_{B,\sigma^2}(\mathrm{D})$. We also determine which cells dominate entropy and distortion, and we relax the requirement that the levels be centroids. Since the derivations in the binary case are similar in spirit to those in the uniform case, we will only state the results and provide no proofs, so as to spare the reader repetitive details.

Theorem 13. *For a binary scalar quantizer with threshold t applied to a Gaussian source with mean zero and variance σ^2*

$$H(\lambda) \;=\; \frac{\log e}{2}\lambda\, G(\lambda)\left[1+o_\lambda\right] \;,$$

where as indicated earlier $\lambda = \frac{t}{\sigma}$, $H(\lambda) = \mathcal{H}(P_{-1}(\lambda), P_1(\lambda))$ and $G(x)$ denotes a zero-mean, unit-variance Gaussian density.

Theorem 14. *For a binary scalar quantizer with threshold t and reconstruction levels at cell centroids applied to a Gaussian source with mean zero and variance σ^2*

$$\frac{\sigma^2 - D(t,\sigma^2)}{\sigma^2} \;=\; \lambda\, G(\lambda)\left[1+o_\lambda\right] \;,$$

where $\lambda = \frac{t}{\sigma}$.

Theorem 15. *In the low resolution region, the operational rate-distortion function of binary scalar quantization for a Gaussian source with variance σ^2 is*

$$R_{B,\sigma^2}(D) \;=\; \frac{\log e}{2}\Big(1-\frac{D}{\sigma^2}\Big)\left[1+o_{_{D\to\sigma^2}}\right] \;.$$

Notice that the expression given in this theorem for binary quantization is precisely the same as that given in Theorem 9 for infinite-level uniform threshold quantization, which in turn matches the Shannon rate-distortion function in the low resolution region. We conclude that binary quantization is another type of quantization that is asymptotically optimal in the low resolution region.

We now comment on the cells that dominate the entropy and distortion. As before, when entropy is small, it is dominated by the cell that has largest probability not close to one, which is S_2. And just as with uniform quantizers, when distortion is close to σ^2, $\sigma^2 - D$ is dominated by the cell whose probability is nearly one, namely, S_1. That is, $\frac{\sigma^2 - D_1}{\sigma^2 - D} \approx 1$.

As with uniform quantizers, the requirement for cell centroids can be relaxed somewhat. The following two theorems and corollary are the direct equivalents of Theorem 10, Theorem 11 and Corollary 12.

Theorem 16. *For a binary scalar quantizer with reconstruction levels $\tilde{r}_{t,1}$, $\tilde{r}_{t,2}$ that are not necessarily centroids, and a Gaussian source with mean zero and variance σ^2, the distortion $\widetilde{D}(t, \sigma^2, \tilde{r}_{t,1}, \tilde{r}_{t,2})$ satisfies*

$$\begin{aligned}\frac{\widetilde{D}(t, \sigma^2, \tilde{r}_{t,1}, \tilde{r}_{t,2}) - D(t, \sigma^2)}{\sigma^2} \;&=\; \lambda\, G(\lambda)\left[\left(\frac{\tilde{r}_{t,2} - t}{t}\right)^2 \big[1 + o_\lambda\big] + \left(\frac{\tilde{r}_{t,2} - t}{t}\right) o_\lambda + o_\lambda\right] \\ &\quad +\; 2\, G(\lambda)\left(\frac{\tilde{r}_{t,1}}{\sigma}\right) \big[1 + o_\lambda\big] + \left(\frac{\tilde{r}_{t,1}}{\sigma}\right)^2 \big[1 + o_\lambda\big]\ ,\end{aligned}$$

where $\lambda = \frac{t}{\sigma}$ and $D(t, \sigma^2)$ is the distortion induced by cell centroids.

Theorem 17. *Consider a family of binary scalar quantizers whose reconstruction levels $\tilde{r}_{t,1}$, $\tilde{r}_{t,2}$ are parameterized by the threshold t, $0 < t < \infty$. Such reconstruction levels are asymptotically low resolution optimal for a Gaussian source with mean zero and variance σ^2 if and only if the following hold:*

$$\begin{aligned} A.\quad \tilde{r}_{t,1} \;&=\; \sqrt{\lambda\, G(\lambda)}\; o_t\ , \\ B.\quad \tilde{r}_{t,2} \;&=\; t\,\big[1 + o_t\big]\ , \end{aligned}$$

where $\lambda = \frac{t}{\sigma}$, and by "asymptotically low resolution optimal" we mean $\lim_{t\to\infty} \frac{\sigma^2 - \widetilde{D}(t, \sigma^2, \tilde{r}_{t,1}, \tilde{r}_{t,2})}{\sigma^2 - D(t, \sigma^2)} = 1$, with $\widetilde{D}(t, \sigma^2, \tilde{r}_{t,1}, \tilde{r}_{t,2})$ and $D(t, \sigma^2)$ denoting distortion assuming parametric and centroid levels, respectively.

Corollary 18. *Consider a family of binary scalar quantizers, as in Theorem 17, whose reconstruction levels $\tilde{r}_{t,1}$, $\tilde{r}_{t,2}$ are parameterized by the threshold t, $0 < t < \infty$. For this family, the rate of the quantizers goes to 0 as $D \to \sigma^2$ with the same slope as that of the operational rate-distortion function of binary scalar quantization if and only if the reconstruction levels satisfy conditions A and B of Theorem 17.*

3.4 Conclusions

We considered infinite-level uniform threshold and binary scalar quantizers in the asymptotic case that the cell sizes go to infinity (for the uniform case) and that the quantizer threshold goes to infinity (for the binary case). In both cases the source of the quantizers was Gaussian. We derived simple formulas for the rate of convergence of entropy to zero and of mean-squared error distortion to the source variance.

The convergence of entropy and distortion as $\lambda \to \infty$ for uniform quantization is not uniform in the offset α. The derived formulas show how entropy and distortion depend on α as well as λ. Specifically, they provide accurate approximations when both $\alpha\lambda$ and $\bar{\alpha}\lambda$ are large.

Using these convergence formulas, the operational rate-distortion of infinite-level uniform threshold and binary scalar quantization was shown to approach zero as $D \to \sigma^2$ with the same slope as that of the Shannon rate-distortion function. This shows that in the low resolution region scalar quantization is asymptotically optimal.

Finally, necessary and sufficient conditions on the reconstruction levels were provided, both in the uniform and binary cases, under which the entropy approaches zero as $D \to \sigma^2$ with the same slope as the operational rate-distortion function of scalar quantization, and in turn the Shannon rate-distortion function.

Appendix

Lemma A1. *For a unit variance continuous source with infinite support, the operational rate-distortion function $R_{U,1,\alpha}(D)$ of infinite-level uniform threshold scalar quantization with a fixed offset $0 < \alpha < 1$ satisfies*

$$\liminf_{D\to 1} \frac{R_{U,1,\alpha}(D)}{1-D} \;\geq\; \liminf_{\lambda\to\infty} \frac{H(\alpha,\lambda)}{1-D(\alpha,\lambda,1)} \;.$$

Proof: To keep notation short, we omit the α and the $\sigma^2 = 1$ from all quantities except $R_{U,1,\alpha}(D)$. For $0 \leq D < 1$, let

$$\lambda_D^* \;\stackrel{\Delta}{=}\; \min\left\{\lambda \geq 0 :\; D(\lambda) \leq D, \text{ and } H(\lambda) = R_{U,1,\alpha}(D)\right\} \;, \tag{A1}$$

with the conventions that $D(0) = 0$ and $H(0) = \infty$. We need to show that the above is well defined, and we will also show that $D \to 1$ implies $\lambda_D^* \to \infty$, from which the result of the lemma will easily follow. In order to do so, we first prove two facts about $R_{U,1,\alpha}(D)$.

We will use without proof the easily seen properties that for $0 \leq \lambda < \infty$, $D(\lambda) < 1$, $H(\lambda) > 0$, $D(\lambda)$ and $H(\lambda)$ are continuous functions of λ, $\lim_{\lambda\to\infty} D(\lambda) = 1$, and $\lim_{\lambda\to\infty} H(\lambda) = 0$.

The first fact is that for $D < 1$, there exists $\lambda \geq 0$ such that $D(\lambda) \leq D$ and $H(\lambda) = R_{U,1,\alpha}(D)$. This is because $R_{U,1,\alpha}(D)$ is defined as the infimum of the continuous function $H(\lambda)$ over the set $\{\lambda \geq 0 : D(\lambda) \leq D\}$, which is not empty (because $D(0) = 0$), closed (because $D(\lambda)$ is continuous), bounded below by 0, and bounded above (because $D(\lambda) \to 1$ implies $D(\lambda) > D$ for all sufficiently large λ).

The second fact is that $D \to 1$ implies $R_{U,1,\alpha}(D) \to 0$. To demonstrate this, let $\lambda_D \stackrel{\Delta}{=} \min\{\lambda \geq 0 : D(\lambda) = D\}$, which is well defined because the set $\{\lambda \geq 0 : D(\lambda) = D\}$ is not empty (because $D(\lambda)$ goes continuously from 0 to 1 as λ ranges

from 0 to ∞), closed (because $D(\lambda)$ is continuous), and bounded (because $\lambda \geq 0$ and $D(\lambda) > D$ for all sufficiently large λ). It now follows straightforwardly from the facts that $D(\lambda)$ is continuous, $D(\lambda) < 1$ for all $0 \leq \lambda < \infty$, and $\lim_{\lambda\to\infty} D(\lambda) = 1$ that $D \to 1$ implies $\lambda_D \to \infty$. Combining this with the property stated earlier, we have that $D \to 1$ implies $H(\lambda_D) \to 0$. Therefore, $D \to 1$ implies $R_{U,1,\alpha}(D) \leq H(\lambda_D) \to 0$.

Returning to λ_D^*, the first of the two facts above shows that the set $\{\lambda \geq 0 : D(\lambda) \leq D, H(\lambda) = R_{U,1,\alpha}(D)\}$ is not empty. It is bounded because it is a subset of $\{\lambda \geq 0 : D(\lambda) \leq D\}$, which was earlier shown to be bounded, and it is closed because $D(\lambda)$ and $H(\lambda)$ are continuous functions. It follows that λ_D^* is well defined.

Since by the definition of λ_D^*, $H(\lambda_D^*) = R_{U,1,\lambda}(D)$, it follows from the second fact above that $D \to 1$ implies $H(\lambda_D^*) \to 0$. As we now show, this implies that $\lambda_D^* \to \infty$. To do so, we prove the contrapositive. Accordingly, if $\lambda_D^* \not\to \infty$, then there exists $\lambda_o < \infty$ such that for all $D < 1$ there exists D', $D \leq D' < 1$ such that $\lambda_{D'}^* \leq \lambda_o$. Consequently, $H(\lambda_{D'}^*) \geq \gamma \stackrel{\Delta}{=} \min_{\lambda\leq\lambda_o} H(\lambda) > 0$, where the second inequality follows from the fact $H(\lambda)$ is continuous and positive on the closed set $\{0 \leq \lambda \leq \lambda_o\}$. It follows that $\liminf_{D\to1} H(\lambda_D^*) \geq \gamma$. Therefore, $\lambda_D^* \not\to \infty$ as $D \to 1$ implies $H(\lambda_D^*) \not\to 0$ as $D \to 1$. The contrapositive together with the fact above that $D \to 1$ implies $H(\lambda_D^*) \to 0$ shows that $D \to 1$ implies $\lambda_D^* \to \infty$.

The statement of the lemma is now demonstrated as follows:

$$\liminf_{D\to1} \frac{R_{U,1,\alpha}(D)}{1-D} \stackrel{(a)}{=} \liminf_{D\to1} \frac{H(\lambda_D^*)}{1-D} \stackrel{(b)}{\geq} \liminf_{D\to1} \frac{H(\lambda_D^*)}{1-D(\lambda_D^*)} \stackrel{(c)}{=} \liminf_{\lambda\to\infty} \frac{H(\lambda)}{1-D(\lambda)},$$

where (a) and (b) follow from the definition of λ_D^*, and (c) follows from having $D \to 1$ imply $\lambda_D^* \to \infty$. □

REFERENCES

[1] R.M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Info. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[2] D. F. Lyons and D. L. Neuhoff, "A coding theorem for low-rate transform codes," *IEEE International Symposium on Information Theory*, p. 333, Jan. 1993.

[3] D.F. Lyons, "Fundamental limits of low-rate transform codes," *Ph.D. Thesis, EECS Department, University of Michigan*, 1992.

[4] J. Ziv, "On universal quantization," *IEEE Trans. Info. Theory*, vol. 31, no. 3, pp. 344–347, May 1985.

[5] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizers," *IEEE Trans. Info. Theory*, vol. 38, no. 2, pp. 428–436, Mar. 1992.

[6] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, John-Wiley & Sons Inc., New York, 1967.

[7] R. G. Gallager, *Information Theory and Reliable Communication*, John-Wiley & Sons Inc., New York, 1968.

CHAPTER IV

Entropy of Highly Correlated Quantized Data

4.1 Introduction

This chapter considers the following question concerning quantization and encoding of oversampled data. The problem of oversampling and and quantization has been widely addressed in the literature, for example [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. But the question posed below has not been answered. Suppose a continuous-time stationary random process X is sampled from the interval $[0, 1]$ every τ seconds, quantized with some fixed scalar quantizer Q, and then encoded with an ideal entropy coder at rate $H_\tau(Q(X))$ bits/sample, where $Q(X)$ denotes the quantized samples over the interval $[0, 1]$, and H_τ denotes the joint entropy of these quantized samples divided by the number of samples (it is subscripted by τ to reflect that the statistics of the quantized samples depend on τ).

As the sampling interval decreases to zero, what happens to the rate R (bits/second) at which bits are produced by the entropy coder? Does it go to zero? Remain finite? Tend to infinity? The answer is not obvious in that $R_\tau = \frac{1}{\tau} H_\tau(Q(X))$ is the product of $1/\tau$, which increases to infinity, and $H_\tau(Q(X))$, which decreases to zero, due to the fact that the samples (and consequently, the quantized samples) become increasingly correlated as τ decreases. Essentially, the question asks if the increasing correlation

can be sufficiently exploited to counteract the increasingly large number of samples. The first result of this paper shows that under very mild conditions, the answer is no. That is, R_τ tends to infinity. In other words, although $H_\tau(Q(X))$ approaches zero, it becomes large relative to τ.

This is not what was anticipated. Instead, it seemed plausible that R_τ would follow the trend of ideal rate-distortion coding and approach a finite value. Specifically, suppose that instead of a scalar quantizing and ideal entropy coding over the finite interval $[0, 1]$, the process were sampled over the interval $(-\infty, \infty)$, and the samples were lossy encoded with an ideal rate-distortion encoder with the same distortion as the scalar quantizer. Then the encoder would produce $\frac{1}{\tau}\mathcal{R}_\tau(D)$ bits/sec, where D is the distortion of the scalar quantizer, and $\mathcal{R}_\tau(D)$ is the rate-distortion function (in bits/sample) of the sampled random process (subscripted to reflect the dependence of the discrete-time process on τ). In the Gaussian case, for example, it is well-known that as τ decreases to zero, $\frac{1}{\tau}\mathcal{R}_\tau(D)$ approaches $\mathcal{R}(D)$, the rate-distortion function of the continuous-time random process X [11, 12]. Since $\mathcal{R}(D)$ is ordinarily finite, we see that restricting the lossy encoder to be a combination of a fixed scalar quantizer and an ideal (sampling-rate adapted) entropy coder over the interval, increases the limiting rate from finite to infinite. This may be surprising because one generally expects scalar quantization with entropy coding to have rate that exceeds the rate-distortion function by at most a constant (c.f. [13, 14, 15]). However, this expectation applies to discrete-time processes, and if the rate $H_\tau(Q(X))$ of scalar quantization with entropy coding exceeds $\mathcal{R}_\tau(D)$ by, say, c bits/sample, then when τ is small, $R_\tau = \frac{1}{\tau}H_\tau(Q(X))$ exceeds $\mathcal{R}(D) \approx \frac{1}{\tau}\mathcal{R}_\tau(D)$ by approximately c/τ, which is large.

We comment that it is not clear whether the restriction to a finite interval or the use of fixed scalar quantization with entropy coding causes the rate to tend to infinity as the sampling interval goes to zero. We believe that both play some role in this behavior. Namely, if an infinite interval together with fixed scalar quantization and entropy coding were considered, then there exist examples (see, for instance, the example given by (6.1) in the Future Work Section of Chapter VI) for which the rate does not tend to infinity, under the same conditions for which it does tend to infinity when a finite interval is considered. However, requiring some additional conditions may eliminate such examples and expose those cases where fixed scalar quantization and entropy coding is the sole reason for the rate tending to infinity.

The question then arises as to how fast R_τ approaches infinity. In general, this is a difficult question. To obtain a partial answer, we make the problem tractable by considering Gaussian sources, uniform threshold quantization, and conditional entropy coding that attains $H\big(Q(X_\tau)\,|\,Q(X_0)\big)$ instead of $H_\tau(Q(X))$. The second result of this paper (Theorem 7) can be applied so as to show that when τ is small

$$R_\tau \;\stackrel{\sim}{\leq}\; \bar{R}_\tau \;\stackrel{\Delta}{=}\; \frac{1}{\tau}H\big(Q(X_\tau)\,|\,Q(X_0)\big) \;\approx\; -\frac{1}{\tau}M\sqrt{1-\rho(\tau)}\,\log_2\sqrt{1-\rho(\tau)}\;, \quad (4.1)$$

where

$$M \;=\; \frac{2\sqrt{2}}{\pi}\sum_{k=0}^{\infty} e^{-\frac{(k+\frac{1}{2})^2\Delta^2}{2\sigma^2}}\;,$$

Δ is the step size of the uniform quantizer, σ^2 is the variance of X, and $\rho(\cdot)$ is its normalized covariance function, i.e. $\rho(\tau)$ is the correlation coefficient between adjacent samples of X. Note that the above expression for $\bar{R}_\tau$ depends only on the behavior of $\rho(\tau)$ for τ near zero, where $\rho(\tau) \approx 1$. This means it does not depend on spectral characteristics of X, such as whether the process is bandlimited or not.

For example, if $\rho(\tau) = e^{-|\tau|}$, which implies X is Markov, and if τ is small, then

$\sqrt{1-\rho(\tau)/\sigma^2} \approx \sqrt{\tau}$, and

$$\bar{R}_\tau \;\approx\; -\frac{M}{2}\frac{\log_2 \tau}{\sqrt{\tau}} \;. \tag{4.2}$$

Or if $\rho(\tau) = e^{-\tau^2}$ and τ is small, then $\sqrt{1-\rho(\tau)/\sigma^2} \approx \tau$, and

$$\bar{R}_\tau \;\approx\; -M \log_2 \tau \;. \tag{4.3}$$

The organization of the chapter is as follows. The chapter is broken into two main parts, each of which introduces relevant notation. The first part is composed of Section 4.2 only, and shows that R_τ tends to infinity as τ goes to zero. The second part, which consists of Sections 4.3 – 4.6, derives the asymptotic formula for conditional entropy. Specifically, Section 4.3 states the result and a corollary, Section 4.4 introduces necessary notation, Section 4.5 proves the result and its corollary, and Section 4.6 provides proofs of lemmas that are used to show the result. In Section 5.5 concluding remarks are offered. Finally, Appendix A contains proofs of certain lemmas, and Appendix B contains a technical discussion regarding separability and measurability of random processes.

4.2 Joint entropy of quantized samples at high sampling rates

Consider a continuous-time, stationary and continuous in probability[1] random process that is sampled every τ seconds. Each sample is quantized using an arbitrary, yet unchanging, scalar quantizer. This section addresses the question of what happens to the joint entropy of the scalar quantized samples from some finite time interval, as the sampling interval τ tends to zero? It will be shown that under a very

[1]Continuity in probability is a very mild technical condition that is needed to allow stationarity to imply that the events $\{\omega : X_t(\omega) \le r,\ \ a < t < b\}$ and $\{\omega : X_t(\omega) \le r,\ \ a+s < t < b+s\}$ have the same probability, and to ensure that when taking the expected value of the time integral of a function of the random process, the expected value can be brought inside the integral.

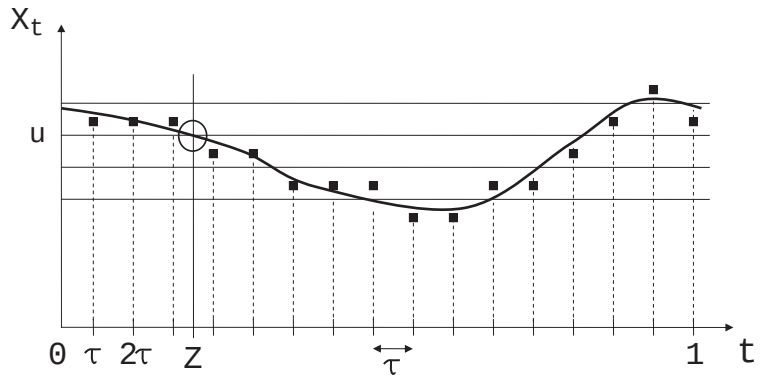

Figure 4.1: A sample path of the random process X_t on the interval $[0, 1]$, which is sampled and quantized. u is the quantization threshold considered, Z is the first crossing time of u, and τ is the sampling interval.

mild condition the joint entropy tends to infinity as $\tau \to 0$. Specifically, we assume that with positive probability the random process crosses (see Definition 2 below) some quantization threshold. This will be our only additional assumption about the random process, aside from stationarity. Next we provide a brief outline of the proof and the issues involved in showing this result.

The key idea[2] is showing that as $\tau \to 0$, one can obtain from the quantized samples an increasingly accurate description of a quantity that has infinite entropy. This in turn will imply that the joint entropy of the quantized samples tends to infinity as $\tau \to 0$. More specifically, we consider the finite time interval $[0, 1]$, and let the approximated quantity be the time of the first crossing in $[0, 1]$ of some quantization threshold, which can be increasingly well approximated from the quantized samples of the random process. It will follow that the joint entropy of the outputs of the quantizers tends to infinity. This is illustrated in Figure 4.1. While the idea is fairly simple, it does involve certain technical hurdles that need to be overcome. Specifically, the following are needed:

[2]The author would like to thank Bruce Hajek for providing this idea.

1. A useful definition of a crossing, whose time can be approximated.
2. A proof that with positive probability a first crossing indeed occurs.
3. The definition of a random variable that is a function of the quantization indices that approximates the time of the first crossing, and a proof that it converges in probability to the time of the first crossing.
4. A proof that the time of the first crossing has infinite entropy. In fact it is shown that it has a property that is equivalent to absolute continuity on $\mathbb{R}$.
5. A proof that 2, 3, and 4 above imply that the entropy of the approximating random variable tends to infinity as the sampling interval goes to zero.

We proceed with some notation and the definition of a crossing, followed by a formal statement of the result of this section and its proof.

Let $\{X_t,\ t \in T\}$, $T = (-\infty, \infty)$, be a continuous-time random process, denoted for short by X, which is defined over the probability space $(\Omega, \mathcal{F}, P)$. Let $\omega \in \Omega$ denote a point in Ω, which corresponds to a sample path denoted $X(\omega)$. We let $X_t(\omega)$ denote the value of the sample path ω at time t.

We now discuss some technical regularity conditions that X must satisfy in order that some basic operations are valid. Formal definitions and discussion are provided in Appendix B. Sets of the form $\{\omega :\ X_t(\omega) \leq r,\quad a < t < b\}$ need not be events unless the process X is *separable*. Furthermore, given separability and stationarity, one also needs to assume *continuity in probability* in order to have that the probability of a shifted version of the above event, namely, $\{\omega :\ X_t(\omega) \leq r,\ \ a+s < t < b+s\}, s \in \mathbb{R}$, remains the same. Lastly, another property that is needed is *measurability*, which ensures that when taking the expected value of the time integral of a function of the random process, the expected value can be brought inside the integral. Thus, we

observe that separability and measurability are technical regularity conditions that, essentially, allow us to perform basic operations on random processes.

We note that for any process $\{X_t,\ t \in T\}$, there exists a process $\{\widetilde{X}_t,\ t \in T\}$ defined on the same probability space, which is separable and *equivalent* to the original process X_t, i.e. whose finite dimensional distributions are the same as those of Y_t [16] (p. 42), [17] (pp. 89). Additionally, for any process $\{X_t,\ t \in T\}$, which is continuous in probability, there exists a process $\{\widetilde{X}_t,\ t \in T\}$ defined on the same probability space, which is measurable and equivalent to X_t. Thus, by requiring that the stationary random process X_t be continuous in probability, we may also assume that it is separable and measurable (for if it is not, we can consider an equivalent random process $\widetilde{X}_t$, which is). Consequently, we shall treat sets of the form $\{\omega :\ X_t(\omega) \leq r,\ \ a < t < b\}$ as events whose shifts have the same probability. For the benefit of those readers who are not well versed in these concepts, we provide a more detailed discussion of separability and measurability in Appendix B.

Next, given some $\tau > 0$, let τ denote a sampling interval and let $I_k^\tau(\omega)$, $k \in \mathbb{Z}$, denote the quantization index of $X_{k\tau}(\omega)$, when quantized by some scalar quantizer q. We define the following.

Definition 1. *Let $u \in \mathbb{R}$, let $\delta > 0$, and let $h : \mathbb{R} \to \mathbb{R}$ be some function. Then*

1. *h has a u-*crossing from below *of size δ at t if $(h(t-s) < u$ and $h(t+s) \geq u)$ or $(h(t-s) \leq u$ and $h(t+s) > u)$ for all $0 < s < \delta$.*
2. *h has a u-*crossing from above *of size δ at t if $(h(t-s) \geq u$ and $h(t+s) < u)$ or $(h(t-s) > u$ and $h(t+s) \leq u)$ for all $0 < s < \delta$.*
3. *h has u-*crossing *of size δ at t if it has a u-crossing from below or from above of size δ at t.*

Definition 2. *Let $u \in R$ and $h : \mathbb{R} \to \mathbb{R}$ be some function. h has a* u-crossing *at t if there exists $\delta > 0$ such that h has a u-crossing of size δ at t.*

We now state the main result of this section.

Theorem 3. *Let X be a continuous-time stationary and continuous in probability random process defined over the probability space $(\Omega, \mathcal{F}, P)$. Let q be a quantizer having a quantization threshold u such that* $\Pr(X$ *has a u-crossing in* $[0,1]) > 0$. *Let also $0 < \tau < 1$, and $N_\tau = \left\lfloor \frac{1}{\tau} \right\rfloor$. Then*

$$\lim_{\tau \to 0} H(I_0^\tau, I_1^\tau, \ldots, I_{N_\tau}^\tau) \;=\; \infty \;.$$

Next, we provide some definitions and some lemmas that will be used to prove Theorem 3. The following is a formal definition of first crossing mentioned earlier.

$$Z_\delta^u(\omega) \;=\; \begin{cases} t, & X(\omega) \text{ has a first } u\text{-crossing in [0,1], it is at } t, \text{ and it is of size } \delta \\ \infty, & \text{else} \end{cases} .$$

Observe that Z_δ^u is an extended-valued random variable. Note further that Z_δ^u might equal ∞ not only due to not having any u-crossing in $[0,1]$, but also due to having an infinite number of crossings with no first crossing, e.g. $\sin \frac{1}{t}$, with $u = 0$.

Let $I(u)$ denote the quantization index with which u is associated. Let $K^{\tau,u}$ denote the position, if there is one, of the first quantization index that is smaller than $I(u)$ such that the next quantization index is greater than or equal to $I(u)$, or vice versa, and if there is no such position, then $K^{\tau,u}$ equals some other value. Specifically, if u is quantized to the cell that is to its right, namely, the cell containing u is of the form $[u, v)$, then given $0 < \tau < 1$ and $N_\tau = \left\lfloor \frac{1}{\tau} \right\rfloor$, we define

$$\begin{aligned} K^{\tau,u}(\omega) \;=\; & \text{smallest } k \in \{0, 1, \ldots, N_\tau - 1\} \text{ such that } \big(I_k^\tau(\omega) < I(u) \text{ and } I_{k+1}^\tau(\omega) \geq I(u)\big) \\ & \text{or } \big(I_k^\tau(\omega) \geq I(u) \text{ and } I_{k+1}^\tau(\omega) < I(u)\big), \text{ or } 2N_\tau \text{ if no such } k \text{ exists.} \end{aligned}$$

If the cell containing u, is of the form $(v, u]$, then change " $\geq$ " into " $>$ " and " $<$ " into " $\leq$ " in the definition of $K^{\tau,u}$ above. Let also

$$\widehat{Z}^{\tau,u}(\omega) \;=\; K^{\tau,u}(\omega)\,\tau \;,$$

which is the approximation of $Z^u_\delta(\omega)$ that can be determined from the quantized samples.

We shall also need the following definitions:

$$\begin{aligned} G^u \;&=\; \{\omega : X \text{ has a } u\text{-crossing in } [0,1]\} \;, \\ G^u_\delta \;&=\; \{\omega : X \text{ has a } u\text{-crossing of size } \delta \text{ in } [0,1]\} \;, \\ G^u_{\delta,t} \;&=\; \{\omega : X \text{ has a } u\text{-crossing of size } \delta \text{ at } t\} \;, \end{aligned}$$

Since u remains fixed throughout, for brevity we omit it from now on. We note that the separability of X ensures that the above sets are events. We now proceed with several lemmas.

