Studies on the Asymptotic Behavior of Parameters in Optimal Scalar Quantization

by

Victoria B. Yee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2010

Doctoral Committee:

Professor David L. Neuhoff, Chair
Professor Jeffrey A. Fessler
Professor Alfred O. Hero III
Emeritus Professor Michael B. Woodroofe

For my family

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Studies on the Asymptotic Behavior of Parameters in Optimal Scalar Quantization
by
Victoria B. Yee

Chair: David L. Neuhoff

The goal in digital device design is to achieve high performance at low cost, and to pursue this goal, all components of the device must be designed accordingly. A principal component common in digital devices is the quantizer, and frequently used is the minimum mean-squared error (MSE) or *optimal*, fixed-rate scalar quantizer. In this thesis, we focus on aids to the design of such quantizers.

For an exponential source with variance σ^2, we estimate the largest finite quantization threshold by providing upper and lower bounds which are functions of the number of quantization levels N. The upper bound is $3\sigma \log N$, $N \geq 1$, and the lower bound is $3\sigma \log N + o_N(1)\,\sigma - 1.46004\,\sigma$, $N > 9$. Using these bounds, we derive an upper bound to the convergence rate of $N^2 D(N)$ to the Panter-Dite constant, where $D(N)$ is the least MSE of any N-level scalar quantizer. Furthermore, we present two, very simple, non-iterative and non-recursive suboptimal quantizer design methods for exponential sources that produce quantizers with good MSE performance.

For an improved understanding of the half steps and quantization thresholds in optimal quantizers as functions of N, we use as inspiration the result by Nitadori [19] where, exploiting a key side effect of the source's memoryless property, he derived an infinite sequence such that for any N, the kth term of the sequence is equal to the kth half step (counting from the right) of the optimal N-level quantizer designed for a unit variance exponential source. In our work, using an asymptotic version of this key side effect which holds for general exponential (GE) sources parameterized by an exponential power p and a utilizing a method of our own devising, we show that for such a source, the kth half step of an optimal N-level quantizer multiplied by the $(p-1)$st power of the kth threshold approaches the kth term of the Nitadori sequence

as N grows to infinity. Thus, the Nitadori sequence asymptotically characterizes the cells of MMSE quantizers for GE-sources, as well as exponential.

CHAPTER I

Introduction

1.1 Motivation and Problem Statement.

We live in a digital world. Everywhere we look, there is plenty of evidence to support the claim that society has certainly embraced the conveniences afforded by using digital devices such as cell phones, PDAs and laptop computers. The pervasiveness of society's use of digital machines is, in large part, due to technological advances which have improved the speed, accuracy and precision with which these machines perform tasks. Lower production and operating costs, also important factors, can often times be brought about by improvements in the design and development of devices as well. All of these factors, improved performance and lower cost, lead to wide availability, affordability and utility for the public which results in the ubiquitous use of these devices in our daily lives.

Justifiably, much of the gain in performance of these devices has been attributed to progress made in the design and development of central processing units (CPUs) and of the algorithms run by CPUs. This makes sense because, as seen in Figure 1.1, each digital device contains a CPU and the CPU utilizes algorithms to manage the process of completing a task. Thus, if the CPU has increased capability (e.g., faster processing speed, lower operating temperature, etc.) to oversee a task, then this gain will manifest itself in the device's performance. Concentrating solely on improving device performance through advances in CPU and algorithm technology, however, limits the gains that can be achieved. This is because the CPU, through its algorithm, acts on binary/numerical input that represents information regarding a real-world event (e.g., outdoor temperature at 5 p.m.). If the numerical input data provided to the CPU through the device's interface components is crude, then the effectiveness of the CPU to manage the task through its output is clearly hindered. Thus, to achieve the best performance gain for the device, all components of the device must be considered for improvement. In this thesis, our focus centers on one of the other components common to digital devices: the quantizer.

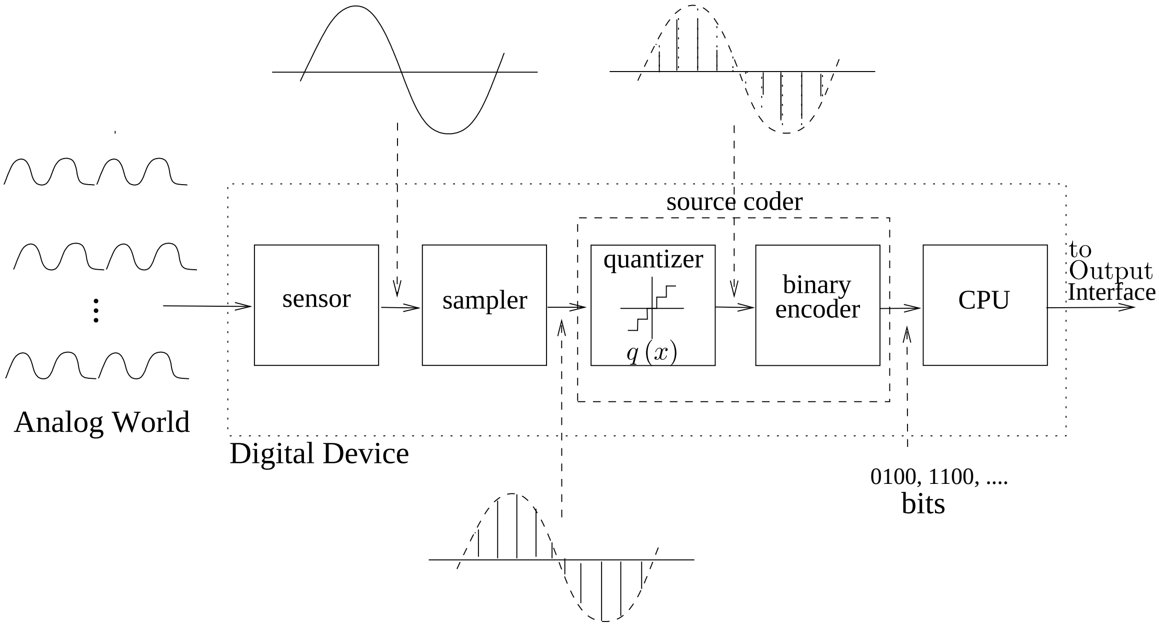


Figure 1.1: Simplified schematic of a general digital device showing its component parts and processing chain.

Housed inside of the source coder (see Figure 1.1), a quantizer takes samples x of an analog input, modeled as a random variable X, and produces a discrete-valued output which is then converted into bits by the binary encoder. One of the most simple, yet commonly used quantizers is the fixed-size scalar quantizer which takes the real value x and maps it to another real value μ_i belonging to a finite subset of N real numbers $\{\mu_i\}_{i=1}^{N}$ called quantization levels.[1] In other words, a quantizer effectively partitions the real numbers into N subsets $\{S_i\}_{i=1}^{N}$ and assigns to each subset $S_i = [t_i, t_{i-1})$, a real-value μ_i so that to each analog input x received by the quantizer, the output μ_i that is produced corresponds to the particular subset S_i that x belongs to, i.e., $x \in S_i$. From this description, it is clear that the output value μ_i produced by the quantizer is generally an approximation of the input value x and that increasing N should serve to improve the approximation. However, increasing N increases the complexity of not only the quantizer's implementation, but also of its output which is seen as an increase in the range of possible output values. Since increased complexity in the quantizer's output requires increased complexity in all of

[1] The quantizer described is known in the literature as a fixed-rate scalar quantizer, and from this point on, any reference to a quantizer will be of this type.

the other device components following it in the processing chain shown in Figure 1.1, it is clear that increasing N translates to a rise in overall cost. Thus, while creating a quantizer involves

- deciding on the size N of the partition,
- deciding how to partition the real numbers into a set of N intervals $[t_i, t_{i-1})$, and
- for each interval $[t_i, t_{i-1})$ of the partition, deciding which value μ_i to assign to it,

it is the job of the quantizer designer to formulate a quantizer that provides a sufficient level of approximation accuracy to the CPU while keeping N as small as possible in order to minimize cost.

Minimum mean-squared error (MMSE) or optimal quantizers. A commonly used measure of accuracy is mean-squared error and N-level minimum mean-squared error (MMSE) quantizers are quantizers which achieve the lowest mean-squared error for a given partition size N. (See Figure 1.2 for a diagram of an optimal quantizer.) The MMSE or *optimal* quantizer is a popular design choice that is favored for several reasons: Performance is mathematically-based and tractable (as opposed to task-dependent performance criteria); there is available the Lloyd-Max algorithm ([13], [15]), a general, iterative algorithm for designing optimal quantizers for a variety of input sources ([7] and [23]); and perhaps most importantly, these quantizers work pretty well for a variety of applications (i.e., low MSE seems to correspond to high task-based accuracy). Therefore, for a given value of N, the optimal quantizer designer must determine the positions of the quantization thresholds t_i which define a partition of the real numbers while also designating the values μ_i (one value for each subset of the partition) so that the mean-squared error produced by the quantizer is minimized.

Our goal. In this thesis, we seek to provide aids to the design of optimal quantizers by investigating the structural relationships that exist between the partition thresholds t_i (called quantization thresholds) and the quantizer output values μ_i (called reconstruction levels) of optimal quantizers. In particular, our study focuses on analyzing the behavior of the difference $\mu_i - t_i$ which we will refer to as the half step of the ith quantization cell as it relates to the corresponding half steps of neighboring quantization cells.

1.2 Background and Known Results.

Optimal quantizer design: The Lloyd-Max algorithm. The Lloyd-Max algorithm ([13], [15]) is a general design method which can be used to design quantizers for a variety of input sources.[2] The method is based on the observation (made independently by Lloyd [13] and by Max [15]) that the thresholds and levels of optimal quantizers must satisfy two optimality conditions: 1) μ_i must be the centroid of the set to which it belongs, and 2) the threshold t_i lying between any two adjacent levels μ_i, μ_{i+1} be equidistant from them. After being initialized with an estimate of the *key parameter* or *support threshold* (the largest finite threshold) of the desired quantizer, the algorithm uses the optimality conditions to iteratively converge to a numerical solution, consisting of the set of thresholds and levels that characterize the requisite quantizer. The significance of starting the algorithm with an accurate support threshold approximation can be seen by observing that fewer algorithmic iterations are required the closer the support threshold estimate is to the optimal value. Thus, the sensitivity of the algorithm (as measured by the number of iterations performed) to the accuracy of the initializing support threshold estimate complicates use of this method since it necessitates some knowledge of the desired quantizer before it has yet been designed.

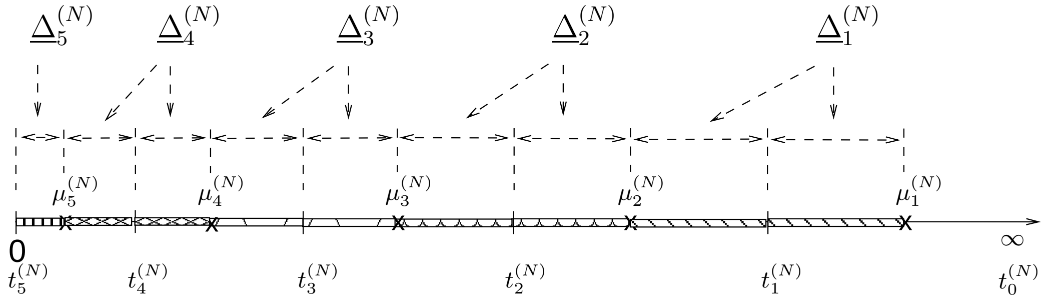

$\underline{\Delta}_5^{(N)}$
$\underline{\Delta}_4^{(N)}$
$\underline{\Delta}_3^{(N)}$
$\underline{\Delta}_2^{(N)}$
$\underline{\Delta}_1^{(N)}$
$\mu_5^{(N)}$
$\mu_4^{(N)}$
$\mu_3^{(N)}$
$\mu_2^{(N)}$
$\mu_1^{(N)}$
0
∞
$t_5^{(N)}$
$t_4^{(N)}$
$t_3^{(N)}$
$t_2^{(N)}$
$t_1^{(N)}$
$t_0^{(N)}$

Figure 1.2: Diagram of a 5-level optimal scalar quantizer, where $t_k^{(N)}$ are the quantization thresholds, $\mu_k^{(N)}$ are the quantization levels and $\underline{\Delta}_k^{(N)}$ are the half steps. Note: The notation used in this figure will be formally defined in Chapter II.

Optimal exponential quantizer design: Nitadori's method. In the special case of optimal quantizer design for exponential sources, there exists a non-iterative,

[2] For more information on this topic, see Chapter II, Section 2.2, under Optimality Conditions (especially the footnote in that discussion), where under certain source conditions, satisfying the optimality criteria becomes sufficient to guarantee that the solution produced by the Lloyd-Max algorithm is both unique and optimal.

recursive design method devised by Nitadori [19] that produces the exact specification for such quantizers. His method [19], which also uses the optimality conditions ([13], [15]), advantageously exploits a special property exclusive to exponential sources, the memoryless property. This property allows for the construction of a generating equation, called the Nitadori generator, which when solved, yields the quantizer specifications as sequence. In more detail, the solution sequence, referred to as the Nitadori sequence [19], is an infinite series of values such that for any N, the kth term of the sequence is equal to the kth half step, defined as the distance between the quantization level μ_k and the lower threshold t_k of the kth quantization cell (see Figure 1.2), of the optimal N-level quantizer designed for a unit variance exponential source. Solving the Nitadori generator in order to determine the sequence values is, however, a recursive process, which generates the next sequence value based on knowing the previous value, that can only be done numerically. Thus, while this method is unlike the Lloyd-Max algorithm in that it is a non-iterative way to design optimal exponential quantizers, the numerically generated solutions it produces obscures, just as is the case when using the Lloyd-Max algorithm, any perception of the relationships that exist between the parameters (t_i, μ_i) of such quantizers.

Also in [19], Nitadori showed that the MSE of an optimal, N-level, unit variance exponential quantizer is exactly equal to the square of the Nth term of the Nitadori sequence.