Lemma 4. *If $P(G) > 0$, then there exists $\delta > 0$ such that*

$$\Pr\left(Z_\delta \in [\frac{\delta}{4}, \frac{\delta}{2}]\right) \;>\; 0 \;.$$

Proof: Let us first observe that

$$\Pr\left(Z_\delta \in [\frac{\delta}{4}, \frac{\delta}{2}]\right) \;=\; \Pr\left(X \text{ has a crossing of size } \delta \text{ in } [\tfrac{\delta}{4}, \tfrac{\delta}{2}]\right) \;.$$

This follows from the fact that if X has a crossing of size δ in the interval $[\frac{\delta}{4}, \frac{\delta}{2}]$, then this crossing is the first crossing of size δ in $[0, 1]$ and vice versa. We will show that the probability of the right-hand term is positive, thus showing the lemma.

Let $\delta_n = \frac{1}{n}$, $n \in \mathbb{Z}^+$. We observe that G_{δ_n} is an increasing sequence of sets whose limit is G. Since $P(G) > 0$, there exists some n_o such that $P(G_{\delta_{n_o}}) \geq \frac{P(G)}{2}$. Next,

set $\delta = \delta_{n_o} = \frac{1}{n_o}$, let $B_i = [\frac{i}{4n_o}, \frac{i+1}{4n_o}] = [\frac{i\delta}{4}, \frac{(i+1)\delta}{4}]$ for $0 \leq i \leq 4n_o - 1$, and define

$$G_\delta(B) \;=\; \{\omega : X(\omega) \text{ has a crossing of size } \delta \text{ in } B\} \; .$$

We now have that

$$\frac{P(G)}{2} \;\leq\; P(G_\delta) \;=\; P(\bigcup_{i=0}^{4n_o-1} G_\delta(B_i)) \;\leq\; \sum_{i=0}^{4n_o-1} P(G_\delta(B_i)) \; .$$

It follows that there must exist $i \in \{0, 1, \ldots, 4n_o - 1\}$ for which

$$P(G_\delta(B_i)) \;\geq\; \frac{1}{4n_o}\frac{P(G)}{2} \;=\; \frac{\delta P(G)}{8} \; .$$

Next, the stationarity of X implies that for any $i, j \in \{0, 1, \ldots, 4n_o - 1\}$, $P(G_\delta(B_i)) = P(G_\delta(B_j))$. In particular, $P(G_\delta(B_1)) \geq \frac{\delta P(G)}{8} > 0$. Recalling that $B_1 = [\frac{\delta}{4}, \frac{\delta}{2}]$ concludes the proof of the lemma. □

Lemma 5. *Let $\delta > 0$. For any $\alpha > 0$, there exists $s > 0$ such that*

$$\Pr(t - s < Z_\delta < t + s) \;<\; \alpha \; ,$$

for any t.

Note that this lemma implies that Z_δ is absolutely continuous on $\mathbb{R}$. (In fact, since probability spaces are finite measure spaces, this lemma is equivalent to absolute continuity on $\mathbb{R}$.)

Proof: The event that $t - s < Z_\delta < t + s$, i.e. that a first crossing of size δ occurs in $(t - s, t + s)$, is a subset of the event that any crossing of size δ occurs in that interval (note that if $t - s > 1$ or $t + s < 0$, then $Z_\delta = \infty$ and, in particular, the event $t - s < Z_\delta < t + s$ has probability zero). We shall in fact upper bound the probability of the latter. We begin by showing that for any t,

$$P(G_{\delta,t}) \;=\; 0 \; . \tag{4.4}$$

To do so, we first make the following two definitions:

$$\Phi_{\delta,t}(\omega) \;=\; \begin{cases} 1, & X(\omega) \text{ has a crossing of size } \delta \text{ at } t \\ 0, & \text{else} \end{cases} ,$$

and

$$A(\omega) \;=\; \{t : X(\omega) \text{ has a crossing of size } \delta \text{ at } t\} \;=\; \{t : \Phi_{\delta,t}(\omega) = 1\} \; . \tag{4.5}$$

Since there can be at most countably many crossings of size δ, it follows that $A(\omega)$ has Lebesgue measure zero. Thus, for any ω, $\int_0^1 \Phi_{\delta,t}(\omega)\, dt \;=\; 0$, and therefore $E\big[\int_0^1 \Phi_{\delta,t}\, dt\big] = 0$, where E denotes expectation. Consequently,

$$\int_0^1 E[\Phi_{\delta,t}]\, dt \;=\; E\Big[\int_0^1 \Phi_{\delta,t}\, dt\Big] \;=\; 0 \; , \tag{4.6}$$

where the swapping of integral and expectation is due to Fubini's theorem [18] (pp. 147-148), which can be used since X is measurable. The stationarity of X implies that $E[\Phi_{\delta,t}]$ is the same for all t. Combining this with (4.6), implies that $E[\Phi_{\delta,t}] = 0$ for all t, which shows (4.4).

Next, define

$$G_{\delta,t,s} \;=\; \{\omega : X(\omega) \text{ has a crossing of size } \delta \text{ in } (t-s, t+s)\} \; .$$

We observe that for any t, as $s \to 0$, $G_{\delta,t,s}$ is a decreasing collection of sets whose limit is $G_{\delta,t}$. Let $\alpha > 0$ be given. Since, as shown, $P(G_{\delta,t}) = 0$, it follows that there exists some s for which $P(G_{\delta,t,s}) < \alpha$. Stationarity of X then implies that

$$P(G_{\delta,t,s}) \;<\; \alpha \quad \text{for any } \; t \; . \tag{4.7}$$

Finally, (4.7) implies that for any $\alpha > 0$, there exists $s > 0$ such that for any t

$$\Pr(t - s < Z_\delta < t + s) \;\leq\; P(G_{\delta,t,s}) \;<\; \alpha \; ,$$

which concludes the proof of the lemma. □

Lemma 6. *Let B be an interval and let Z be a random variable such that* $\Pr(Z \in B) > 0$. *Let Z also have the property that for any $\alpha > 0$, there exists $s > 0$ such that for any t,* $\Pr(t - s < Z < t + s) < \alpha$. *Finally, let $\{Z_n\}$ be a sequence of discrete random variables such that for any $\varepsilon > 0$,* $\lim_{n\to\infty} \Pr(|Z - Z_n| < \varepsilon \,|\, Z \in B) = 1$. *Then,*

$$\lim_{n\to\infty} H(Z_n) = \infty \ .$$

Proof: We begin by writing

$$\begin{aligned} H(Z_n) \ &\geq \ H(Z_n|Z \in B) \Pr(Z \in B) + H(Z_n|Z \in B^c) \Pr(Z \in B^c) \\ &\geq \ H(Z_n|Z \in B) \Pr(Z \in B) \ = \ \bar{p} \sum_k p(z_{n,k}|Z \in B) \log \frac{1}{p(z_{n,k}|Z \in B)} \ , \end{aligned} \tag{4.8}$$

where $\{z_{n,k}\}$ are the values that Z_n may assume, and $\bar{p} \stackrel{\Delta}{=} \Pr(Z \in B)$.

Let $\varepsilon > 0$ be given. Set $A_{n,\varepsilon} = \{|Z - Z_n| < \varepsilon\}$. By assumption, $\lim_{n\to\infty} \Pr(A_{n,\varepsilon} \,|\, Z \in B) = 1$ and $\lim_{n\to\infty} \Pr(A^c_{n,\varepsilon} \,|\, Z \in B) = 0$, where c denotes set complement.

Next, let $\varepsilon_m = \frac{1}{m}$, $m \in \mathbb{Z}^+$. It follows from the fact that for any $\alpha > 0$, there exists $s > 0$ such that for any t, $\Pr(t - s < Z < t + s) < \alpha$, that there exists a sequence $\{s_m\}_{m=1}^{\infty}$ such that for any m, $\Pr(t - s_m < Z < t + s_m) \leq \varepsilon_m$, for any t.

We now proceed by upper bounding $p(z_{n,k} \,|\, Z \in B)$.

$$\begin{aligned} p(z_{n,k} \,|\, Z \in B) \ &= \ \Pr(Z_n = z_{n,k}, A_{n,s_m} \,|\, Z \in B) + \Pr(Z_n = z_{n,k}, A^c_{n,s_m} \,|\, Z \in B) \\ &\leq \ \Pr(Z_n = z_{n,k}, A_{n,s_m} \,|\, Z \in B) + \Pr(A^c_{n,s_m} \,|\, Z \in B) \\ &= \ \frac{\Pr(Z_n = z_{n,k}, |Z - Z_n| < s_m, Z \in B)}{\Pr(Z \in B)} + \Pr(A^c_{n,s_m} \,|\, Z \in B) \\ &\leq \ \frac{\Pr(|Z - z_{n,k}| < s_m)}{\bar{p}} + \Pr(A^c_{n,s_m} \,|\, Z \in B) \\ &\leq \ \frac{\varepsilon_m}{\bar{p}} + \Pr(A^c_{n,s_m} \,|\, Z \in B) \ \stackrel{\Delta}{=} \ b_{n,m} \ . \end{aligned}$$

Substituting the above into (4.8), we obtain

$$H(Z_n) \;\geq\; \bar{p}\sum_{k=1}^{\infty} p(z_{n,k}\,|\,Z\in B)\log\frac{1}{b_{n,m}} \;=\; \bar{p}\log\frac{1}{b_{n,m}}\;.$$

We now observe that $\lim_{n\to\infty} b_{n,m} < \frac{2\varepsilon_m}{\bar{p}}$ because $\Pr(A^c_{n,s_m}\,|\,Z\in B)\to 0$ as $n\to\infty$. Therefore,

$$\lim_{n\to\infty} H(Z_n) \;\geq\; \lim_{n\to\infty}\bar{p}\log\frac{1}{b_{n,m}} \;>\; \bar{p}\log\frac{\bar{p}}{2\varepsilon_m} \;=\; \bar{p}\log\frac{m\bar{p}}{2}\;.$$

Finally, since m is arbitrary, it follows that $\lim_{n\to\infty} H(Z_n)=\infty$, as we needed to show. □

Proof of Theorem 3:

We claim the following three facts:

Fact A: There exists δ_o such that $\Pr\left(Z_{\delta_o}\in\left[\frac{\delta_o}{4},\frac{\delta_o}{2}\right]\right)>0$.

Fact B: For any $\alpha>0$, there exists $s>0$ such that $\Pr(t-s<Z_{\delta_o}<t+s)<\alpha$, for any t.

Fact C: For any $\varepsilon>0$, $\lim_{\tau\to 0}\Pr\left(|Z_{\delta_o}-\widehat{Z}^{\tau}|<\varepsilon\,|\,Z_{\delta_o}\in\left[\frac{\delta_o}{4},\frac{\delta_o}{2}\right]\right) \;=\; 1$.

Fact A follows from the fact that X has a crossing with positive probability, thus stationarity implies that with positive probability it has a crossing in the interval $[0,1]$. Consequently, $P(G)>0$ and applying Lemma 4 shows Fact A. Fact B is due to Lemma 5. Fact C can be derived as follows: From the definition of a crossing of size δ_o we have that if there is crossing from below (above) of size δ_o in $[\frac{\delta_o}{4},\frac{\delta_o}{2}]$, then all samples of X taken at times $t\in[0,\frac{\delta_o}{4})$ will lie below (above) u and all samples of X taken at times $t\in(\frac{\delta_o}{2},\delta_o]$ will lie above (below) u. Thus, it follows that for any $\varepsilon>0$, if the sampling interval is sufficiently small, specifically, if $\tau<\min\{\frac{\delta_o}{4},\varepsilon\}$, then $\Pr(|Z_{\delta_o}-\widehat{Z}^{\tau}|<\varepsilon\,|\,Z_{\delta_o}\in[\frac{\delta_o}{4},\frac{\delta_o}{2}]) \;=\; 1$, which implies Fact C.

Using these three facts, it now follows via Lemma 6, with $B = \left[\frac{\delta_o}{4}, \frac{\delta_o}{2}\right]$ in the lemma, that

$$H(I_1^\tau, I_2^\tau, \ldots, I_{N_\tau}^\tau) \;\geq\; H(K^\tau) \;\geq\; H(\widehat{Z}^\tau) \;\longrightarrow \infty \;\text{ as } \tau \longrightarrow 0 \;, \tag{4.9}$$

which concludes the proof of the theorem. □

4.3 Asymptotic formula for conditional entropy

The following is the main result of this section.

Theorem 7. *Let X_1 and X_2 be jointly Gaussian random variables with zero mean, variance σ^2, and correlation coefficient ρ that are quantized with an infinite-level uniform threshold quantizer with step size Δ, whose k^{th} cell is $[(k-\frac{1}{2})\Delta, (k+\frac{1}{2})\Delta)$. Let $\lambda = \frac{\Delta}{\sigma}$. If I_1 and I_2 denote the integers representing the quantization indices associated with X_1 and X_2, respectively, then*

$$\lim_{\rho \to 1} \frac{H(I_2|I_1)}{-M_\lambda \sqrt{1-\rho}\, \log \sqrt{1-\rho}} \;=\; 1 \;,$$

where

$$M_\lambda \;=\; \frac{2\sqrt{2}}{\pi} \sum_{k=0}^{\infty} e^{-\frac{(k+\frac{1}{2})^2 \lambda^2}{2}}$$

is a positive constant that depends on the ratio $\frac{\Delta}{\sigma}$.

Corollary 8.

$$\lim_{\lambda \to 0} \lim_{\rho \to 1} \frac{H(I_2|I_1)}{-\frac{2}{\sqrt{\pi}} \frac{1}{\lambda} \sqrt{1-\rho}\, \log \sqrt{1-\rho}} \;=\; 1 \;.$$

Letting $\lambda \to 0$ means the quantizers become high resolution. Intuitively this should increase the conditional entropy. Indeed, the corollary shows that asymptotically, as $\lambda \to 0$, the increase in conditional entropy is of the order of $\frac{1}{\lambda}$.

We comment that (4.1) uses Theorem 7 to upper bound the rate at which the joint entropy in Theorem 3 tends to infinity as $\tau \to 0$ in the case of infinite-level

uniform scalar quantizers and a stationary Gaussian process. Equations (4.2) and (4.3) provide explicit expressions for this upper bound in the case of exponential and Gaussian autocorrelation functions.

4.4 Notation

The following notation is used throughout the remainder of the Chapter. X_1 and X_2 denote jointly Gaussian random variables with zero mean and variance σ^2. Their correlation coefficient is denoted by ρ.

Let q denote the quantization rule of an infinite-level uniform threshold scalar quantizer with step size Δ such that $S_k \stackrel{\Delta}{=} [t_k, t_{k+1})$ is the k^{th} cell of the quantizer, where $t_k \stackrel{\Delta}{=} (k - \frac{1}{2})\Delta$ is the left threshold of the k^{th} cell. We let I_1 and I_2 denote the quantization indices representing the quantized values of X_1 and X_2, respectively, where $I_1 = k$ if X_1 lies in S_k, and similarly for I_2 and X_2. Let also $\lambda \stackrel{\Delta}{=} \frac{\Delta}{\sigma}$, and let $\sigma_\rho^2 \stackrel{\Delta}{=} \sigma^2(1-\rho^2)$ be the conditional variance of X_2 given X_1.

Let $P_k \stackrel{\Delta}{=} \Pr(I_i = k)$, where $i \in \{1, 2\}$, and $P_{l|k} \stackrel{\Delta}{=} \Pr(I_2 = l | I_1 = k)$. Observe that $P_{l|k}$ depends on ρ, but to keep notation short, we do not write this dependency explicitly. Let $\mathcal{H}$ denote the entropy function, i.e. $\mathcal{H}(..., z_{-1}, z_0, z_1, ..) = \sum_{k=-\infty}^{\infty} -z_k \log z_k$, where the z_k's are a finite or countably infinite set of nonnegative numbers that need not sum to one, and where $-0 \log 0$ is taken to be 0. All logarithms are base 2, unless otherwise stated. Let also $H_{l|k} = \mathcal{H}(P_{l|k})$. Thus, for example with this notation, $H(I_2|I_1 = k) = \sum_{l=-\infty}^{\infty} H_{l|k} = \sum_{l=-\infty}^{\infty} \mathcal{H}(P_{l|k})$, Let also $H_q(f) = \mathcal{H}(\ldots, P_{-1}(f), P_0(f), P_1(f), \ldots)$, where $P_i(f) = \int_{S_i} f(x)\, dx$. (Note that f need not be a probability density function (pdf), but it will be nonnegative).

We let $\mathcal{N}_{\mu,\sigma^2}(x)$ denote a Gaussian pdf with mean μ and variance σ^2 at point x. $Q(\cdot)$ denotes the standard "Q function", namely, $Q(\alpha) \stackrel{\Delta}{=} \int_\alpha^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$. We define

$g(\alpha) \stackrel{\Delta}{=} \frac{Q(\alpha)}{\frac{1}{2}e^{-\frac{\alpha^2}{2}}}$ for all $\alpha \geq 0$. We refer to g as the "correction factor" to the upper bound $\frac{1}{2}e^{-\frac{\alpha^2}{2}}$ of the Q function (see Fact 1 below).

The following is a list of elementary facts about the Q, g and $\mathcal{H}$ functions that will be useful throughout.

Fact 1: $Q(x) \leq \frac{1}{2}e^{-\frac{x^2}{2}}$, for any $x \geq 0$.

Fact 2: $Q(x) < \frac{1}{\sqrt{2\pi}x}e^{-\frac{x^2}{2}}$ for any $x > 0$.

Fact 3: $Q(x) > \frac{1}{\sqrt{2\pi}x}(1 - \frac{1}{x^2})e^{-\frac{x^2}{2}}$ for any $x > 0$.

Fact 4: $Q(x\lambda) - Q((x+1)\lambda) > \frac{1}{2\sqrt{2\pi}x\lambda}e^{-\frac{x^2\lambda^2}{2}}$ for all $x > \max\{\frac{2}{\lambda}, \frac{2}{\lambda^2}\}$, where $\lambda > 0$.

Fact 5: $\frac{Q((a+1)z)}{Q(az)} < 2e^{-\frac{z^2}{2}}$ for all $az > \frac{1}{2}$, where $a, z > 0$.

Fact 6: $g(0) = 1, g(\infty) = 0$ and $g(x)$ is a strictly decreasing function of x, for all $x \geq 0$.

Fact 7: $-p\log p$ is concave and attains its maximum at $p = \frac{1}{e}$.

Fact 8: $p < p' < \frac{1}{e}$ implies $-p\log p < -p'\log p'$, and $p > p' > \frac{1}{e}$ implies $-p\log p < -p'\log p'$.

Fact 9: For any $\{a_k\} \in \mathbb{R}^+$, $\mathcal{H}(\sum_k a_k) < \sum_k \mathcal{H}(a_k)$.

Facts 1, 2 and 3 are shown in [19] (pp. 82-83); Facts 4 and 5 can be shown straightforwardly using Facts 2 and 3. The first two parts of Fact 6 are immediate, and the monotonicity part follows from having $g'(x) < 0$ for all $x \geq 0$, which can be shown using Fact 2; Fact 7 is well-known; Fact 8 is a direct consequence of Fact 7; and Fact 9 is a result of the concavity of $\mathcal{H}$ and can also be seen by showing that $-\sum_k a_k \log a_k + \sum_k (a_k \log \sum_k a_k) > 0$, or equivalently, $\sum_k a_k \log \frac{\sum_k a_k}{a_k} > 0$, which is clearly true.

4.5 Proofs of Theorem 7 and Corollary 8

Proof of Theorem 7:

First let us observe that finding a conditional entropy of the form $H(I_2|X_1)$ is not trivial. Since the conditional distribution of X_2 given X_1 is Gaussian, we have, in fact, solved such a problem in Chapter III, where the output entropy of a uniform quantizer with a Gaussian source has been evaluated. For the case examined here, namely, finding the conditional entropy $H(I_2|I_1)$, the situation worsens significantly in terms of derivation difficulty. The reason for this stems from the fact that I_1 is the quantized version of X_1, which in turn makes the conditional distribution of X_2 given $I_1 = k$ no longer be Gaussian. Consequently, finding the conditional entropy of I_2 given I_1 becomes quite difficult.

The key to finding this conditional entropy lies in understanding the behavior of the conditional distribution of X_2 given that $I_1 = k$. To do so, we first write an expression for $f_{X_2|I_1}(x|k)$. (Note that in the sequel we will use an alternate expression, which will in all cases but Lemma 20, be more useful.)

$$\begin{aligned} f_{X_2|I_1}(x|k) &= \int_{t_k}^{t_{k+1}} f_{X_2|X_1}(x|y)\frac{f_{X_1}(y)}{P_k}\,dy \\ &= \frac{1}{P_k}\int_{t_k}^{t_{k+1}} \frac{1}{\sqrt{2\pi}\sigma_\rho}\, e^{-\frac{(x-\rho y)^2}{2\sigma_\rho^2}} \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{y^2}{2\sigma^2}}\,dy\ . \end{aligned} \tag{4.10}$$

As can be seen from the equation above the conditional density of X_2 given that X_1 lies in the k^{th} cell is a weighted average of conditional densities of X_2 given that X_1 equals particular values in S_k, where the weighing function is Gaussian. The latter conditional densities are Gaussian with mean ρ times the value of X_1, and variance σ_ρ^2, i.e. when ρ is close to one, these conditional densities are very narrow Gaussians, whose mean is close to the value of X_1. We observe that since the weighing function is Gaussian, if k is large, then X_1 values that lie in the beginning of S_k, i.e. closer to the

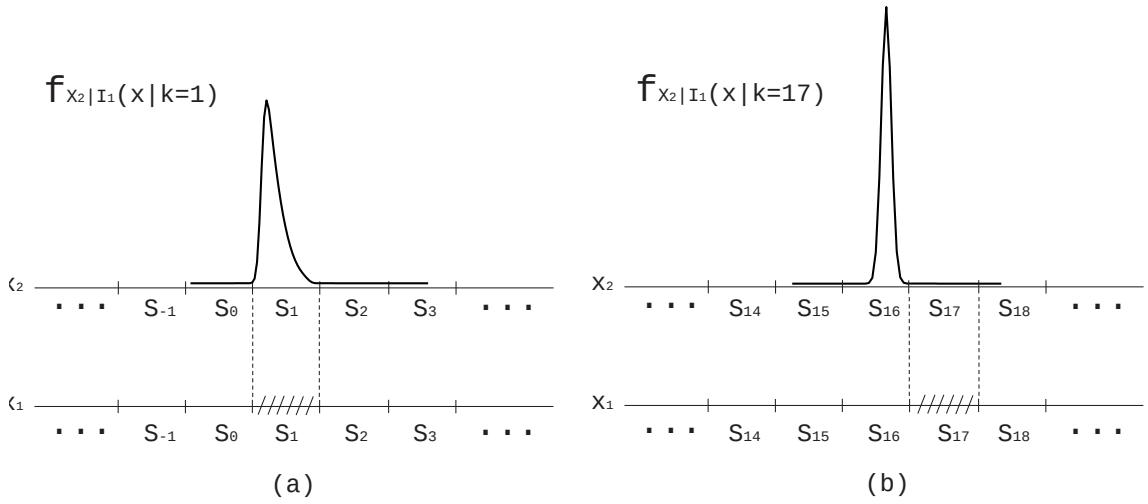

Figure 4.2: The conditional pdf of X_2 given that X_1 lies in the k^{th} quantization cell. The used parameters are $\Delta = 2$, $\sigma = 1$ and $\rho = 0.99$ (a) $k = 1$. (b) $k = 17$.

origin, are more significant in determining $f_{X_2|I_1}(x|k)$. Consequently, the larger the value of k, the narrower $f_{X_2|I_1}(x|k)$ becomes. This, among other things, is illustrated in Figure 4.2.

Next, let us consider in more detail the influence of k, i.e. the quantization cell in which X_1 lies, on $f_{X_2|I_1}(x|k)$. Suppose that ρ is very close to one, thus, X_2 and X_1 are very correlated, which means that if X_1 lies in quantization cell k, we would expect that with high probability X_2 would lie in the same cell. This intuitive notion is illustrated in Figure 4.2a, where $\rho = 0.99$ and $k = 1$. However, we observe that if k is is sufficiently large, then a seemingly strange thing happens, namely, X_2 lies with high probability in a different cell than the one in which X_1 lies. Figure 4.2b illustrates this phenomenon for the case that $\rho = 0.99$ and $k = 17$, for which X_2 is lies with high probability in cell 16 rather than 17. While this may indeed seem somewhat bizarre at first glance, the reason for this shifting towards the origin phenomenon is quite simple. Specifically, if $X_1 = x_1$, then $EX_2 = \rho x_1$. Thus, no matter how close

to one ρ may be, the distance between the value of X_1, i.e. x_1, and the mean of X_2, i.e. ρx_1, equals $(1-\rho)x_1$, which can be made arbitrarily large, (i.e. it can equal the length of many quantization cells) by letting x_1 be sufficiently large. The take away message from this discussion should be that for any value of ρ, if k is sufficiently small, then X_2 lies in the k^{th} cell with high probability, and if k is too large, then this is no longer true.

Due to this behavior of the conditional distribution of X_2 given $I_1 = k$, the main idea of the proof is to first approximate $H(I_2|I_1) = \sum_{k=-\infty}^{\infty} H(I_2|I_1 = k)P_k$, by a truncated sum that only considers sufficiently small k values. Once such an approximation is established, each term $H(I_2|I_1 = k)$ can be dealt with more easily, since X_2 would be guaranteed to lie in the same cell as X_1 with high probability. The second step would then be to approximate $H(I_2|I_1 = k) = \sum_{l=\infty}^{\infty} \mathcal{H}(P_{l|k})$ by a truncated sum, as well. The details are given below.

The proof is composed of five main steps, each showing that one term on the right-hand side of the following equation converges to one as $\rho \to 1$:

$$\begin{aligned}\frac{H(I_2|I_1)}{M_\lambda \mathcal{H}(\sqrt{1-\rho}\,)} &= \frac{H(I_2|I_1)}{\sum_{|k|\le N(\rho)} H(I_2|I_1=k)P_k} \times \frac{\sum_{|k|\le N(\rho)} H(I_2|I_1=k)P_k}{\sum_{|k|\le N(\rho)} \big(H_{k-1|k} + H_{k|k} + H_{k+1|k}\big)P_k} \\ &\times \frac{\sum_{|k|\le N(\rho)} \big(H_{k-1|k} + H_{k|k} + H_{k+1|k}\big)P_k}{\sum_{|k|\le N(\rho)} \big(H_{k-1|k} + H_{k+1|k}\big)P_k} \\ &\times \frac{\sum_{|k|\le N(\rho)} \big(H_{k-1|k} + H_{k+1|k}\big)P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)\sum_{|k|\le N(\rho)} \big(M_{L,\lambda}(k) + M_{R,\lambda}(k)\big)} \\ &\times \frac{\mathcal{H}(\sqrt{1-\rho^2}\,)\sum_{|k|\le N(\rho)} \big(M_{L,\lambda}(k) + M_{R,\lambda}(k)\big)}{M_\lambda \mathcal{H}(\sqrt{1-\rho}\,)}\,, \end{aligned} \tag{4.11}$$

where $M_{L,\lambda}(k) = \frac{1}{2\pi} e^{-\frac{(k-\frac{1}{2})^2\lambda^2}{2}}$, $M_{R,\lambda}(k) = \frac{1}{2\pi} e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}}$, and where $N(\rho)$ is a carefully chosen integer function of ρ that goes to infinity as $\rho \to 1$. Specifically, the idea is to choose $N(\rho)$ so that on the one hand it grows slowly enough so that for $|k| \le N(\rho)$, $f_{X_2|I_1}(x|k)$ will be so concentrated on S_k that $P_{k|k} \approx 1$ and $H(I_2|I_1 = k) \approx H_{k-1|k} +$

$H_{k|k} + H_{k+1|k}$, while on the other hand it grows fast enough that the first term on the right hand side above tends to one as $\rho \to 1$. A choice of $N(\rho)$ that satisfies these two competing requirements is

$$N(\rho) \;=\; \Big\lfloor \big(\ln \frac{1}{1-\rho} \big)^{\frac{3}{4}} - \frac{1}{2} \Big\rfloor \,.$$

As mentioned above, our goal is to show that each of the five terms in (4.11) goes to one as $\rho \to 1$. To simplify matters slightly, we use the symmetry of the Gaussian pdf and the uniform quantizers to focus on nonnegative k's. Thus, we rewrite (4.11) as follows:

$$\begin{aligned}
\frac{H(I_2|I_1)}{M_\lambda \mathcal{H}(\sqrt{1-\rho}\,)} \;=\; & \frac{H(I_2|I_1)}{2\sum_{k=1}^{N(\rho)} H(I_2|I_1=k)P_k + H(I_2|I_1=0)P_0} \\
& \times \frac{2\sum_{k=1}^{N(\rho)} H(I_2|I_1=k)P_k + H(I_2|I_1=0)P_0}{2\sum_{k=1}^{N(\rho)} \big(H_{k-1|k} + H_{k|k} + H_{k+1|k}\big)P_k + \big(H_{-1|0} + H_{0|0} + H_{1|0}\big)P_0} \\
& \times \frac{2\sum_{k=1}^{N(\rho)} \big(H_{k-1|k} + H_{k|k} + H_{k+1|k}\big)P_k + \big(H_{0|0} + H_{0|0} + H_{1|0}\big)P_0}{2\sum_{k=1}^{N(\rho)} \big(H_{k-1|k} + H_{k+1|k}\big)P_k + \big(H_{-1|0} + H_{1|0}\big)P_0} \\
& \times \frac{2\sum_{k=1}^{N(\rho)} \big(H_{k-1|k} + H_{k+1|k}\big)P_k + \big(H_{-1|0} + H_{1|0}\big)P_0}{\mathcal{H}(\sqrt{1-\rho^2}\,)\Big[2\sum_{k=1}^{N(\rho)} \big(M_{L,\lambda}(k) + M_{R,\lambda}(k)\big) + M_{L,\lambda}(0) + M_{R,\lambda}(0)\Big]} \\
& \times \frac{\mathcal{H}(\sqrt{1-\rho^2}\,)\Big[2\sum_{k=1}^{N(\rho)} \big(M_{L,\lambda}(k) + M_{R,\lambda}(k)\big) + M_{L,\lambda}(0) + M_{R,\lambda}(0)\Big]}{M_\lambda \mathcal{H}(\sqrt{1-\rho}\,)} \,.
\end{aligned} \tag{4.12}$$

Before proceeding with the proof, we investigate and establish some properties of $f_{X_2|I_1}(x|k)$, $P_{l|k}$ and other important quantities that we will need. Lemmas will be provided as needed; their proofs will be provided in Section 4.6. We begin by finding an expression for the conditional pdf:

$$\begin{aligned}
f_{X_2|I_1}(x|k) \;&=\; \frac{\Pr(I_1 = k \,|\, X_2 = x) f_{X_2}(x)}{\Pr(I_1 = k)} \\
&=\; \frac{1}{P_k} f_{X_2}(x) \left[Q\big(\frac{t_k - \rho x}{\sigma_\rho}\big) - Q\big(\frac{t_{k+1} - \rho x}{\sigma_\rho}\big) \right] \,.
\end{aligned} \tag{4.13}$$

For tractability we would like to drop the smaller of the two Q function terms in the above expression, which leads us to define $\tilde{f}_{X_2|I_1}(x|k)$ as follows:

$$\tilde{f}_{X_2|I_1}(x|k) \triangleq \begin{cases} \frac{1}{P_k} f_{X_2}(x) Q\big(\frac{t_k - \rho x}{\sigma_\rho}\big), & x < \frac{t_k}{\rho} \\ \frac{1}{P_k} f_{X_2}(x) \Big[Q\big(\frac{t_k - \rho x}{\sigma_\rho}\big) - Q\big(\frac{t_{k+1} - \rho x}{\sigma_\rho}\big) \Big], & \frac{t_k}{\rho} \le x \le \frac{t_{k+1}}{\rho} \\ \frac{1}{P_k} f_{X_2}(x) Q\big(\frac{\rho x - t_{k+1}}{\sigma_\rho}\big), & x > \frac{t_{k+1}}{\rho} \end{cases} , \tag{4.14}$$

It is easy to see that

$$f_{X_2|I_1}(x|k) \le \tilde{f}_{X_2|I_1}(x|k) \quad \text{for all } x \text{ and } k \, .$$

The following lemma shows that $\tilde{f}_{X_2|I_1}(x|k)$ is an asymptotically tight upper bound to $f_{X_2|I_1}(x|k)$.

Lemma 9.

$$\frac{f_{X_2|I_1}(x|k)}{\tilde{f}_{X_2|I_1}(x|k)} \longrightarrow 1 \;\; \textit{as } \rho \longrightarrow 1, \; \textit{uniformly in } k \textit{ and } x.$$

and consequently, for any $0 \le \gamma < 1$*, there exists* $\rho_\gamma < 1$ *such that for all* $\rho > \rho_\gamma$*,* $f_{X_2|I_1}(x|k) \ge \gamma \tilde{f}_{X_2|I_1}(x|k)$ *for all* k *and* x*.*

We will find it useful to use the following alternate representation of $\tilde{f}_{X_2|I_1}(x|k)$:

$$\tilde{f}_{X_2|I_1}(x|k) = \begin{cases} L_k(\rho) \mathcal{N}_{t_k\rho, \sigma_\rho^2}(x) g\big(\frac{t_k - \rho x}{\sigma_\rho}\big), & x < \frac{t_k}{\rho} \\ \frac{\mathcal{N}_{0,\sigma^2}(x)}{P_k} - L_k(\rho) \mathcal{N}_{t_k\rho, \sigma_\rho^2}(x) g\big(\frac{\rho x - t_k}{\sigma_\rho}\big) & \\ \qquad - R_k(\rho) \mathcal{N}_{t_{k+1}\rho, \sigma_\rho^2}(x) g\big(\frac{t_{k+1} - \rho x}{\sigma_\rho}\big), & \frac{t_k}{\rho} \le x \le \frac{t_{k+1}}{\rho} \\ R_k(\rho) \mathcal{N}_{t_{k+1}\rho, \sigma_\rho^2}(x) g\big(\frac{\rho x - t_{k+1}}{\sigma_\rho}\big), & x > \frac{t_{k+1}}{\rho} \end{cases} , \tag{4.15}$$

where

$$L_k(\rho) \triangleq \frac{1}{2} \frac{1}{P_k} e^{-\frac{t_k^2}{2\sigma^2}} \sqrt{1 - \rho^2} \qquad \text{and} \qquad R_k(\rho) \triangleq \frac{1}{2} \frac{1}{P_k} e^{-\frac{t_{k+1}^2}{2\sigma^2}} \sqrt{1 - \rho^2} \, .$$

(4.15) is obtained by manipulating exponentials. For example, for the region $x < \frac{t_k}{\rho}$,

$$\begin{aligned}
\frac{1}{P_k} f_{X_2}(x) Q\big(\frac{t_k - \rho x}{\sigma_\rho}\big) &= \frac{1}{P_k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{2} e^{-\frac{(t_k-\rho x)^2}{2\sigma_\rho^2}} g\big(\frac{t_k - \rho x}{\sigma_\rho}\big) \\
&= \frac{1}{2}\frac{1}{P_k}\sqrt{1-\rho^2}\,\frac{1}{\sqrt{2\pi\sigma^2(1-\rho^2)}}\, e^{-\frac{x^2-\rho^2x^2+t_k^2-2t_k\rho x+\rho^2x^2}{2\sigma^2(1-\rho^2)}} g\big(\frac{t_k - \rho x}{\sigma_\rho}\big) \\
&= \frac{1}{2}\frac{1}{P_k}\sqrt{1-\rho^2}\,\frac{1}{\sqrt{2\pi}\sigma_\rho}\, e^{-\frac{x^2-2t_k\rho x+t_k^2\rho^2}{2\sigma^2(1-\rho^2)}} e^{-\frac{t_k^2-t_k^2\rho^2}{2\sigma^2(1-\rho^2)}} g\big(\frac{t_k - \rho x}{\sigma_\rho}\big) \\
&= \frac{1}{2}\frac{1}{P_k} e^{-\frac{t_k^2}{2\sigma^2}} \sqrt{1-\rho^2}\,\frac{1}{\sqrt{2\pi}\sigma_\rho}\, e^{-\frac{(x-t_k\rho)^2}{2\sigma_\rho^2}} g\big(\frac{t_k - \rho x}{\sigma_\rho}\big) \\
&= L_k(\rho)\mathcal{N}_{t_k\rho,\sigma_\rho^2}(x)\, g\big(\frac{t_k - \rho x}{\sigma_\rho}\big)\ .
\end{aligned}$$

Similar algebraic steps can be used for the regions $\frac{t_k}{\rho} \le x \le \frac{t_{k+1}}{\rho}$ and $x > \frac{t_{k+1}}{\rho}$. The representation of $\tilde{f}_{X_2|I_1}(x|k)$ in (4.15) will be useful since it will turn out that $P_{k-1|k}$ and $P_{k+1|k}$ converge to multiples of $L_k(\rho)$ and $R_k(\rho)$, respectively. This motivates us to examine these quantities. Specifically, the following lemma provides upper and lower bounds to $L_k(\rho)$ and $R_k(\rho)$.

Lemma 10. *For all ρ the following is true,*

$$\begin{aligned}
&A.\quad L_k(\rho) < \sqrt{1-\rho^2}\,\sqrt{2\pi}k\lambda \qquad \textit{for}\quad k > N_\lambda\ , \\
&B.\quad L_k(\rho) > \sqrt{1-\rho^2} \qquad \textit{for}\quad k \ge 1\ , \\
&C.\quad R_k(\rho) < \sqrt{1-\rho^2}\,\sqrt{\frac{\pi}{2}}\frac{1}{\lambda} \qquad \textit{for}\quad k \ge 0\ , \\
&D.\quad R_k(\rho) > \sqrt{1-\rho^2}\, e^{-k\lambda^2} \qquad \textit{for}\quad k \ge 1\ ,
\end{aligned}$$

where $N_\lambda = \frac{1}{2} + \max\{\frac{2}{\lambda}, \frac{1}{\lambda^2}\}$.

Next, when considering the second, third and fourth terms on the right-hand side of (4.12), $P_{k-1|k}$ and $P_{k+1|k}$ for $0 \le k \le N(\rho)$, will play a key role; hence we shall find expressions for these. (It will turn out that finding an expression for $P_{k|k}$ can be avoided.) From Lemma 9 it will follow that it will be enough to

use $\tilde{f}_{X_2|I_1}(x|k)$ rather than $f_{X_2|I_1}(x|k)$, i.e. $\int_{t_{k+1}}^{t_{k+2}} \tilde{f}_{X_2|I_1}(x|k)\,dx$ will be shown to be a sufficiently good approximation to $P_{k+1|k}$ as $\rho \to 1$, and similarly it will follow that $\int_{t_{k-1}}^{t_k} \tilde{f}_{X_2|I_1}(x|k)\,dx$ is a sufficiently good approximation to $P_{k-1|k}$ as $\rho \to 1$. However, since we are considering $k \geq 0$, one can see from (4.14) and (4.15) that the expression for $\tilde{f}_{X_2|I_1}(x|k)$ in the interval $[t_{k+1}, \frac{t_{k+1}}{\rho}]$ is rather complicated. Therefore, for analytical tractability we define

$$P^*_{k+1|k} = \int_{\frac{t_{k+1}}{\rho}}^{t_{k+2}} f_{X_2|I_1}(x|k)\,dx\ .$$

From its definition, it is easy to see that $P^*_{k+1|k} < P_{k+1|k}$ for all $k \geq 0$. We will show, however, that $P^*_{k+1|k}$ is a sufficiently good approximation of $P_{k+1|k}$, and consequently we will be able to use $P^*_{k+1|k}$ instead of $P_{k+1|k}$.

We observe that since $N(\rho)$ grows as $\rho \to 1$, the number of summands in the numerators and denominators of the second and third terms in (4.12) grow as well. Consequently, when making convergence statements regarding $P_{k-1|k}$, $P_{k+1|k}$ and $P^*_{k+1|k}$, we will require a kind of uniform convergence. To this end we define the following limit notation, which captures the uniformity we require.

Definition 11. *Let* $\lim^*_{\rho\to 1} f(k,\rho) = c$ *mean that* $\lim_{\rho\to 1} \sup_{0\leq k\leq N(\rho)} |f(k,\rho) - c| = 0$, *where* $f(k,\rho)$ *is a function of* k *and* ρ, *and* c *is some constant. In other words,* $\lim^*_{\rho\to 1} f(k,\rho) = c$ *if and only if* $\lim_{\rho\to 1} \sup_{0\leq k\leq N(\rho)} f(k,\rho) = c$ *and* $\lim_{\rho\to 1} \inf_{0\leq k\leq N(\rho)} f(k,\rho) = c$.

The following lemma shows that certain properties of standard limits hold for $\lim^*$, as well.

Lemma 12.

A. *(Dominated Convergence Theorem) If* $\sup_{0\le k\le N(\rho)} |f(k,\rho,x)| \le G(x)$ *a.e. for some integrable function* G*, and* $\lim^*_{\rho\to 1} f(k,\rho,x)$ *exists a.e., then* $\lim^*_{\rho\to 1} \int f(k,\rho,x)\,dx = \int \lim^*_{\rho\to 1} f(k,\rho,x)\,dx$.

B. *If* $\lim^*_{\rho\to 1} a_k(\rho)$ *and* $\lim^*_{\rho\to 1} b_k(\rho)$ *exist, then* $\lim^*_{\rho\to 1} \big(a_k(\rho) + b_k(\rho)\big) = \lim^*_{\rho\to 1} a_k(\rho) + \lim^*_{\rho\to 1} b_k(\rho)$.

C. *If* $\lim^*_{\rho\to 1} a_k(\rho)$ *and* $\lim^*_{\rho\to 1} b_k(\rho)$ *exist, then* $\lim^*_{\rho\to 1} a_k(\rho)\, b_k(\rho) = \lim^*_{\rho\to 1} a_k(\rho) \lim^*_{\rho\to 1} b_k(\rho)$.

D. *Let* $b_k(\rho)$ *be positive for all* ρ *and* $0 \le k \le N(\rho)$*. If* $\lim^*_{\rho\to 1} \frac{a_k(\rho)}{b_k(\rho)} = 1$*, then* $\lim_{\rho\to 1} \frac{\sum_{k=0}^{N(\rho)} a_k(\rho)}{\sum_{k=0}^{N(\rho)} b_k(\rho)} = 1$.

E. *Let* $c_k(\rho)$ *and* $d_k(\rho)$ *be positive for all* ρ *and* $0 \le k \le N(\rho)$*. If* $\lim^*_{\rho\to 1} \frac{a_k(\rho)}{c_k(\rho)} = 1$ *and* $\lim^*_{\rho\to 1} \frac{b_k(\rho)}{d_k(\rho)} = 1$*, then* $\lim^*_{\rho\to 1} \frac{a_k(\rho)+b_k(\rho)}{c_k(\rho)+d_k(\rho)} = 1$.

F. *If* $\lim_{z\to z_o} G(z) = c$ *and* $\lim^*_{\rho\to 1} a_k(\rho) = z_o$*, then* $\lim^*_{\rho\to 1} G\big(a_k(\rho)\big) = c$.

G. *If* $\lim_{\frac{x}{y}\to z} G(x,y) = c$*,* $\lim^*_{\rho\to 1} \frac{a_k(\rho)}{b_k(\rho)} = z$*, then* $\lim^*_{\rho\to 1} G\big(a_k(\rho), b_k(\rho)\big) = c$.

Next we provide five additional lemmas. Lemma 13 links $L_k(\rho)$ and $R_k(\rho)$ to $P_{k-1|k}$ and $P^*_{k+1|k}$, respectively, which will illuminate the importance of $L_k(\rho)$ and $R_k(\rho)$ and, hence, the reason for decomposing $\tilde{f}_{X_2|I_1}(x|k)$ according to (4.15). Lemma 14 shows that $P^*_{k+1|k}$ is a sufficiently good approximation of $P_{k+1|k}$, Lemma 15 shows that $P_{k-1|k}$ and $P_{k+1|k}$ converge to zero in the $\lim^*$ sense, Lemma 16 shows how rapidly $P_l(f)$ decays, where f is a Gaussian density, and Lemma 17 shows how rapidly $P_{k+l|k}$ and $P_{k-l|k}$ decay in terms of $P^*_{k+1|k}$ for $0 \le k \le N(\rho)$ and $l \ge 2$.

Lemma 13.

$$A. \quad \lim_{\rho\to 1}{}^{*} \frac{P_{k-1|k}}{\frac{1}{\pi}L_k(\rho)} \;=\; 1 \;,$$

$$B. \quad \lim_{\rho\to 1}{}^{*} \frac{P^{*}_{k+1|k}}{\frac{1}{\pi}R_k(\rho)} \;=\; 1 \;,$$

and consequently,

C. for all ρ sufficiently close to one, $P^{*}_{k+1|k} < \frac{1}{2}R_k(\rho) < \frac{\sqrt{2\pi}}{4\lambda}\sqrt{1-\rho^2}$ for $0 \le k \le N(\rho)$,

D. for ρ sufficiently close to one, $P^{*}_{k+1|k} > \frac{1}{5}R_k(\rho)$ for $0 \le k \le N(\rho)$,

where the second inequality in C follows from Lemma 10 (part C).

Lemma 14.

$$\lim_{\rho\to 1}{}^{*} \frac{P_{k+1|k}}{P^{*}_{k+1|k}} \;=\; 1 \;.$$

Lemma 15.

$$A. \quad \lim_{\rho\to 1}{}^{*} P_{k-1|k} \;=\; 0 \;,$$

$$B. \quad \lim_{\rho\to 1}{}^{*} P_{k+1|k} \;=\; 0 \;.$$

Lemma 16. *Let $f = \mathcal{N}_{\mu,\sigma^2}$ be a Gaussian density, and let j be such that $\mu \in S_j$. Then*

$$A. \quad P_{j+l+1}(f) \;<\; \frac{1}{2}P_{j+l}(f) \qquad \textit{for} \quad l \;\ge\; B+2 \;,$$

$$B. \quad P_{j-l-1}(f) \;<\; \frac{1}{2}P_{j-l}(f) \qquad \textit{for} \quad l \;\ge\; B \;,$$

where $B \stackrel{\Delta}{=} \lceil \frac{\sigma^2}{\Delta^2}\ln 3 \rceil$.

Lemma 17. *For all ρ sufficiently close to one the following holds for $0 \le k \le N(\rho)$,*

$$A. \quad P_{k+l|k} < (P^{*}_{k+1|k})^{l} \qquad \textit{for} \quad l \;\ge\; 2 \;,$$

$$B. \quad P_{k-l|k} < (P^{*}_{k+1|k})^{l} \qquad \textit{for} \quad l \;\ge\; 2 \;.$$

Now that we have established basic properties of $P_{k-1|k}$, $P_{k+1|k}$ and $P^*_{k+1|k}$, and the rate of decay of certain conditional probabilities for arbitrary k values, we are ready to proceed with the main part of the proof and show that each of the five terms in (4.12) converges to one as $\rho \to 1$. We consider the terms in the following order: second, first, third, fourth and fifth.

The following lemma shows the second term converges to one as $\rho \to 1$.

Lemma 18.

$$\lim_{\rho \to 1} \frac{2 \sum_{k=1}^{N(\rho)} H(I_2|I_1 = k) P_k + H(I_2|I_1 = 0) P_0}{2 \sum_{k=1}^{N(\rho)} \big(H_{k-1|k} + H_{k|k} + H_{k+1|k}\big) P_k + \big(H_{-1|0} + H_{0|0} + H_{1|0}\big) P_0} \;=\; 1 \;.$$

The following two lemmas are needed in order to show that the first term in (4.12) converges to one.

Lemma 19. *Let $f = \mathcal{N}_{\mu,\sigma^2}$ be a Gaussian density. Then,*

$$H_q(f) \;<\; 2 \left\lceil \frac{\sigma^2}{\Delta^2} \ln 3 \right\rceil + 10 \;.$$

Lemma 20. *There exists a constant M such that for all ρ and k,*

$$H(I_2|I_1 = k) \;<\; M \;.$$

The next lemma shows that the first term in (4.12) converges to one as $\rho \to 1$.

Lemma 21.

$$\lim_{\rho \to 1} \frac{H(I_2|I_1)}{2 \sum_{k=1}^{N(\rho)} H(I_2|I_1 = k) P_k + H(I_2|I_1 = 0) P_0} \;=\; 1 \;.$$

Finally, the next three lemmas show that the third, fourth and fifth terms, respectively, converge to one, and conclude the proof of the theorem.

Lemma 22.

$$\lim_{\rho \to 1} \frac{2 \sum_{k=1}^{N(\rho)} \big(H_{k-1|k} + H_{k|k} + H_{k+1|k}\big) P_k + \big(H_{0|0} + H_{0|0} + H_{1|0}\big) P_0}{2 \sum_{k=1}^{N(\rho)} \big(H_{k-1|k} + H_{k+1|k}\big) P_k + \big(H_{-1|0} + H_{1|0}\big) P_0} \;=\; 1 \;.$$

Lemma 23.

$$\lim_{\rho\to 1} \frac{2\sum_{k=1}^{N(\rho)} \big(H_{k-1|k} + H_{k+1|k}\big)P_k + \big(H_{-1|0} + H_{1|0}\big)P_0}{\mathcal{H}(\sqrt{1-\rho^2}\,)\Big[2\sum_{k=1}^{N(\rho)} \big(M_{L,\lambda}(k) + M_{R,\lambda}(k)\big) + \big(M_{L,\lambda}(0) + M_{R,\lambda}(0)\big)\Big]} \;=\; 1\;.$$

Lemma 24.

$$\lim_{\rho\to 1} \frac{\mathcal{H}(\sqrt{1-\rho^2}\,)\Big[2\sum_{k=1}^{N(\rho)} \big(M_{L,\lambda}(k) + M_{R,\lambda}(k)\big) + \big(M_{L,\lambda}(0) + M_{R,\lambda}(0)\big)\Big]}{M_\lambda \mathcal{H}(\sqrt{1-\rho}\,)} \;=\; 1\;.$$

□

Proof of Corollary 8:

From Theorem 7 $\lim_{\rho\to 1} \frac{H(I_2|I_1)}{-M_\lambda\sqrt{1-\rho}\,\log\sqrt{1-\rho}} = 1$, where $M_\lambda = \frac{2\sqrt{2}}{\pi}\sum_{k=0}^{\infty} e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}}$. What needs to be shown is that $\lim_{\lambda\to 0} \frac{M_\lambda}{\frac{2}{\sqrt{\pi}}\frac{1}{\lambda}} = 1$. We have

$$\sum_{k=0}^{\infty} e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}} \;=\; \sqrt{2\pi}\frac{1}{\lambda}\sum_{k=0}^{\infty}\frac{\lambda}{\sqrt{2\pi}}e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}} \;>\; \sqrt{2\pi}\frac{1}{\lambda}\int_{\frac{1}{2}}^{\infty}\mathcal{N}_{0,\frac{1}{\lambda^2}}(x)\,dx \;=\; \sqrt{2\pi}\frac{1}{\lambda}Q\big(\frac{\lambda}{2}\big)\,.$$

Similarly, we have

$$\sum_{k=0}^{\infty} e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}} \;=\; \sqrt{2\pi}\frac{1}{\lambda}\sum_{k=0}^{\infty}\frac{\lambda}{\sqrt{2\pi}}e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}} \;<\; \sqrt{2\pi}\frac{1}{\lambda}\int_{-\frac{1}{2}}^{\infty}\mathcal{N}_{0,\frac{1}{\lambda^2}}(x)\,dx \;=\; \sqrt{2\pi}\frac{1}{\lambda}Q\big(-\frac{\lambda}{2}\big)\,.$$

Since $\lim_{\lambda\to 0} Q\big(\frac{\lambda}{2}\big) = \lim_{\lambda\to 0} Q\big(-\frac{\lambda}{2}\big) = \frac{1}{2}$, it follows that $\lim_{\lambda\to 0} \frac{M_\lambda}{\frac{2}{\sqrt{\pi}}\frac{1}{\lambda}} = 1$, concluding the proof of the corollary. □

4.6 Lemma Proofs

Proof of Lemma 9:

We notice from (4.14) that there are three regions to consider: $x < \frac{t_k}{\rho}$, $\frac{t_k}{\rho} \le x \le \frac{t_{k+1}}{\rho}$ and $x > \frac{t_{k+1}}{\rho}$. Observe that for any k, $\tilde{f}_{X_2|I_1}(x|k) = f_{X_2|I_1}(x|k)$, for $\frac{t_k}{\rho} \le x \le \frac{t_{k+1}}{\rho}$. Thus, the lemma holds trivially in this region. Next, we will show that the lemma holds for $x < \frac{t_k}{\rho}$. The result for $x > \frac{t_{k+1}}{\rho}$ can be shown in a similar way.

$$\frac{f_{X_2|I_1}(x|k)}{\tilde{f}_{X_2|I_1}(x|k)} \;=\; \frac{\frac{1}{P_k}f_{X_2}(x)\left[Q\big(\frac{t_k-\rho x}{\sigma_\rho}\big) - Q\big(\frac{t_{k+1}-\rho x}{\sigma_\rho}\big)\right]}{\frac{1}{P_k}f_{X_2}(x)\,Q\big(\frac{t_k-\rho x}{\sigma_\rho}\big)} \;=\; 1 - \frac{Q\big(\frac{t_{k+1}-\rho x}{\sigma_\rho}\big)}{Q\big(\frac{t_k-\rho x}{\sigma_\rho}\big)}\,.$$

We now focus on showing that the fraction on the right-hand side above approaches zero. Specifically,

$$\frac{Q\left(\frac{t_{k+1}-\rho x}{\sigma_\rho}\right)}{Q\left(\frac{t_k-\rho x}{\sigma_\rho}\right)} = \frac{\frac{1}{2}e^{-\frac{(t_{k+1}-\rho x)^2}{2\sigma_\rho^2}} g\left(\frac{t_{k+1}-\rho x}{\sigma_\rho}\right)}{\frac{1}{2}e^{-\frac{(t_k-\rho x)^2}{2\sigma_\rho^2}} g\left(\frac{t_k-\rho x}{\sigma_\rho}\right)} < \frac{e^{-\frac{(t_{k+1}-\rho x)^2}{2\sigma_\rho^2}}}{e^{-\frac{(t_k-\rho x)^2}{2\sigma_\rho^2}}} = e^{\frac{t_k^2-t_{k+1}^2+2(t_{k+1}-t_k)\rho x}{2\sigma_\rho^2}}$$
$$< e^{\frac{t_k^2-t_{k+1}^2+2(t_{k+1}-t_k)t_k}{2\sigma_\rho^2}} = e^{-\frac{(t_{k+1}-t_k)^2}{2\sigma_\rho^2}} = e^{-\frac{\Delta^2}{2\sigma_\rho^2}} = e^{-\frac{\lambda^2}{2(1-\rho^2)}} ,$$

where the first inequality follows from $\frac{t_{k+1}-\rho x}{\sigma_\rho} > \frac{t_k-\rho x}{\sigma_\rho}$ and the fact that g is monotonically decreasing (Fact 6). The second inequality follows from the facts that $t_{k+1} - t_k > 0$ and $x < \frac{t_k}{\rho}$. Finally, it follows that

$$\lim_{\rho\to 1} \sup_{k, x<\frac{t_k}{\rho}} \left| \frac{Q\left(\frac{t_{k+1}-\rho x}{\sigma_\rho}\right)}{Q\left(\frac{t_k-\rho x}{\sigma_\rho}\right)} \right| < \lim_{\rho\to 1} \sup_{k, x<\frac{t_k}{\rho}} \left| e^{-\frac{\lambda^2}{2(1-\rho^2)}} \right| = 0 ,$$

which concludes the proof of the lemma. □

Proof of Lemma 10:

We begin by showing Part A.

$$\frac{L_k(\rho)}{\sqrt{1-\rho^2}} = \frac{1}{2}\frac{1}{Q\left((k-\frac{1}{2})\lambda\right) - Q\left((k+\frac{1}{2})\lambda\right)} e^{-\frac{t_k^2}{2\sigma^2}}$$
$$< \sqrt{2\pi}(k-\frac{1}{2})\lambda e^{\frac{(k-\frac{1}{2})^2\lambda^2}{2}} e^{-\frac{(k-\frac{1}{2})^2\lambda^2}{2}} < \sqrt{2\pi}k\lambda ,$$

where the first inequality follows from Fact 4 and having $k - \frac{1}{2} > \max\{\frac{2}{\lambda}, \frac{1}{\lambda^2}\}$.

Next, Parts B and C are shown as follows:

$$\frac{L_k(\rho)}{\sqrt{1-\rho^2}} = \frac{1}{2}\frac{1}{Q(\frac{t_k}{\sigma}) - Q(\frac{t_{k+1}}{\sigma})} e^{-\frac{t_k^2}{2\sigma^2}} > \frac{1}{2}\frac{1}{\frac{1}{2}e^{-\frac{t_k^2}{2\sigma^2}}} e^{-\frac{t_k^2}{2\sigma^2}} = 1 ,$$

where the inequality follows from dropping the smaller of the two Q terms and using Fact 1 to upper bound the remaining Q term (since $k \geq 1$, Fact 1 can be used).

$$\frac{R_k(\rho)}{\sqrt{1-\rho^2}} = \frac{1}{2}\frac{1}{\int_{t_k}^{t_{k+1}} \mathcal{N}_{0,\sigma^2}(x)\,dx} e^{-\frac{t_{k+1}^2}{2\sigma^2}} < \frac{1}{\frac{2}{\sqrt{2\pi}\sigma}e^{-\frac{t_{k+1}^2}{2\sigma^2}}\Delta} e^{-\frac{t_{k+1}^2}{2\sigma^2}} = \sqrt{\frac{\pi}{2}}\frac{1}{\lambda} ,$$

where the inequality follows from substituting the Gaussian pdf with its value at the upper bound of the integral (recalling that $k \geq 0$) over the whole integration region.

Finally, Part D follows in a similar way to Part B. Specifically,

$$\frac{R_k(\rho)}{\sqrt{1-\rho^2}} \;=\; \frac{1}{2}\frac{1}{Q(\frac{t_k}{\sigma}) - Q(\frac{t_{k+1}}{\sigma})} e^{-\frac{t_{k+1}^2}{2\sigma^2}} \;>\; \frac{1}{2}\frac{1}{\frac{1}{2}e^{-\frac{t_k^2}{2\sigma^2}}} e^{-\frac{t_{k+1}^2}{2\sigma^2}} \;=\; e^{-k\lambda^2} \;,$$

where the inequality follows from dropping the small Q function and using Fact 1 to upper bound the large Q function (recalling that $k \geq 1$, thus Fact 1 can be used). $\square$

Proof of Lemma 12:

We show the statements of the lemma in the following order: G, F, C, B, D, E and A. Consider Part G. Let $\varepsilon > 0$ be given. Then by assumption there exists $\delta > 0$ such that if $\left|\frac{x}{y} - z\right| < \delta$, then $|G(x,y) - c| < \varepsilon$. Similarly, by assumption, there exists ρ_o such that for all $\rho > \rho_o$, $\sup_{0\leq k\leq N(\rho)} \left|\frac{a_k(\rho)}{b_k(\rho)} - z\right| < \delta$. Combining the last two statements it follows that for all $\rho > \rho_o$, $\sup_{0\leq k\leq N(\rho)} \left|G\big(a_k(\rho), b_k(\rho)\big) - c\right| < \varepsilon$. Since ε is arbitrary, the result follows.