Asymptotic quantization theory and asymptotically optimal quantizers. In addition to directly studying the structure of optimal quantizers, asymptotic quantization theory, which considers the case when the number of levels N is large in a quantizer, along with the study of asymptotically optimal quantizers, quantizers whose N-level MSE performance divided by the least achievable N-level MSE performance converges to 1 as the number of levels N increases to infinity, have together yielded important relationships related to the number of levels N in such quantizers. The combined results by Bennett [2], and Panter and Dite [20] in these areas give:

- An implementation method for realizing asymptotically optimal quantizers that utilizes companding (asymptotically optimal companding system)
- An expression (in N) for the support threshold of asymptotically optimal companding systems (which does not yet exist for optimal quantizers)
- A limiting relationship that indicates how the MSE of optimal quantizers decays as a function of N

- An asymptotically optimal compressor function (part of an asymptotically optimal companding system) which can be interpreted as a distribution-like function which "predicts" the distribution of quantization levels as a function of N according to a specific location along the real axis (optimal point densities)

Since the connection between the MSE performance of asymptotically optimal quantizers and the MSE performance of optimal quantizers is asymptotic in N, it would seem natural that the structure of asymptotically optimal quantizers and the structure of optimal quantizers (in terms of thresholds t_i, levels μ_i and half steps $\mu_i - t_i$) should also resemble each other asymptotically in N as well. This idea that quantizers with similar performance should have similar structure (in terms of the placement of thresholds and levels) is the seed for some of the following results discussed next.

Known support threshold estimation results. Work on developing support threshold estimation techniques has mostly relied on adhoc methods to produce estimates. Bucklew and Gallagher[3] used the support threshold of an asymptotically optimal companding system to approximate the support threshold for an optimal quantizer. Lu and Wise [14] constructed an estimator of the form $c_1 \log N + c_2$ via curve-fitting, by first computing the exact support thresholds of optimal quantizers (for Gaussian, Laplacian and Rayleigh sources) with $N = 4$ to 64 levels. The constants c_1 and c_2 were then determined using least squares and extrapolation was used to generate estimates for values $N > 64$. Na and Neuhoff [17] proposed a numerical method of estimating the support threshold that is based on minimizing an approximation (from asymptotically optimal quantization theory) to the MSE of optimal quantizers. Na and Neuhoff [18] also developed several other adhoc (non-numerical) estimators (for the quantization of generalized gamma sources). All of these estimators possessed the same dominant function (in N, according to the source being quantized) which indicated the likelihood that their estimators had captured the "correct" growth rate (in N) of the optimal support threshold function. Numerical evaluation of these estimators against the true support threshold values appear to confirm this.

Rigorous theoretical work on support threshold estimation for optimal quantizers has been more difficult to come by. For optimal quantizers, only the following results related to optimal, unit variance Laplacian quantizers (Na [16]) have been reported:

1. The ratio of the support threshold over the function $\frac{3}{\sqrt{2}} \log N$ (which is the estimator given in [3]) goes to 1 as $N \to \infty$.

2. The $\limsup$ (in N) of the difference between the support threshold and $\frac{3}{\sqrt{2}} \log N$ is less than a small constant 0.0669.

In regards to support threshold estimation of N-level *uniform scalar quantizers*, a suboptimal class of quantizers in which quantizers use N equal-sized intervals to partition the real numbers, along with output values which are the midpoints of those intervals, Hui and Neuhoff [10] rigorously derived the support threshold functions (in N) of such quantizers for the family of generalized gamma sources.

1.3 Contributions.

Chapter III: Aids to the design of optimal exponential quantizers. Since our primary interest is to study the relationships that govern the placement of thresholds and levels in optimal quantizers in general so as to improve our understanding of how to design such quantizers, we focus on the exponential case because, even though the Nitadori design method eliminates the need to find another procedure for creating optimal exponential quantizers, it still remains unclear how the thresholds and levels are related to one another. Any information we ascertain, we hope to generalize as aid in the construction of optimal quantizers designed for other sources. Thus, in a way, the exponential case is a natural starting point from which to glean intuition on these relationships since there already exists an expression (the Nitadori generator, albeit an obscure one) that relates the half steps of neighboring quantization cells to one another.

In this chapter, we focus our work primarily on the problem of rigorous support threshold estimation in optimal exponential quantizers. From our efforts, we were able to achieve the following results:

- We have derived theoretical bounds to the support threshold function of optimal, fixed-rate quantizers designed for one-sided, exponential sources with variance σ^2, where the difference between the bounds converges to a constant that depends only on σ^2. From this result, it is clear that the ratio of the bounds converges to 1 as $N \to \infty$, as was shown by Na [16]. The upper bound we provide, $3\sigma \log N$, for $N \geq 1$, is tighter than the $\limsup$ bound stated in [16], and the lower bound given, $3\sigma \log N + o_N(1)\,\sigma - 1.46004\,\sigma$, for $N > 9$, is the first to be reported. In addition, our result analytically confirms that the support threshold grows logarithmically in N and that the estimates given in [3] and [18] are essentially correct.

- Regarding the well-known fact established by Panter and Dite [20] that $N^2 D(N) \to \frac{\beta}{12}$ as $N \to \infty$, where $D(N)$ is the least MSE of an N-level quantizer and β (a source dependent constant) is the Panter-Dite constant [20], we have, in addition, used our approach to support threshold estimation to derive a result which concerns when it is effective to use the Panter-Dite formula [20] to estimate the mean-squared error performance $D(N)$ of optimal exponential quantizers. More specifically, we have derived an upper bound to the convergence rate of $N^2 D(N)$ to $\frac{\beta}{12}$ as a function of N. This is the first such bound.

- From our approach to constructing the support threshold lower bound, we present two, non-iterative and non-recursive, Nitadori-like suboptimal quantizer design methods for exponential sources. These methods are so simple to use that designing one of these quantizers does not require the use of a computer or a huge database of tabulated values. They can be designed by-hand, knowing only the first eight values of the Nitadori sequence. Moreover, quantizers produced in this manner yield good (low) mean-squared error performance. (Example: For a quantizer designed using the sequence $v_k = \underline{\eta}_k$, $1 \leq k \leq 8$, and $v_k = \frac{3}{2k}$, $k \geq 9$, and $N = 64$, the ratio of the MSE for the v_k quantizer over the MSE of an optimal quantizer with 64 levels is less than 1.0014. By comparison, for the same number of levels ($N = 64$), the ratio of the MSE for an asymptotically optimal companding system over the MSE of an optimal quantizer is greater than 1.0169.)

Chapter IV: An extension of Nitadori's sequence in relation to optimal quantization of sources other than exponential. The desire to continue our pursuit to amass information on the ties that exist between the parameters of MMSE quantizers with the goal of providing aid the problem of optimal quantizer design provided the main impetus behind the work described in Chapter IV. In this chapter, we set out to find a way to generalize the result by Nitadori [19] to optimal quantizers designed for sources other than exponential. Since Nitadori's method depends on the memoryless property which is unique to the exponential source, a direct generalization was not revealed in our studies. However, our investigation into the asymptotic behavior of the parameters of optimal quantizers (half steps and quantization thresholds, refer to Figure 1.2) designed for general exponential (GE) sources, parameterized by an exponential power p, led to a surprising result: We show that for a GE-source, the kth half step of an optimal N-level quantizer multiplied by the $(p-1)$st power of the kth threshold approaches the kth term of the Nitadori sequence as N grows to infinity. Thus, the Nitadori sequence not only provides the specifications for optimal exponential quantizers, it also asymptotically characterizes the cells of MMSE quantizers for GE-sources.

CHAPTER II

Quantization Review

This chapter provides a brief review of minimum mean-squared error (MMSE) or optimal scalar quantization. We only describe the results from the literature that are relevant to the work reported in this thesis. During this discussion, we will be introducing much of the notation and symbols used to indicate quantization parameters of interest. For the sake of clarity, we only discuss the quantization of one-sided, non-negative, finite variance, real-valued sources that have support equal to the contiguous half open interval $[0, \infty)$. We remark that the extension to two-sided, real-valued sources with finite variance and contiguous support over all of the real numbers is a fairly straightforward extension of the one-sided case. Also, whenever we refer to quantization, we mean scalar quantization.

2.1 Scalar Quantization.

Let X be a non-negative, real-valued, scalar random variable with finite mean, finite variance and infinite support. Suppose we want to represent the values we observe from the random variable X and suppose we will only be able to distinguish between at most N different values. Questions concerning this scenario arise quickly: Which N non-negative real numbers would best represent observed/sample values of the source X and how should the sample values of X be assigned to these N representative real numbers? Before addressing possible answers to the question of what the "best representation" for X would be, we first clarify the functionality of a "machine" who's input is $X \geq 0$ and produces an output representation of the sample values of X, the scalar quantizer.

An N-level, scalar quantizer is defined by three objects:

- A set of *quantization thresholds* t_k, $k = 0, 1, \ldots, N$, where $t_0 \triangleq +\infty$ and $t_0 >$

$t_1 \geq \cdots \geq t_N$

- A set of *reconstruction levels* $\mu_k \geq 0$, $k = 1, 2, \ldots, N$
- A *quantization rule* $q : [0, \infty) \rightarrow [0, \infty)$ such that $q(x) = \mu_k$, if $x \in [t_k, t_{k-1})$ for all $x \geq 0$

In other words, it is clear that for any observed value $x \geq 0$, a quantizer, with thresholds $\{t_i\}_{i=0}^N$, levels $\{\mu_i\}_{i=1}^N$ and quantization rule q, maps x to the value of μ_k if x belongs to the kth quantization cell $[t_k, t_{k-1})$.

Before leaving this brief discussion of the definition of a scalar quantizer, we point out that the indexing scheme used in the definition begins with the quantization cell that is farthest away from the origin, and so quantization cell indexing begins from the right and moves left toward the origin as the index is increased. This particular indexing scheme is a great aid to our work and this is why we call attention to it. This indexing scheme is shown in Figure 2.1.

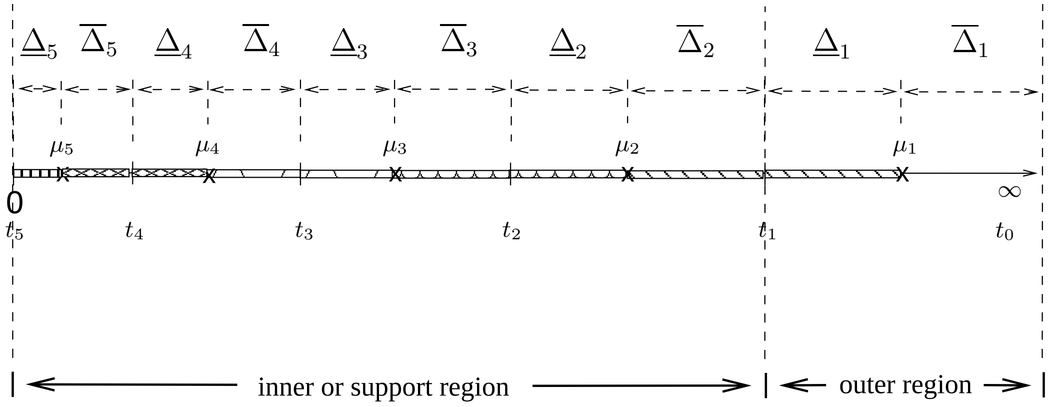


Figure 2.1: Illustration of a 5-level scalar quantizer.

Traditionally, the support region of the random variable X, in our case, the non-negative reals, is divided into two quantization regions: The *inner* or *support region* of the quantizer $[0, t_1)$ is defined to be the distance from the *support threshold* t_1 to the origin, and the *outer* or *overload region* of the quantizer $[t_1, \infty)$ is defined to be region of the reals that is greater than or equal to the support threshold t_1. This quantization region should not be confused with what we refer to as the *tail region of the source* which is the region where $x >> 0$ or when x is *far* from the origin, though often times, the outer region and tail region of the source may, in fact, coincide, as is the case when $t_1 >> 0$.

2.2 Quantizer Performance: Mean-Squared Error.

Returning to the question, "What is the best representation of X?", we arrive at the topic of quantizer performance.

The performance of a quantizer is assessed by how well accuracy is achieved for a given cost or constraint. Accuracy is determined by choosing a fidelity criteria (or goodness measure) by which the effectiveness of the quantizer's ability to present the values of X will be measured. For our work, we have chosen the popular and tractable mean-squared error (MSE) criterion as our "goodness" measure and we have chosen to track the number of levels N in a quantizer as our cost variable. Thus comparing the performance of two different N-level quantizers designed for the same source becomes simply a matter of looking at the MSE corresponding to each quantizer and the quantizer with the lower MSE is deemed to be superior. Note that in the quantization literature, the number of levels N is related to the rate $R = \log_2 N$ or bits used by the quantizer to index/distinguish between quantization cells, and thus we could use rate R as our cost variable, but for our work, using N is preferable.

Assuming that X has a *probability density function* (pdf) $f(x)$, $x \geq 0$ (this is an assumption that we will use throughout this thesis), the MSE or *MSE distortion* of an N-level quantizer designed for X is

$$E\left[(X - q(X))^2\right] \triangleq D(q) \triangleq \sum_{k=1}^{N} \int_{t_k}^{t_{k-1}} (x - \mu_k)^2 f(x)\, dx. \tag{2.2.1}$$

This is a static statistic that can be compared against the MSE of other N-level quantizers designed for the same random variable. An N-level quantizer with the lowest or minimum mean-squared error (MMSE) is judged to be *optimal.* (Throughout this thesis, whenever we refer to an optimal quantizer, we mean a quantizer that minimizes mean-squared error.)

Thus, with a fixed number of levels N, the goal in MMSE quantizer design is to find a quantizer that performs with mean-squared error equal to

$$\min_{\{t_i\}_{i=0}^{N},\, \{\mu_i\}_{i=1}^{N}} E\left[(X - q(X))^2\right].$$

In the quantization literature, the optimum performance theoretically attainable (OPTA) function, which gives the least MSE distortion of any scalar quantizer with not more than N levels, is

$$D(N) \overset{\triangle}{=} \min_{q:N(q)\leq N} D(q), \tag{2.2.2}$$

where $N(q)$ equals the number of level in the quantizer q.[1] Thus, when designing with MMSE performance in mind, the goal of MMSE quantizer design becomes a search to find the thresholds $\{t_i^{(N)}\}_{i=0}^{N}$ and reconstruction levels $\{\mu_i^{(N)}\}_{i=1}^{N}$ in an N-level quantizer that attains the MSE given by $D(N)$.

Optimality conditions: An aid to MMSE quantizer design. Given a fixed source X with pdf $f(x)$, MMSE quantization design is simplified by knowing that the thresholds and reconstruction levels of an MMSE quantizer must adhere to the optimality conditions [13], [15]:

1. **Centroid condition.** The reconstruction levels $\mu_k^{(N)}$, $k = 1, 2, \ldots, N$, are equal to the centroids of the quantization cells to which they belong

$$\mu_k^{(N)} = \frac{\int_{t_k^{(N)}}^{t_{k-1}^{(N)}} uf(u)\,du}{\int_{t_k^{(N)}}^{t_{k-1}^{(N)}} f(w)\,dw}, \tag{2.2.3}$$

   where it is clear that the centroid of kth quantization cell is equal to the conditional mean of X given $X \in [t_k^{(N)}, t_{k-1}^{(N)})$.

2. **Nearest Neighbor condition.** The quantization threshold $t_k^{(N)}$ lying between any two adjacent finite reconstruction levels $\mu_k^{(N)}$, $\mu_{k+1}^{(N)}$ lies halfway between the two thresholds, i.e.,

$$\mu_k^{(N)} - t_k^{(N)} = t_k^{(N)} - \mu_{k+1}^{(N)} \tag{2.2.4}$$

   for all $k = 1, 2, \ldots, N-1$.

The optimality conditions are, in general, only necessary conditions for optimality, i.e., for a quantizer to minimize mean-squared error. However, for exponential sources and sources with strictly log convex pdfs, the thresholds and reconstruction levels for

[1] The OPTA is generally defined using "inf" instead of "min" in the literature. However, for finite dimensional quantization, the OPTA is always achievable at each value of N, and thus, we state the OPTA definition in (2.2.2) using "min".

an N-level MSE quantizer that satisfy the optimality conditions are unique,[2] i.e., the optimality conditions are sufficient for determining the single N-level quantizer that minimizes MSE.

2.3 Quantizer Re-parametrization Using Half Steps.

As shown in Figure 2.1, an alternative, yet equivalent characterization of a scalar quantizer that is more useful for our work discussed in Chapters III and IV specifies the positions of the quantization thresholds and reconstruction levels of an N-level scalar quantizer in terms of the *lower half steps* or just *half steps* $\underline{\Delta}_k \overset{\triangle}{=} \mu_k - t_k$ and *upper half steps* $\overline{\Delta}_k \overset{\triangle}{=} t_{k-1} - \mu_k$ where $k = 1, 2, \ldots, N$ is the quantization cell index of the quantizer. With the step size of the kth quantization cell defined to be $\Delta_k \overset{\triangle}{=} t_{k-1} - t_k$, $k = 1, 2, \ldots, N$, the $1-1$ correspondence between a set of half steps and a set of quantization thresholds and reconstruction levels is made clear by the following relationships:

1. $t_0 = \infty$, $t_N = 0$,
2. $t_k = \sum_{i=k+1}^{N} \Delta_i = \sum_{i=k+1}^{N} \underline{\Delta}_i + \overline{\Delta}_i$, $k = 1, 2, \ldots, N$,
3. $\mu_k = t_k + \underline{\Delta}_k$, $k = 1, 2, \ldots, N$.

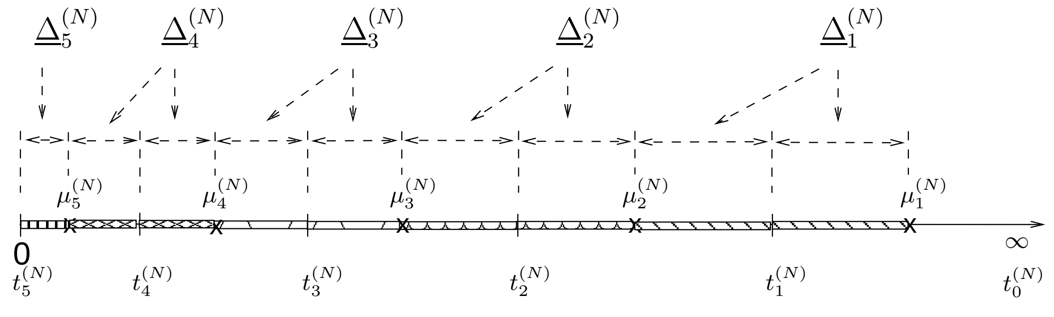
For an N-level, optimal scalar quantizer, the definition of the upper half step $\overline{\Delta}_k^{(N)}$, $k = 1, 2, \ldots, N$, is actually not necessary since adherence to the nearest neighbor condition causes $\overline{\Delta}_k^{(N)} = \underline{\Delta}_{k-1}^{(N)}$, $k = 2, 3, \ldots, N$. In this case, the $1-1$ correspondence between the set of (lower) half steps $\underline{\Delta}_k^{(N)}$, $k = 1, 2, \ldots, N$, of an N-level optimal scalar quantizer and its set of quantization thresholds $\{t_i^{(N)}\}_{i=0}^{N}$ and reconstruction levels $\{\mu_i^{(N)}\}_{i=1}^{N}$ becomes

1. $t_0^{(N)} = \infty$, $t_N^{(N)} = 0$,
2. $t_k^{(N)} = \underline{\Delta}_N^{(N)} + \underline{\Delta}_k^{(N)} + \sum_{i=k+1}^{N-1} 2\underline{\Delta}_i^{(N)}$, $k = 1, 2, \ldots, N$,
3. $\mu_k^{(N)} = t_k^{(N)} + \underline{\Delta}_k^{(N)}$, $k = 1, 2, \ldots, N$.

Thus we see that if the set of thresholds and levels for an N-level optimal scalar quantizer is unique, then so is the set of half steps for the same quantizer. Since knowing the set of half steps $\{\underline{\Delta}_i^{(N)}\}_{i=1}^{N}$ belonging to an N-level optimal quantizer is

[2] Uniqueness for strictly log convex sources was proved by Fleischer in [7]. Fleischer [7] also proposed uniqueness for exponential sources, but his argument was later corrected by Trushkin [23].

sufficient to knowing $\{t_i^{(N)}\}_{i=0}^N$ and $\{\mu_i^{(N)}\}_{i=1}^N$ (as shown in Figure 2.2), in our work, we focus on determining the values of $\underline{\Delta}_k$, $k = 1, 2 \ldots, N$ when trying to define N-level optimal quantizers.



$\underline{\Delta}_5^{(N)}$
$\underline{\Delta}_4^{(N)}$
$\underline{\Delta}_3^{(N)}$
$\underline{\Delta}_2^{(N)}$
$\underline{\Delta}_1^{(N)}$
$\mu_5^{(N)}$
$\mu_4^{(N)}$
$\mu_3^{(N)}$
$\mu_2^{(N)}$
$\mu_1^{(N)}$
0
∞
$t_5^{(N)}$
$t_4^{(N)}$
$t_3^{(N)}$
$t_2^{(N)}$
$t_1^{(N)}$
$t_0^{(N)}$

Figure 2.2: A 5-level optimal scalar quantizer with half step parametrization shown.

Before proceeding, we remark briefly that the use of half steps in gauging quantizer performance is also possible and indeed, in the case of exponential optimal quantization, half steps provide a means of determining the exact MSE performance of such quantizers [19]. We leave, however, this discussion for Chapters III and IV.

2.4 The Lloyd-Max (LM) Design Algorithm for MMSE Scalar Quantizers and Support Threshold Estimation.

Historically, the first published algorithm for designing MMSE quantizers was independently formulated by Lloyd [13] in 1957 and Max [15] in 1960. In short, the algorithm produces a set of thresholds and reconstruction levels that adheres to the optimality conditions (see (2.2.3) and (2.2.4)) based on an initial estimate of the support threshold or *key parameter* $t_1^{(N)}$. Since the algorithm's performance, as measured by the number of iterations it takes to produce an acceptable design solution, relies heavily on the accuracy of the support threshold estimate that the algorithm is initialized with (e.g., for a unit variance Laplacian source, an accuracy criteria of $\delta = 1 \times 10^{-5}$, and $N = 1000$, the algorithm initialized with the support threshold estimator proposed by Na and Neuhoff in [17] required 37.5% fewer iterations than the number required by the algorithm initialized with the Bucklew and Gallagher estimator given in [3]), there is interest in finding methods for accurate support threshold estimation as a function of N. To shed more light on the impact of the initial support threshold estimate on algorithm performance, we give the following overview:

> Suppose the goal is to design an N-level quantizer for a given source that is optimal or, in reality, close to optimal. Since the LM algorithm

is iterative, a stopping criteria which indicates how close to optimal a quantizer generated after the ith iteration is must be chosen. Examples of stopping criteria are: Stop when the decrease in MSE between consecutive design iterations falls below a prescribed value or stop when the estimated threshold for $t_N^{(N)}$ is within a preset distance of the origin.

Using i as the iteration index, at the outset of each iteration of the LM algorithm, an estimate of the support threshold $t_1^{est,i}$ is made. Using $t_1^{est,i}$, along with the knowledge that all thresholds and reconstruction levels in optimal quantizers must adhere to the optimality constraints, the remaining estimated values of the quantization thresholds $t_k^{est,i}$, $k = 2, 3, \ldots, N$ and reconstruction levels $\mu_k^{est,i}$, $k = 1, 2, \ldots, N$ are computed. To see how this is accomplished, with ${t_0}^{est,1} = \infty$, for each $k = 1, 2, \ldots, N - 1$,

1. ${\mu_k}^{est,1}$ is the centroid of $[{t_k}^{est,1}, {t_{k-1}}^{est,1})$.
2. ${\mu_{k+1}}^{est,1} = {t_k}^{est,1} - {\mu_k}^{est,1}$ since ${\mu_{k+1}}^{est,1}$ satisfies the nearest neighbor optimality condition.
3. Since ${\mu_{k+1}}^{est,1}$ is the centroid of $[{t_{k+1}}^{est,1}, {t_k}^{est,1})$, solve for ${t_{k+1}}^{est,1}$.

Once all $N+1$ thresholds and N reconstruction levels have been estimated, the stopping condition is checked, and if it is satisfied, the algorithm halts and the quantizer design process is complete. If the stopping condition has not been satisfied, the iteration index is increased $(i \to i + 1)$ and an updated estimate of the support threshold $t_1^{est,i+1}$ is created so that the algorithm can run again, producing an updated set of $N-1$ thresholds $t_k^{est,i+1}$, $k = 2, 3, \ldots, N$ and N reconstruction levels $\mu_k^{est,i+1}$, $k = 1, 2, \ldots, N$, and another check against the stopping criteria is made.

It is noted that since the resulting MSE produced by the quantizers generated at the end of each iteration of the algorithm decreases, the number of completed iterations is finite, with the final iteration being the one that produces a quantizer with MSE or an estimate for $t_N^{(N)}$ that meets the stopping criteria.

It should be clear from the discussion above that an accurate initial estimate $t_1^{est,1}$ for $t_1^{(N)}$ would reduce the number of iterations that the algorithm must complete in order to produce an acceptable design solution, and hence the need for accurate support threshold estimation. For optimal quantization of exponential sources, the topic of support threshold estimation is discussed in Chapter III.

2.5 Quantizer Implementation and Design: Companding.

Companding systems or companders are one method of implementing a quantizer design. The popularity of companding as a means of quantizer implementation stems from two facts: i) any N-level scalar quantizer can be realized through use of a companding system, and ii) companding systems utilize the simpler-to-execute N-level *uniform scalar quantizer* (USQ), a scalar quantizer that has quantization cells all of the same size and reconstruction levels set to the midpoints of the quantization cells. Moreover, companding system performance analysis in the case when the number of levels N is large has yielded some important facts and relationships. We defer that discussion for later. For now, a quick review of companding is presented.

N-level companding: Overview and optimal quantization. As shown in Figure 2.3, for a finite number of levels N, a compander (C, q_N^{USQ}, C^{-1}) is a quantization system that consists of three components sequentially applied to the input source X:

1. A non-linear, strictly increasing, continuous, differentiable *compressor* or *compressing function* $C : [0, \infty) \to [0, 1]$ which maps the source's support region onto $[0, 1]$

2. An N-level uniform scalar quantizer q_N^{USQ} with finite support over $[0, 1]$ that consists of N quantization cells, all of which have the same step size equal to $\Delta_{USQ,N} \triangleq \frac{1}{N}$, with thresholds $t_i = 1 - \frac{i}{N}$, $i = 0, 1, \ldots, N$, and reconstruction levels $\mu_i = \frac{t_i + t_{i-1}}{2} = 1 - \frac{i}{N} + \frac{1}{2N}$, $i = 1, 2, \ldots, N$

3. A non-linear decompressing function $C^{-1} : [0, 1] \to [0, \infty)$ which is the inverse to C

Thus the output from a compander that operates on X is $q_N(X) = C^{-1}\left(q_N^{USQ}(C(x))\right)$ and the MSE performance of the overall quantizer q_N is equal to

$$E\left[(X - q_N(X))^2\right] = E\left[\left(X - C^{-1}\left(q_N^{USQ}(C(x))\right)\right)^2\right]. \qquad (2.5.5)$$

Since an N-level USQ is used in every companding system that realizes an N-level quantizer, clearly, it is the compressing function C that determines the "identity" of the overall quantizer q_N.

By virtue of the fact that *any* N-level quantizer can be realized by a companding system and thus, existence of a function C which corresponds to a companding system realizing a quantizer is assured, it is important to understand how to construct a compressor function for a companding system. Given a set of levels μ_k, $k = 1, 2, \ldots, N$, and thresholds t_k, $k = 0, 1, \ldots, N$, the compressor function C must satisfy two requirements:

- For each $k = 0, 1, \ldots, N$, C must map t_k to $1 - \frac{k}{N}$, the k-th threshold of the USQ.
- For each $k = 1, 2, \ldots, N$, C must map μ_k to $1 - \frac{k}{N} + \frac{1}{2N}$, the kth reconstruction level of the USQ.

We point out that the two necessary conditions stated are not sufficient to uniquely identify C, because for any $x \in [t_k, t_{k-1})$, $q_N(x) = \mu_k$, there exist many compressor functions C and hence companding systems, that can realize the same N-level quantizer. Thus, while there may be a single, unique optimal scalar quantizer for a given value of N, there are many companding systems that achieve the lowest possible MSE performance. Nevertheless, we still consider C to characterize the identity of an N-level quantizer, and thus, for a companding system that realizes an N-level optimal quantizer, its compressor function C is referred to as an *optimal compressing function.* In this case, since the reconstruction thresholds and levels depend on the source pdf $f(x)$, so does C, which reinforces the observation that C is the identifying component of a companding system of an optimal quantizer.