Part F is a special case of Part G with, for example, $x = z_o$, $y = 1$, $\lim^*_{\rho\to_1} a_k(\rho) = z_o$ and $b_k(\rho) = 1$ for all ρ and for $0 \leq k \leq N(\rho)$.

Next, we show Part C. Let $\lim^*_{\rho\to 1} a_k(\rho) = a$, $\lim^*_{\rho\to 1} b_k(\rho) = b$. Then,

$$\lim_{\rho\to 1} \sup_{0\leq k\leq N(\rho)} a_k(\rho)\, b_k(\rho) \;\leq\; \lim_{\rho\to 1} \sup_{0\leq k\leq N(\rho)} a_k(\rho) \lim_{\rho\to 1} \sup_{0\leq k\leq N(\rho)} b_k(\rho) \;=\; ab \;,$$

and

$$\lim_{\rho\to 1} \inf_{0\leq k\leq N(\rho)} a_k(\rho)\, b_k(\rho) \;\geq\; \lim_{\rho\to 1} \inf_{0\leq k\leq N(\rho)} a_k(\rho) \lim_{\rho\to 1} \inf_{0\leq k\leq N(\rho)} b_k(\rho) \;=\; ab \;,$$

where the equalities in the two equations above follow from the definition of $\lim^*$. Combining the two equations above and using the definition of $\lim^*$, we obtain that $\lim^*_{\rho\to 1} a_k(\rho)\, b_k(\rho) = ab$, which is what was needed to show. Part B can be shown in a similar way.

We proceed with Part D. Since $\lim^*_{\rho\to 1} \frac{a_k(\rho)}{b_k(\rho)} = 1$, it follows that for any $\varepsilon > 0$, there exists ρ_o such that for all $\rho > \rho_o$ and for $0 \le k \le N(\rho)$, $1 - \varepsilon < \frac{a_k(\rho)}{b_k(\rho)} < 1 + \varepsilon$. Consequently, for all such ρ, $1 - \varepsilon < \frac{\sum_{k=0}^{N(\rho)} a_k(\rho)}{\sum_{k=0}^{N(\rho)} b_k(\rho)} < 1 + \varepsilon$. Since ε is arbitrary, the result follows. Part E can be shown in a similar way.

Finally, we show Part A as follows:

$$\begin{aligned}\lim_{\rho\to 1} \sup_{0\le k\le N(\rho)} \int f(k,\rho,x)\,dx \;&\le\; \lim_{\rho\to 1} \int \sup_{0\le k\le N(\rho)} f(k,\rho,x)\,dx \\ &\overset{(a)}{=} \int \lim_{\rho\to 1} \sup_{0\le k\le N(\rho)} f(k,\rho,x)\,dx \;\overset{(b)}{=}\; \int \lim^*_{\rho\to 1} f(k,\rho,x)\,dx\ ,\end{aligned} \tag{4.16}$$

where (a) follows from having $|\sup_{0\le k\le N(\rho)} f(k,\rho,x)| \le \sup_{0\le k\le N(\rho)} |f(k,\rho,x)| \le G(x)$ a.e. and having $\lim_{\rho\to 1} \sup_{0\le k\le N(\rho)} f(k,\rho,x)$ exist a.e. (due to the fact that if $\lim^*_{\rho\to 1} f(k,\rho,x)$ exists, it equals $\lim_{\rho\to 1} \sup_{0\le k\le N(\rho)} f(k,\rho,x)$), and applying the dominated convergence theorem[3] [20] (p. 209), and (b) is due to the just mentioned fact that if $\lim^*_{\rho\to 1} f(k,\rho,x)$ exists, it equals $\lim_{\rho\to 1} \sup_{0\le k\le N(\rho)} f(k,\rho,x)$. In a similar way we also have

$$\begin{aligned}\lim_{\rho\to 1} \inf_{0\le k\le N(\rho)} \int f(k,\rho,x)\,dx \;&\ge\; \lim_{\rho\to 1} \int \inf_{0\le k\le N(\rho)} f(k,\rho,x)\,dx \\ &\overset{(a)}{=} \int \lim_{\rho\to 1} \inf_{0\le k\le N(\rho)} f(k,\rho,x)\,dx \;\overset{(b)}{=}\; \int \lim^*_{\rho\to 1} f(k,\rho,x)\,dx\ ,\end{aligned} \tag{4.17}$$

where (a) follows from having $|\inf_{0\le k\le N(\rho)} f(k,\rho,x)| \le \sup_{0\le k\le N(\rho)} |f(k,\rho,x)| \le G(x)$ a.e. and having $\lim_{\rho\to 1} \inf_{0\le k\le N(\rho)} f(k,\rho,x)$ exist a.e. (due to the fact that if $\lim^*_{\rho\to 1} f(k,\rho,x)$ exists, it equals $\lim_{\rho\to 1} \inf_{0\le k\le N(\rho)} f(k,\rho,x)$), and applying the dominated convergence theorem, and (b) is due to the just mentioned fact that if $\lim^*_{\rho\to 1} f(k,\rho,x)$ exists, it equals $\lim_{\rho\to 1} \inf_{0\le k\le N(\rho)} f(k,\rho,x)$. Combining (4.16) and

[3]It is easily shown that the theorem applies when the integrand is parameterized by some t converging continuously to some t_o, rather than some integer n converging to ∞.

(4.17) concludes the proof of Part A and of the lemma. □

Proof of Lemma 13:

We begin by showing Part B of the lemma. From Lemma 9 and the definition of $P^*_{k+1|k}$ and $\tilde{f}_{X_2|I_1}(x|k)$ we have that for all ρ and k,

$$P^*_{k+1|k} \;<\; R_k(\rho) \int_{\frac{t_{k+1}}{\rho}}^{t_{k+2}} \mathcal{N}_{t_{k+1}\rho,\sigma_\rho^2}(x) g\big(\frac{\rho x - t_{k+1}}{\sigma_\rho}\big)\, dx \;, \tag{4.18}$$

and that for any $0 \le \gamma < 1$, there exists $\rho_\gamma < 1$ such that for all $\rho > \rho_\gamma$ and all k,

$$P^*_{k+1|k} \;>\; \gamma R_k(\rho) \int_{\frac{t_{k+1}}{\rho}}^{t_{k+2}} \mathcal{N}_{t_{k+1}\rho,\sigma_\rho^2}(x) g\big(\frac{\rho x - t_{k+1}}{\sigma_\rho}\big)\, dx \;. \tag{4.19}$$

Assuming $k \ge 0$, which is sufficient for the purpose of this lemma, we evaluate the integrals above, as follows:

$$\begin{aligned}
\int_{\frac{t_{k+1}}{\rho}}^{t_{k+2}} \mathcal{N}_{t_{k+1}\rho,\sigma_\rho^2}(x) g\big(\frac{\rho x - t_{k+1}}{\sigma_\rho}\big)\, dx \;&=\; \int_{\frac{\frac{t_{k+1}}{\rho} - t_{k+1}\rho}{\sigma_\rho}}^{\frac{t_{k+2} - t_{k+1}\rho}{\sigma_\rho}} \mathcal{N}_{0,1}(x)\, g\big(\frac{\rho\sigma_\rho x + t_{k+1}\rho^2 - t_{k+1}}{\sigma_\rho}\big)\, dx \\
&=\; \int_{\frac{t_{k+1}(1-\rho^2)}{\rho\sigma_\rho}}^{\frac{t_{k+1}(1-\rho)+\Delta}{\sigma_\rho}} \mathcal{N}_{0,1}(x)\, g\big(\rho x - \frac{t_{k+1}(1-\rho^2)}{\sigma_\rho}\big)\, dx \\
&=\; \int_0^\infty \mathcal{N}_{0,1}(x)\, g\big(\rho(x - S_{k,\rho})\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)\, dx \;,
\end{aligned} \tag{4.20}$$

where $S_{k,\rho} \triangleq \frac{t_{k+1}(1-\rho^2)}{\rho\sigma_\rho}$, $T_{k,\rho} \triangleq \frac{t_{k+1}(1-\rho)+\Delta}{\sigma_\rho}$, and $I_F(x)$ denotes the indicator function of the event F. Next, we would like to apply Lemma 12 (Part A) to the right-hand side above. To justify using this lemma, we first observe that for all $x \in [0,\infty)$, the integrand is dominated by an integrable function because

$$\sup_{0\le k\le N(\rho)} \Big|\mathcal{N}_{0,1}(x)\, g\big(\rho(x - S_{k,\rho})\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)\Big| \;\le\; \mathcal{N}_{0,1}(x) \;.$$

Secondly, we need to show that show that $\lim^*_{\rho\to 1} \mathcal{N}_{0,1}(x)\, g\big(\rho(x - S_{k,\rho})\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)$

exists for almost all $x \in [0,\infty)$. For $x \in (0,\infty)$,

$$\begin{aligned}\lim_{\rho\to 1} \sup_{0\le k\le N(\rho)} \mathcal{N}_{0,1}(x)\, g\big(\rho(x-S_{k,\rho})\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x) \le\ & \mathcal{N}_{0,1}(x) \lim_{\rho\to 1} \sup_{0\le k\le N(\rho)} g\big(\rho(x-S_{k,\rho})\big) \\ \overset{(a)}{=}\ & \mathcal{N}_{0,1}(x) \lim_{\rho\to 1} g\big(\rho(x-S_{N(\rho),\rho})\big) \\ \overset{(b)}{=}\ & \mathcal{N}_{0,1}(x)\, g(x)\ , \end{aligned} \tag{4.21}$$

where (a) follows from the monotonicity of g (Fact 6), and (b) follows from having $x > 0$ and $\lim_{\rho\to 1} S_{N(\rho),\rho} = 0$. We also have for $x > 0$,

$$\begin{aligned}&\lim_{\rho\to 1} \inf_{0\le k\le N(\rho)} \mathcal{N}_{0,1}(x)\, g\big(\rho(x-S_{k,\rho})\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x) \\ &\quad\ge\ \mathcal{N}_{0,1}(x) \lim_{\rho\to 1} \inf_{0\le k\le N(\rho)} g\big(\rho(x-S_{k,\rho})\big) \lim_{\rho\to 1} \inf_{0\le k\le N(\rho)} I_{(S_{k,\rho},T_{k,\rho})}(x) \\ &\quad\overset{(a)}{=}\ \mathcal{N}_{0,1}(x) \lim_{\rho\to 1} g\big(\rho(x-S_{0,\rho})\big) \lim_{\rho\to 1} I_{(S_{N(\rho),\rho},T_{0,\rho})}(x) \\ &\quad\overset{(b)}{=}\ \mathcal{N}_{0,1}(x)\, g(x)\ , \end{aligned} \tag{4.22}$$

where (a) follows from the monotonicity of g (Fact 6), and (b) is due to having $x > 0$, $\lim_{\rho\to 1} S_{0,\rho} = 0$, $\lim_{\rho\to 1} S_{N(\rho),\rho} = 0$, and $\lim_{\rho\to 1} T_{0,\rho} = \infty$. Equations (4.21) and (4.22) now imply that $\lim^*_{\rho\to 1} \mathcal{N}_{0,1}(x)\, g\big(\rho(x-S_{k,\rho})\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x) = \mathcal{N}_{0,1}(x)\, g(x)$ for all $x \in (0,\infty)$. Consequently, we may apply Lemma 12 (Part A) to the right-hand side of (4.20) and obtain

$$\begin{aligned}&\lim^*_{\rho\to 1} \int_0^\infty \mathcal{N}_{0,1}(x)\, g\big(\rho(x-S_{k,\rho})\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)\, dx \\ &\quad=\ \int_0^\infty \lim^*_{\rho\to 1} \mathcal{N}_{0,1}(x)\, g\big(\rho(x-S_{k,\rho})\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)\, dx \\ &\quad=\ \int_0^\infty \mathcal{N}_{0,1}(x)\, g(x)\, dx\ =\ \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{Q(x)}{\frac{1}{2}e^{-\frac{x^2}{2}}}\, dx \\ &\quad=\ \sqrt{\frac{2}{\pi}} \int_0^\infty Q(x)\, dx\ =\ \frac{1}{\pi}\ , \end{aligned} \tag{4.23}$$

Finally, combining the fact that γ in (4.19) can be chosen arbitrarily close to

one, together with (4.18), (4.20) and (4.23), it follows that $\lim^*_{\rho\to 1} \frac{P^*_{k+1|k}}{\frac{1}{\pi}R_k(\rho)} = 1$, which concludes the proof of Part B.

Next, we show Part A, the proof of which is similar to the proof of Part B, up to a point. Assume first that $k \geq 1$. Then, $t_k < \frac{t_k}{\rho}$, and using Lemma 9 and the definition of $\tilde{f}_{X_2|I_1}(x|k)$ we have that for all ρ,

$$P_{k-1|k} \;<\; L_k(\rho) \int_{t_{k-1}}^{t_k} \mathcal{N}_{t_k\rho,\sigma_\rho^2}(x) g\big(\frac{t_k - \rho x}{\sigma_\rho}\big)\, dx \;, \tag{4.24}$$

and that for any $0 \leq \gamma < 1$, there exists $\rho_\gamma < 1$ such that for all $\rho > \rho_\gamma$,

$$P_{k-1|k} \;>\; \gamma L_k(\rho) \int_{t_{k-1}}^{t_k} \mathcal{N}_{t_k\rho,\sigma_\rho^2}(x) g\big(\frac{t_k - \rho x}{\sigma_\rho}\big)\, dx \;. \tag{4.25}$$

Next, we evaluate the integrals above as follows:

$$\begin{aligned}
\int_{t_{k-1}}^{t_k} \mathcal{N}_{t_k\rho,\sigma_\rho^2}(x) g\big(\frac{t_k - \rho x}{\sigma_\rho}\big)\, dx \;&=\; \int_{\frac{t_{k-1}-t_k\rho}{\sigma_\rho}}^{\frac{t_k-t_k\rho}{\sigma_\rho}} \mathcal{N}_{0,1}(x)\, g\big(\frac{t_k - \rho(x\sigma_\rho + t_k\rho)}{\sigma_\rho}\big)\, dx \\
&=\; \int_{\frac{t_k(1-\rho)-\Delta}{\sigma_\rho}}^{\frac{t_k(1-\rho)}{\sigma_\rho}} \mathcal{N}_{0,1}(x)\, g\big(\frac{t_k(1-\rho^2)}{\sigma_\rho} - \rho x\big)\, dx \\
&=\; \int_{-\infty}^{0} \mathcal{N}_{0,1}(x)\, g\big(T_{k,\rho}(1+\rho) - \rho x\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)\, dx \;,
\end{aligned} \tag{4.26}$$

where $S_{k,\rho} \stackrel{\Delta}{=} \frac{t_k(1-\rho)-\Delta}{\sigma_\rho}$ and $T_{k,\rho} \stackrel{\Delta}{=} \frac{t_k(1-\rho)}{\sigma_\rho}$. We now introduce a modified version of the $\lim^*$ operator. Specifically, we let $\widehat{\lim}^*$ be the same operator as $\lim^*$ except that k is greater than or equal to one rather than zero. Namely, if $\widehat{\lim}^* f_k(\rho) = c$ for some function $f_k(\rho)$, then it means that $\lim_{\rho\to 1} \sup_{1\leq k\leq N(\rho)} |f_k(\rho) - c| = 0$.

Next, we would like to apply Lemma 12 (Part A) to the right-hand side of (4.26). Clearly, the result of the lemma holds for the $\widehat{\lim}^*$ operator as well. To justify using this lemma, we first observe that for all $x \in (-\infty, 0]$, the integrand is dominated by an integrable function because

$$\sup_{1\leq k\leq N(\rho)} \Big|\mathcal{N}_{0,1}(x)\, g\big(T_{k,\rho}(1+\rho) - \rho x\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)\Big| \;\leq\; \mathcal{N}_{0,1}(x) \;.$$

Secondly, we need to show that show that $\widehat{\lim}^{*}_{\rho\to 1}\mathcal{N}_{0,1}(x)\, g\big(T_{k,\rho}(1+\rho)-\rho x\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)$ exists for almost all $x \in (-\infty, 0]$. For $x \in (-\infty, 0]$,

$$\begin{aligned}
&\lim_{\rho\to 1} \sup_{1\le k\le N(\rho)} \mathcal{N}_{0,1}(x)\, g\big(T_{k,\rho}(1+\rho)-\rho x\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x) \\
&\quad\le\ \mathcal{N}_{0,1}(x) \lim_{\rho\to 1} \sup_{1\le k\le N(\rho)} g\big(T_{k,\rho}(1+\rho)-\rho x\big) \\
&\quad\overset{(a)}{=}\ \mathcal{N}_{0,1}(x) \lim_{\rho\to 1} g\big(T_{1,\rho}(1+\rho)-\rho x\big) \\
&\quad\overset{(b)}{=}\ \mathcal{N}_{0,1}(x)\, g(-x)\ , \qquad (4.27)
\end{aligned}$$

where (a) follows from the monotonicity of g (Fact 6), and (b) follows from having $x \le 0$ and $\lim_{\rho\to 1} T_{1,\rho} = 0$. We also have for $x \le 0$,

$$\begin{aligned}
&\lim_{\rho\to 1} \inf_{1\le k\le N(\rho)} \mathcal{N}_{0,1}(x)\, g\big(T_{k,\rho}(1+\rho)-\rho x\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x) \\
&\quad\ge\ \mathcal{N}_{0,1}(x) \lim_{\rho\to 1} \inf_{1\le k\le N(\rho)} g\big(T_{k,\rho}(1+\rho)-\rho x\big) \lim_{\rho\to 1} \inf_{1\le k\le N(\rho)} I_{(S_{k,\rho},T_{k,\rho})}(x) \\
&\quad\overset{(a)}{=}\ \mathcal{N}_{0,1}(x) \lim_{\rho\to 1} g\big(T_{N(\rho),\rho}(1+\rho)-\rho x\big) \lim_{\rho\to 1} I_{(S_{N(\rho),\rho},T_{1,\rho})}(x) \\
&\quad\overset{(b)}{=}\ \mathcal{N}_{0,1}(x)\, g(-x)\ , \qquad (4.28)
\end{aligned}$$

where (a) follows from the monotonicity of g (Fact 6), and (b) is due to having $x \le 0$, $\lim_{\rho\to 1} T_{N(\rho),\rho} = 0$, $\lim_{\rho\to 1} S_{N(\rho),\rho} = -\infty$, and $T_{1,\rho} \ge 0$ for all ρ. Equations (4.27) and (4.28) now imply that $\widehat{\lim}^{*}_{\rho\to 1}\mathcal{N}_{0,1}(x)\, g\big(T_{k,\rho}(1+\rho)-\rho x\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x) = \mathcal{N}_{0,1}(x)\, g(-x)$ for all $x \in (-\infty, o]$. Consequently, we may apply Lemma 12 (Part A), which, as mentioned, holds for the $\widehat{\lim}^{*}$ operator as well, to the right-hand side of (4.26) and obtain

$$\begin{aligned}
&\widehat{\lim}^{*}_{\rho\to 1} \int_{-\infty}^{0} \mathcal{N}_{0,1}(x)\, g\big(T_{k,\rho}(1+\rho)-\rho x\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)\, dx \\
&\quad=\ \int_{-\infty}^{0} \widehat{\lim}^{*}_{\rho\to 1} \mathcal{N}_{0,1}(x)\, g\big(T_{k,\rho}(1+\rho)-\rho x\big)\, I_{(S_{k,\rho},T_{k,\rho})}(x)\, dx \\
&\quad=\ \int_{-\infty}^{0} \mathcal{N}_{0,1}(x)\, g(-x)\, dx\ =\ \int_{0}^{\infty} \mathcal{N}_{0,1}(x)\, g(x)\, dx\ =\ \frac{1}{\pi}\ , \qquad (4.29)
\end{aligned}$$

where the last equality follows from (4.23).

Next, combining the fact that γ in (4.25) can be chosen arbitrarily close to one, together with (4.24), (4.26) and (4.29), it follows that

$$\widehat{\lim}^{*}_{\rho\to 1} \frac{P_{k-1|k}}{\frac{1}{\pi}L_k(\rho)} \;=\; 1\;. \tag{4.30}$$

It remains to consider the case that $k = 0$. We observe that by symmetry $P_{-1|0} = P_{1|0}$. Now, from Lemma 14 we have that $\lim_{\rho\to 1} \frac{P_{1|0}}{P^*_{1|0}} = 1$, where we comment that although Lemma 14 uses in its proof the current lemma, i.e. Lemma 13, there is no circularity. The reason is that Lemma 14 uses Part D of the current lemma, which in turn depends on Part B, but there is no dependence on Part A, which is considered here. By Part B, $\lim_{\rho\to 1} \frac{P^*_{1|0}}{\frac{1}{\pi}R_0(\rho)} = 1$. Combining this with the facts $R_0(\rho) = L_0(\rho)$, $\lim_{\rho\to 1} \frac{P_{1|0}}{P^*_{1|0}} = 1$, and $P_{-1|0} = P_{1|0}$, we get that $\lim_{\rho\to 1} \frac{P_{-1|0}}{\frac{1}{\pi}L_0(\rho)} = 1$. This together with (4.30) complete the proof of Part B and of the lemma as a whole. □

Proof of Lemma 14:

Let $k \geq 0$, which is sufficient for the purpose of the lemma. We write the following.

$$P_{k+1|k} \;=\; \int_{t_{k+1}}^{\frac{t_{k+1}}{\rho}} f_{X_2|I_1}(x|k)\,dx + P^*_{k+1|k} \;<\; \int_{t_{k+1}}^{\frac{t_{k+1}}{\rho}} \frac{\mathcal{N}_{0,\sigma^2}(x)}{P_k}\,dx + P^*_{k+1|k}\;,$$

where the inequality follows from the fact that $f_{X_2|I_1}(x|k) \leq \tilde{f}_{X_2|I_1}(x|k) < \frac{\mathcal{N}_{0,\sigma^2}(x)}{P_k}$ for $t_{k+1} \leq x \leq \frac{t_{k+1}}{\rho}$, as seen by (4.15). Combining this with the fact that for all ρ sufficiently close to one, $P^*_{k+1|k} > \frac{1}{5}R_k(\rho)$ for $0 \leq k \leq N(\rho)$, which is shown in Lemma 13 (Part D), it follows that it suffices to show that

$$\lim^{*}_{\rho\to 1} \frac{\int_{t_{k+1}}^{\frac{t_{k+1}}{\rho}} \frac{\mathcal{N}_{0,\sigma^2}(x)}{P_k}\,dx}{\frac{1}{5}R_k(\rho)} \;=\; 0\;. \tag{4.31}$$

Simplifying the above expression, we obtain

$$\begin{aligned}
\sup_{0\le k\le N(\rho)} \frac{\int_{t_{k+1}}^{\frac{t_{k+1}}{\rho}} \frac{\mathcal{N}_{0,\sigma^2}(x)}{P_k}\,dx}{\frac{1}{5}R_k(\rho)} &= \sup_{0\le k\le N(\rho)} \frac{\frac{1}{P_k}\int_{t_{k+1}}^{\frac{t_{k+1}}{\rho}} \mathcal{N}_{0,\sigma^2}(x)\,dx}{\frac{1}{5}\frac{1}{2}\frac{1}{P_k}e^{-\frac{t_{k+1}^2}{2\sigma^2}}\sqrt{1-\rho^2}} \\
&\overset{(a)}{<} \sup_{0\le k\le N(\rho)} \frac{10\frac{1}{\sqrt{2\pi}\sigma}\,e^{-\frac{t_{k+1}^2}{2\sigma^2}}\left[t_{k+1}\frac{(1-\rho)}{\rho}\right]}{e^{-\frac{t_{k+1}^2}{2\sigma^2}}\sqrt{1-\rho^2}} \\
&\overset{(b)}{<} \frac{10\lambda}{\sqrt{2\pi}\rho}\left(\left\lfloor\left(\ln\frac{1}{1-\rho}\right)^{\frac{3}{4}}-\frac{1}{2}\right\rfloor+\frac{1}{2}\right)\sqrt{1-\rho} \\
&< -\frac{10\lambda}{\sqrt{2\pi}\rho}\sqrt{1-\rho}\big(\ln(1-\rho)\big)^{\frac{3}{4}} \longrightarrow 0 \text{ as } \rho\longrightarrow 1\ ,
\end{aligned}$$

where (a) follows from substituting the Gaussian pdf with its value at the lower limit of the integral, and (b) is obtained by recalling that $t_{k+1} = (k+\frac{1}{2})\Delta$, substituting $N(\rho)$ for k, and noting that $1 < \sqrt{1+\rho}$. This shows (4.31) and concludes the proof of the lemma. □

Proof of Lemma 15:

Part B: Lemmas 12 (Part C), 13 (Part B) and 14 imply that $\lim^*_{\rho\to 1}\frac{P_{k+1|k}}{\frac{1}{\pi}R_k(\rho)} = 1$. Lemma 10 (Part C) shows that $R_k(\rho) < \sqrt{\frac{\pi}{2}}\frac{1}{\lambda}\sqrt{1-\rho^2}$ for all $k \ge 0$, so that $\lim^*_{\rho\to 1} R_k(\rho) = 0$. It now follows from Lemma 12 (Part C) that $\lim^*_{\rho\to 1} P_{k+1|k} = 0$, which is Part B of the lemma.

Part A: We begin by considering $N_\lambda < k \le N(\rho)$, where N_λ is the constant given in Lemma 10 (Part A). Using Lemma 10 (Part A), we obtain

$$\begin{aligned}
L_k(\rho) &< \sqrt{2\pi}k\lambda\sqrt{1-\rho^2} \le \sqrt{2\pi}N(\rho)\lambda\sqrt{1-\rho^2} \\
&= \sqrt{2\pi}\left\lfloor\left(\ln\frac{1}{1-\rho}\right)^{\frac{3}{4}}-\frac{1}{2}\right\rfloor\lambda\sqrt{1-\rho^2} < -2\sqrt{\pi}\lambda\sqrt{1-\rho}\left(\ln(1-\rho)\right)^{\frac{3}{4}}\ .
\end{aligned} \tag{4.32}$$

Next, consider $0 \le k \le N_\lambda$. Since $\frac{1}{2}\frac{1}{P_k}e^{-\frac{t_k^2}{2\sigma^2}}$ can assume at most N_λ values, $L_k(\rho) = \frac{1}{2}\frac{1}{P_k}e^{-\frac{t_k^2}{2\sigma^2}}\sqrt{1-\rho^2} \to 0$ as $\rho \to 1$ for $0 \le k \le N_\lambda$. Combining this with

(4.32), it follows that $\lim^*_{\rho\to 1} L_k(\rho) = 0$. Finally, using Lemma 13 (Part A) together with Lemma 12 (Part C) shows that $\lim^*_{\rho\to 1} P_{k-1|k} = 0$, which concludes the proof of Part A and of the lemma as a whole. □

Proof of Lemma 16:

We need to show that

$$\begin{aligned} A.\quad & P_{j+l+1}(f) \;<\; \frac{1}{2}\, P_{j+l}(f) \qquad \text{for} \quad l \;\geq\; B+2\ , \\ B.\quad & P_{j-l-1}(f) \;<\; \frac{1}{2}\, P_{j-l}(f) \qquad \text{for} \quad l \;\geq\; B\ , \end{aligned}$$

where $B = \lceil \frac{\sigma^2}{\Delta^2} \ln 3 \rceil$.

To keep notation short, we omit the parameter f from $P_l(f)$ throughout this lemma, and comment that P_l should not be confused with the general term P_k that represents the probability of a Gaussian random variable with *zero* mean and variance σ^2 of lying in S_k.

We begin by showing Part A. Using the notation of Lemma A1 we have that

$$P_{j+l+1} \;=\; \int_{t_{j+l+1}}^{t_{j+l+2}} \mathcal{N}_{\mu,\sigma^2}(x)\, dx \;=\; P_{\mu,\sigma^2,\Delta}(t_{j+l}+\Delta)\ ,$$

and

$$\frac{1}{2}P_{j+l} \;=\; \frac{1}{2}\int_{t_{j+l}}^{t_{j+l+1}} \mathcal{N}_{\mu,\sigma^2}(x)\, dx \;=\; \frac{1}{2}\, P_{\mu,\sigma^2,\Delta}(t_{j+l})\ .$$

Thus, we need to show that

$$P_{\mu,\sigma^2,\Delta}(t_{j+l}+\Delta) \;<\; \frac{1}{2}\, P_{\mu,\sigma^2,\Delta}(t_{j+l})\ .$$

Applying Lemma A1 (Part A) with $s = \frac{1}{2}$, we obtain that the above holds if $t_{j+l} \geq \mu + \Delta[1 + \frac{\sigma^2}{\Delta^2} \ln 3]$. Since $\mu \in S_j$, it follows that $t_j \leq \mu < t_{j+1}$. Thus, it suffices to have $t_{j+l} \geq t_{j+1} + \Delta[1 + \frac{\sigma^2}{\Delta^2} \ln 3]$. Equivalently, we need to have $l \geq \frac{\sigma^2}{\Delta^2} \ln 3 + 2$, which holds by definition. This shows Part A.