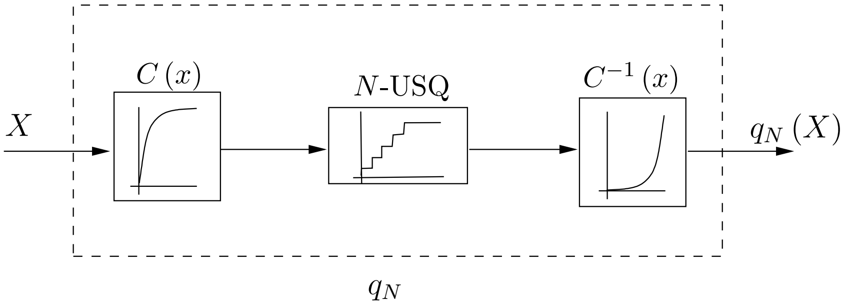


Figure 2.3: Schematic of a compander that realizes the quantizer q_N, where C is the compressor function, N-USQ denotes an N-level uniform scalar quantizer and C^{-1} is the inverse to the compressor function C.

2.6 Asymptotically Optimal Quantization.

Designing optimal scalar quantizers becomes more difficult when the number of levels N is large. In general, when N is increased, the MSE distortion decreases and, by itself, it is of interest to understand how MSE distortion changes as the number of quantization levels is increased in optimal scalar quantization. Aside from pure intellectual interest, however, knowledge of the relationship between MSE performance and N yields methods of designing high level or *high rate* quantizers that have near-optimal performance with fewer design costs. For these reasons, asymptotically optimal scalar quantization has been a much studied topic in the quantization literature.

Consider two sequences of MSE quantizers, all of which have been designed for the same random variable X: $q_{1,N},\ N = 1, 2, \ldots,$ and $q_{2,N},\ N = 1, 2, \ldots,$ where N is the number of reconstruction levels in a quantizer. Suppose the first sequence of quantizers $q_{1,N}$ consists of optimal quantizers and suppose the second sequence $q_{2,N}$ consists of quantizers that are not necessarily optimal. We say the sequence of quantizers $q_{2,N}$ is *asymptotically optimal* if

$$\lim_{N\to\infty} \frac{E\left[\left(X - q_{2,N}\left(X\right)\right)^2\right]}{E\left[\left(X - q_{1,N}\left(X\right)\right)^2\right]} = \lim_{N\to\infty} \frac{D\left(q_{2,N}\right)}{D\left(q_{1,N}\right)} = 1,$$

i.e., the ratio of the MSE performance of $q_{2,N}$ to the MSE performance of $q_{1,N}$ goes to 1 as the number of levels N increases. Clearly, a sequence of optimal N-level quantizers is also asymptotically optimal by definition.

Note: The topic of asymptotically optimal quantization is also referred to as *high resolution quantization theory*, where the term "high resolution" refers to the case when the number of levels N is large, the cells in the support region of the quantizer are small and the support threshold is sufficiently large so that the distortion or MSE due to $x \in [t_1, \infty)$ is small compared to the MSE distortion of the quantizer (over all quantization cells).

Asymptotically optimal quantizer design - Companding revisited. The study of companding systems that realize high resolution quantizers has yielded much insight into the relationship between N and both the design (placement of quantization thresholds and levels) and the MSE performance of these quantizers. Here, we recount two well-known results/theorems that are relevant to our work.

Performance of companding systems when N is large.

Theorem II.1. ***(Bennett's Theorem [2]****. From [4].) Let X be a source such that $E\left[X^{2+\epsilon}\right] < \infty$ for some $\epsilon > 0$. Then for any quantizer q_N designed for X that is realized by a companding system (C, q_N^{USQ}, C^{-1}), when the number of levels N is large* *and the distortion due to the overload region is negligible compared to the distortion from the inner region,* **Bennett's integral**

$$\frac{1}{N^2}B\left(C\right) \triangleq \frac{1}{12N^2}\int\limits_0^\infty \frac{f\left(x\right)}{c^2\left(x\right)}dx$$

provides an approximation to $E\left[\left(X - q_N\left(X\right)\right)^2\right]$ *in the sense that*

$$\lim_{N\to\infty}\frac{\frac{1}{12N^2}\int_0^\infty \frac{f(x)}{c^2(x)}dx}{E\left[\left(X - q_N\left(X\right)\right)^2\right]} = \lim_{N\to\infty}\frac{\frac{1}{12N^2}\int_0^\infty \frac{f(x)}{c^2(x)}dx}{E\left[\left(X - C^{-1}\left(q_N^{USQ}\left(C\left(x\right)\right)\right)\right)^2\right]} = 1, \quad (2.6.6)$$

where $c\left(x\right) \triangleq \frac{d}{dx}C\left(x\right)$*.*

In other words, Theorem II.1 says that the approximation

$$E\left[\left(X - q_N\left(X\right)\right)^2\right] \approx \frac{1}{12N^2}\int\limits_0^\infty \frac{f\left(x\right)}{c^2\left(x\right)}dx$$

holds when N is large if (C, q_N^{USQ}, C^{-1}) is a realization of q_N. Since there always exists a companding system that realizes a given N-level quantizer, Bennett's integral provides a general method of approximating the MSE performance of any quantizer with a large number of levels. Thus, it is clear that if $q_{1,N}$ is an optimal N-level quantizer, then

$$E\left[\left(X - q_{1,N}\left(X\right)\right)^2\right] \approx \frac{1}{12N^2}B\left(C_{1,N}\right)$$

if $(C_{1,N}, q_N^{USQ}, {C_{1,N}}^{-1})$ is a companding system that corresponds to $q_{1,N}$ and N is large.

Gauging optimal and asymptotically optimal quantizer performance with Bennett's Integral. Here, we summarize a connection between optimal quantization performance, asymptotically optimal quantization performance and Bennett's integral. Since Bennett's integral is intimately connected to a compressing system realization of a given quantizer through the system's compressing function C, we first consider an *asymptotically optimal compressing function* [20] which is defined to be a compressor C^* that minimizes the Bennett integral[3]

$$C^* \triangleq \underset{C:c=\frac{d}{dx}C}{\operatorname{argmin}} \frac{1}{12N^2} \int\limits_0^\infty \frac{f(x)}{c^2(x)} dx. \tag{2.6.7}$$

We point out that, in order to determine a suitable compressing function C^*, the minimization of Bennett's integral occurs over all possible compressor functions C and that this minimization is performed irrespective of the value of N. In other words, C^* minimizes Bennett's integral for all values of N. So, how does C^* or $B(C^*)$ relate to the performance of optimal quantizers and/or asymptotically optimal quantizers?

To address the question, first consider a sequence of optimal N-level quantizers $q_{1,N}$ realized by $(C_{1,N}, q_N^{USQ}, C_{1,N}^{-1})$. Then for each value of $N \geq 1$,

$$\frac{1}{N^2} B(C^*) = \frac{1}{12N^2} \int\limits_0^\infty \frac{f(x)}{{c^*}^2(x)} dx \leq \frac{1}{12N^2} \int\limits_0^\infty \frac{f(x)}{{c_{1,N}}^2(x)} dx = \frac{1}{N^2} B(C_{1,N}),$$

since C^* minimizes Bennett's integral and $C_{1,N}$ is just one particular compressing function that may not necessarily minimize Bennett's integral. Dividing by $D(q_{1,N})$, the inequality just stated becomes

$$\frac{\frac{1}{N^2} B(C^*)}{D(q_{1,N})} = \frac{\frac{1}{12N^2} \int_0^\infty \frac{f(x)}{{c^*}^2(x)} dx}{D(q_{1,N})} \leq \frac{\frac{1}{12N^2} \int_0^\infty \frac{f(x)}{{c_{1,N}}^2(x)} dx}{D(q_{1,N})} = \frac{\frac{1}{N^2} B(C_{1,N})}{D(q_{1,N})}$$

for every N, and thus, using (2.6.6), we have

$$\limsup_{N\to\infty} \frac{\frac{1}{N^2} B(C^*)}{D(q_{1,N})} \leq \lim_{N\to\infty} \frac{\frac{1}{N^2} B(C_{1,N})}{D(q_{1,N})} = 1. \tag{2.6.8}$$

Now, consider the sequence of quantizers $q_{2,N}$ (indexed by N) realized by $(C^*, q_N^{USQ}, C^{*-1})$, a sequence of companding systems (also indexed by N) that utilizes an asymptotically optimal compressing function C^*. Comparing the performance of an optimal N-level quantizer $q_{1,N}$ to the performance of an N-level quantizer built using C^*, it is clear that

$$\frac{D(q_{2,N})}{D(q_{1,N})} = \frac{\frac{1}{N^2} B(C^*)}{D(q_{1,N})} \cdot \frac{D(q_{2,N})}{\frac{1}{N^2} B(C^*)}$$

[3] Existence of C^* is guaranteed by the Panter-Dite Theorem to be discussed shortly.

and thus,

$$\limsup_{N\to\infty} \frac{D\left(q_{2,N}\right)}{D\left(q_{1,N}\right)} = \limsup_{N\to\infty} \frac{\frac{1}{N^2}B\left(C^*\right)}{D\left(q_{1,N}\right)} \cdot \lim_{N\to\infty} \frac{D\left(q_{2,N}\right)}{\frac{1}{N^2}B\left(C^*\right)}$$
$$\overset{(2.6.6)}{=} \limsup_{N\to\infty} \frac{\frac{1}{N^2}B\left(C^*\right)}{D\left(q_{1,N}\right)} \overset{(2.6.8)}{\leq} 1. \tag{2.6.9}$$

Since $q_{1,N}$ is a sequence of optimal quantizers, we know that

$$\frac{D\left(q_{2,N}\right)}{D\left(q_{1,N}\right)} \geq 1$$

for all $N \geq 1$, and this fact implies

$$\liminf_{N\to\infty} \frac{D\left(q_{2,N}\right)}{D\left(q_{1,N}\right)} \geq 1.$$

Therefore, existence of $\lim_{N\to\infty} \frac{D\left(q_{2,N}\right)}{D\left(q_{1,N}\right)}$ is assured and

$$\lim_{N\to\infty} \frac{D\left(q_{2,N}\right)}{D\left(q_{1,N}\right)} = 1. \tag{2.6.10}$$

Thus it is clear that the sequence of quantizers $q_{2,N}$, constructed about an asymptotically optimal compressor C^*, is a sequence of asymptotically optimal quantizers. Moreover, from (2.6.6), since the performance of this sequence of asymptotically optimal quantizers can be approximated by the Bennett integral evaluated at C^* when N is large, it is clear that the OPTA function $D\left(N\right)$ can also be approximated by the Bennett integral evaluated at C^* in the sense that

$$\lim_{N\to\infty} \frac{D\left(N\right)}{\frac{1}{N^2}B\left(C^*\right)} = 1, \tag{2.6.11}$$

since for each $N \geq 1$, the OPTA function $D\left(N\right) = D\left(q_{1,N}\right)$. We remind ourselves at this point that the relationships in (2.6.10) and (2.6.11) depend on the fact that C^* exists and is well-defined. The following theorem guarantees the existence of C^*.

Asymptotic performance results from companding: The Panter-Dite Theorem. The theorem stated below connects/links Bennett's integral approximation for MSE quantizer performance to the performance of both asymptotically optimal and optimal quantizers.

Theorem II.2. *(**Panter-Dite Theorem** [20]. From [4].) Let X be a random variable with pdf f and suppose $E\left[x^{2+\epsilon}\right] < \infty$ for some $\epsilon > 0$. Then*

$$\lim_{N\to\infty} N^2 D\left(N\right) = \frac{1}{12}\sigma^2\beta,$$

where $D\left(N\right)$ is the OPTA function for the source X and

$$\beta \triangleq \frac{1}{\sigma^2}\left(\int\limits_0^\infty f^{\frac{1}{3}}\left(x\right)dx\right)^3 \tag{2.6.12}$$

is the **Panter-Dite constant**[4] *for the source X.*

While not transparent in this review, when solving for β in (2.6.12), the proof of this theorem also guarantees not only the existence of an asymptotically optimal compressing function C^* (defined in (2.6.7)) that minimizes the Bennett integral, but also yields the form of it

$$C^*\left(x\right) = \int\limits_0^x \frac{f^{\frac{1}{3}}\left(u\right)}{\int_0^\infty f^{\frac{1}{3}}\left(w\right)dw}du \tag{2.6.13}$$

or equivalently (and more familiarly),

$$c^*\left(x\right) = \frac{f^{\frac{1}{3}}\left(x\right)}{\int_0^\infty f^{\frac{1}{3}}\left(w\right)dw}.$$

We note that β depends on the source X through its pdf $f\left(x\right)$ but is otherwise independent of the source variance σ^2. Since we will only compare quantizer performance across quantizers that have been designed for the same source, the actual value of β being discussed should be clear and we feel it is unnecessary to notate explicitly the dependence of β on X.

Thus, if q_N is a sequence of quantizers that is realized by $(C^*, q_N^{USQ}, C^{*-1})$, then the sequence is asymptotically optimal (from the argument in the previous section)

[4] In the literature, $\frac{\beta}{12}$ is also referred to as the Panter-Dite constant, with β defined as in (2.6.12). We will use both conventions in our text, but we will include the explicit formula that we are referring to in order to make clear our intent. Also in the literature, $\frac{\beta}{12N^2}$ is referred to as the Panter-Dite formula.

and

$$\lim_{N\to\infty} \frac{N^2 D\left(q_N\right)}{\sigma^2 \frac{\beta}{12}} = 1.$$

Another approach to asymptotically optimal quantization - Point densities. For an alternate, yet related point of view on high resolution quantization, we review the use of point densities in the design and performance of optimal and nearly optimal quantizers when N is large. When designing an N-level quantizer for a source X, instead of focusing on locating the exact placement of reconstruction levels (or equivalently, the reconstruction thresholds), it is beneficial to think of the reconstruction levels or *reconstruction points* of the quantizer as being distributed throughout the support region of the source X. Using this notion, asymptotically optimal design can be construed as determining the asymptotic distribution of reconstruction points in the support region of X. We characterize this distribution of points as a density which we call an asymptotically *optimal point density* of X.

Informally, the concept of a point density can be described when N is large: Consider a sequence of quantizers q_N indexed by N. Fix $\Delta > 0$ small. When N is large, suppose there exists a function $\lambda_N\left(x\right)$ such that

$$\text{fraction of reconstruction levels in } [x, x+\Delta) \approx \lambda_N\left(x\right)\Delta$$

for all $x \geq 0$. If $\lambda_N\left(x\right)$ is an integrable function, then for an arbitrary half open interval $[a, b)$ with $b > a \geq 0$,

$$\text{fraction of reconstruction levels in } [a, b) \approx \sum_{i=0}^{\lfloor\frac{b-a}{\Delta}\rfloor} \lambda_N\left(a + i\cdot\Delta\right)\cdot\Delta \approx \int\limits_a^b \lambda_N\left(x\right)dx.$$

Now suppose that for this sequence of quantizers q_N, the corresponding sequence $\lambda_N\left(x\right)$ converges to a well-defined limiting function $\lambda\left(x\right)$ as $N \to \infty$ in *some* sense. In this case, the limiting function $\lambda\left(x\right)$ is referred to as the *point density* belonging to the sequence of quantizers q_N.

More formally, a *point density* $\lambda_N\left(x\right)$ for a sequence of quantizers q_N (indexed by N) over $[0, \infty)$ is defined to be a function that satisfies

1. $\lambda_N\left(x\right) \geq 0$ for all $x \geq 0$.

2. $\int_0^\infty \lambda_N\left(x\right)dx = 1$.

3. For any $[s,t)$, $t > s \geq 0$,

$$\lim_{N\to\infty} \frac{\int_s^t \lambda_N(x)\,dx}{\text{fraction of reconstruction levels or points in } [s,t)} = 1.$$

A point density that satisfies the criteria just stated is generally referred to as a *normalized* point density definition. In the literature, there is also an *unnormalized* point density which is similarly defined, with the exception that the unnormalized point density gives a count of the number of levels rather than the fraction of levels in any particular interval. Throughout this thesis, however, reference to a point density will always be consistent with the normalized definition.

Point densities and companding. Connecting the idea of a point density to companding systems and their associated compressor functions, consider a quantizer q_N realized by the companding system (C, q_N^{USQ}, C^{-1}). From the definition of a compressor function, it is clear that $c(x) = \frac{d}{dx}C(x)$ is a point density for the quantizer q_N. To see why this is so, we note that the compressing function C can be viewed as a cumulative distribution function (CDF) of reconstruction points for q_N over $[0,\infty)$ since it satisfies all of the following properties:

1. $\lim_{x\to 0} C(x) = 0$.

2. $\lim_{x\to\infty} C(x) = 1$.

3. C is right-continuous.

4. C is monotone increasing.

5. The CDF interpretation of C is reinforced by the observation that if the number of levels N is large, then for any $x \geq 0$, $C(x)$ approximately equals the fraction of levels to the left of the value x.

We remark that the first four properties stated come directly from the definition of a CDF, and that the fifth property draws the connection between C as a CDF and the quantizer q_N, through q_N's set of reconstruction points. Since it has been established that C is a CDF and hence, its derivative is a pdf (which is well-defined by C's definition), it is clear that $c(x) = \frac{d}{dx}C(x)$ can, in turn, also be interpreted as a point density for q_N, where, for any given half open interval $[s,t)$, $t > s \geq 0$, the fraction of levels or reconstruction points in $[s,t)$ is approximately equal to $C(t) - C(s) = \int_s^t c(x)\,dx$.

Bennett's Integral – The point density version. Since every compressor function C is a CDF and any quantizer can be realized by a companding system, Bennett's asymptotic approximation to the MSE distortion of a quantizer can be re-expressed in terms of point densities. For a sequence of quantizers q_N (indexed by N) for whom λ is the corresponding point density for each value of N, we have

$$\lim_{N\to\infty} \frac{E\left[\left(X - q_N\left(X\right)\right)^2\right]}{\frac{1}{N^2}B\left(\Lambda\right)} = 1, \tag{2.6.14}$$

where

$$\frac{1}{N^2}B\left(\Lambda\right) \overset{\triangle}{=} \frac{1}{12N^2}\int\limits_0^\infty \frac{f\left(x\right)}{\lambda^2\left(x\right)}dx$$

is the point density version of the Bennett integral and $\Lambda\left(x\right) \overset{\triangle}{=} \int_0^x \lambda\left(u\right)du$.

Panter-Dite Theorem – Remarks from the point density perspective. An alternative proof of the Panter-Dite Theorem [20] using point densities, in particular, using the point density version of Bennett's integral, can also be shown. In this proof, the *optimal point density function* λ^*, a point density that minimizes the point density version of Bennett's integral,

$$\lambda^*\left(x\right) \overset{\triangle}{=} \frac{f^{\frac{1}{3}}\left(x\right)}{\int_0^\infty f^{\frac{1}{3}}\left(u\right)du}, \tag{2.6.15}$$

is used instead of the optimal compressor function C^* (and the companding version of Bennett's integral), and the same results are obtained. (See Figure 2.4 for an illustration of λ^*.)

While it may appear that the point density perspective is no different than the compressor function perspective, point densities are a more direct, yet general way of thinking about asymptotically optimal quantizer design and analysis than compressor functions because companding systems are a method of implementation, albeit a universal one.

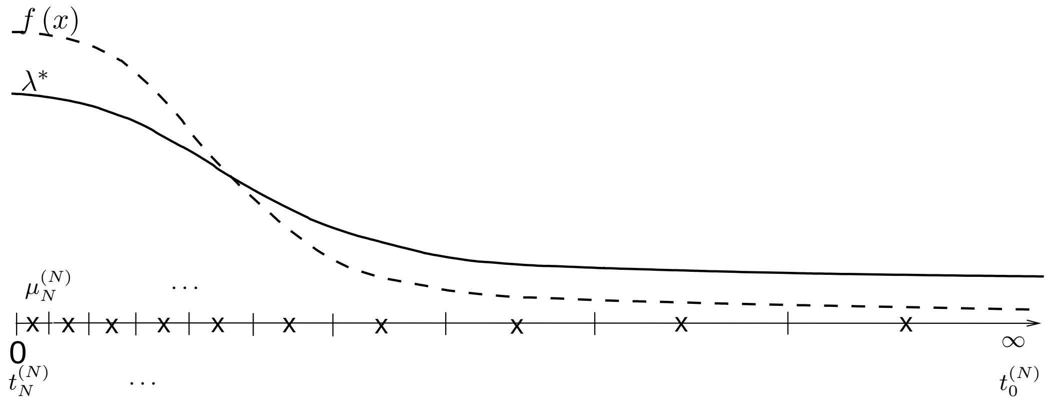


Figure 2.4: An optimal quantizer q_N shown with an illustration of its optimal point density λ^*.

2.7 Nitadori's Two Results for Optimal Exponential Quantization.

In his 1965 paper, Nitadori [19] derived two results for optimal quantization of Laplacian sources, but since Laplacian sources are just two-sided exponential sources and in this thesis, we are limiting our discussion to one-sided exponential sources in order to simplify the discourse, we state his result for the one-sided, unit variance case. First, he solved the N-level optimal exponential quantizer design problem by demonstrating the existence of a unique real-valued sequence $\underline{\eta}_k$, $k = 1, 2, \ldots$, which provides a complete description for the quantizer. His solution

$$\mu_k^{(N)} = t_k^{(N)} + \underline{\eta}_k = \underline{\eta}_N + \sum_{i=k}^{N-1} 2\underline{\eta}_i$$
$$t_k^{(N)} = \mu_{k+1}^{(N)} + \underline{\eta}_k = \underline{\eta}_N + \sum_{i=k+1}^{N-1} 2\underline{\eta}_i + \underline{\eta}_k$$

where $\underline{\eta}_k$, which equals $\underline{\Delta}_k^{(N)}$, is determined from $\underline{\eta}_{k-1}$ by solving

$$-\left(1-\underline{\eta}_k\right)e^{-\left(1-\underline{\eta}_k\right)} = -\left(1+\underline{\eta}_{k-1}\right)e^{-\left(1+\underline{\eta}_{k-1}\right)} \qquad (2.7.16)$$

with $\underline{\eta}_1 = 1$. (2.7.16) is referred to as the *Nitadori recursion* and the sequence of values $\underline{\eta}_k$ is referred to as the *Nitadori sequence.* Solving for $\underline{\eta}_k$ given $\underline{\eta}_{k-1}$ amounts to evaluating the *principal branch of the Lambert W function* W_0 [6] at

$-\left(1+\underline{\eta}_{k-1}\right)e^{-\left(1+\underline{\eta}_{k-1}\right)}$, or more precisely,

$$\underline{\eta}_k = W_0\left(-\left(1+\underline{\eta}_{k-1}\right)e^{-\left(1+\underline{\eta}_{k-1}\right)}\right)+1,$$

where $W_0 : \left[-\frac{1}{e},\infty\right) \to [-1,\infty)$ is defined to be the inverse to the function $z_0 : [-1,\infty) \to \left[-\frac{1}{e},\infty\right)$ given by

$$z_0\left(W\right) \triangleq We^{W}.$$

Figure 2.5 illustrates the recursive process of solving for values of the Nitadori sequences using the inverse to the principal branch of the Lambert W function.

Nitadori derived (2.7.16) by taking partial derivatives of the MSE expression for the optimal quantizer and solving the following equations

$$\frac{\partial}{\partial\mu_k^{(N)}}\sum_{i=1}^{N}\int_{t_i^{(N)}}^{t_{i-1}^{(N)}}\left(x-\mu_i^{(N)}\right)^2 e^{-x}dx = 0$$

$$\frac{\partial}{\partial t_k^{(N)}}\sum_{i=1}^{N}\int_{t_i^{(N)}}^{t_{i-1}^{(N)}}\left(x-\mu_i^{(N)}\right)^2 e^{-x}dx = 0$$

for $k = 1, 2, \ldots, N$. In Chapter IV, however, we will describe an alternate way to derive (2.7.16) that is different from the method used by Nitadori in that no derivatives are used. This alternate method is fundamental as a guide to the way in which we are able to generalize Nitadori's first result to optimal quantization of sources other than exponential, which is the focus of that chapter.

For the second result reported in his paper [19], Nitadori used his solution sequence $\underline{\eta}_k$ in the expression for MSE to show that the MSE produced by an optimal quantizer can be expressed explicitly in terms of his solution sequence $\underline{\eta}_k$

$$D\left(N\right) = \left(\underline{\eta}_N\right)^2. \tag{2.7.17}$$

Figure 2.6 shows a plot of the first 64 values of $\underline{\eta}_k$ and Table 3.1 (in Chapter III) lists these values.

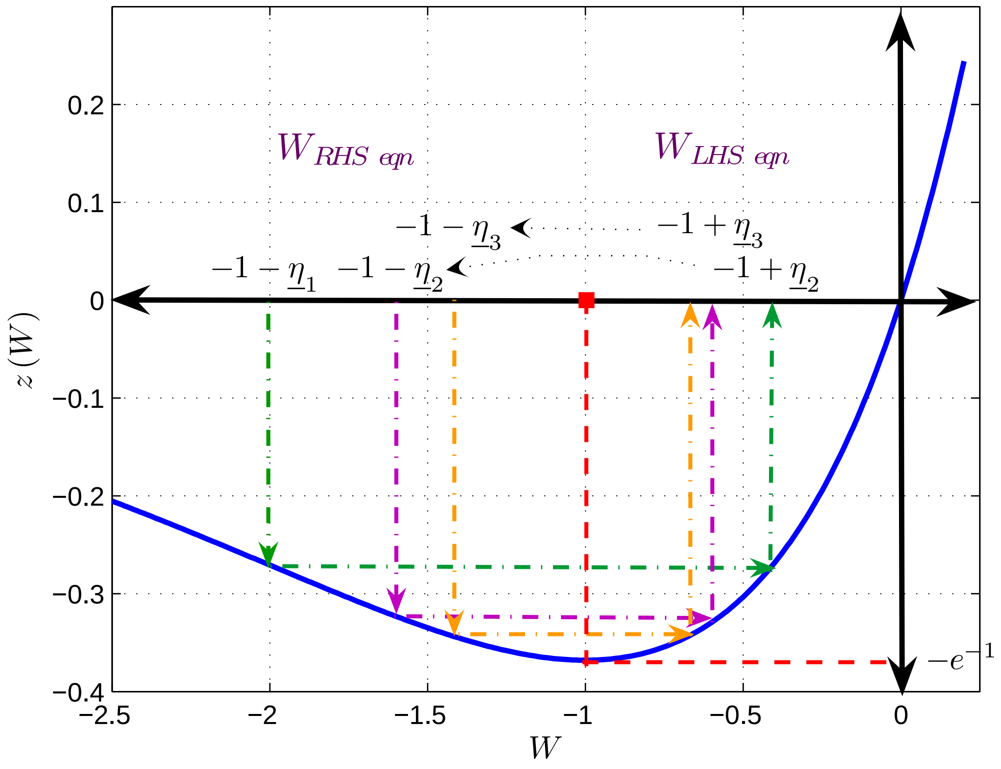


Figure 2.5: Illustration of the function $z\left(W\right) = We^{W}$. Also shown are the values of W corresponding to the values of $\underline{\eta}_1$, $\underline{\eta}_2$, $\underline{\eta}_3$.

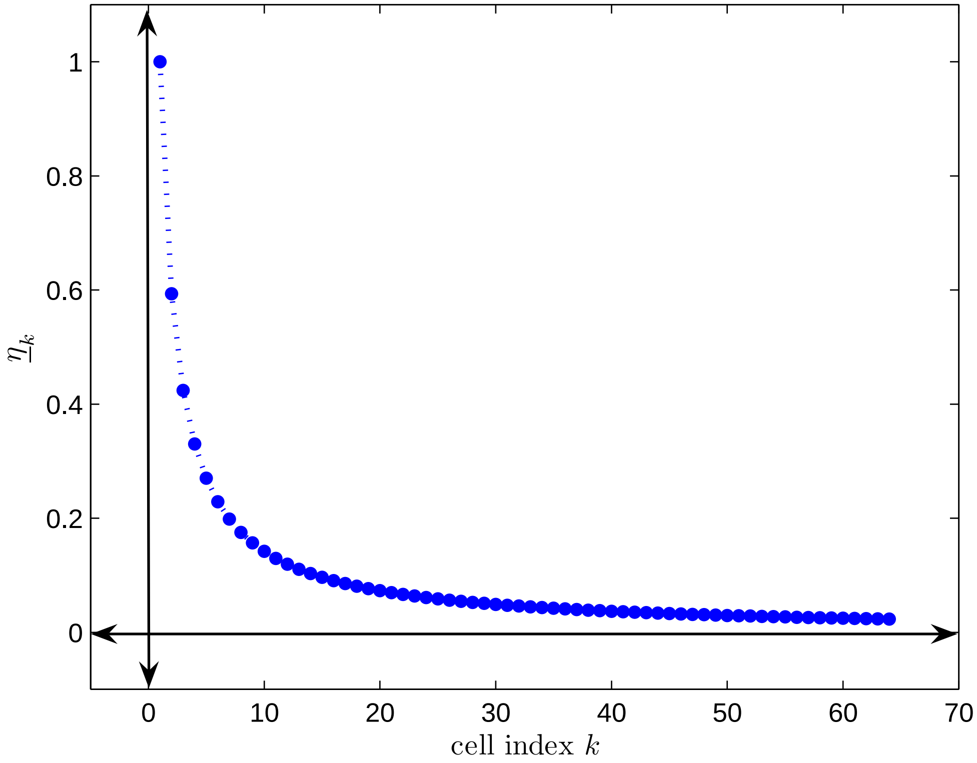


Figure 2.6: The first 64 values of the Nitadori sequence $\underline{\eta}_k$.