Next, we consider Part B. The derivation is very similar to Part A. Specifically, using the notation of Lemma A1 we have that

$$P_{j-l-1} \;=\; \int_{t_{j-l-2}}^{t_{j-l-1}} c\mathcal{N}_{\mu,\sigma^2}(x)\,dx \;=\; c\,P_{\mu,\sigma^2,\Delta}(t_{j-l-1}-\Delta)\;,$$

and

$$\frac{1}{2}P_{j-l} \;=\; \frac{1}{2}\int_{t_{j-l-1}}^{t_{j-l}} c\mathcal{N}_{\mu,\sigma^2}(x)\,dx \;=\; \frac{1}{2}\,c\,P_{\mu,\sigma^2,\Delta}(t_{j-l-1})\;.$$

Thus, we need to show that

$$P_{\mu,\sigma^2,\Delta}(t_{j-l-1}-\Delta) \;<\; \frac{1}{2}\,P_{\mu,\sigma^2,\Delta}(t_{j-l-1})\;.$$

Applying Lemma A1 (Part B) with $s=\frac{1}{2}$, we obtain that the above holds if $t_{j-l-1} \le \mu - \Delta[1+\frac{\sigma^2}{\Delta^2}\ln 3]$. Since $t_j \le \mu < t_{j+1}$, it suffices to have $t_{j-l-1} \le t_j - \Delta[1+\frac{\sigma^2}{\Delta^2}\ln 3]$. Equivalently, we need to have $l \ge \frac{\sigma^2}{\Delta^2}\ln 3$, which holds by definition. This concludes the proof of Part B and of the lemma as a whole. □

Proof of Lemma 17:

We need to show that for all ρ sufficiently close to one and for $0 \le k \le N(\rho)$, the following holds:

$$A.\quad P_{k+l|k} \;<\; (P^*_{k+1|k})^l \qquad \text{for} \quad l \ge 2\;,$$

$$B.\quad P_{k-l|k} \;<\; (P^*_{k+1|k})^l \qquad \text{for} \quad l \ge 2\;.$$

We begin with the proof of Part A. Specifically, we show the following two steps, from which A easily follows.

A1: $P_{k+l+1|k} \;<\; P_{k+l|k}P^*_{k+1|k} \quad \text{for} \;\; l \ge 2$

A2: $P_{k+2|k} \;<\; (P^*_{k+1|k})^2$

Step A1: To keep notation compact, we rewrite what needs to be shown as: For all ρ sufficiently close to one and for $0 \le k \le N(\rho)$, $P_{l+1|k} < P_{l|k}P^*_{k+1|k}$ for

$l \geq k+2$. We observe that $t_l > \frac{t_{k+1}}{\rho}$ when $0 \leq k \leq N(\rho)$ and $l \geq k+2$. Thus, since $f_{X_2|I_1}(x|k) < \tilde{f}_{X_2|I_1}(x|k)$ for all $x > \frac{t_{k+1}}{\rho}$, it follows from (4.15) and the monotonicity of g (Fact 6) that for all ρ and $0 \leq k \leq N(\rho)$

$$\begin{aligned} P_{l+1|k} &< R_k(\rho) \int_{t_{l+1}}^{t_{l+2}} \mathcal{N}_{t_{k+1}\rho,\sigma_\rho^2}(x)\, g\big(\frac{\rho x - t_{k+1}}{\sigma_\rho}\big)\, dx \\ &< R_k(\rho)\, g\big(\frac{\rho t_{l+1} - t_{k+1}}{\sigma_\rho}\big) Q\Big(\frac{t_{l+1} - t_{k+1}\rho}{\sigma_\rho}\Big) \ . \end{aligned} \tag{4.33}$$

Next, using Lemma 9 with $\gamma = \frac{1}{2}$, (4.15), and the monotonicity of g (Fact 6), we obtain that for all ρ sufficiently close to one and $0 \leq k \leq N(\rho)$

$$\begin{aligned} P_{l|k} &> \frac{1}{2} R_k(\rho) \int_{t_l}^{t_{l+1}} \mathcal{N}_{t_{k+1}\rho,\sigma_\rho^2}(x)\, g\big(\frac{\rho x - t_{k+1}}{\sigma_\rho}\big)\, dx \\ &> \frac{1}{2} R_k(\rho)\, g\big(\frac{\rho t_{l+1} - t_{k+1}}{\sigma_\rho}\big) \left[Q\Big(\frac{t_l - t_{k+1}\rho}{\sigma_\rho}\Big) - Q\Big(\frac{t_{l+1} - t_{k+1}\rho}{\sigma_\rho}\Big) \right] \ . \end{aligned} \tag{4.34}$$

It now follows from (4.33) and (4.34) that it suffices to show that for all ρ sufficiently close to one and $0 \leq k \leq N(\rho)$

$$Q\Big(\frac{t_{l+1} - t_{k+1}\rho}{\sigma_\rho}\Big) < \frac{1}{2} \left[Q\Big(\frac{t_l - t_{k+1}\rho}{\sigma_\rho}\Big) - Q\Big(\frac{t_{l+1} - t_{k+1}\rho}{\sigma_\rho}\Big) \right] P^*_{k+1|k} \ ,$$

or alternatively, since $P^*_{k+1|k} \leq 1$, it suffices to have

$$\frac{3Q\big(\frac{t_{l+1} - t_{k+1}\rho}{\sigma_\rho}\big)}{Q\big(\frac{t_l - t_{k+1}\rho}{\sigma_\rho}\big)} < P^*_{k+1|k} \ . \tag{4.35}$$

Next, we observe that the argument of the Q function in the numerator can be written as $\frac{t_{l+1} - t_{k+1}\rho}{\sigma_\rho} = \frac{t_l - t_{k+1}\rho + \Delta}{\sigma_\rho} = \frac{\Delta(\frac{t_l - t_{k+1}\rho}{\Delta} + 1)}{\sigma_\rho}$. Using Fact 5, with $a = \frac{t_l - t_{k+1}\rho}{\Delta}$ and $z = \frac{\Delta}{\sigma_\rho} = \frac{\lambda}{\sqrt{1-\rho^2}}$, where we notice that $l \geq k+2$ implies that $a > 1$, it follows that for all ρ sufficiently close to one

$$\frac{3Q\big(\frac{t_{l+1} - t_{k+1}\rho}{\sigma_\rho}\big)}{Q\big(\frac{t_l - t_{k+1}\rho}{\sigma_\rho}\big)} < 6e^{-\frac{\lambda^2}{2(1-\rho^2)}} \ . \tag{4.36}$$

To obtain a lower bound to $P^*_{k+1|k}$, we apply Lemmas 10 (Part D) and 13 (Part D), and obtain that for all ρ sufficiently close to one and $1 \leq k \leq N(\rho)$

$$\begin{aligned} P^*_{k+1|k} &> \frac{1}{5}\sqrt{1-\rho^2}\, e^{-k\lambda^2} \geq \frac{1}{5}\sqrt{1-\rho^2}\, e^{-\lfloor (\ln \frac{1}{1-\rho})^{\frac{3}{4}} - \frac{1}{2} \rfloor \lambda^2} > \frac{1}{5}\sqrt{1-\rho}\, e^{-(\ln \frac{1}{1-\rho})\lambda^2} \\ &= \frac{1}{5}(1-\rho)^{\lambda^2+\frac{1}{2}} , \end{aligned} \tag{4.37}$$

where the second inequality derives from substituting $N(\rho)$ for k, and the third inequality uses $1 < 1+\rho$. If $k = 0$, then from Lemma 13 (Part D), $P^*_{1|0} > \frac{1}{5}\frac{1}{2}\frac{1}{P_0}\, e^{-\frac{t_1}{2\sigma^2}}\sqrt{1-\rho^2}$, where we notice that P_0 and $e^{-\frac{t_1}{2\sigma^2}}$ do not depend on ρ. Combining this and (4.37) with (4.36) it is easy to see that (4.35) holds for all ρ sufficiently close to one and $0 \leq k \leq N(\rho)$, which completes Step A1.

Step A2: Applying (4.33) with $l = k+1$, we obtain that for all ρ and $0 \leq k \leq N(\rho)$

$$P_{k+2|k} < R_k(\rho)\, g\Big(\frac{\rho t_{k+2} - t_{k+1}}{\sigma_\rho}\Big) Q\Big(\frac{t_{k+2} - t_{k+1}\rho}{\sigma_\rho}\Big) .$$

Using the same ideas (4.34) with $l = k+1$, we get that for all ρ sufficiently close to one and $0 \leq k \leq N(\rho)$

$$\begin{aligned} P^*_{k+1|k} &> \frac{1}{2} R_k(\rho)\, g\Big(\frac{\rho t_{k+2} - t_{k+1}}{\sigma_\rho}\Big) \left[Q\Big(\frac{\frac{t_{k+1}}{\rho} - t_{k+1}\rho}{\sigma_\rho}\Big) - Q\Big(\frac{t_{k+2} - t_{k+1}\rho}{\sigma_\rho}\Big) \right] \\ &> \frac{1}{2} R_k(\rho)\, g\Big(\frac{\rho t_{k+2} - t_{k+1}}{\sigma_\rho}\Big) \left[\frac{1}{4} - Q\Big(\frac{t_{k+2} - t_{k+1}\rho}{\sigma_\rho}\Big) \right] , \end{aligned} \tag{4.38}$$

where the second inequality follows from observing that $\frac{\frac{t_{k+1}}{\rho} - t_{k+1}\rho}{\sigma_\rho} > 0$ and $\lim^*_{\rho \to 1} \frac{\frac{t_{k+1}}{\rho} - t_{k+1}\rho}{\sigma_\rho} = 0$, and, therefore, the left Q function above goes to $\frac{1}{2}$ (from below) as ρ goes to one, uniformly for $0 \leq k \leq N(\rho)$. Consequently, for all ρ sufficiently close to one, the left Q function is larger than $\frac{1}{4}$.

The last two equations imply that it suffices to show that for all ρ sufficiently close to one and $0 \leq k \leq N(\rho)$

$$12 Q\Big(\frac{t_{k+2} - t_{k+1}\rho}{\sigma_\rho}\Big) < P^*_{k+1|k} . \tag{4.39}$$

Finally, using Fact 1, we observe that for all ρ sufficiently close to one and $0 \leq k \leq N(\rho)$

$$12Q\big(\frac{t_{k+2} - t_{k+1}\rho}{\sigma_\rho}\big) \;<\; 12\frac{1}{2}\, e^{-\frac{(t_{k+2}-t_{k+1}\rho)^2}{2\sigma_\rho^2}} \;<\; 6\, e^{-\frac{\lambda^2}{2(1-\rho^2)}} \;<\; P^*_{k+1|k}\ ,$$

where the last inequality was shown in Step A1. This concludes Step A2 and the proof of Part A.

Next, we consider Part B, whose proof is similar to that of Part A, yet not identical. The difference lies in the fact that B has $P^*_{k+1|k}$ in its expression instead of $P_{k-1|k}$. Part B follows easily from the following two steps.

B1: $P_{k-l-1|k} \;<\; P_{k-l|k}P^*_{k+1|k} \quad \text{for} \;\; l \geq 2$

B2: $P_{k-2|k} \;<\; (P^*_{k-1|k})^2$

Step B1: To keep notation compact, we rewrite what needs to be shown as: For all ρ sufficiently close to one and for $0 \leq k \leq N(\rho)$, $P_{l-1|k} < P_{l|k}P^*_{k+1|k}$ for $l \leq k-2$. We observe that $t_{l+1} < \frac{t_k}{\rho}$ when $\rho > \frac{1}{3}$, $0 \leq k \leq N(\rho)$, and $l \leq k-2$ (we require $\rho > \frac{1}{3}$ for the case that $k = 0$). Thus, since $f_{X_2|I_1}(x|k) < \tilde{f}_{X_2|I_1}(x|k)$ for all $x < \frac{t_k}{\rho}$, it follows using derivations similar to those of (4.33) that for all $\rho > \frac{1}{3}$ and $0 \leq k \leq N(\rho)$

$$P_{l-1|k} \;<\; L_k(\rho)\, g\big(\frac{t_k - \rho t_l}{\sigma_\rho}\big) Q\Big(\frac{t_k\rho - t_l}{\sigma_\rho}\Big)\ . \tag{4.40}$$

Additionally, using a derivation similar to that of (4.34) we obtain that for all ρ sufficiently close to one and $0 \leq k \leq N(\rho)$

$$P_{l|k} \;>\; \frac{1}{2}L_k(\rho)\, g\big(\frac{t_k - \rho t_l}{\sigma_\rho}\big) \left[Q\Big(\frac{t_k\rho - t_{l+1}}{\sigma_\rho}\Big) - Q\Big(\frac{t_k\rho - t_l-}{\sigma_\rho}\Big)\right]\ .$$

Using the above two equations together with similar steps to those used to obtain (4.35), it follows that it suffices to show

$$\frac{3Q\big(\frac{t_k\rho - t_l}{\sigma_\rho}\big)}{Q\big(\frac{t_k\rho - t_{l+1}}{\sigma_\rho}\big)} \;<\; P^*_{k+1|k}\ . \tag{4.41}$$

Following the same derivation as that which was used to obtain (4.36), we have that

$$\frac{3Q\big(\frac{t_k\rho-t_l}{\sigma_\rho}\big)}{Q\big(\frac{t_k\rho-t_{l+1}}{\sigma_\rho}\big)} \;<\; 6e^{-\frac{\lambda^2}{2(1-\rho^2)}} \;.$$

Combining this with (4.41) and with the lower bound to $P^*_{k+1|k}$ given in (4.37) and with the derivation thereafter, completes the proof of Step B1.

Step B2: Applying (4.40) with $l = k-1$, we obtain that for all $\rho > \frac{1}{3}$ and $0 \le k \le N(\rho)$

$$\begin{aligned} P_{k-2|k} \;&<\; L_k(\rho)\, g\big(\frac{t_k - t_{k-1}\rho}{\sigma_\rho}\big) Q\Big(\frac{t_k\rho - t_{k-1}}{\sigma_\rho}\Big) \\ &=\; L_k(\rho)\, g\big(\frac{(k-\frac{1}{2})\Delta(1-\rho)+\Delta\rho}{\sigma_\rho}\big) Q\Big(\frac{t_k\rho - t_{k-1}}{\sigma_\rho}\Big) \;. \end{aligned} \tag{4.42}$$

From (4.38) we have that for all ρ sufficiently close to one and $0 \le k \le N(\rho)$

$$\begin{aligned} P^*_{k+1|k} \;&>\; \frac{1}{2}R_k(\rho)\, g\big(\frac{\rho t_{k+2} - t_{k+1}}{\sigma_\rho}\big) \left[\frac{1}{4} - Q\Big(\frac{t_{k+2} - t_{k+1}\rho}{\sigma_\rho}\Big)\right] \\ &\overset{(a)}{>}\; \frac{1}{16}R_k(\rho)\, g\big(\frac{\rho t_{k+2} - t_{k+1}}{\sigma_\rho}\big) \\ &=\; \frac{1}{16}R_k(\rho)\, g\big(\frac{-(k+\frac{1}{2})\Delta(1-\rho)+\Delta\rho}{\sigma_\rho}\big) \;, \end{aligned} \tag{4.43}$$

where (a) follows from the fact that for the Q term tends to zero in the $\lim^*$ sense, thus, in particular, it is less than $1/8$ for all ρ sufficiently close to one. Next, since

$$g\big(\frac{(k-\frac{1}{2})\Delta(1-\rho)+\Delta\rho}{\sigma_\rho}\big) \;\le\; g\big(\frac{-(k+\frac{1}{2})\Delta(1-\rho)+\Delta\rho}{\sigma_\rho}\big) \;,$$

due to the monotonicity of g (Fact 6), it follows from (4.42) and (4.43) that it suffices to show

$$16\frac{L_k(\rho)}{R_k(\rho)}Q\Big(\frac{t_k\rho - t_{k-1}}{\sigma_\rho}\Big) \;<\; P^*_{k+1|k} \;. \tag{4.44}$$

Using the definitions of $L_k(\rho)$ and $R_k(\rho)$ we obtain that for all ρ sufficiently close

to one,

$$
\begin{aligned}
16\frac{L_k(\rho)}{R_k(\rho)}\,Q\Big(\frac{t_k\rho - t_{k-1}}{\sigma_\rho}\Big) \;&=\; 16\frac{\frac{1}{2}\frac{1}{P_k}e^{\frac{-t_k^2}{2\sigma_\rho^2}}\sqrt{1-\rho^2}}{\frac{1}{2}\frac{1}{P_k}e^{\frac{-t_{k+1}^2}{2\sigma_\rho^2}}\sqrt{1-\rho^2}}\,Q\Big(\frac{\big[(k-\frac{1}{2})\rho-(k-\frac{3}{2})\big]\Delta}{\sigma\sqrt{1-\rho^2}}\Big)\\
&=\; 16\,e^{k\lambda^2}\,Q\Big(\frac{\big(1-(k-\frac{1}{2})(1-\rho)\big)\lambda}{\sqrt{1-\rho^2}}\Big)\\
&\overset{(a)}{<}\; 16\,e^{k\lambda^2}\,Q\Big(\frac{\lambda}{2\sqrt{1-\rho^2}}\Big) \;\overset{(b)}{\leq}\; 16\,e^{k\lambda^2}\frac{1}{2}e^{-\frac{\lambda^2}{8(1-\rho^2)}}\\
&\overset{(c)}{<}\; 8\,e^{N(\rho)\lambda^2}e^{-\frac{\lambda^2}{8(1-\rho^2)}} \;<\; e^{(\ln\frac{1}{1-\rho})^{\frac{3}{4}}\lambda^2}\,e^{-\frac{\lambda^2}{8(1-\rho^2)}}\\
&<\; e^{(\ln\frac{1}{1-\rho})\lambda^2}\,e^{-\frac{\lambda^2}{8(1-\rho^2)}} \;=\; \big(\frac{1}{1-\rho}\big)^{\lambda^2}e^{-\frac{\lambda^2}{8(1-\rho^2)}}\;,
\end{aligned}
$$

where (a) follows from having $\lim^*(k-\frac{1}{2})(1-\rho)=0$, (b) is due to Fact 1, and (c) is obtained by substituting $N(\rho)$ for k.

Finally, combining the above equation with (4.44) and (4.37) we obtain that it suffices to show that for all ρ sufficiently close to one,

$$
\big(\frac{1}{1-\rho}\big)^{\lambda^2}e^{-\frac{\lambda^2}{8(1-\rho^2)}} \;<\; \frac{1}{5}(1-\rho)^{\lambda^2+\frac{1}{2}}\;,
$$

or equivalently,

$$
e^{-\frac{\lambda^2}{8(1-\rho^2)}} \;<\; \frac{1}{5}(1-\rho)^{2\lambda^2+\frac{1}{2}}\;,
$$

which is easily seen to be true for all ρ sufficiently close to one. This concludes Step B2 and the proof of the lemma. □

Proof of Lemma 18:

It follows from Lemma 12 (Part D) that it suffices to show

$$
\lim_{\rho\to 1}{}^*\frac{H(I_2|I_1=k)}{H_{k-1|k}+H_{k|k}+H_{k+1|k}} \;=\; 1\;.
$$

Combining the above with the fact that $H(I_2|I_1=k)=\big(H_{k-1|k}+H_{k|k}+H_{k+1|k}\big)+\sum_{l=-\infty}^{-2}H_{k+l|k}+\sum_{l=2}^{\infty}H_{k+l|k}$, and with Lemma 12 (Part B), it follows that it suffices

to show

$$\lim_{\rho\to 1}{}^{*}\frac{\sum_{l=-\infty}^{-2} H_{k+l|k} + \sum_{l=2}^{\infty} H_{k+l|k}}{H_{k-1|k} + H_{k|k} + H_{k+1|k}} \;=\; 0\;. \tag{4.45}$$

We proceed by upper bounding the second sum in the numerator above. In a similar way the same expression can be shown to be an upper bound for the first term. First, let us write the second sum more explicitly as $\sum_{l=2}^{\infty} H_{k+l|k} = -\sum_{l=2}^{\infty} P_{k+l|k}\log P_{k+l|k}$. Next, recall from Lemma 13 (Part C) that for all ρ sufficiently close to one, $P^*_{k+1|k} < \frac{\sqrt{2\pi}}{4\lambda}\sqrt{1-\rho^2}$ for $0 \le k \le N(\rho)$. Thus, for all ρ sufficiently close to one, $P^*_{k+1|k} < \frac{1}{e}$ for $0 \le k \le N(\rho)$. We will assume for the rest of the proof that $P^*_{k+1|k} < \frac{1}{e}$. Combining this with Fact 8 and with the fact that $P_{k+l|k} < (P^*_{k+1|k})^l$ for $l \ge 2$ and $0 \le k \le N(\rho)$, as shown in Lemma 17 (Part A), it follows that $-P_{k+l|k}\log P_{k+l|k} < -(P^*_{k+1|k})^l \log (P^*_{k+1|k})^l$ for $l \ge 2$ and $0 \le k \le N(\rho)$. Therefore,

$$\begin{aligned}
-\sum_{l=2}^{\infty} P_{k+l|k}\log P_{k+l|k} \;&<\; -\sum_{m=0}^{\infty}(P^*_{k+1|k})^{m+2}\log (P^*_{k+1|k})^{m+2} \\
&=\; -(P^*_{k+1|k})^2\big(\log P^*_{k+1|k}\big)\sum_{m=0}^{\infty}(m+2)(P^*_{k+1|k})^m \\
&=\; -(P^*_{k+1|k})^2\big(\log P^*_{k+1|k}\big)\left[\frac{P^*_{k+1|k}}{(1-P^*_{k+1|k})^2} + \frac{2}{1-P^*_{k+1|k}}\right] \\
&<\; -P^*_{k+1|k}\big(\log P^*_{k+1|k}\big)\,\frac{4P^*_{k+1|k}}{1-P^*_{k+1|k}}\;,
\end{aligned}$$

where the last equality follows from having $\frac{P^*_{k+1|k}}{(1-P^*_{k+1|k})^2} < \frac{2}{1-P^*_{k+1|k}}$, since $P^*_{k+1|k} < \frac{1}{e}$. As mentioned, $-\sum_{l=-\infty}^{-2} P_{k+l|k}\log P_{k+l|k}$ can be upper bounded by the same expression (using Lemma 17 (Part B)). Thus it follows that for all ρ sufficiently close to one and $0 \le k \le N(\rho)$

$$-\sum_{l=-\infty}^{-2} P_{k+l|k}\log P_{k+l|k} - \sum_{l=2}^{\infty} P_{k+l|k}\log P_{k+l|k} \;<\; -P^*_{k+1|k}(\log P^*_{k+1|k})\,\frac{8P^*_{k+1|k}}{1-P^*_{k+1|k}}\;.$$

Using this upper bound to the numerator of (4.45), we obtain that for all ρ sufficiently close to one

$$\begin{aligned}
\frac{\sum_{l=-\infty}^{-2} H_{k+l|k} + \sum_{l=2}^{\infty} H_{k+l|k}}{H_{k-1|k} + H_{k|k} + H_{k+1|k}} &< \sum_{k=0}^{N(\rho)} \frac{-P^*_{k+1|k}\big(\log P^*_{k+1|k}\big)\frac{8P^*_{k+1|k}}{1-P^*_{k+1|k}}}{-\sum_{l=k-1}^{k+1} P_{l|k}\log P_{l|k}} \\
&\overset{(a)}{<} \sum_{k=0}^{N(\rho)} \frac{-P^*_{k+1|k}\big(\log P^*_{k+1|k}\big)\frac{8P^*_{k+1|k}}{1-P^*_{k+1|k}}}{-P^*_{k+1|k}\log P^*_{k+1|k}} = \sum_{k=0}^{N(\rho)} \frac{8P^*_{k+1|k}}{1-P^*_{k+1|k}} \\
&\overset{(b)}{<} 16\sum_{k=0}^{N(\rho)} P^*_{k+1|k} \overset{(c)}{<} 16\sum_{k=0}^{N(\rho)} \frac{\sqrt{2\pi}}{4\lambda}\sqrt{1-\rho^2} \\
&= \left(\left\lfloor\left(\ln\frac{1}{1-\rho}\right)^{\frac{3}{4}} - \frac{1}{2}\right\rfloor + 1\right)\frac{4\sqrt{2\pi}}{\lambda}\sqrt{1-\rho^2} \\
&< \frac{8\sqrt{\pi}}{\lambda}\sqrt{1-\rho}\left[\left(\ln\frac{1}{1-\rho}\right)^{\frac{3}{4}} + \frac{1}{2}\right] \longrightarrow 0 \ \text{ as } \rho \longrightarrow 1 \ ,
\end{aligned}$$

where (a) derives from the fact that for all ρ sufficiently close to one and $0 \leq k \leq N(\rho)$, $-P^*_{k+1|k}\log P^*_{k+1|k} < -P_{k+1|k}\log P_{k+1|k}$, which follows from Fact 8, given that $P^*_{k+1|k} < P_{k+1|k}$ and $P_{k+1|k} < \frac{1}{e}$, as shown by Lemmas 14 and 13 (Part C). (b) is due to having $P^*_{k+1|k} < \frac{1}{e}$ and so $\frac{1}{(1-P^*_{k+1|k})^2} < 2$. Finally, (c) is obtained using Lemma 13 (Part C). This shows that (4.45) holds, and completes the proof of the lemma. $\square$

Proof of Lemma 19:

Let j be the cell in which lies the mean of the density f, i.e. $\mu \in S_j$. We rewrite f as follows:

$$f(x) = \begin{cases} f^L(x), & x < t_j \\ f^C(x), & t_j \leq x < t_{j+1} \\ f^R(x), & x \geq t_{j+1} \end{cases} .$$

where the functions $f^L(x), f^C(x)$ and $f^R(x)$ are zero outside of their respective regions. We now have that

$$H_q(f) \ \leq \ H_q(f^L) + H_q(f^C) + H_q(f^R) \ \leq \ H_q(f^L) + H_q(f^R) + 1 \ , \qquad (4.46)$$

where the first inequality follows from the fact that for those quantization cells where two of the three functions f^L, f^C and f^R are nonzero (i.e. the quantization cells containing t_j and t_{j+1}), the right-hand side is larger as shown by Fact 9. The second inequality follows by upper bounding $H_q(f^C)$ by $-\frac{1}{e}\log\frac{1}{e} < 1$.

Next, we upper $H_q(f^R)$. To keep notation short, we omit the parameter f from $P_l(f)$ throughout this lemma, and stress that P_l should not be confused with the general term P_k that represents the probability of a Gaussian random variable with *zero* mean and variance σ^2 of lying in S_k.

From Lemma 16 (Part A) we have that $P_{j+l+1} < \frac{1}{2}P_{j+l}$ for $l \geq B+2$, where $B = \lceil \frac{\sigma^2}{\Delta^2}\ln 3 \rceil$. Since $P_{j+B+2} < 1$, it follows that $P_{j+l} < \frac{1}{e}$ for $l \geq B+4$. Combining this with Fact 8 implies that $-P_{j+l+1}\log P_{j+l+1} < -\frac{1}{2}P_{j+l}\log(\frac{1}{2}P_{j+l})$ for $l \geq B+4$. We use this fact in the following:

$$\begin{aligned}
-\sum_{l=B+4}^{\infty} P_{j+l}\log P_{j+l} &< -\sum_{l=0}^{\infty}(\frac{1}{2})^l P_{j+B+4}\log\big((\frac{1}{2})^l P_{j+B+4}\big) \\
&< P_{j+B+4}\left[-\sum_{l=0}^{\infty} l\big(\frac{1}{2}\big)^l\log\frac{1}{2} - \sum_{l=0}^{\infty}\big(\frac{1}{2}\big)^l\log P_{j+B+4}\right] \\
&= P_{j+B+4}\left[2 - 2\log P_{j+B+4}\right] \\
&< 2 - 2\frac{1}{e}\log\frac{1}{e}\ , \qquad (4.47)
\end{aligned}$$

where the last inequality follows from the fact that $-p\log p$ is maximized at $p = \frac{1}{e}$ (Fact 7) and that $P_{j+B+4} < 1$. Using the above we obtain that

$$\begin{aligned}
H_q(f^R) &= -\sum_{l=1}^{\infty} P_{j+l}\log P_{j+l} \\
&= -\sum_{l=1}^{B+3} P_{j+l}\log P_{j+l} - \sum_{l=B+4}^{\infty} P_{j+l}\log P_{j+l} \\
&< -(B+3)\frac{1}{e}\log\frac{1}{e} + \big(2 - 2\frac{1}{e}\log\frac{1}{e}\big) \ < \ B+5\ . \qquad (4.48)
\end{aligned}$$

Finally, we upper bound $H_q(f^L)$. The derivation is very similar to the case of $H_q(f^R)$, and so some of the details will be skipped. From Lemma 16 (Part B) we have that $P_{j-l-1} < \frac{1}{2}P_{j-l}$ for $l \geq B$. Since $P_{j-B} < 1$, it follows that $P_{j-l} < \frac{1}{e}$ for $l \geq B+2$. Combining this with Fact 8 implies that $-P_{j-l-1}\log P_{j-l-1} < -\frac{1}{2}P_{j-l}\log(\frac{1}{2}P_{j-l})$ for $l \geq B+2$. Using this fact and an almost identical derivation to that of (4.47) shows that $-\sum_{l=B+2}^{\infty} P_{j-l}\log P_{j-l} < 2 - 2\frac{1}{e}\log\frac{1}{e}$. Consequently,

$$\begin{aligned} H_q(f^L) &\leq -\sum_{l=1}^{\infty} P_{j-l}\log P_{j-l} \\ &= -\sum_{l=1}^{B+1} P_{j-l}\log P_{j-l} - \sum_{l=B+2}^{\infty} P_{j-l}\log P_{j-l} \\ &< -(B+1)\frac{1}{e}\log\frac{1}{e} + \big(2 - 2\frac{1}{e}\log\frac{1}{e}\big) \;<\; B+4\;. \end{aligned}$$

Combining (4.46), (4.48), and the above concludes the proof of the lemma. □

Proof of Lemma 20:

We start by recalling from (4.10) that

$$f_{X_2|I_1}(x|k) \;=\; \int_{t_k}^{t_{k+1}} f_{X_2|X_1}(x|y)\frac{f_{X_1}(y)}{P_k}\,dy\;,$$

where $f_{X_2|X_1}(x|y)$ is a Gaussian density whose mean is ρy. Thus, it is not hard to see that for $x < t_k\rho$,

$$f_{X_2|I_1}(x|k) \;<\; \int_{t_k}^{t_{k+1}} f_{X_2|X_1}(x|t_k)\frac{f_{X_1}(y)}{P_k}\,dy \;=\; f_{X_2|X_1}(x|t_k) \stackrel{\Delta}{=} \bar{f}_k^L(x)\;.$$

Note that for tractability we let $\bar{f}_k^L(x)$ be defined as above for all x and not just $x < t_k\rho$. Similarly, for $x > t_{k+1}\rho$, we have

$$f_{X_2|I_1}(x|k) \;<\; \int_{t_k}^{t_{k+1}} f_{X_2|X_1}(x|t_{k+1})\frac{f_{X_1}(y)}{P_k}\,dy \;=\; f_{X_2|X_1}(x|t_{k+1}) \stackrel{\Delta}{=} \bar{f}_k^R(x)\;.$$

As before, we let $\bar{f}_k^R(x)$ be defined as above for all x and not just $x > t_{k+1}\rho$.