CHAPTER III

Support Threshold Estimation and Related Asymptotic Results for Exponential Optimal Quantizers

3.1 Introduction.

Three areas related to the design of quantizers for an exponential source are addressed in this chapter. The first (our primary subject) deals with key parameter or support threshold estimation of optimal quantizers, which improves the understanding of how a support threshold varies as a function of the number of levels in an optimal quantizer and is useful practically since it can be used with the Lloyd-Max algorithm ([13], [15]) to create optimal quantizers with reduced computational cost. The second area is related to the use of the Panter-Dite formula (Theorem II.2, Chapter II) as an estimator for the MSE of optimal quantizers. The final topic concerns a new, suboptimal quantizer design method which simplifies Nitadori-style optimal quantizer design (Chapter II, Section 2.7) yet still offers good MSE performance. This basis for this design method, as well as the result regarding the Panter-Dite formula, comes out of the work completed for the support threshold estimates to be discussed.

As stated in Chapter II, the computation time required when using the Lloyd-Max algorithm (Chapter II, Section 2.4) is highly dependent on the accuracy of the initial support threshold estimate. A good estimate serves to reduce the number of iterations required to arrive at a solution to the MMSE quantizer design problem and this reduction translates to reduced computation time. With the advent of fast digital computing readily available nowadays, it may not be clear what actually is the impact of this reduction in computation time. One application where the impact of the time savings resulting from a reduction in the number of iterations of the algorithm performed would be evident is in adaptive quantization where the design of a quantizer is frequently updated.

To gain a better sense of how the accuracy of the initializing support threshold affects the runtime of the Lloyd-Max algorithm for the specific case of designing optimal exponential quantizers, it turns out that the system of equations (that comes from using the optimality conditions, see (2.2.3) and (2.2.4)) to be solved at each iteration of the Lloyd-Max algorithm can be pre-determined prior to running the algorithm: For the unit variance exponential source, at each iteration i, beginning with $k = 1$, followed by $k = 2$, up to $k = N$, the estimated reconstruction levels and thresholds for the kth quantization cell are determined (in order) by

$$1) \quad \mu_k^{est,i} = \frac{t_k^{est,i} e^{-t_k^{est,i}} - t_{k-1}^{est,i} e^{-t_{k-1}^{est,i}}}{P\left[t_k^{est,i}, t_{k-1}^{est,i}\right)} + 1 \tag{3.1.1}$$

$$2) \quad t_{k+1}^{est,i} = \mu_k^{est,i} - \left(\mu_k^{est,i} - t_k^{est,i}\right), \tag{3.1.2}$$

where $t_0^{est,i} \overset{\triangle}{=} \infty$. With the relationships in (3.1.1) and (3.1.2), the sensitivity of the algorithm to the accuracy of $t_1^{est,1}$ is made more tangible since it is easy to see how any inaccuracy in $t_1^{est,1}$ trickles down through all of the other estimates for the reconstruction levels and thresholds during the first iteration, which ultimately affects all subsequent estimates in the iterations that follow.

To estimate the support threshold, since summing the half steps of an optimal quantizer yields the quantizer's support threshold, we adopt the half step parameterization view for quantizers, and look at ways to discern information regarding the half steps of optimal quantizers. To do this, we begin by considering the half steps of asymptotically optimal quantizers realized by companding, and this examination leads to an upper bound on the support threshold. Next, to construct a lower bound on the support threshold, we study the structural relationships between the half steps of adjacent quantization cells of optimal quantizers through careful examination of the way in which the values of these half steps are produced (numerically) by the Nitadori design method for optimal exponential quantizers. While it may seem odd to turn to a method which, at first glance, appears to obviate the need for further study of the exponential design case (because it gives the exact specification of such quantizers non-iteratively), our intent is to achieve further comprehension of the structure of optimal quantizers (with the idea that "optimal" structure leads to optimal performance), and it is to this end that we focus on the Nitadori recursion ((2.7.16) in Chapter II). Since the recursion relates the half steps of adjacent quantization cells to each other (through the Lambert W function, a function that cannot be expressed in terms of finite combinations of elementary functions[1]), the insight we gain on the ties between neighboring half steps through close examination of it can be applied to the construction of a lower bound to the support threshold.

3.2 Optimal Support for Exponential Sources - Main Result.

Definition of exponential source with variance σ^2 (or parameter σ). In this work, an exponential source X with variance σ^2 (or parameter σ) has a pdf

$$f(x) \triangleq \frac{1}{\sigma} e^{-\frac{x}{\sigma}}, \quad x \geq 0,$$

where

$$E[X] = \sigma$$
$$Var[X] = \sigma^2.$$

The main results of this section are the following theorem and corollaries, where Corollary III.2 addresses the use of the Panter-Dite formula to approximate the MSE of optimal quantizers. The discussion on simplified quantizer design methods is presented later in Section 3.4.3.

Theorem III.1. *For an exponential source with variance σ^2, the optimal support threshold $t_1^{(N)}$ has the following bounds:*

a) $\frac{t_1^{(N)}}{\sigma} \leq 3\log N$ *when $N \geq 1$, where equality is achieved when $N = 1$.*

b) $\frac{t_1^{(N)}}{\sigma} > 3\log N + \delta(N) - 1.46004$ *when $N > 9$, where*

$$\delta(N) \triangleq \frac{3}{N-1} - \frac{21}{4}\frac{1}{N^2} + \frac{3}{2N}\left(1 - \frac{2}{N} + \frac{8}{N^2}\right)^{\frac{1}{2}}. \tag{3.2.3}$$

Corollary III.2. *For an exponential source with variance σ^2 and pdf $f(x) = \frac{1}{\sigma}e^{-\frac{x}{\sigma}}$, $x \geq 0$, if $N \geq 3$,*

$$\left| N^2 \left(\frac{D(N)}{\sigma^2} \right) - \frac{\beta}{12} \right| \leq \frac{9}{4}\left(\frac{2}{N} - \frac{8}{N^2} \right),$$

[1] Elementary functions include e^x, $\log x$ and nth roots.

where $\beta = \frac{1}{\sigma^2}\left(\int_0^\infty f^{\frac{1}{3}}(x)\,dx\right)^3 = 27$ *is the Panter-Dite constant, and* $D(N)$ *is defined to be the least MSE distortion of any scalar quantizer with not more than* N *levels (see* (2.2.2)*).*

For a Laplacian source, which is a two-sided exponential source, with pdf

$$p(x) = \frac{1}{\sqrt{2}\sigma}\exp^{-\frac{\sqrt{2}}{\sigma}x},$$

we have the following corollary to Theorem III.1:

Corollary III.3. *Given an* M*-level optimal quantizer designed for a Laplacian source with variance* σ^2*.*

1. *The optimal support threshold* $t_1^{(N)}$ *has the following bounds:*

   a) *(Upper bound.) If* M *is even (*$M = 2N$*) and* $M \geq 4$ *(or equivalently,* $N \geq 2$*), then*

   $$\frac{t_1^{(N)}}{\sigma} < \frac{3}{\sqrt{2}}\log\left\lfloor\frac{M}{2}\right\rfloor = \frac{3}{\sqrt{2}}\log N. \tag{3.2.4}$$

   If M *is odd (*$M = 2N+1$*) and* $M \geq 7$ *(*$N \geq 3$*), then* (3.2.4) *still holds.*

   b) *(Lower bound.) For* $M > 18$ *and either* $M = 2N$ *or* $M = 2N+1$*,*

   $$\begin{aligned}\frac{t_1^{(N)}}{\sigma} &> \frac{3}{\sqrt{2}}\log\left\lfloor\frac{M}{2}\right\rfloor + \xi\left(\left\lfloor\frac{M}{2}\right\rfloor\right) - 1.03241\\ &= \frac{3}{\sqrt{2}}\log N + \xi(N) - 1.03241,\end{aligned}$$

   where

   $$\begin{aligned}\xi\left(\left\lfloor\frac{M}{2}\right\rfloor\right) &\triangleq \frac{3}{\sqrt{2}}\frac{1}{\left\lfloor\frac{M}{2}\right\rfloor - 1} - \frac{21}{4\sqrt{2}}\frac{1}{\left\lfloor\frac{M}{2}\right\rfloor^2} + \frac{3}{2\sqrt{2}\left\lfloor\frac{M}{2}\right\rfloor}\left(1 - \frac{2}{\left\lfloor\frac{M}{2}\right\rfloor} + \frac{8}{\left\lfloor\frac{M}{2}\right\rfloor^2}\right)^{\frac{1}{2}}\\ &= \frac{3}{\sqrt{2}}\frac{1}{N-1} - \frac{21}{4\sqrt{2}}\frac{1}{N^2} + \frac{3}{2\sqrt{2}N}\left(1 - \frac{2}{N} + \frac{8}{N^2}\right)^{\frac{1}{2}}.\end{aligned}$$

2. *An upper bound to the convergence rate of* $M^2 D(M)$ *to the Panter-Dite constant*

$\frac{\beta}{12}$[2] is

$$\left|M^2\left(\frac{D(M)}{\sigma^2}\right)-\frac{\beta}{12}\right|\le\frac{9}{2}\left(\frac{4}{M}-\frac{32}{M^2}\right),$$

when $M > 18$, where $\beta = \frac{1}{\sigma^2}\left(\int_0^\infty p^{\frac{1}{3}}(x)\,dx\right)^3 = 54$.

Comments on Corollary III.2. The connection between Corollary III.2 and Theorem III.1 and Corollary III.2 is made more clear if we let $M = 2N$, for some $N = 1, 2, \ldots$ so that M is even. In this case, the support threshold $t_1^{(M)} = t_1^{(2N)}$ for an optimal M-level quantizer designed for a unit variance Laplacian source is equal to the support threshold $\frac{1}{\sqrt{2}}t_1^{(N)}$ for an optimal N-level quantizer designed for the corresponding one-sided exponential source with variance $\frac{1}{2}$. Thus, since it is easy to see how the results of Theorem III.1 and Corollary III.2 for exponential sources can be translated to the case where the source is Laplacian and M is even, the proofs for Corollary III.2, Parts 1*a*) (the even case), 1*b*), and 2, will not be presented, aside from the following remarks: For Part 1*b*) of Corollary III.2, when M is odd, the lower bound to $t_1^{(M)}$ shown holds because $t_1^{(M)} > t_1^{(M-1)}$ and thus $t_1^{(M)}$ is greater than any lower bound to $t_1^{(M-1)}$. With regards to Part 2 of Corollary III.2, the restriction on M being even is because the result relies on Nitadori's MSE distortion expression for optimal exponential (one-sided) quantizers, and this expression does not extend to optimal Laplacian quantizers with an odd number of levels. For Part 1 of Corollary III.2, which deals with the upper bound to the support threshold $t_1^{(M)}$, Appendix A gives a brief discussion of the proof.

The chapter is organized as follows: We first present the proofs to Theorem III.1 and Corollary III.2. Following this, we discuss applications of these facts. We finish the chapter with some concluding remarks and offer suggestions for future work.

3.3 Theorem III.1 and Corollary III.2 Proofs.

To facilitate the readability and reduce the notation used through out the discussion, we will present the proofs to Theorem III.1 and Corollary III.2 for the case when the source has unit variance. The proofs for the non-unit variance exponential source can be derived directly from the unit variance case proofs by inserting the appropriate scale factors.

[2] As stated in a footnote in Chapter II, both $\frac{\beta}{12}$ and β are commonly referred to as the Panter-Dite constant, where β is defined as in (2.6.12).

3.3.1 Useful Relationships Regarding the Centroid of $[0, \Delta)$ for an Exponential Source.

Before proving Theorem III.1 and Corollary III.2, we will introduce some important facts regarding centroids for the one-sided exponential source with unit variance. Recall that the effect of the memoryless property of the exponential source on the centroid of an arbitrary cell $[t, t + \Delta)$, $t \geq 0$, $\Delta > 0$, is

$$E\left[X \mid X \in [t, t + \Delta)\right] = t + E\left[X \mid X \in [0, \Delta)\right],$$

i.e., the centroid of an arbitrary cell $[t, t + \Delta)$ equals the lower threshold value plus the centroid of the cell $[0, \Delta)$. Thus, it is of interest to us to focus some attention on $E\left[X \mid X \in [0, \Delta)\right]$.

For a (general) source with pdf $f\left(x\right)$, $x \geq 0$, we define the centroid of $[0, \Delta)$, $\Delta > 0$, to be

$$\Delta_l\left(\Delta\right) \stackrel{\triangle}{=} \frac{1}{P_{[0,\Delta)}} \int\limits_{0}^{\Delta} x f\left(x\right) dx,$$

when $P_{[0,\Delta)} > 0$, where

$$P_{[0,\Delta)} \stackrel{\triangle}{=} \int\limits_{0}^{\Delta} f\left(x\right) dx,$$

and we define the distance between the centroid and the upper threshold of $[0, \Delta)$ to be

$$\Delta_u\left(\Delta\right) \stackrel{\triangle}{=} \Delta - \Delta_l\left(\Delta\right).$$

Remark 1. It is clear from the definitions of $\Delta_l\left(\Delta\right)$ and $\Delta_u\left(\Delta\right)$ that both functions are continuous and differentiable in Δ.

Remark 2. The above definitions are general and not tied to a quantization scheme. However, if we were discussing an N-level quantizer, and if that quantizer utilized centroids for its reconstruction levels, then for the cell at the origin $[0, \Delta_N)$, we would have

$$\Delta_l\left(\Delta_N\right) = \underline{\Delta}_N = \mu_N$$

since $t_N = 0$. (See Chapter II, where $\underline{\Delta}_k$ is defined for arbitrary quantizers and $\underline{\Delta}_k^{(N)}$ is defined for optimal quantizers.)

Specializing to the case of the unit variance exponential source, the expressions for $\Delta_l\left(\Delta\right)$ and $\Delta_u\left(\Delta\right)$ are

$$\Delta_l\left(\Delta\right) = 1 - \frac{\Delta e^{-\Delta}}{1-e^{-\Delta}} \tag{3.3.5}$$

$$\Delta_u\left(\Delta\right) = \frac{\Delta}{1-e^{-\Delta}} - 1. \tag{3.3.6}$$

For the derivation of (3.3.5) and (3.3.6), with $f\left(x\right) = e^{-x}$, $x \geq 0$, we have

$$P_{[0,\Delta)} = \int\limits_0^\Delta e^{-x} dx = -e^{-x}\Big|_0^\Delta = 1 - e^{-\Delta}$$

and from integration by parts,

$$\begin{aligned} u &= x & du &= dx \\ v &= -e^{-x} & dv &= e^{-x} dx, \end{aligned}$$

we have

$$\begin{aligned} \Delta_l\left(\Delta\right) &= \frac{1}{P_{[0,\Delta)}} \int\limits_0^\Delta x f\left(x\right) dx = \frac{1}{P_{[0,\Delta)}} \left[-xe^{-x}\Big|_0^\Delta + \int\limits_0^\Delta f\left(x\right) dx \right] \\ &= \frac{1}{P_{[0,\Delta)}} \left[-xe^{-x}\Big|_0^\Delta + P_{[0,\Delta)} \right] = \frac{1}{P_{[0,\Delta)}} \left[-\Delta e^{-\Delta} \right] + 1 = 1 - \frac{\Delta e^{-\Delta}}{1-e^{-\Delta}} \end{aligned}$$

and

$$\Delta_u\left(\Delta\right) = \Delta - 1 + \frac{\Delta e^{-\Delta}}{1-e^{-\Delta}} = \Delta\left(1 + \frac{e^{-\Delta}}{1-e^{-\Delta}}\right) - 1 = \frac{\Delta}{1-e^{-\Delta}} - 1.$$

We are now ready to begin the proof of Theorem III.1.

3.3.2 The Proof of Theorem III.1.

To improve readability and to help convey the main points of the proof, we prove Theorem III.1 for a unit variance exponential source, remarking that extension of the proof below to the case where the variance is given by $\sigma^2 > 0$ is not difficult.

The proof is presented in two halves: The first half will show the upper bound to $t_1^{(N)}$ and second half will derive the lower bound to $t_1^{(N)}$ in Theorem III.1.

3.3.2.1 Proof of upper bound to $t_1^{(N)}$ in Theorem III.1.

To prove Part a) of Theorem III.1, we first design a special, companding-based quantizer q_N^c (see Chapter II for a discussion on companding systems) with support equal to $3 \log N$. It turns out that for any $N \geq 2$, the half steps of q_N^c are always greater than or equal to the half steps of the corresponding N-level optimal quantizer q_N^*. From this fact, it follows easily that the support of the q_N^c quantizer is greater than or equal to the support threshold of q_N^*.

Our companding system for a 1-sided exponential source. Fix $N \geq 2$. Let q_N^c be the quantizer whose thresholds are induced by the asymptotically optimal compressor function

$$C^* (x) \overset{\triangle}{=} 1 - \exp^{-\frac{x}{3}}, \quad x \geq 0, \tag{3.3.7}$$

(see Chapter II, Section 2.6, (2.6.7) and (2.6.13)) followed by an N-level quantizer defined over $[0, 1]$ with uniform cell size $\Delta_{k,c}^{(N)} = \frac{1}{N}$, for $k = 1, 2, \cdots, N$ (*uniform step-size quantizer*), followed by the inverse companding function C^{*-1}, creating thresholds $t_{k,c}^{(N)} = C^{*-1}\left(1 - \frac{k}{N}\right)$, for $k = 0, 1, \cdots, N$.

In a departure from the standard convention where the quantization levels of a companding quantization system are derived from the midpoints of the uniform step-size quantizer (in this case, the uniform step-size quantizer becomes a USQ), which we will refer to as a uniform scalar quantizer companding system (USQC), our companding system q_N^c has quantization levels set at each cell's centroid and we will refer to our companding system as a uniform threshold companding system with centroid reconstruction levels (UTCC). Then the step widths, thresholds and levels of q_N^c are given by

$$\Delta_{k,c}^{(N)} = -3\log\left(1-\frac{1}{k}\right), \qquad \text{for } k = 1, 2, \cdots, N \tag{3.3.8}$$

$$t_{k,c}^{(N)} = -3\log\left(\frac{k}{N}\right), \qquad \text{for } k = 0, 1, \cdots, N \tag{3.3.9}$$

$$\mu_{k,c}^{(N)} = t_{k,c}^{(N)} + \underline{\Delta}_{k,c}^{(N)}, \qquad \text{for } k = 1, 2, \cdots, N \tag{3.3.10}$$

where the derivation of $\underline{\Delta}_{k,c}^{(N)} = \Delta_l\left(\Delta_{k,c}^{(N)}\right)$ and $\mu_{k,c}^{(N)}$ use the memoryless property of the source. We first note that the support threshold of this companding system is $t_{1,c}^{(N)} = 3\log N$ and this is the upper bound in Theorem III.1. Next, we note that, as in the case of the Nitadori sequence which is the sequence of optimal half steps [19], the sequence of step widths $\Delta_{k,c}^{(N)}$ (and hence the sequence of half steps) depends on N only for the length of the sequence, while the thresholds and the levels of q_N^c depend directly on N.

Revisiting $\Delta_l(\Delta)$: Centroids for our companding system. Using (3.3.5) and (3.3.6) with (3.3.8) for q_N^c, we have

$$\begin{aligned}\underline{\Delta}_{k,c}^{(N)} &= 1 - \frac{\Delta_{k,c}^{(N)} e^{-\Delta_{k,c}^{(N)}}}{1-e^{-\Delta_{k,c}^{(N)}}} = 1 - \frac{\Delta_{k,c}^{(N)} e^{3\log\left(1-\frac{1}{k}\right)}}{1-e^{3\log\left(1-\frac{1}{k}\right)}} \\ &= 1 - \frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\end{aligned} \tag{3.3.11}$$

and

$$\begin{aligned}\overline{\Delta}_{k,c}^{(N)} &= \Delta_{k,c}^{(N)} - \underline{\Delta}_{k,c}^{(N)} = \Delta_{k,c}^{(N)} - \left[1 - \frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right] \\ &= \frac{\Delta_{k,c}^{(N)}\left[1-\left(1-\frac{1}{k}\right)^3\right] + \Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3} - 1 \\ &= \frac{\Delta_{k,c}^{(N)}}{1-\left(1-\frac{1}{k}\right)^3} - 1.\end{aligned} \tag{3.3.12}$$

Lemmas used in the proof of the upper bound to Theorem III.1.

Lemma III.4. *For $k \geq 1$, $\overline{\Delta}_{k+1,c}^{(N)} > \underline{\Delta}_{k,c}^{(N)}$.*

Proof. Goal: To prove that for $k \geq 1$,

$$\frac{\overline{\Delta}^{(N)}_{k+1,c}}{\underline{\Delta}^{(N)}_{k,c}} > 1, \tag{3.3.13}$$

we will find a lower bound to the numerator in (3.3.13) and an upper bound to the denominator in (3.3.13) and show that the ratio of these bounds is greater than 1.

Constructing $\overline{\Delta}^{(N)}_{k+1,c_{LB}}$ and $\underline{\Delta}^{(N)}_{k,c\ UB}$. From (3.3.11) and (3.3.12)

$$\underline{\Delta}^{(N)}_{k,c} = 1 - \frac{\Delta^{(N)}_{k,c}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3} \tag{3.3.14}$$

$$\overline{\Delta}^{(N)}_{k+1,c} = \frac{\Delta^{(N)}_{k+1,c}}{1-\left(1-\frac{1}{k+1}\right)^3} - 1, \tag{3.3.15}$$

it is clear that if we replace $\Delta^{(N)}_{k,c}$ by a lower bound in (3.3.14) and if we replace $\Delta^{(N)}_{k+1,c}$ with a lower bound in (3.3.15), then we will be creating an upper bound to $\underline{\Delta}^{(N)}_{k,c}$ and a lower bound to $\overline{\Delta}^{(N)}_{k+1,c}$. Using a power series expansion for $\log(1-x)$ when $0 \leq x < 1$ [11]

$$\log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \ldots,$$

for $k \geq 1$, we create a lower bound to $\Delta^{(N)}_{k,c}$

$$\begin{aligned}
\Delta^{(N)}_{k,c} &= -3\log\left(1-\frac{1}{k}\right) \\
&= 3\left[\frac{1}{k} + \frac{1}{2}\frac{1}{k^2} + \frac{1}{3}\frac{1}{k^3} + \frac{1}{4}\frac{1}{k^4} + \ldots\right] \qquad (3.3.16) \\
&> 3\left[\frac{1}{k} + \frac{1}{2}\frac{1}{k^2} + \frac{1}{3}\frac{1}{k^3} + \frac{1}{4}\frac{1}{k^4}\right] \qquad (3.3.17)
\end{aligned}$$

and likewise, a lower bound to $\Delta^{(N)}_{k+1,c}$

$$\Delta^{(N)}_{k+1,c} > 3\left[\frac{1}{k+1} + \frac{1}{2}\frac{1}{(k+1)^2} + \frac{1}{3}\frac{1}{(k+1)^3} + \frac{1}{4}\frac{1}{(k+1)^4}\right], \tag{3.3.18}$$

and using these bounds, we have

$$\begin{aligned}
\frac{\overline{\Delta}_{k+1,c}^{(N)}}{\underline{\Delta}_{k,c}^{(N)}} &= \frac{\frac{\Delta_{k+1,c}^{(N)}}{1-\left(1-\frac{1}{k+1}\right)^3} - 1}{1 - \frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}} > \frac{\frac{3\left[\frac{1}{k+1}+\frac{1}{2}\frac{1}{(k+1)^2}+\frac{1}{3}\frac{1}{(k+1)^3}+\frac{1}{4}\frac{1}{(k+1)^4}\right]}{1-\left(1-\frac{1}{k+1}\right)^3} - 1}{1 - \frac{3\left[\frac{1}{k}+\frac{1}{2}\frac{1}{k^2}+\frac{1}{3}\frac{1}{k^3}+\frac{1}{4}\frac{1}{k^4}\right]\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}} \\
&= \frac{3(3k^2-3k+1)k^4(6k^2+12k+7)}{(18k^5-18k^4+3k^3+3k^2-5k+3)(k+1)(3k^2+3k+1)} \\
&= \frac{54k^8+54k^7-27k^6-27k^5+21k^4}{54k^8+54k^7-27k^6-27k^5-3k^4-6k^3+k^2+7k+3} > 1
\end{aligned}$$

when $k \geq 1$. Thus we have shown that (3.3.13) is true for $k \geq 1$. ■

Remarks. From Lemma III.4, it is clear that the thresholds $t_{k,c}^{(N)}$ relative to the chosen levels of the companding-based q_N^c do not satisfy the nearest neighbor (n.n.) condition (Chapter II). Furthermore, this lemma shows that for any adjacent quantization cells, the upper half step of the quantization cell on the left is always larger than the lower half step of the cell to the right. This is a key fact that will be used later.

Now, we prove some general facts regarding the centroid/half step function $\Delta_l(\Delta)$.

Lemma III.5. *Suppose f is a pdf and let $\Delta > 0$. Then:*

1. *$0 < \Delta_l(\Delta) < \Delta$ (when $P_{[0,\Delta)} > 0$).*
2. *If $f > 0$ a.e. (in its domain $[0,\infty)$) then $\Delta_l(\Delta)$ is a strictly increasing function of Δ.*
3. *If f is strictly decreasing, then $\Delta_u(\Delta)$ is a strictly increasing function of Δ.*

Proof.

1. Suppose $P_{[0,\Delta)} > 0$ (because if not, $\Delta_l(\Delta)$ is not defined). Since f is a pdf and $P_{[0,\Delta)} = \int_0^\Delta f(x)\,dx > 0$, the set $S_{[0,\Delta)} \triangleq$ $\{x \in [0,\Delta) : f(x) > 0 \ \ a.e.\} \subseteq [0,\Delta)$ has Lebesgue measure greater than zero, i.e., $m\left(S_{[0,\Delta)}\right) > 0$ where m denotes Lebesgue measure. Then to show $\Delta_l(\Delta) >$

0,

$$\Delta_l(\Delta) = \frac{1}{P_{[0,\Delta)}} \int_0^{\Delta} x f(x)\,dx = \frac{1}{P_{[0,\Delta)}} \left(\int_{S_{[0,\Delta)}} x f(x)\,dx + \int_{[0,\Delta)\setminus S_{[0,\Delta)}} x f(x)\,dx \right)$$
$$= \frac{1}{P_{[0,\Delta)}} \int_{S_{[0,\Delta)}} x f(x)\,dx > 0.$$

where the inequality is due to the fact that $xf(x) > 0$ a.e. in $S_{[0,\Delta)}$.

Similarly, to show $\Delta_l(\Delta) < \Delta$,

$$\Delta - \Delta_l(\Delta) = \frac{1}{P_{[0,\Delta)}} \int_0^{\Delta} (\Delta - x)\, f(x)\,dx = \frac{1}{P_{[0,\Delta)}} \int_{S_{[0,\Delta)}} (\Delta - x)\, f(x)\,dx > 0$$

since $(\Delta - x) f(x) > 0$ a.e. in $S_{[0,\Delta)}$. ■

2. Fix $\epsilon > 0$. Since i) $f > 0$ a.e. and ii) $P_{[0,\Delta)},\ P_{[0,\Delta+\epsilon)} > 0$, $\Delta_l(\Delta)$ and $\Delta_l(\Delta+\epsilon)$ are well-defined. Then

$$\begin{aligned}\Delta_l(\Delta+\epsilon) &= \frac{1}{P_{[0,\Delta+\epsilon)}} \int_0^{\Delta+\epsilon} x f(x)\,dx \\ &= \frac{1}{P_{[0,\Delta+\epsilon)}} \left[\int_0^{\Delta} x f(x)\,dx + \int_{\Delta}^{\Delta+\epsilon} x f(x)\,dx \right] \\ &= \frac{P_{[0,\Delta)}}{P_{[0,\Delta+\epsilon)}} \times \Delta_l(\Delta) + \frac{P_{[\Delta,\Delta+\epsilon)}}{P_{[0,\Delta+\epsilon)}} \times E\left[X \mid \Delta < X < \Delta+\epsilon\right] \\ &> \frac{P_{[0,\Delta)}}{P_{[0,\Delta+\epsilon)}} \times \Delta_l(\Delta) + \frac{P_{[\Delta,\Delta+\epsilon)}}{P_{[0,\Delta+\epsilon)}} \times \Delta_l(\Delta) \\ &= \Delta_l(\Delta)\end{aligned}$$

where the inequality comes from the fact that since $P_{[\Delta,\Delta+\epsilon)} > 0$ (because $f > 0$ a.e. is assumed), $E[X \mid \Delta < X < \Delta+\epsilon]$ is well-defined, and $E[X \mid \Delta < X < \Delta+\epsilon] \geq \Delta > \Delta_l(\Delta)$ by Part 1. ■

3. Since $\Delta_l(\Delta)$ is differentiable and since f strictly decreasing implies $f > 0$

everywhere, we have

$$\begin{aligned}\frac{d}{d\Delta}\Delta_l(\Delta) &= \frac{d}{d\Delta}\frac{1}{P_{[0,\Delta)}}\int_0^{\Delta} uf(u)\,du \\ &= -\frac{1}{\left(P_{[0,\Delta)}\right)^2}\times f(\Delta)\times\int_0^{\Delta} uf(u)\,du + \frac{\Delta f(\Delta)}{P_{[0,\Delta)}} \\ &= \frac{f(\Delta)}{P_{[0,\Delta)}}\left[\Delta - \Delta_l(\Delta)\right] \qquad (3.3.19) \\ &> 0,\end{aligned}$$

where the inequality comes from Part 1.