Letting $f_k^C(x) = f_{X_2|I_1}(x|k)$ for $t_k\rho \leq x \leq t_{k+1}\rho$, and zero for all other x's, we define

$$\bar{f}_k(x) \;=\; \begin{cases} \bar{f}_k^L(x), & x < t_k\rho \\ \bar{f}_k^C(x), & t_k\rho \leq x \leq t_{k+1}\rho \\ \bar{f}_k^R(x), & t_{k+1}\rho < x \end{cases} .$$

We rewrite $f_{X_2|I_1}(x|k)$ in a similar manner. Specifically,

$$f_{X_2|I_1}(x|k) \;=\; \begin{cases} f_k^L(x), & x < t_k\rho \\ f_k^C(x), & t_k\rho \leq x \leq t_{k+1}\rho \\ f_k^R(x), & t_{k+1}\rho < x \end{cases} ,$$

where f_k^L, f_k^C and f_k^R are zero outside their respective regions.

Next, we have that

$$H(I_2|I_1 = k) \;=\; H_q(f_{X_2|I_1=k}) \;<\; H_q(f_k^L) + H_q(f_k^C) + H_q(f_k^R) \;, \tag{4.49}$$

where the inequality follows from the fact that for those quantization cells where two of the three functions f_k^L, f_k^C and f_k^R are nonzero (i.e. the quantization cells containing $t_k\rho$ and $t_{k+1}\rho$), the right-hand side is larger as shown by Fact 9. Similarly,

$$H_q(\bar{f}_k) \;<\; H_q(\bar{f}_k^L) + H_q(\bar{f}_k^C) + H_q(\bar{f}_k^R) \;, \tag{4.50}$$

where the inequality follows from the same reason as in (4.49). We notice, however, that since $\bar{f}_k^L$ and $\bar{f}_k^R$ are nonzero outside the regions $x < t_k\rho$ and $x > t_{k+1}\rho$, respectively, it follows that for all cells there are at least two functions that are nonzero, and in the cells containing $t_k\rho$ and $t_{k+1}\rho$ three functions are nonzero. In all cases, Fact 9 implies that the right hand side is larger.

Clearly $H_q(f_k^C) = H_q(\bar{f}_k^C)$, since the two functions are equal. We claim that $H_q(f_k^L) < H_q(\bar{f}_k^L) - 3\frac{1}{e}\log\frac{1}{e}$ and that $H_q(f_k^R) < H_q(\bar{f}_k^R) - 3\frac{1}{e}\log\frac{1}{e}$. We will show the

first claim and note that the second claim follows via same arguments. By definition $f_k^L < \bar{f}_k^L$. Thus, letting n be such that $t_k\rho \in S_n$, and letting $\bar{P}_{l|k} \stackrel{\Delta}{=} \int_{S_l} \bar{f}_k(x)\,dx$, we have that $P_{l|k} < \bar{P}_{l|k}$, for any $l < n$. For those l's for which $\bar{P}_{l|k} < \frac{1}{e}$, Fact 8 implies that $P_{l|k}\log P_{l|k} < \bar{P}_{l|k}\log\bar{P}_{l|k}$. Since $\bar{f}_k^L$ is a pdf, it follows that there can be at most two cells for which $\bar{P}_{l|k} > \frac{1}{e}$. Thus, for all $l < n$, except for at most two cells, we have $P_{l|k}\log P_{l|k} < \bar{P}_{l|k}\log\bar{P}_{l|k}$. The contribution to $H_q f_k^L$ of those cells for which $P_{l|k}\log P_{l|k} \geq \bar{P}_{l|k}\log\bar{P}_{l|k}$ is upper bounded by $-2\frac{1}{e}\log\frac{1}{e}$, where $\max_p\{-p\log p\} = \frac{1}{e}\log\frac{1}{e}$. Lastly, the contribution of cell n is upper bounded by $-\frac{1}{e}\log\frac{1}{e}$. Consequently, we obtain that $H_q f_k^L < H_q\bar{f}_k^L - 3\frac{1}{e}\log\frac{1}{e}$. Therefore, we may write

$$H(I_2|I_1=k) \;<\; H_q(\bar{f}_k) - 6\frac{1}{e}\log\frac{1}{e}\;. \tag{4.51}$$

We proceed by showing that $H_q(\bar{f}_k) = H_q(\bar{f}_k^L) + H_q(\bar{f}_k^C) + H_q(\bar{f}_k^R)$ is uniformly bounded in k. We upper bound each of the three terms. Consider first $H_q(\bar{f}_k^C)$. Since $\bar{f}_k^C$ is non zero in at most two cells, it follows that

$$H_q(\bar{f}_k^C) \;\leq\; -2\frac{1}{e}\log\frac{1}{e} \;<\; 2\;. \tag{4.52}$$

Next, since $\bar{f}_k^L$ and $\bar{f}_k^R$ are Gaussian densities, we may apply Lemma 19 and get that that $H_q(\bar{f}_k^L) < 2B + 10$ and $H_q(\bar{f}_k^R) < 2B + 10$, where $B = \lceil\frac{\sigma^2(1-\rho^2)}{\Delta^2}\ln 3\rceil < \lceil\frac{\ln 3}{\lambda^2}\rceil < \frac{\ln 3}{\lambda^2} + 1$. Thus, we have that for all ρ, $H_q(\bar{f}_k^L) < \frac{2\ln 3}{\lambda^2} + 12$ and $H_q(\bar{f}_k^R) < \frac{2\ln 3}{\lambda^2} + 12$. Combining this with (4.52) and with (4.50), we obtain that

$$H_q(\bar{f}_k) \;<\; \frac{2\ln 3}{\lambda^2} + 12 + 2 + \frac{2\ln 3}{\lambda^2} + 12 \;=\; \frac{4\ln 3}{\lambda^2} + 26\;,$$

The above together with (4.51) concludes the proof of the lemma. □

Proof of Lemma 21:

We begin by lower and upper bounding the expression in the lemma statement.

$$\begin{aligned} 1 &< \frac{H(I_2|I_1)}{2\sum_{k=1}^{N(\rho)} H(I_2|I_1=k)P_k + H(I_2|I_1=0)P_0} \\ &= 1 + \frac{2\sum_{k=N(\rho)+1}^{\infty} H(I_2|I_1=k)P_k}{2\sum_{k=1}^{N(\rho)} H(I_2|I_1=k)P_k + H(I_2|I_1=0)P_0} \\ &< 1 + \frac{2\sum_{k=N(\rho)+1}^{\infty} H(I_2|I_1=k)P_k}{H(I_2|I_1=0)P_0} \ . \end{aligned} \tag{4.53}$$

It suffices to show that $\lim_{\rho\to 1} \frac{2\sum_{k=N(\rho)+1}^{\infty} H(I_2|I_1=k)P_k}{H(I_2|I_1=0)P_0} = 0$. To show this, we upper bound the numerator and lower bound the denominator. Consider the numerator first. For all ρ sufficiently close to one,

$$\begin{aligned} \sum_{k=N(\rho)+1}^{\infty} 2H(I_2|I_1=k)P_k &\overset{(a)}{<} 2M \sum_{k=N(\rho)+1}^{\infty} P_k \ = \ 2MQ\big(\frac{t_{N(\rho)+1}}{\sigma}\big) \\ &< M\, e^{-\frac{\left(\left\lfloor (\ln\frac{1}{1-\rho})^{\frac{3}{4}} - \frac{1}{2}\right\rfloor + \frac{1}{2}\right)^2 \Delta^2}{2\sigma^2}} \ < \ M\, e^{-\frac{\left(\left(\ln\frac{1}{1-\rho}\right)^{\frac{3}{4}} - 1\right)^2 \lambda^2}{2}} \\ &\overset{(b)}{<} M\, e^{-\frac{\left(\frac{1}{2}\left(\ln\frac{1}{1-\rho}\right)^{\frac{3}{4}}\right)^2 \lambda^2}{2}} \ = \ M(1-\rho)^{\frac{\lambda^2\sqrt{\ln\frac{1}{1-\rho}}}{8}} \ , \end{aligned} \tag{4.54}$$

where (a) follows from Lemma 20 (with M being the constant given in the lemma), and (b) uses the fact that for all ρ sufficiently close to one, $\frac{1}{2}\big(\ln\frac{1}{1-\rho}\big)^{\frac{3}{4}} < \big(\ln\frac{1}{1-\rho}\big)^{\frac{3}{4}} - 1$.

Next, we lower bound the denominator. Specifically, for all ρ sufficiently close to one,

$$\begin{aligned} H(I_2|I_1=0)P_0 &= \big(-\sum_{l=-\infty}^{\infty} P_{l|0}\log P_{l|0}\big)P_0 \ > \ \big(-P_{1|0}\log P_{1|0}\big)P_0 \\ &\overset{(a)}{>} \big(-P^*_{1|0}\log P^*_{1|0}\big)P_0 \overset{(b)}{>} \Big(\frac{1}{5}R_0(\rho)\log\big(\frac{1}{5}R_0(\rho)\Big)P_0 \\ &= \left[-\frac{1}{5}\frac{1}{2}\frac{1}{P_0}e^{-\frac{(\frac{1}{2})^2\lambda^2}{2}}\sqrt{1-\rho^2}\log\left(\frac{1}{5}\frac{1}{2}\frac{1}{P_0}e^{-\frac{(\frac{1}{2})^2\lambda^2}{2}}\sqrt{1-\rho^2}\right)\right]P_0 \\ &\overset{(c)}{>} \frac{1}{10}e^{-\frac{\lambda^2}{8}}\sqrt{1-\rho}\ , \end{aligned} \tag{4.55}$$

where (a) follows from having $P^*_{1|0} < P_{1|0}$ and $P_{1|0} < \frac{1}{e}$ for all ρ sufficiently close to one, which imply (using Fact 8) that $-P^*_{1|0}\log P^*_{1|0} < -P_{1|0}\log P_{1|0}$; (b) is due to

Lemma 13 (Part D); and (c) derives from having $-\log\left(\frac{1}{5}\frac{1}{2}\frac{1}{P_0}e^{-\frac{(\frac{1}{2})^2\lambda^2}{2}}\sqrt{1-\rho^2}\right) > 1$, for all ρ sufficiently close to one, and having $1 < \sqrt{1+\rho}$.

Finally, plugging the upper bound given by (4.54) and the lower bound given by (4.55) into the last fraction in (4.53), we obtain that for all ρ sufficiently close to one,

$$\frac{2\sum_{k=N(\rho)+1}^{\infty} H(I_2|I_1=k)P_k}{H(I_2|I_1=0)P_0} \;<\; \frac{M(1-\rho)^{\frac{\lambda^2\sqrt{\ln\frac{1}{1-\rho}}}{8}}}{\frac{1}{10}\,e^{-\frac{\lambda^2}{8}}\sqrt{1-\rho}} \;=\; 10M\,e^{\frac{\lambda^2}{8}}(1-\rho)^{\frac{\lambda^2\sqrt{\ln\frac{1}{1-\rho}}}{8}-\frac{1}{2}}$$
$$\longrightarrow 0 \;\text{ as } \rho \longrightarrow 1\;, \tag{4.56}$$

which concludes the proof of the lemma. □

Proof of Lemma 22:

It follows from Lemma 12 (Part D) that it suffices to show

$$\lim_{\rho\to 1}{}^{*}\,\frac{H_{k-1|k}+H_{k|k}+H_{k+1|k}}{H_{k-1|k}+H_{k+1|k}} \;=\; 1\;.$$

Next, using Lemma 12 (Part B) and Fact 9, we obtain that it is enough to show

$$\lim_{\rho\to 1}{}^{*}\frac{\mathcal{H}(P_{k|k})}{\mathcal{H}(P_{k-1|k}+P_{k+1|k})} \;=\; 0\;. \tag{4.57}$$

In order to show that (4.57) holds, we write $P_{k|k} = 1 - P_{k-1|k} - P_{k+1|k} - P_{t|k} = 1 - \widetilde{P}_k - P_{t|k}$, where $\widetilde{P}_k \stackrel{\Delta}{=} P_{k-1|k} + P_{k+1|k}$ and $P_{t|k} \stackrel{\Delta}{=} \sum_{l=-\infty}^{-2} P_{k+l|k} + \sum_{l=2}^{\infty} P_{k+l|k}$. With this notation (4.57) becomes

$$\lim_{\rho\to 1}{}^{*}\frac{\mathcal{H}(1-\widetilde{P}_k-P_{t|k})}{\mathcal{H}(\widetilde{P}_k)} \;=\; 0\;. \tag{4.58}$$

We proceed by upper bounding the numerator in (4.58). To do so, we will upper bound $P_{t|k}$ in terms of $\widetilde{P}_k$ and then use Fact 8. From Lemma 17 we have that for all ρ sufficiently close to one, and for $0 \le k \le N(\rho)$,

$$P_{t|k} \;=\; \sum_{l=-\infty}^{k-2} P_{l|k} + \sum_{l=k+2}^{\infty} P_{l|k} \;<\; 2\sum_{l=0}^{\infty}(P^{*}_{k+1|k})^{l+2} \;=\; \frac{2(P^{*}_{k+1|k})^2}{1-P^{*}_{k+1|k}}\;,$$

and consequently,

$$\frac{P_{t|k}}{\widetilde{P}_k} \;<\; \frac{\frac{2(P^*_{k+1|k})^2}{1-P^*_{k+1|k}}}{P^*_{k+1|k}} \;=\; \frac{2P^*_{k+1|k}}{1-P^*_{k+1|k}} \;.$$

Since Lemma 13 (Part C) implies that $\lim^*_{\rho\to 1} P^*_{k+1|k} = 0$, it follows that for all ρ sufficiently close to one, and for $0 \le k \le N(\rho)$, $P_{t|k} < \frac{1}{2}\widetilde{P}_k$.

Next, we would like to use Fact 8 in order to get an upper bound for the numerator in (4.58). We now show that the conditions required by Fact 8 are indeed met. Lemma 12 (Part B) and Lemma 15 imply that $\lim^*_{\rho\to 1} \widetilde{P}_k = 0$. Therefore, for all ρ sufficiently close to one, and for $0 \le k \le N(\rho)$, $1 - \widetilde{P}_k - P_{t|k} > 1 - \widetilde{P}_k - \frac{1}{2}\widetilde{P}_k > \frac{1}{e}$. Consequently, it follows from Fact 8 that $\mathcal{H}(1-\widetilde{P}_k - P_{t|k}) < \mathcal{H}(1-\frac{3}{2}\widetilde{P}_k)$. Combining this with (4.58), it follows that it is enough to show

$$\lim_{\rho\to 1}{}^* \frac{\mathcal{H}(1-\frac{3}{2}\widetilde{P}_k)}{\mathcal{H}(\widetilde{P}_k)} = 0 \;. \tag{4.59}$$

Finally, (4.59) can be seen to hold using the fact, shown earlier, that $\lim^*_{\rho\to 1} \widetilde{P}_k = 0$, together with Lemmas A2 and 12 (Part F). This concludes the proof of the lemma. □

Proof of Lemma 23:

It follows from Lemma 12 (Part D) that it suffices to show

$$\lim_{\rho\to 1}{}^* \frac{(H_{k-1|k} + H_{k+1|k})P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)\big(M_{L,\lambda}(k) + M_{R,\lambda}(k)\big)} \;=\; 1 \;.$$

Next, applying Lemma 12 (Part E) it follows that it suffices to show the following.

$$A.\quad \lim_{\rho\to 1}{}^* \frac{H_{k-1|k}\,P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)\,M_{L,\lambda}(k)} \;=\; 1 \;,$$

$$B.\quad \lim_{\rho\to 1}{}^* \frac{H_{k+1|k}\,P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)\,M_{R,\lambda}(k)} \;=\; 1 \;.$$

To show A we write

$$\frac{H_{k-1|k}\,P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)M_{L,\lambda}(k)} \;=\; \frac{H_{k-1|k}\,P_k}{\mathcal{H}\big(\frac{1}{\pi}L_k(\rho)\,P_k\big)}\;\frac{\mathcal{H}\big(\frac{1}{\pi}L_k(\rho)\,P_k\big)}{\mathcal{H}(\sqrt{1-\rho^2}\,)\,M_{L,\lambda}(k)} \;.$$

Lemmas 13 (Part A), A3, and 12 (Part G) (with $G(x,y) = \frac{\mathcal{H}(x)}{\mathcal{H}(y)}$) show that the first term on the right-hand side has $\lim^*$ equal one. Therefore by Lemma 12 (Part C), it suffices to show

$$\lim_{\rho\to 1}^{*} \frac{\mathcal{H}\big(\frac{1}{\pi}L_k(\rho)\big)P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)\,M_{L,\lambda}(k)} \;=\; 1\;. \tag{4.60}$$

Similarly, to show B we write

$$\frac{H_{k+1|k}\,P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)M_{R,\lambda}(k)} \;=\; \frac{H_{k+1|k}\,P_k}{\mathcal{H}\big(\frac{1}{\pi}R_k(\rho)\,P_k\big)}\;\frac{\mathcal{H}\big(\frac{1}{\pi}R_k(\rho)\,P_k\big)}{\mathcal{H}(\sqrt{1-\rho^2}\,)\,M_{R,\lambda}(k)}\;.$$

Lemmas 12 (Parts C and G), 13 (Part B), 14, and A3 show that the first term on the right-hand side has $\lim^*$ equal one. Therefore by Lemma 12 (Part C), it suffices to show

$$\lim_{\rho\to 1}^{*} \frac{\mathcal{H}\big(\frac{1}{\pi}R_k(\rho)\big)P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)\,M_{R,\lambda}(k)} \;=\; 1\;. \tag{4.61}$$

Our focus is now on showing (4.60) and (4.61), which will prove the lemma. We begin with (4.60). To show (4.60), we observe that by the definitions of $M_{L,\lambda}(k)$ and $L_k(\rho)$ one can show that

$$\frac{\mathcal{H}\big(\frac{1}{\pi}L_k(\rho)\big)P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)\,M_{L,\lambda}(k)} \;=\; \frac{-\frac{1}{\pi}L_k(\rho)\log\big(\frac{1}{\pi}L_k(\rho)\big)}{-\frac{1}{\pi}L_k(\rho)\log\sqrt{1-\rho^2}} \;=\; \frac{\log\big(\frac{1}{\pi}L_k(\rho)\big)}{\log\sqrt{1-\rho^2}} \;=\; 1+\frac{\log\big(\frac{1}{\pi}\frac{L_k(\rho)}{\sqrt{1-\rho^2}}\big)}{\log\sqrt{1-\rho^2}}\;.$$

We now show that the fraction on the right-hand side above converges to zero as $\rho\to 1$ in the $\lim^*$ sense, from which (4.60) will follow via Lemma 12 (Part B).

Consider first $N_\lambda < k \le N(\rho)$, where N_λ is as defined in Lemma 10. For all ρ sufficiently close to one,

$$\begin{aligned}\frac{\log\big(\frac{1}{\pi}\frac{L_k(\rho)}{\sqrt{1-\rho^2}}\big)}{\big|\log\sqrt{1-\rho^2}\,\big|} &\overset{(a)}{<} \frac{\log\big(\sqrt{\tfrac{2}{\pi}}k\lambda\big)}{\big|\log\sqrt{1-\rho^2}\,\big|} \overset{(b)}{<} \frac{\log\big(\lambda\sqrt{\tfrac{2}{\pi}}\big\lfloor\big(\ln\frac{1}{1-\rho}\big)^{\frac{3}{4}}-\frac{1}{2}\big\rfloor\big)}{\big|\log\sqrt{2(1-\rho)}\,\big|}\\ &\overset{(c)}{<} \frac{\log\big(\lambda\sqrt{\frac{2}{\pi}}\big)}{\big|\frac{1}{2}\log\left[2(1-\rho)\right]\big|} + \frac{\frac{3}{4}\log\big(\ln\frac{1}{1-\rho}\big)}{\big|\frac{1}{2}\log\left[2(1-\rho)\right]\big|} \;\longrightarrow 0 \;\text{ as }\rho\longrightarrow 1\;,\end{aligned} \tag{4.62}$$

where (a) is due to Lemma 10 (Part A) and observing that since $k > N_\lambda > \frac{2}{\lambda}$ the logarithm in the numerator is positive for all such k's, (b) is by upper bounding k by

$N(\rho)$, and $\sqrt{1-\rho^2}$ by $\sqrt{2(1-\rho)}$, where the latter is less than one for all ρ sufficiently close to one, and (c) is from having $\lfloor \big(\ln \frac{1}{1-\rho} \big)^{\frac{3}{4}} - \frac{1}{2} \rfloor > 1$ for all ρ sufficiently close to one. We also have from Lemma 10 (Part B), which can be used since $k > N_\lambda \geq 1$, that $\frac{1}{\pi} \frac{L_k(\rho)}{\sqrt{1-\rho^2}} > \frac{1}{\pi}$. Thus,

$$\frac{\log \big(\frac{1}{\pi} \frac{L_k(\rho)}{\sqrt{1-\rho^2}} \big)}{\big| \log \sqrt{1-\rho^2} \, \big|} \;\geq\; \frac{\log \frac{1}{\pi}}{\big| \log \sqrt{1-\rho^2} \, \big|} \;\longrightarrow 0 \;\text{ as } \rho \longrightarrow 1 \;. \tag{4.63}$$

Consider now $0 \leq k \leq N_\lambda$. For such k's, $\log \big(\frac{1}{\pi} \frac{L_k(\rho)}{\sqrt{1-\rho^2}} \big) = \log \big(\frac{1}{2\pi} \frac{1}{P_k} e^{-\frac{t_k^2}{2\sigma^2}} \big)$ can assume only finitely many values. Therefore, it follows that

$$\frac{\log \big(\frac{1}{\pi} \frac{L_k(\rho)}{\sqrt{1-\rho^2}} \big)}{\log \sqrt{1-\rho^2}} \;\longrightarrow 0 \;\text{ as } \rho \longrightarrow 1 \quad \text{uniformly for } \; 0 \leq k \leq N_\lambda \;.$$

Equations (4.62) and (4.63) show that

$$\frac{\log \big(\frac{1}{\pi} \frac{L_k(\rho)}{\sqrt{1-\rho^2}} \big)}{\log \sqrt{1-\rho^2}} \;\longrightarrow 0 \;\text{ as } \rho \longrightarrow 1 \quad \text{uniformly for } \; N_\lambda < k \leq N(\rho) \;.$$

Together, the two previous equations imply

$$\lim_{\rho \to 1}{}^{*} \frac{\log \big(\frac{1}{\pi} \frac{L_k(\rho)}{\sqrt{1-\rho^2}} \big)}{\log \sqrt{1-\rho^2}} \;=\; 0 \;,$$

which completes the proof of (4.60).

It remains to show (4.61). We observe that by the definitions of $M_{R,\lambda}(k)$ and $R_k(\rho)$ one can show that

$$\frac{\mathcal{H}\big(\frac{1}{\pi} R_k(\rho)\big) P_k}{\mathcal{H}(\sqrt{1-\rho^2}\,)\, M_{R,\lambda}(k)} \;=\; \frac{-\frac{1}{\pi} R_k(\rho) \log \big(\frac{1}{\pi} R_k(\rho) \big)}{-\frac{1}{\pi} R_k(\rho) \log \sqrt{1-\rho^2}} \;=\; \frac{\log \big(\frac{1}{\pi} R_k(\rho) \big)}{\log \sqrt{1-\rho^2}} \;=\; 1 + \frac{\log \big(\frac{1}{\pi} \frac{R_k(\rho)}{\sqrt{1-\rho^2}} \big)}{\log \sqrt{1-\rho^2}} \;.$$

We now show that the fraction on the right-hand side above converges to zero as $\rho \to 1$ in the $\lim^{*}$ sense, from which (4.61) will follow via Lemma 12 (Part B).

On the one hand we have from Lemma 10 (Part C) that $\frac{1}{\pi} \frac{R_k(\rho)}{\sqrt{1-\rho^2}} < \frac{1}{\lambda \sqrt{2\pi}}$, for $k \geq 0$. On the other hand Lemma 10 (Part D) implies that for $1 \leq k \leq N(\rho)$,

$$\frac{1}{\pi} \frac{R_k(\rho)}{\sqrt{1-\rho^2}} \;>\; \frac{1}{\pi} e^{-k\lambda^2} \;\geq\; \frac{1}{\pi} e^{-N(\rho)\lambda^2} \;=\; \frac{1}{\pi} e^{-\lfloor (\ln \frac{1}{1-\rho})^{\frac{3}{4}} - \frac{1}{2} \rfloor \lambda^2} \;>\; \frac{1}{\pi} (1-\rho)^{\lambda^2 (\ln \frac{1}{1-\rho})^{-\frac{1}{4}}} \;.$$

Therefore, it follows that

$$\frac{\log\left(\frac{1}{\pi}\frac{R_k(\rho)}{\sqrt{1-\rho^2}}\right)}{\log\sqrt{1-\rho^2}} \;\longrightarrow 0 \;\text{ as } \rho\longrightarrow 1 \quad \text{uniformly for } \; 1\leq k\leq N(\rho)\;.$$

Since the above also hold for $k=0$, it follows that

$$\lim_{\rho\to 1}{}^{*}\frac{\log\left(\frac{1}{\pi}\frac{R_k(\rho)}{\sqrt{1-\rho^2}}\right)}{\log\sqrt{1-\rho^2}} \;=\; 0\;,$$

which completes the proof of (4.61) and concludes the proof of the lemma. □

Proof of Lemma 24:

It needs to be shown that

$$\lim_{\rho\to 1}\frac{\mathcal{H}(\sqrt{1-\rho^2}\,)\Big[2\sum_{k=1}^{N(\rho)}\big(M_{L,\lambda}(k)+M_{R,\lambda}(k)\big)+\big(M_{L,\lambda}(0)+M_{R,\lambda}(0)\big)\Big]}{M_\lambda\mathcal{H}(\sqrt{1-\rho}\,)} \;=\; 1\;.$$

First we observe that

$$\frac{\mathcal{H}(\sqrt{1-\rho^2}\,)}{\mathcal{H}(\sqrt{1-\rho}\,)} \;=\; \frac{-\sqrt{1-\rho^2}\,\log\sqrt{1-\rho^2}}{-\sqrt{1-\rho}\,\log\sqrt{1-\rho}} \;\longrightarrow\sqrt{2}\;\text{ as }\rho\longrightarrow 1\;.$$

Therefore, it needs to be shown that

$$\lim_{\rho\to 1}\sqrt{2}\Big[2\sum_{k=1}^{N(\rho)}\big(M_{L,\lambda}(k)+M_{R,\lambda}(k)\big)+\big(M_{L,\lambda}(0)+M_{R,\lambda}(0)\big)\Big] \;=\; M_\lambda\;.$$

Recalling that $M_{L,\lambda}(k) = \frac{1}{2\pi}e^{-\frac{(k-\frac{1}{2})^2\lambda^2}{2}}$, $M_{R,\lambda}(k) = \frac{1}{2\pi}e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}}$, and $M_\lambda = \frac{2\sqrt{2}}{\pi}\sum_{k=0}^{\infty}e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}}$, the above can straightforwardly be shown in the following way:

$$\begin{aligned}
&2\sqrt{2}\sum_{k=1}^{N(\rho)}\big(M_{L,\lambda}(k)+M_{R,\lambda}(k)\big)+\sqrt{2}\big(M_{L,\lambda}(0)+M_{R,\lambda}(0)\big)\\
&\qquad= \;2\sqrt{2}\sum_{k=1}^{N(\rho)}\frac{1}{2\pi}\big(e^{-\frac{(k-\frac{1}{2})^2\lambda^2}{2}}+e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}}\big)+\frac{2\sqrt{2}}{2\pi}\,e^{-\frac{\lambda^2}{8}}\\
&\qquad= \;\sum_{k=0}^{N(\rho)}\frac{\sqrt{2}}{\pi}\,e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}}+\sum_{k=1}^{N(\rho)}\frac{\sqrt{2}}{\pi}\,e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}}+\frac{\sqrt{2}}{\pi}\,e^{-\frac{\lambda^2}{8}}\\
&\qquad= \;\frac{2\sqrt{2}}{\pi}\sum_{k=0}^{N(\rho)}e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}}\;\stackrel{\rho\to 1}{\longrightarrow}\;\frac{2\sqrt{2}}{\pi}\sum_{k=0}^{\infty}e^{-\frac{(k+\frac{1}{2})^2\lambda^2}{2}} \;=\; M_\lambda\;.
\end{aligned}$$

□

4.7 Conclusions

This chapter considered the behavior of the entropy of highly correlated quantized data. The chapter consisted of two parts. In the first part we examined the case that a stationary random process is sampled over some finite interval, and each sample was separately quantized with arbitrary, yet identical, scalar quantizers. The question that was raised is what happens to the joint entropy of these quantized samples as the sampling interval goes to zero? The answer is not obvious, since the joint entropy of N quantized samples can be written as $H(I_1^\tau, I_2^\tau, \ldots, I_{N_\tau}^\tau) = N_\tau \frac{H(I_1^\tau, I_2^\tau, \ldots, I_{N_\tau}^\tau)}{N_\tau}$, where the first term in the product tends to infinity as the sampling interval τ goes to zero, and the second term tends to zero. Hence, the answer to the posed question could conceivably be zero, some other finite value, or infinity. Theorem 3 showed that the latter is the case when the random process is stationary and crosses a quantization threshold with positive probability.

The second part of the paper was concerned with establishing an upper bound to the rate at which the joint entropy above tends to infinity as the sampling interval goes to zero. This upper bound was obtained by upper bounding high order conditional entropies by first order conditional entropy, namely, $H(I_k^\tau | I_1^\tau, I_2^\tau, \ldots, I_k^\tau) \leq H(I_1^\tau | I_0^\tau)$. A simple asymptotic formula was derived (see Theorem 7) for the first order conditional entropy in the case of infinite-level uniform threshold quantizers and a stationary Gaussian random process whose mean lies at a midpoint of some quantization cell. This formula holds for bandlimited and non-bandlimited processes alike. Indeed, the convergence rate of the first order conditional entropy was shown to depend only on the behavior of the autocorrelation function near the origin, thus it is of no consequence whether it is bandlimited or not. As examples, we considered

two autocorrelation functions: An exponential and a Gaussian. In the former case, the upper bound to the joint entropy was shown to go to infinity at rate $\frac{\log \tau}{\tau}$, while in the latter case the rate was shown to be $\log \tau$.

Appendix A

Lemma A1. *Let* $P_{\mu,\sigma^2,\Delta}(x) \stackrel{\Delta}{=} \int_x^{x+\Delta} \mathcal{N}_{\mu,\sigma^2}(t)\,dt$. *Let* $0 < s < 1$ *be some constant. Then*

$$\begin{aligned} &A. \quad P_{\mu,\sigma^2,\Delta}(x+\Delta) \;<\; s \cdot P_{\mu,\sigma^2,\Delta}(x) \qquad \textit{for all} \quad x \geq \mu + \Delta\Big[1 + \frac{\sigma^2}{\Delta^2} \ln \frac{1+s}{s}\Big] \;, \\ &B. \quad P_{\mu,\sigma^2,\Delta}(x-\Delta) \;<\; s \cdot P_{\mu,\sigma^2,\Delta}(x) \qquad \textit{for all} \quad x \leq \mu - \Delta\Big[1 + \frac{\sigma^2}{\Delta^2} \ln \frac{1+s}{s}\Big] \;. \end{aligned} \tag{A1}$$

Proof: We show A only. B can be shown in a similar way. Let us write

$$\begin{aligned} P_{\mu,\sigma^2,\Delta}(x+\Delta) \;&=\; \int_{x+\Delta}^{x+2\Delta} \mathcal{N}_{\mu,\sigma^2}(t)\,dt \;<\; Q\big(\frac{x-\mu+\Delta}{\sigma}\big) \;, \\ P_{\mu,\sigma^2,\Delta}(x) \;&=\; \int_{x}^{x+\Delta} \mathcal{N}_{\mu,\sigma^2}(t)\,dt \;=\; Q\big(\frac{x-\mu}{\sigma}\big) - Q\big(\frac{x-\mu+\Delta}{\sigma}\big) \;. \end{aligned}$$

It follows from above that it suffices to show that for $x \geq \mu + \Delta\big[1 + \frac{\sigma^2}{\Delta^2} \ln \frac{1+s}{s}\big]$, $Q\big(\frac{x-\mu+\Delta}{\sigma}\big) < s\big[Q\big(\frac{x-\mu}{\sigma}\big) - Q\big(\frac{x-\mu+\Delta}{\sigma}\big)\big]$, or equivalently $\frac{Q\left(\frac{x-\mu}{\sigma}\right)}{Q\left(\frac{x-\mu+\Delta}{\sigma}\right)} > \frac{1+s}{s}$. We simplify the left-hand side as follows:

$$\begin{aligned} \frac{Q\left(\frac{x-\mu}{\sigma}\right)}{Q\left(\frac{x-\mu+\Delta}{\sigma}\right)} \;&\stackrel{(a)}{=}\; \frac{\frac{1}{2}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, g\big(\frac{x-\mu}{\sigma}\big)}{\frac{1}{2}\, e^{-\frac{(x-\mu+\Delta)^2}{2\sigma^2}}\, g\big(\frac{x-\mu+\Delta}{\sigma}\big)} \;\stackrel{(b)}{>}\; \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{e^{-\frac{(x-\mu+\Delta)^2}{2\sigma^2}}} \;=\; e^{\frac{2(x-\mu)\Delta+\Delta^2}{2\sigma^2}} \\ &=\; e^{(\frac{(x-\mu)}{\Delta}+\frac{1}{2})\frac{\Delta^2}{\sigma^2}} \;\stackrel{(c)}{>}\; \frac{1+s}{s} \;, \end{aligned}$$

where (a) is due to having $x \geq \mu + \Delta$ and so the arguments of both g functions are nonnegative and hence, well defined. (b) follows from the monotonicity of g (Fact 6), and (c) follows from $x > \mu + \Delta\big[\frac{\sigma^2}{\Delta^2} \ln \frac{1+s}{s} - \frac{1}{2}\big]$. □

Lemma A2. *Let* $\alpha \in \mathbb{R}$ *be given. Then*

$$\lim_{p \to 0} \frac{\mathcal{H}(1-\alpha p)}{\mathcal{H}(p)} \;=\; 0 \;.$$

This is a similar version to Lemma 2 of Chapter III.

Proof: We need to show that $\lim_{p\to 0} \frac{-(1-\alpha p)\ln(1-\alpha p)}{-p\ln p} = 0$. The following string of equalities proves the lemma.

$$\begin{aligned}\frac{-(1-\alpha p)\ln(1-\alpha p)}{-p\ln p} &= \left[\frac{-(1-\alpha p)\ln(1-\alpha p)}{(1-\alpha p)\alpha p}\right]\left[\frac{(1-\alpha p)\alpha p}{-p\ln p}\right] \\ &= \left[\frac{\ln(1-\alpha p)}{-\alpha p}\right]\left[(1-\alpha p)\alpha\,\frac{1}{-\ln p}\right] \longrightarrow 0 \text{ as } p \longrightarrow 0\ ,\end{aligned}$$

where we used the well-known fact that $\lim_{x\to 0}\frac{\ln(1-x)}{-x} = 1$. □

Lemma A3. *Let $a(s)$ and $b(s)$ be positive functions on $\mathbb{R}$ such that $\lim_{s\to s_o}\frac{a(s)}{b(s)} = 1$ and $\lim_{s\to s_o} b(s) = b_o \neq 1$. Then*

$$\lim_{s\to s_o}\frac{\mathcal{H}(a(s))}{\mathcal{H}(b(s))} = 1\ .$$

This lemma is a slightly weaker version of Lemma 4 of Chapter III.

Proof: To keep notation short, we omit the parameter s from $a(s)$ and $b(s)$. The following string of equalities proves the lemma.

$$\begin{aligned}\frac{\mathcal{H}(a)}{\mathcal{H}(b)} = \frac{-a\log a}{-b\log b} = \frac{a}{b}\frac{\log\left[\frac{a}{b}\,b\right]}{\log b} = \frac{a}{b}\left[\frac{\log\frac{a}{b}}{\log b}+1\right] &= \frac{a}{b}+\frac{\frac{a}{b}\log\frac{a}{b}}{\log b} \\ \overset{s\to s_o}{\longrightarrow} 1+\frac{1\log 1}{\log b_o} &= 1\ .\end{aligned}$$

□

Appendix B

The following discussion appears in [16] (pp. 41-45) and in [17] (pp. 86-92). A random process $\{X_t,\ t\in T\}$ is said to be separable if there exists a countable set $S\subseteq T$ and a fixed null event Λ such that for any closed set $K\subseteq[-\infty,\infty]$ and any open interval I, the two sets

$$\{\omega : X_t(\omega)\in K,\ t\in I\cap T\} \quad \text{and} \quad \{\omega : X_t(\omega)\in K,\ t\in I\cap S\}$$

differ by a subset of Λ. The countable set S is called a *separating set* or *separant*.

It follows from the definition of separability that when the underlying probability space is complete, for any $a \in \mathbb{R}$ the set $\{\omega : X_t(\omega) < a,\ t \in I \cap T\}$ is an event and has the same probability as the event $\{\omega : X_t(\omega) < a,\ t \in I \cap S\}$, where S is a separating set.

It would be desirable to have stationarity imply that the two events $\{\omega :\ X_t(\omega) \leq r,\ \ a < t < b\}$ and $\{\omega :\ X_t(\omega) \leq r,\ \ a + s < t < b + s\}$ have the same probability. Using separability, one can find a countable set, which is dense in (a, b) and is a separating set, and use it to compute the probability of the first event. However, when shifting this set by s, it is no longer guaranteed that the shifted set, which is of course dense in $(a + s, b + s)$, is a separating set for the interval $(a + s, b + s)$. Thus, separability alone is not enough to allow us to use stationarity in this way. As mentioned in Section 4.2, in addition to separability we need continuity in probability. Specifically, we cite the following result:

Let $\{X_t,\ t \in T\}$ be a separable process, which is continuous in probability, and let T be an interval. Then every countable set dense in T is a separating set.

Therefore, whenever the process is both continuous in probability and separable, the probability of an event involving an uncountable number of X_t can be computed using *any* countable subset that is dense in T, and consequently shifting events of the above form does not alter their probabilities.

Next, let us consider measurability. A random process $\{X_t,\ t \in T\}$ is said to be measurable if $X_t(\omega)$ is a (t, ω) function measurable with respect to $\mathcal{B} \otimes \mathcal{F}$, where $\mathcal{B}$ is the σ-algebra of Lebesgue measurable sets in T, and $\mathcal{F}$ is the σ-algebra of events in the probability space $(\Omega, \mathcal{F}, P)$, i.e. for any $x \in (\infty, \infty)$,

$$\{(t, \omega) :\ X_t(\omega) \leq x\} \in \mathcal{B} \otimes \mathcal{F}\ .$$

Suppose now that Φ_X is some indicator function of the random process X, and let I be some Lebesgue measurable set, for example, an interval. Then it is desirable that the following hold:

$$E\big[\int_I \Phi_x(t)\,dt\big] \;=\; \int_I E[\Phi_x(t)]\,dt\ .$$

The swapping above of integration and expectation is permitted if the function Φ_X is a measurable function with respect to $\mathcal{B}\otimes\mathcal{F}$, as Fubini's theorem shows. Thus, the measurability of the process X ensures that the swapping above is indeed correct.

Finally, we cite the following important result:

Let $\{X_t,\ t\in T\}$ be a continuous in probability process, and let T be an interval. Then there exists a random process $\{\widetilde{X}_t,\ t\in T\}$ defined on the same probability space, which is equivalent to X_t and is separable and measurable.

It follows from this last result that the assumption of continuity in probability suffices to guarantee that basic operations on stationary continuous-time random processes such as preservation of probability under shifting and exchange of integration and expectation can be performed.

REFERENCES

[1] S. K. Tewksbury and R. W. Hallock, "Oversampled, linear predictive and noise-shaping coders of order $N > 1$," *IEEE Trans. Cir. Sys.*, vol. 25, no. 7, pp. 436–447, July 1978.

[2] M. W. Hauser, "Principles of oversampling A/D conversion," *J. Audio Eng. Soc.*, vol. 39, no. 1/2, pp. 3–26, Jan./Feb. 1991.

[3] K. Benhenni and S. Cambanis, "The effect of quantization on the performance of sampling designs," *IEEE Trans. Info. Theory*, vol. 44, no. 5, pp. 1981–1992, Sep. 1998.

[4] J. Tuqan and P. P. Vaidynathan, "Oversampling PCM techniques and optimum noise shapers for quantizing a class of nonbandlimited signals," *IEEE Trans. Signal Proccessing*, vol. 47, no. 2, pp. 389–407, Feb. 1999.

[5] Z. Cvetkovic and M. Vetterli, "Error-rate characteristics of oversampled analog-to-digital conversion," *IEEE Trans. Info. Theory*, vol. 44, no. 5, pp. 1961–1964, Sep. 1998.

[6] Z. Cvetkovic and M. Vetterli, "On simple oversampled A/D conversion in $L^2(\mathbb{R})$," *IEEE Trans. Info. Theory*, vol. 47, no. 1, pp. 146–154, Jan. 2001.

[7] Z. Cvetkovic and I. Daubechies, "Single-bit oversampled A/D conversion with exponential accuracy in the bit-rate," *Data Compression Conference, DCC, Snowbird, UT*, pp. 343–352, Mar. 2000.

[8] N. T. Thao and M. Vetterli, "Reduction of the MSE in R-times oversampled A/D conversion $O(1/R)$ to $O(1/R^2)$," *IEEE Trans. Signal Processing*, vol. 42, no. 1, pp. 200–203, Jan. 1994.

[9] N. T. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Trans. Signal Processing*, vol. 42, no. 3, pp. 519–531, Mar. 1994.

[10] N. T. Thao and M. Vetterli, "Lower bound on the mean-squared error in oversampled quantization of periodic signals using vector quantization analysis," *IEEE Trans. Info. Theory*, vol. 42, no. 2, pp. 469–479, Mar. 1996.

[11] T. Berger, *Rate Distortion Theory*, Prentice-Hall, Englewood Cliffs, 1971.

[12] R. G. Gallager, *Information Theory and Reliable Communication*, John-Wiley & Sons Inc., New York, 1968.

[13] H. Gish and J. N. Pierce, "Asymptotically efficient quantization," *IEEE Trans. Info. Theory*, vol. 14, pp. 676–683, Sep. 1968.

[14] J. Ziv, "On universal quantization," *IEEE Trans. Info. Theory*, vol. 31, no. 3, pp. 344–347, May 1985.

[15] R.M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Info. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[16] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, second edition, 1985.

[17] J. Neveu, *Mathematical foundations of the calculus of probability*, Holden-Day Inc., San Fransisco, 1965.

[18] P. R. Halmos, *Measure Theory*, Reprint, Springer-Verlag, 1974.

[19] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, John-Wiley & Sons Inc., New York, 1967.

[20] P. Billingsley, *Probability and Measure*, Wiley & Sons, New York, third edition, 1995.

CHAPTER V

Field-Gathering Sensor Networks[1]

5.1 Introduction

In this chapter we consider sensor networks for field-gathering, which is one of the proposed uses for sensor networks, e.g. [2, 1, 3, 4, 5, 6]. More specifically, we use the results of Chapter IV to examine the capability of large-scale sensor networks to measure and transport a two-dimensional field. Although our analysis is for one-dimension, we have no doubt that the results hold for two dimensions as well (see "Future Work" Section in Chapter VI). We consider a data-gathering wireless sensor network in which densely deployed sensors, with identical scalar quantizers, take periodic samples of the sensed field, and then separately scalar quantize, encode and transmit them to a central location, referred to as the collector, where snapshot images of the sensed field are reconstructed. The network operates in slotted time steps to transport bits from the sensors to the collector. The main question to be addressed is how many time slots does it take to transport the quantized data that corresponds to one snapshot of the field from all the sensors to the collector? We call this number of time slots *slot usage*. There are two factors that determine the slot usage: The ability of the sensors' encoders to compress their data and the *many-to-*

[1]This work is a slightly modified subset of [1].

one transport capacity of the network, which is the average number of bits a sensor can send to the collector per time slot. The better the compression, the smaller the slot usage. Ordinarily, the quality requirements for the reconstructed field play an important role in determining slot usage. The better the desired quality, the finer the quantizers and consequently the greater the number of bits that the sensors' encoders produce, which in turn increases slot usage.

More specifically, we shall be interested in determining the behavior of slot usage as sensor density increases to infinity. When this happens, more sensors send data to the collector. However, the data is more correlated, and the encoder at each sensor can do more compression. Thus, given a constraint on the quality of the reconstructed snapshots, what determines the asymptotic behavior of slot usage are the scaling laws of the many-to-one transport capacity and the average number of bits generated by each sensor per snapshot. If these are of the same order, then slot usage saturates at some finite value, if the latter decreases faster than the former, then slot usage tends to zero, and finally, if the latter decreases faster than the former, then slot usage tends to infinity. Equivalently, one can compare the scaling laws of the *total many-to-one transport capacity*, which is the total number of bits the collector can receive from the sensors per time slot, and the average total number of bits generated by all the sensors per snapshot. This is the approach taken here.

Under the network model considered, the total many-to-one transport capacity remains constant as sensor density increases. Therefore, the question boils down to whether or not the total number of bits generated by all the sensors per snapshot, remains bounded or not. The first result of Chapter IV shows that if all the sensors utilize identical scalar quantizers, then this total number of bits increases to infinity regardless of the scheme used by the sensors' lossless encoders. Hence, we conclude

that for the given scenario (i.e. the network communication model used – see next section – and identical scalar quantization), even though the correlation between sensor data increases as density increases, no scheme can transport the required amount of data for a given quality in a bounded number of time slots, i.e. slot usages tends to infinity. The second result of Chapter IV is used to upper bound the rate at which slot usage tends to infinity, for the special case of a Gaussian field.

The remainder of this chapter is organized as follows. Section 5.2 describes the network model. In Section 5.3 a detailed problem formulation is provided. Section 5.4 presents the results. Finally, Section 5.5 summarizes and concludes.

5.2 Sensor Network Model

Given a fixed positive number D, the goal of the sensor network is to sample, quantize, encode, transport and reconstruct/reproduce (we shall use these terms interchangeably) snapshots of the field with distortion D or less, with fewest number of time slots per snapshot. We will assume that the transport system and compression system of the network are designed and operated separately.

Next, we provide a description of the operation of the sensor network, the model for the sensed field, the fidelity criterion, the transport system, the compression system, and the decoder at the collector.

5.2.1 Sensor Network Operation

At regular time intervals, each sensor in the network measures the field value at its location; then quantizes its value and losslessly encodes it with bits. (The field at a given sampling time is called a snapshot). These are transported to the *collector*, which is a processing unit, where decoding of the quantized data is performed, and a reconstruction of the field snapshot, corresponding to the time of the quantized

data, is produced. The performance of the network is measured by the mean-squared error of its reproductions with respect to the original corresponding field snapshots, and the frequency with which the quantized data representing one snapshot can be transported to the collector, or equivalently, the number of time slots needed to transport this quantized data.

More specifically, the sensor network consists of N sensors that are uniformly spaced over a finite geographical region of interest G. All sensors use identical scalar quantizers. We comment that one may consider a network where other forms of quantization are performed. For example, temporal vector quantization, i.e. vector quantization of samples at successive time instances may be used. Another possibility is to use dithering, e.g. [7, 8]. But as a first step, we consider the simple scheme of identical scalar quantization, both in space and time, with no dithering.

5.2.2 Sensed Field Model

The sensed field, i.e. the snapshot, is modeled as a stationary and continuous in probability[2], two-dimensional random field $X(u, v)$. That is, $X(u, v)$ is a real-valued random variable representing the field value at Euclidean coordinates (u, v), where u and v vary continuously. We make no assumption as to whether the random field is bandlimited or not (bandlimited refers to spatial frequency content). We let G denote the geographical region of interest, which is the region over which the network is deployed. As we shall discuss in the next subsection, the sensors will use identical scalar quantizers. Thus, with respect to these quantizers, the only requirement we have from the field (aside from stationarity), in order to obtain the first result, i.e. that the total number of bits per snapshot tends to infinity as sensor density

[2]Since continuity in probability is a very mild technical condition, we shall omit it in the sequel and not mention it explicitly.

increases, is that it have a quantization threshold crossing with positive probability in G. In other words, we require that the probability that each $X(u,v)$ in the entire region of interest would lie in the same quantization cell is less than one (notice, of course, that not all $X(u,v)$ are quantized, but rather only the X's at the sensor locations – see shortly). This is a benign assumption, because if it does not hold, i.e. if with probability one all X's lies in the same quantization cell, then clearly the quantizer is too coarse to be of use. We notice that this requirement precludes, for example, the possibility of the field being constant, even if the constant is random.

We define a *snapshot* of the field to consist of all values $X(u,v)$, $u,v \in G$. We assume that successive snapshots are independent. That is, each snapshot is modeled as a random field that is independent of the random fields modeling other snapshots.

A principal characteristic of the random field is its autocorrelation function $R(\tau_1,\tau_2)$, which indicates the correlation between values of X separated horizontally and vertically by distances τ_1 and τ_2, respectively. For instance, $R(\tau_1,\tau_2) = \exp\left\{ -\sqrt{\tau_1^2+\tau_2^2} \right\}$ is an example of an isotropic autocorrelation function that decays exponentially with Euclidean distance.

In order to obtain the second result, namely, an upper bound to the rate at which the total number of bits per snapshot goes to infinity, we will require that the autocorrelation function of the random field is continuous at the origin, which implies quadratic mean continuity, and also implies that as the distance between samples decreases to zero, their correlation coefficient tends to one. We comment, that although the latter requirement is used in our proof that the number of bits per sensor per snapshot can be made to go to zero as sensor density increases, it might not actually be necessary for the result to hold.

5.2.3 Fidelity Criterion

Let $(u_1, v_1), (u_2, v_2), ..., (u_N, v_N)$ denote the locations of the sensors. The field values at these locations are sampled, quantized, encoded and transported to the collector. The collector creates a reconstruction/reproduction (these terms will be used interchangeably) $\widehat{X}(u, v)$, $(u, v) \in G$ as a reproduction of the original snapshot $X(u, v)$, $(u, v) \in G$.

The fidelity criterion of the reproduction $\widehat{X}$ relative to the original field value is mean-squared error (MSE), and is given by the following expression:

$$\text{MSE} \;=\; \frac{1}{|G|} \int_G E\Big(X(u,v) - \widehat{X}(u,v)\Big)^2 \, du \, dv \;,$$

where E denotes expected value with respect to the random field, the integral is taken over the region G, and $|G|$ denotes its area.

5.2.4 Transport System

The wireless network has a transceiver at each sensor and operates in slotted time steps to transport the bits generated by the sensors' encoders to the collector. Multiple hops may be required. There is a number W such that each sensor can transmit or receive at most W bits in one slot.

When the collector has received from each sensor the encoded quantized value corresponding to a particular sampling time, i.e. corresponding to one complete snapshot, it forms a reconstruction of that snapshot. The sampling and data transport are pipelined in the sense that further snapshots may be taken by the sensors and their transport may begin before the network has finished transporting prior snapshots to the collector.

5.2.5 Compression System

Since the field is continuous-valued the compression system is lossy. As mentioned, the sensors take samples of the random field at locations denoted $(u_1, v_1), (u_2, v_2), ..., (u_N, v_N)$. Since a sensor value is known only at its own location, the quantization and lossless encoding at each sensor is done separately, where by "quantization" we mean the mapping of a field value to a quantization index, i.e. every quantizer maps a sensor value $X(u_i, v_i)$ to an integer that indexes the possible quantization cells. This index is then encoded in some lossless fashion. The encoding is done separately at each sensor, namely, the quantization indices representing quantized values of several sensors are not jointly encoded.

It is assumed that the encoders know the correlation structure of the field and the location of all other sensors, in other words, the encoders know the correlation between any two sensor samples of the field. Consequently, the encoders can use Slepian-Wolf distributed lossless coding [9]. This is a coding method that permits lossless coders to separately encode the data from correlated sources (such as the data produced by neighboring sensors) as efficiently as if each encoder could see the values produced by the other data sources. To illustrate this, consider the following simple example. Let X and Y be two correlated discrete-time discrete-valued random processes that are encoded separately by two encoders. That is, each encoder knows the sample values of the random process it encodes only. The encoded samples are the input to a decoder, whose output are the original sample values of X and Y. This is depicted in Figure 5.1a. The question is at what rate (i.e. average number of bits per sample) can the two encoders encode so that at the decoder, the original sample values can be reconstructed with arbitrarily small error probability? Clearly, the first can have rate $R_X = H(X)$ to encode X and the second can have rate

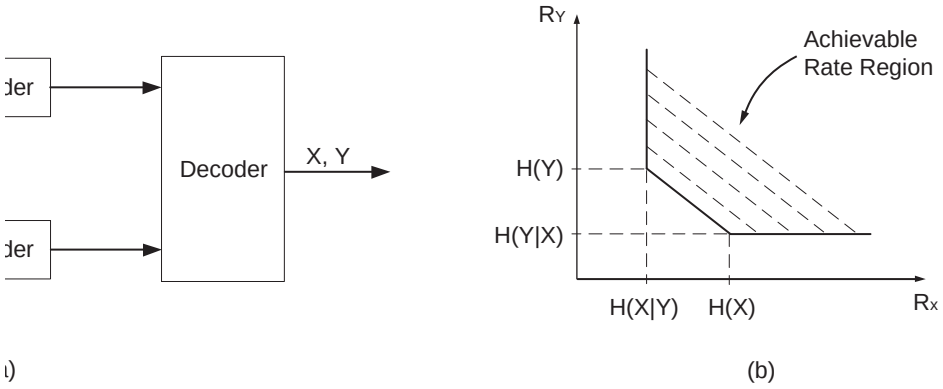

Figure 5.1: Slepian-Wolf coding. (a) Two separate encoders. (b) Achievable rate region.

$R_Y = H(Y)$ to encode Y (e.g. by using entropy coding, where we ignore the fact that $H(X)$ and $H(Y)$ might not be attained exactly). It seems plausible that since neither encoder has knowledge of the other encoder's input, $H(X)$ and $H(Y)$ would be the absolute minimum rates the encoders could have. Surprisingly, however, this is not the case. By allowing both encoders to know the joint statistics of X and Y, one can do better by using Slepian-Wolf coding. Specifically, if the first encoder uses rate $R_X = H(X)$ to encode X, the second encoder can use rate $R_Y = H(Y|X)$ to encode Y, where $H(Y|X)$ is the conditional entropy of Y given X, i.e. the second encoder can encode Y at the same rate as if it knew the value of X. Figure 5.1b shows the achievable rate-region, i.e. the pairs of rates that the encoders can use so that the sample values can be reconstructed losslessly. (In fact, Slepian-Wolf coding is an *almost lossless* coding technique that has arbitrarily small probability of error, and is normally regarded as lossless.)

We, note that Slepian-Wolf coding entails the simultaneous encoding of a block of successive outputs from the quantizer of a given sensor. Thus, the encoders are allowed to encode blocks of quantization indices at a time.

5.2.6 Collector Decoder

First and foremost the decoder at the collector is designed to match the sensors' Slepian-Wolf distributed encoding. Thus, it is assumed that the decoder obtains a lossless description of all the quantization indices from all the sensors. The next task of the decoder is to produce a reproduction of the field. To do so, an interpolation is performed. For simplicity we assume that simple linear interpolation (e.g. sample and hold) is used.

When N is large and, consequently, the sensors are closely spaced, the component of MSE due to interpolation error is negligible, and for ordinary random fields, the MSE is well approximated simply by the average MSE between the N sensor samples and their reconstructions. That is,

$$\text{MSE} \;\approx\; \frac{1}{N}\sum_{i=1}^{N} E\left(X(u_i,v_i)-\widehat{X}(u_i,v_i)\right)^2 \;=\; E\left(X(u_1,v_1)-\widehat{X}(u_1,v_1)\right)^2 \;. \quad (5.1)$$

where $\widehat{X}(u_i,v_i)$ denotes the quantized version of the field value at location (u_i,v_i). That is, with the quantizer fixed, as $N\to\infty$, MSE approaches the distortion of the scalar quantizer.

We note that it is an open question whether there exists a nonlinear interpolation method for the given scalar quantizers whose MSE is less or even tends to zero as the density of quantized values increases. Such a method could perhaps be constructed using level crossing theory [10] in conjunction with random sampling theory [11]. If the answer is that MSE can indeed be made to go to zero, then the results presented in the sequel need to be reevaluated. If, however, MSE can be made smaller, but cannot be made to go to zero, then the results that follow are still valid.

Remark: The results presented in Section 5.4 will be for one-dimensional random fields, however, we have no doubt that they extend to two-dimensional random fields as well (see the "Future Work" Section of Chapter VI).

5.3 Performance

As described in the previous section, we wish to investigate the ability of a dense wireless sensor network, to measure and transport independent snapshots of a two-dimensional field to a central location, i.e. a collector, where reconstructions of these field snapshots are formed.

The principal question to be addressed is how frequently can a new snapshot be taken and transported successfully to the collector. If new snapshots can be received by the collector every u slots, then we say the network has a *slot usage* of u slots per snapshot, and a *throughput* of $1/u$ snapshots per slot. Clearly small slot usage and large throughput are desired.

One might also ask how much time must transpire between the time the snapshot is taken by the sensors and the time the collector has the data needed for its reconstruction. This *delay* will not be discussed here, except to say that due to pipelining the slot usage is at most as large as the delay, and usually substantially smaller.

We are particularly interested in how the network slot usage and throughput of an optimized system vary as N, the number of sensors, increases. Of course, the sensor spacing decreases with N, and the sensor density increases with N. Must the slot usage increase with N? If so, does it saturate at some finite value? Or does it increase without bound?

To answer these questions, given that we wish to use a scheme for which the transport system and the compression system are designed separately, one must an-

swer a *compressibility* question and a *capacity* question: How many bits must be generated by each sensor's quantizer/encoder per snapshot? And how many bits can be transported on the average by the network to the collector per sensor per slot? (Here, we only count new bits generated at the sensors – not bits relayed by the sensors.) This is the many-to-one transport capacity. Suppose the answer to the compressibility question is b_N, i.e. b_N is the minimum number of bits per sensor per snapshot that must be generated for a network of size N, and suppose the answer to the capacity question is c_N, i.e. c_N is the many-to-one transport capacity, namely, it is the maximum average number of bits that can be transported to the collector per sensor per slot. (c_N is less than W – usually much less.) Then the smallest possible slot usage is $u_N = b_N/c_N$ slots/snapshot. Equivalently, the maximum possible throughput is $t_N = c_N/b_N$ snapshots/slot.

Duarte-Melo and Liu [4] answered the capacity question. They adopted a transmission and interference model similar to the protocol model of Gupta and Kumar [12], except it considered many-to-one communication instead of peer-to-peer communication. They showed that

$$c_N = \Theta\Big(\frac{1}{N}\Big) \text{ bits/sensor/slot} , \tag{5.2}$$

where $\Theta(\frac{1}{N})$ means there exist constants a_1 and a_2 such that $\frac{a_1}{N} \leq c_N \leq \frac{a_2}{N}$ for sufficiently large N. Notice that the fact that the number of bits per slot that the collector can receive is bounded by W, easily implies that c_N at most of the order of $\frac{1}{N}$. The result in [4] shows, however, that this upper bound is attainable. In comparison, Gupta and Kumar [12] found the *peer-to-peer* transport capacity of a similar network to be $c_N = \Theta\left(\frac{1}{\sqrt{N \log N}}\right)$.

We comment that Gupta and Kumar had a second transmission and interference model in [12], called the physical model, whose capacity results agree with those shown for the protocol model, and so they would also agree with those shown in [4]. Under this model, the attenuation is assumed to be of the form $a(x) = \frac{1}{x^\gamma}$, where $\gamma > 0$. Clearly, this attenuation model is optimistically nonrealistic when x is very small. Alternatively, we could use the attenuation model $a(x) = \frac{1}{(1+x)^\gamma}$, as used in [13], for example. However, since the results presented here are of negative nature, i.e. they show how throughput tends to zero and slot usage to infinity, it is reasonable to assume an optimistic model.

Our focus from now on is on the compressibility question. Given the models for the random field and the fidelity measure that were provided in Section 5.2, and given a fixed target MSE D, then as shown in Section 5.4, it is possible to have $b_N \to 0$ as $N \to \infty$, where b_N is the minimum number of encoded bits per sensor per snapshot that must be transported to the collector to attain MSE less than or equal to D. The idea is that as N increases, the sensors become closer, the correlation between the field values sampled by nearby sensors increases, and it is possible to exploit this correlation using, for example, Slepian-Wolf distributed lossless coding on the quantizer outputs, to make $b_N \to 0$. On the other hand, although $b_N \to 0$, we also show in Section 5.4, using the result shown in Chapter IV concerning the joint entropy of quantized samples over a finite interval as the sampling interval goes to zero, that no matter how the lossless coding is done, b_N does not decrease as rapidly as $1/N$. That is,

$$Nb_N \longrightarrow \infty \text{ as } N \longrightarrow \infty \, . \tag{5.3}$$

Note that Nb_N is the total number of bits coming from the quantizers/encoders of all sensors. Note also that the above result is quite general and is not limited to a

particular lossless coding scheme.

Combining (5.3) with the many-to-one transport capacity result (5.2), we find that the smallest slot usage for which the MSE can be D or less is

$$U(N,D) \;=\; \frac{b_N}{c_N} \;=\; \frac{Nb_N}{Nc_N} \;\longrightarrow \infty \text{ as } N \longrightarrow \infty \,. \tag{5.4}$$

This indicates that to obtain a given MSE D, the number of slots per snapshot must grow without bound as N increases.

It must be said that this is somewhat disappointing, as it had been hoped that as N increases, the inter-sensor correlation would increase sufficiently rapidly to make Nb_N (and $U(N,D)$) saturate at a finite value, rather than approach infinity. Note, however, that this result does not say that sensor networks cannot do the desired job of measuring and transporting a two-dimensional field. Rather it says that the efficiency with which it does so, as expressed by the slot usage or throughput, degrades as the density of the sensors becomes very large.

It should be noted that the efficiency also degrades when N becomes too small. Specifically, there is some threshold value N_o such that for $N < N_o$, the interpolation error by itself exceeds D. Thus, there is no quantization-encoding-transport scheme that attains MSE D. Moreover, as N approach N_o from above, the quantizer must have increasingly fine resolution, which causes $b_N \to \infty$. And since in this case $N > N_o$, we also have $Nb_N \to \infty$. Thus as in (5.4), $U(N,D) \to \infty$ as $N \searrow N_o$. We conclude that given a target MSE D and a random field model, there is an optimum value of N. This is the value for which Nb_N is smallest. This conclusion applies to bandlimited and non-bandlimited fields alike. For bandlimited fields the optimum value of N is not necessarily the value that leads to Nyquist sampling.

Based on the above analysis, an alternative strategy, to be pursued in future work, is to fix the number of sensors at the value of N that minimizes Nb_N, and then to permit there to be an additional set of transceivers at locations between the sensors. This is equivalent to having a network of $N' > N$ sensors, and putting all but N of them to sleep, while keeping all transceivers active.

We assert that the result in (5.3) is not at all obvious. Indeed, the limiting behavior of Nb_N has been a long standing question in the theory of sampling and quantization, which has only recently been resolved in the work presented in Chapter IV of this thesis. Furthermore, Nb_N is a quantity of basic interest in source coding, because it shows whether scalar quantizers with entropy coding can be used efficiently to encode continuous-parameter random processes. If Nb_N were to remain "close" to the information theoretic rate-distortion function for continuous-parameter processes (i.e. if Nb_N remained bounded), then scalar quantization with entropy coding would be a viable technique for quantizing and encoding continuous-parameter random processes. Equation (5.3) shows that this is not the case.

The question regarding the behavior of Nb_N is quite delicate, and as mentioned in Section 5.4, rate-distortion theory shows that if *ideal* lossy coding were used instead of scalar quantization plus binary lossless coding, then Nb_N would not increase to infinity. However, the sensor network requires that coding be done separately at each sensor. This is why we use scalar quantization, rather than say vector or predictive quantization across sensors. However, we are certain (see "Future Work" Section of Chapter VI) that even if one were allowed to use vector quantization, unless the dimension of the quantizer increases with N, Nb_N would still grow without bound.

Having shown that Nb_N grows to infinity, the question arises as to how fast it grows. Using the asymptotic formula for conditional entropy of highly correlated

Gaussian random variables shown in Chapter IV, we find in Section 5.4 the rate with which Nb_N increases, for the special case of a Gaussian random field and a particular form of Slepian-Wolf coding. This also leads to a result on how fast $U(N, D)$ grows in this special case. For example, for a one-dimensional Gaussian field with exponential autocorrelation, it is shown that $U(N, D) \to \infty$ at rate $\Theta(\sqrt{N} \log N)$.

In addition to the many-to-one transport capacity, one may also consider the *all-to-all* transport capacity, which is the maximum average number of bits per sensor per slot that can be transported from each sensor to every other sensor. In [1] it was shown that the all-to-all transport capacity is

$$c_N \;=\; \Theta\left(\frac{1}{N}\right) \text{ bits/sensor/slot} \;,$$

which is the same as the many-to-one transport capacity. Thus the behavior of a network operating in all-to-all fashion, e.g. the asymptotic slot usage $U(N, D)$, is the same as the behavior of a network operating in many-to-one fashion.

5.4 Results

We need to assess the minimum number of bits that an encoder could produce when encoding a quantized sensor value, when sensors are densely placed, and consequently, their values are highly correlated. We will summarize and use the results of Chapter IV.

As stated in Section 5.2, we view the sensors as taking uniformly spaced samples of a stationary two-dimensional random field over a finite geographical region. The collection of all samples taken at one time instance form a snapshot. Successive snapshots are assumed to be independent.

Though the field is two-dimensional, the basic ideas are more readily apparent and simpler to describe in one dimension. Therefore, we will focus on the case that

N sensors are uniformly spaced on a straight line of length 1. In this case, let $X(s)$, $0 \leq s \leq 1$ denote the field value at location s. $X(s)$ is assumed to be a continuous parameter stationary random process. Let $(X_1, \ldots, X_N)$ denote the N sensor values taken at a spacing of $d = 1/N$. Let $(I_1, \ldots, I_N)$ denote the integers resulting from quantizing $(X_1, \ldots, X_N)$ with some fixed scalar quantizer.

5.4.1 $b_N \longrightarrow 0$

From basic information theory we know that no lossless compression technique could compress the output of the quantizer with fewer than

$$H(I_1, \ldots, I_N) \text{ bits.} \tag{5.5}$$

Equivalently, it requires on average at least

$$\frac{1}{N} H(I_1, \ldots, I_N) \text{ bits per sample}$$

to losslessly encode each quantized sensor value.

The lower bound in (5.5) can in fact be attained using Slepian-Wolf distributed lossless coding. This requires every sensor to simultaneously encode a block of, say, M successive outputs from *its* quantizer. Observe that the block of outputs is a temporal block rather than a spatial one. Temporal blocks are needed in order for the encoder, at each sensor, to operate at rate close to some conditional entropy value (these conditional entropies will be stated shortly). Spatial blocks, however, are not used since every sensor knows only its own values and so the quantization and encoding must be done separately at each sensor.

The lower bound in (5.5) is attained in the following way. Let all sensors quantize their values separately. Let sensor 1 losslessly encode its block of M successive quantizer outputs into approximately $MH(I_1)$ bits using conventional block loss-

less coding[3], where $H(I_1)$ denotes the entropy of one of its quantizer outputs, and where the independence of successive outputs has been used. Let sensor 2 encode its values using Slepian-Wolf style coding with respect to sensor 1. Then, it losslessly encodes its block of M successive quantizer outputs into approximately $MH(I_2|I_1)$ bits, where $H(I_2|I_1)$ denotes the conditional entropy of an output of sensor 2 given an output of sensor 1 in the same snapshot. (The decoder will already have decoded the I_1's, before decoding the I_2's.) Similarly, sensor 3 uses Slepian-Wolf coding with respect to sensors 1 and 2, thus mapping its M quantizer outputs into approximately $MH(I_3|I_2, I_1)$ bits. And so on. It follows that for the k^{th} sensor, the number of bits per snapshot generated by its quantizer/encoder is approximately $b_N(k) = H(I_k|I_1, \ldots, I_{k-1})$. It is well known that $b_N(k)$ decreases monotonically with k. Thus, for large N, most of the $b_N(k)$'s are approximately the same. That is, there is a value b_N such that $b_N(k) \approx b_N$ for most k. It is this value to which Section 5.3 refers when prescribing the number of bits per sensor per slot produced by each sensor's quantizer/encoder.

It also follows that the total number of bits B_N produced by all the sensors is given by:

$$\begin{aligned} B_N &= \sum_{k=1}^{N} b_N(k) \\ &= H(I_1) + H(I_2|I_1) + \ldots + H(I_N|I_{N-1}, I_{N-2}, \ldots, I_1) \\ &= H(I_1, \ldots, I_N) \ , \end{aligned}$$

where the last equality is an elementary property of entropy. This shows that the

[3]This and subsequent similar approximations can be made arbitrarily tight by choosing M large. Moreover, this and subsequent block encodings are *nearly* rather than *perfectly* lossless, meaning that there is a nonzero probability that the decoder output does not match the encoder input. However, such decoding error probabilities can be made arbitrarily small by choosing M large, thereby having negligible effect on the overall MSE.

Slepian-Wolf approach does indeed attain the lower bound in (5.5).

We now show $b_N \to 0$ as $N \to \infty$. Using elementary information theory relations,

$$\begin{aligned} b_N &= \frac{1}{N}\sum_{k=1}^{N} b_N(k) \\ &= \frac{1}{N}\sum_{k=2}^{N} H(I_k|I_{k-1}, I_{k-2}, \ldots, I_1) \\ &\leq \frac{1}{N}\Big[H(I_1) + \sum_{k=2}^{N} H(I_k|I_{k-1})\Big] \\ &= \frac{H(I_1)}{N} + \frac{(N-1)}{N} H(I_2|I_1) \\ &\longrightarrow H(I_2|I_1) \ \text{ as } N \longrightarrow \infty \ . \end{aligned}$$

As N increases the sensors become closer and closer. Consequently their correlation increases. Specifically, as $N \to \infty$, the distance between sensors 1 and 2 goes to zero. Thus their sample values become essentially identical resulting in $H(I_2|I_1) \to 0$, which in turn implies that $b_N \to 0$.

5.4.2 $Nb_N \longrightarrow \infty$

Theorem 3 of Chapter IV shows that $H(I_1, \ldots, I_N) \to \infty$ as $N \to \infty$. Since from (5.5) $B_N = Nb_N$ can be no smaller than $H(I_1, \ldots, I_N)$, we see that $Nb_N \to \infty$.

We comment that Theorem 3 of Chapter IV is for a one-dimensional random process. However, this theorem can no doubt be generalized to the case of a two-dimensional random field, which will be the focus of future work. We note further that if the snapshots of the field were dependent, we strongly believe that using an encoding scheme that encodes based on previous snapshots will do no better. This, too will be the focus of future work.

5.4.3 The Growth of Rate for a Gaussian Random Field

As mentioned, although the encoding of the sensor value X_i must be done without knowledge of the other sensor values with which it is correlated, one could nevertheless losslessly encode it with approximately $b_N = \frac{1}{N}H(I_1, \ldots, I_N)$ bits, assuming, for example, Slepian-Wolf distributed coding is used. A suboptimal but easier to analyze case is where Slepian-Wolf coding is used to encode each sensor value with approximately $b_N = H(I_2|I_1)$ bits by encoding each I_i assuming I_{i-1} is known to the decoder. Since $B_N = Nb_N$, we can now apply (4.1) from Chapter IV, with B_N playing the role of $\bar{R}$, to obtain an asymptotic, as $N \to \infty$, expression for B_N. We note that (4.1) is based on Theorem 7 of Chapter IV, which holds for the case of infinite-level uniform scalar quantizers, a stationary Gaussian source whose mean is at the midpoint of some quantization cell, and small spacing between adjacent samples, i.e. $\tau = \frac{1}{N}$ is small. Applying (4.1) we obtain

$$B_N \;\approx\; -\frac{1}{\tau}M_\lambda\sqrt{1-\rho(\tau)}\,\log_2\sqrt{1-\rho(\tau)}\;,$$

where σ^2 is the variance of the source, $\lambda = \frac{\Delta}{\sigma}$ with Δ being the step size of the uniform quantizers, and $\rho(\cdot)$ is the normalized covariance function of the source, i.e. $\rho(\tau)$ is the correlation coefficient between adjacent samples of X.

Equations (4.2) and (4.3) of Chapter IV provide examples of B_N for the special cases that the Gaussian random process has an exponential autocorrelation function and a Gaussian autocorrelation function, respectively. It follows from these equations that in the former case

$$B_N \;\approx\; \frac{M_\lambda}{2}\sqrt{N}\log N \;\longrightarrow\infty \;\text{ as } N\longrightarrow\infty\;, \tag{5.6}$$

i.e. B_N increases as $\sqrt{N}\log N$, and in the latter case

$$B_N \;\approx\; M_\lambda \log N \;\longrightarrow\infty \;\text{ as } N\longrightarrow\infty\;. \tag{5.7}$$

i.e. B_N increases as $\log N$.

In light of the previous discussion that the total number of bits must increase to infinity as N increases, it should not be surprising that (5.6) and (5.7) increase without bound as $N \to \infty$. Note that in these examples the number of bits per sensor $b_N = B_N/N$ goes to 0.

5.5 Conclusions

In this chapter we characterized the amount of data required to sample, quantize, encode, and reconstruct a field densely deployed with wireless sensors, which use identical scalar quantization. We showed that as the number of sensors increases to infinity, the total number of bits generated by all the sensors for every snapshot also goes to infinity for every system with distortion D. At the same time, the total many-to-one transport capacity remains constant on the order of one. Similarly, while the number of bits per sensor per snapshot could be made to go to zero, it will do so strictly slower than $\frac{1}{N}$, whereas the transport capacity per sensor (i.e. the many-to-one transport capacity) is of the order $\frac{1}{N}$, as shown in [4]. Therefore the number of bits required for a fixed MSE cannot be transported with bounded slot usage as N increases. We would like to emphasize that this result holds for both a bandlimited and non-bandlimited random field, regardless of the encoding scheme used.

We showed that in the special case of a one-dimensional Gaussian random field with two example autocorrelation functions, there exists a compression system for which the number of bits per sensor per snapshot is on the order of $\frac{\log N}{\sqrt{N}}$ and $\frac{\log N}{N}$. Since the achievable transport capacity per sensor is on the order of $\frac{1}{N}$, it follows that in this special case the slot usage is $\Theta(\sqrt{N}\log N)$ and $\Theta(\log N)$, respectively.

We also discussed that since the required number of slots per snapshot must increase with the number of sensors, there should exist an optimal number of sensors that minimizes the number of slots per snapshot. We do not know what this optimum is, but if we did, it would place a limit on how densely sensors should be deployed, beyond which one should *suppress* sensors, e.g. put sensors to sleep, to prevent oversampling.

REFERENCES

[1] D. Marco, E. Duarte-Melo, M. Liu, and D. L. Neuhoff, "On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data," *Workshop on Information Processing in Sensor Networks (IPSN), Palo Alto, CA.*, pp. 1–16, Apr. 2003.

[2] A. Scaglione and S. D. Servetto, "On the interdependence of routing and data compression in multi-hop sensor networks," *International Conference on Mobile Computing and Networking (MobiCom), Atlanta, GA.*, pp. 140–147, Sep. 2002.

[3] S. D. Servetto, "Sensing lena – massively distributed compression of sensor images," *IEEE International Conference on Image Processing (ICIP), Barcelona, Spain.*, Sep. 2003.

[4] E. J. Duarte-Melo and M. Liu, "Data-gathering wireless sensor networks: Organization and capacity," *Special Issue Computer Networks (Elsevier) on Wireless Sensor Networks*, vol. 43, no. 4, pp. 519–537, Nov. 2003.

[5] R. Cristescu, "Efficient decentralized communications in sensor networks," *Ph.D. Thesis, EPFL.*, Mar. 2004.

[6] D. Ganesan, R. Cristescu, and B. Beferull-Lozano, "Power-efficient sensor placement and transmission structure for data gathering under distortion constraints," *Workshop on Information Processing in Sensor Networks (IPSN), Berkeley, CA.*, pp. 142–150, Apr. 2004.

[7] P. Ishwar, A. Kumar, and K. Ramchandran, "Distributed sampling for dense sensor networks: A 'bit-conservation principle'," *Workshop on Information Processing in Sensor Networks (IPSN), Palo Alto, CA.*, pp. 17–31, Apr. 2003.

[8] A. Kumar, P. Ishwar, and K. Ramchandran, "On distributed sampling of smooth non-bandlimited fields," *Workshop on Information Processing in Sensor Networks (IPSN), Berkeley, CA.*, pp. 89–98, Apr. 2004.

[9] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Info. Theory*, vol. 19, pp. 471–480, Jul. 1973.

[10] F. A. Marvasti, *A unified approach to zero-crossings and nonuniform sampling of single and multidimensional signals and systems*, Kings College, London, 1987.

[11] E. Masry, "Polynomial interpolation and prediction of continuous-time processes from random samples," *IEEE Trans. Info. Theory*, vol. 43, no. 2, pp. 776–783, Mar. 1997.

[12] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Info. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[13] O. Arpacioglu and Z. J. Haas, "On the scalability and capacity of wireless networks with omnidirectional antennas," *Workshop on Information Processing in Sensor Networks (IPSN), Berkeley, CA.*, pp. 169–177, Apr. 2004.

CHAPTER VI

Summary and Future Work

In this final chapter, we summarize the contributions of this dissertation and point to future research issues that remain unsolved.

6.1 Summary

In this dissertation we considered three asymptotic scalar quantization problems, the last of which was applied to the sensor networks setting. Next, we provide a brief description of the results shown in each chapter.

In Chapter II, the widely used additive noise model for high resolution uniform threshold scalar quantizers was rigorously demonstrated for input densities that are continuous and satisfy several other mild conditions. Specifically, it was shown that as step size decreases, the correlation between input and quantization error becomes asymptotically negligible relative to the mean-squared error. Furthermore, this model was shown to be valid even when the input density has a discontinuity at the origin, but discontinuities elsewhere might prevent the correlation from being negligible. In such cases, a formula for the correlation was derived in terms of the step size and the heights and positions of the discontinuities. Furthermore, for input densities with finite support, it has been shown that various noise models can be

attained by appropriately matching the support of a finite-level uniform quantizer.

In Chapter III we showed that the operational rate-distortion function of scalar quantization, for stationary memoryless Gaussian sources approaches zero with the same slope as that of the Shannon rate-distortion function. From which we concluded that scalar quantization is an asymptotically, as distortion tends to the source variance, optimal coding technique for such sources. This result was demonstrated using uniform scalar quantizers and binary quantizers. Thus, it not only shows that scalar quantization can be optimal in general, but rather it provides specific quantizers that achieve such optimality.

Chapter IV was concerned with the entropy of highly correlated quantized samples. Two results have been demonstrated. First we examined the case that a stationary random process is sampled over some finite interval, and each sample is separately quantized with arbitrary, yet identical, scalar quantizers. A fundamental question is what happens to the joint entropy of the quantized samples as the sampling interval goes to zero. This question has been answered, and specifically it was shown that if the random process crosses some quantization threshold with positive probability, then the joint entropy tends to infinity as the sampling interval goes to zero.

The second result provided an upper bound to the rate at which the joint entropy above tends to infinity, for the case of a stationary Gaussian process that is quantized with identical infinite-level uniform threshold scalar quantizers, such that the mean of the Gaussian process lies at a midpoint of some quantization cell. Specifically, an asymptotic formula was derived for the conditional entropy of one quantized sample conditioned on another quantized sample.

Finally, Chapter V applied the first two results of Chapter IV to field-gathering sensor networks, whose sensors use identical scalar quantizers. The purpose of these networks is to convey snapshots of the quantized field values at the sensors' locations to a central location, i.e. collector, where reconstructed snapshots of the field are produced. The question that was raised is what happens to the frequency, equivalently throughput, with which snapshots of the field can be transported and reconstructed at the collector, subject to a fidelity constraint. It was shown, using the joint entropy result from Chapter IV, that when the field is stationary and crosses a quantization threshold with positive probability, as the density of the sensors increases to infinity, the frequency (throughput) above tends to zero.

Furthermore, using the asymptotic formula, from Chapter IV, for the conditional entropy of quantized Gaussian random variables, an upper was found to the rate at which the throughput of such networks degrades to zero in the case of infinite-level uniform quantizers and a Gaussian field.

6.2 Future Work

There are several questions that are still open with regard to the problems discussed in some of the chapters of this dissertation. Next, we list these possible avenues for future research.

It would be of interest to generalize the low resolution problem, discussed in Chapter III, to source densities other than Gaussian, for example, Laplacian. In fact, it might be possible to obtain tail conditions on the source density that may insure optimal performance of scalar quantization, i.e. conditions that insure that the operational rate-distortion of scalar quantization go to zero with the slope as the rate-distortion function. We point out, however, that such tail conditions might

not be sufficient. Indeed, they would probably determine the asymptotic behavior of the entropy of the quantizer, but getting a handle on the behavior of distortion may require some further regularity conditions. This is due to the fact that the rate of convergence of distortion to the source variance is dominated by the cell containing the origin, and the location of the reconstruction level within this cell.

More open problems arise when considering the work of Chapter IV. For example, since the joint entropy of quantized samples over a finite interval has been shown to go to infinity, as the sampling interval goes to zero, a reasonable question to ask is whether, under similar conditions, the same happens to the entropy-rate of the quantized process $H_\infty(I)$ divided by the sampling interval τ. We notice that the answer to this question does not follow from the finite interval result. Specifically, let the joint entropy in the finite interval case be written as $H(I_1, I_2, \ldots, I_N) = NH_N(I)$, where I denotes the quantized process and N is the number of samples in the finite interval considered. We observe that $H_\infty(I)$, the entropy-rate of the process I, might be significantly smaller than $H_N(I)$, the N^{th} order entropy of the process I. Consequently, it is conceivable that when multiplying $H_\infty(I)$, which plays the role of $H_N(I)$ in the finite interval case, by one over the sampling interval, which plays the role of N in the finite interval case, the product might tend to a finite value or even zero, as the sampling interval goes to zero. The following is an example of a process for which the product is in fact zero for all τ.

Let X_t be a stationary random process constructed from a random variable W that is uniform on $[0, 1]$ in the following way:

$$X_t = \begin{cases} 1, & t \in [k - 1 + W, k + W\,] \quad \text{for } k \text{ even} \\ -1, & \text{else} \end{cases} . \tag{6.1}$$

For this process $H_{\tau,\infty}(I)$ is zero for any τ, where τ is the sampling interval. (We subscript $H_{\tau,\infty}(I)$ by τ to reflect the dependence on τ.) Consequently, $\lim_{\tau\to 0}\frac{1}{\tau}H_{\tau,\infty}(I) = 0$. Here is a brief sketch of why $H_{\tau,\infty}(I) = 0$ for any τ. Given some τ, with probability one the sample path of the random process is such that the time of the zero crossing in $[0, 1]$ is an irrational multiple of τ. Consequently, each quantized sample adds some information about the time of the zero crossing in $[0, 1]$. Specifically, one can use the quantized samples in the interval $(-\infty, 0)$, of which there is a countable number, to obtain the time of the zero crossing in the interval $[0, 1]$ (and hence all future crossings) perfectly. This in turn implies that $H_{\tau,\infty}(I) = 0$. We notice that a countable number of, say, bits can be used to describe perfectly a real value, since real values can be represented by their binary expansions.

Future research is also warranted with regard to the asymptotic formula for conditional entropy of quantized Gaussian random variables. Specifically, it would be interesting to extend this formula to arbitrary order of conditioning. Beyond the intellectual challenge in doing so, such an extension can be used to determine the rate at which the joint entropy of quantized Gaussian samples over a finite interval tends to infinity as the sampling interval goes to zero. We believe, however, that deriving such an extended formula would be very difficult using the same method that was used to show the current formula for conditional entropy.

Additionally, it is interesting to ask whether for a Gauss-Markov process, the current formula also holds for arbitrary order of conditioning. While the fact that the process is Markov does not necessarily imply that the quantized process is Markov, it is not unreasonable to expect that perhaps the quantized process is sufficiently close to being Markov, in the sense that further conditioning does not alter the rate at which the conditional entropy tends to zero.

Furthermore, it would be interesting to obtain such an asymptotic formula for the case that the uniform quantizers have offset other than half, i.e. for the case that the mean of the Gaussian random variables does not lie at the midpoint of a quantization cell. Additionally, as in the low resolution case, it is natural to consider source densities other than Gaussian and seek asymptotic conditional entropy formulas for them.

Another issue deserving further consideration is the extension to higher dimensional random fields and to vector quantization of the results concerning the convergence of joint entropy of quantized samples over a finite interval, and of the entropy rate divided by the sampling interval, to infinity as the sampling interval goes to zero. We are quite certain the results shown for one-dimensional random process with scalar quantization, extend naturally to higher dimensions both of the random process and the quantization. If so, it would imply, for example, that the joint entropy of quantized samples of scalar quantizers with a finite period dither, which is a type of vector quantization, would also tend to infinity as the sampling interval goes to zero.

Finally with regard to Chapter V, an avenue of exploration is to examine whether nonlinear interpolation affects the conclusion that rate must go to infinity as sensor density goes to infinity to attain a fixed target MSE with an encoder that uses identical scalar quantizers and entropy coding. In Chapter V we argued that with simple linear interpolation fixing D is equivalent to fixing the quantizer. But potentially, nonlinear interpolation with a fixed quantizer might attain distortion that tends to zero. If so, then with a fixed target distortion, we could change the quantizer with the sampling interval τ (making it coarser and hence having less output entropy) and attain distortion D, thus, perhaps, attaining a rate that does not go to infinity

as τ goes to zero.

We comment that the process given in (6.1) is an example for which linear interpolation can be used to obtain distortion that tends to zero as $\tau \to 0$. However, this process is degenerate, in the sense that its rate-distortion function is zero.