Then using (3.3.19), we have

$$\frac{d}{d\Delta}\Delta_u(\Delta) = \frac{d}{d\Delta}\left[\Delta - \Delta_l(\Delta)\right] = 1 - \frac{f(\Delta)}{P_{[0,\Delta)}}\left[\Delta - \Delta_l(\Delta)\right]$$

$$> 1 - \frac{f(\Delta)}{\Delta f(\Delta)}\left[\Delta - \Delta_l(\Delta)\right] = 1 - \frac{\Delta - \Delta_l(\Delta)}{\Delta} > 0,$$

where in the first inequality, we have used the fact that f is strictly decreasing and the second inequality comes from Part 1. ■

Corollary III.6. *Related to Lemma III.5, parts 2 and 3: Let $\Delta > 0$ and suppose f is a pdf. Then:*

1. *Δ is a strictly increasing function of Δ_l when $f > 0$ everywhere.*

2. *Δ is a strictly increasing function of Δ_u when f is strictly decreasing.*

Proof. These properties follow from the Inverse Function Theorem. ■

We are now ready to finish the proof of the upper bound in Theorem III.1.

Lemma III.7. *For any $N \geq 2$, $\Delta_{k,c}^{(N)} > \Delta_k^{(N)}$, $k = 2, 3, \cdots, N$.*

Proof. By induction. Fix $N \geq 2$.

Let $k = 2$. By straightforward calculation using (3.3.8) and (3.3.12),

$$\overline{\Delta}_{2,c}^{(N)} = \frac{8}{7} \cdot 3 \log 2 - 1 \approx 1.3765,$$

and by the translation invariance of the lower half step of optimal quantizers and because optimal quantizers satisfy n.n. optimality,

$$\overline{\Delta}_{2}^{(N)} = \underline{\Delta}_{1}^{(N)} = \underline{\eta}_1 = 1.$$

Since $\overline{\Delta}_{2,c}^{(N)} > \overline{\Delta}_{2}^{(N)}$ and because Δ is strictly increasing in Δ_u (Corollary III.6), we see that $\Delta_{2,c}^{(N)} > \Delta_{2}^{(N)}$.

Now, assume that $\Delta_{k,c}^{(N)} > \Delta_{k}^{(N)}$ for all $2 \leq k \leq m < N$.

Let $k = m + 1$. Then we know the following:

$$\begin{aligned} \overline{\Delta}_{m+1,c}^{(N)} &> \underline{\Delta}_{m,c}^{(N)} \quad \text{(Lemma III.4)} \\ &> \underline{\Delta}_{m}^{(N)} \quad \text{(previous assumption and from Corollary III.6)} \\ &= \overline{\Delta}_{m+1}^{(N)} \quad \text{(n.n. condition satisfied by half steps of optimal quantizers).} \end{aligned}$$

Since Δ is strictly increasing in Δ_u (Corollary III.6), it is clear that $\Delta_{m+1,c}^{(N)} > \Delta_{m+1}^{(N)}$.
Thus for all $k = 2, 3, \ldots, N$, $\Delta_{k,c}^{(N)} > \Delta_{k}^{(N)}$. ■

With Lemma III.7, proof of Part 1 of Theorem III.1 is a simple matter of comparing the sums of the step sizes of q_N^* and q_N^c:

$$t_1^{(N)} = \sum_{k=2}^{N} \left(\Delta_k^{(N)} \right) < \sum_{k=2}^{N} \Delta_{k,c}^{(N)} = t_{1,c}^{(N)}.$$

Since $t_{1,c}^{(N)} = 3 \log N$, we conclude that $t_1^{(N)} < 3 \log N$.
This is the end of the proof to the upper bound in Theorem III.1. ■

Remarks. We remark that since for any N, $\Delta_{k,c}^{(N)} > \Delta_{k}^{(N)}$ for all $k \geq 2$, the difference between $t_{1,c}^{(N)}$ and $t_1^{(N)}$ is strictly increasing in N.

We also remark that the sequence of step sizes $\Delta_{k,c}^{(N)}$ from the companding-based quantization system in q_N^c is, like the Nitadori sequence, a fixed set of values that do not change as the number of levels N is altered. This phenomenon comes from two facts:

1. The structure of the companding system requires that each quantization cell of such an N-level quantizer has what we will call *measure* equal to

$$M_{k,N,c} = \frac{1}{N}, \tag{3.3.20}$$

where the *measure of the kth quantization cell* of an N-level companding system is

$$M_{k,N,c} \triangleq \int\limits_{t_{k,c}^{(N)}}^{t_{k-1,c}^{(N)}} \lambda\left(x\right) dx \tag{3.3.21}$$

and $\lambda\left(x\right) = \frac{1}{3} e^{-\frac{x}{3}}$ is the point density of the companding system.

2. Memoryless property with respect to $M_{k,N,c}$. The point density of the companding system, being equal to the derivative of the compressor function $C^*\left(x\right)$, has the form of an exponential pdf (with mean $E\left[X\right] = 3$ and variance $\sigma^2 = 9$) over $[0, \infty)$. Expanding the definition of our measure, we define the *measure of reconstruction points in* $[t, t + \Delta)$ *for* $t \geq 0$ *and* $\Delta > 0$ to be

$$M\left([t, t+\Delta)\right) \triangleq \int\limits_{t}^{t+\Delta} \lambda\left(x\right) dx.$$

We can think of the measure M as characterizing a random variable Y in the sense that, for an arbitrary interval $[t, t + \Delta)$,

$$P_{Y \in [t, t+\Delta)} = M\left([t, t+\Delta)\right).$$

Thus Y has a memoryless property which, using the convention $M(Y{\in}[t,t{+}\Delta)){=}$ $M([t, t + \Delta))$, can be expressed as

$$M\left(Y \geq s + t \left| Y \geq s\right.\right) = M\left(Y \geq t\right) \tag{3.3.22}$$

Using (3.3.22), we have

$$\begin{aligned} M\left(Y \in [t, t+\Delta)\right) &= M\left(Y \geq t\right) \cdot M\left(t \leq Y < t + \Delta \,|\, Y \geq t\right) \\ &= M\left(Y \geq t\right) \cdot M\left(0 \leq Y < \Delta\right). \end{aligned} \tag{3.3.23}$$

Fix $N \geq 1$. Using (3.3.20) and applying (3.3.23) to (3.3.21), we have

$$\begin{aligned} \frac{1}{N} = M_{k,N,c} = \int\limits_{t_{k,c}^{(N)}}^{t_{k,N}^{(N)}+\Delta_{k,c}^{(N)}} \lambda\left(x\right) dx = M\left(Y \in \left[t_{k,c}^{(N)}, t_{k,c}^{(N)} + \Delta_{k,c}^{(N)}\right)\right) \\ = M\left(Y \geq t_{k,c}^{(N)}\right) \cdot M\left(0 \leq Y < \Delta_{k,c}^{(N)}\right) \end{aligned} \tag{3.3.24}$$

Since

$$M\left(Y \geq t_{k,c}^{(N)}\right) = M\left(Y \in \bigcup_{j=1}^{k} \left[t_{j,c}^{(N)}, t_{j-1,c}^{(N)}\right)\right) = \sum_{j=1}^{k} M_{j,N,c} = \sum_{j=1}^{k} \frac{1}{N} = \frac{k}{N}, \tag{3.3.25}$$

(3.3.24) becomes

$$\frac{1}{N} = \frac{k}{N} \cdot M\left(0 \leq Y < \Delta_{k,c}^{(N)}\right)$$

or equivalently,

$$M\left(0 \leq Y < \Delta_{k,c}^{(N)}\right) = \frac{1}{k}, \tag{3.3.26}$$

which shows that the measure of $\left[0, \Delta_{k,c}^{(N)}\right)$ is independent of N. From (3.3.26), the expression for $\Delta_{k,c}^{(N)}$ can be determined

$$\left(-e^{-\frac{1}{3}\Delta_{k,c}^{(N)}} + 1\right) = \frac{1}{k}$$

or equivalently,

$$e^{-\frac{1}{3}\Delta_{k,c}^{(N)}} = 1 - \frac{1}{k}$$

or finally,

$$\Delta_{k,c}^{(N)} = -3\log\left(1 - \frac{1}{k}\right).$$

Thus we see that since the measure of $\Delta_{k,c}^{(N)}$ is independent of N, we can conclude that the step sizes $\Delta_{k,c}^{(N)}$ also do not depend on the value of N. Furthermore, the fact that the step sizes $\Delta_{k,c}^{(N)}$ depend only on the value of k (which is less than or equal to N) is in contrast to the fact that the thresholds $t_{k,c}^{(N)}$ are determined by the relationship in (3.3.25) that is a function of both k and N. As a final observation, in contrast to the Nitadori sequence $\underline{\eta}_k$ and optimal quantizers, where the sequence of half steps $\underline{\Delta}_k^{(N)} = \underline{\eta}_k$ is fixed, irrespective of N, for these companding systems, it is the sequence of step sizes (not half step sizes) that is fixed and does not depend on N.

3.3.2.2 Proof of lower bound to $t_1^{(N)}$ in Theorem III.1.

To show

$$t_1^{(N)} > 3\log N + \delta(N) - 1.46004 \tag{3.3.27}$$

where $\delta(N)$ has been defined in (3.2.3), we divide the analysis into two steps:

1. Show there exists a sequence s_k that is a term-by-term lower bound to the Nitadori sequence, i.e.,

$$s_k \leq \underline{\eta}_k \tag{3.3.28}$$

   for all $k \geq 1$, which subsequently, implies that s_k satisfies

$$t_1^{(N)} = \underline{\eta}_1 + \underline{\eta}_N + 2\sum_{k=2}^{N-1} \underline{\eta}_k \geq s_1 + s_N + 2\sum_{k=2}^{N-1} s_k. \tag{3.3.29}$$

2. Show that for $N > 9$

$$s_1 + s_N + 2\sum_{k=2}^{N-1} s_k > 3\log N + \delta(N) - 1.46004 \tag{3.3.30}$$

(3.3.27) follows from Steps 1 and 2.

Execution of Step 1. Define the sequence s_k as

$$s_k = \underline{\eta}_k, \ \text{for } k = 1, 2, \ldots, 8, \qquad \text{and} \qquad s_k = \frac{3}{2k}\Omega_k\,, \quad k \geq 9, \tag{3.3.31}$$

where for $k \geq 9$,

$$\Omega_k \triangleq \left(1 - \frac{2}{k} + \frac{8}{k^2}\right)^{\frac{1}{2}}. \tag{3.3.32}$$

As seen in Table 3.1, when $1 \leq k \leq 64$, it appears that $s_k \leq \underline{\eta}_k$, and this observation supports (3.3.28). However, proving (3.3.28) when $k \geq 9$ is difficult to do directly since $\underline{\eta}_k$ is not easily expressed in a closed form that would facilitate a term-by-term comparison. In fact, $\underline{\eta}_k$ is defined recursively via

$$\underline{\eta}_{k+1} = G\left(\underline{\eta}_k\right) + 1, \quad \text{for } k \geq 1, \tag{3.3.33}$$

where G is the Nitadori generator function (refer to Chapter II) and $\underline{\eta}_1 \triangleq 1$. Thus we take an indirect approach to arrive at (3.3.28). We will construct a function F that satisfies

$$s_{k+1} \leq F\left(s_k\right) + 1 \leq \underline{\eta}_{k+1} \tag{3.3.34}$$

for $k \geq 1$, i.e., the function F essentially produces a sequence which is sandwiched between s_k and $\underline{\eta}_k$. The following lemma lists a sufficient set of properties that, if possessed by F, ensures that (3.3.34) is true.

Lemma III.8. *Suppose a function F satisfies*

a. F is a lower bound to G, where G is the Nitadori sequence generator function.

b. F is increasing.

c. For all $k \geq 1$, $s_{k+1} \leq F\left(s_k\right) + 1$, where the sequence s_k, $k \geq 1$, is defined as shown in (3.3.31)*.*

Then (3.3.34) *is true for all $k \geq 1$.*

Proof. By induction. Suppose F is a function with properties a, b, c. Let $k = 1$. Since $s_1 = \underline{\eta}_1$ (from the definition in (3.3.31)), then

$$s_2 \overset{c}{\leq} F\left(s_1\right) + 1 \overset{a}{\leq} G\left(\underline{\eta}_1\right) + 1 = \underline{\eta}_2.$$

Thus for $k = 1$, (3.3.34) is true. Assume that for $k = n - 1$, (3.3.34) is also true, i.e.,

$$s_n \leq F\left(s_{n-1}\right) + 1 \leq \underline{\eta}_n. \tag{3.3.35}$$

Now let $k = n$. If $n \leq 8$, then

$$s_{n+1} \overset{c}{\leq} F\left(s_n\right) + 1 \overset{a}{\leq} G\left(\underline{\eta}_n\right) + 1 = \underline{\eta}_{n+1}$$

Table 3.1: The first 64 values of the sequence s_k and the Nitadori sequence $\underline{\eta}_k$.

index	s_k	$\underline{\eta}_k$	index	s_k	$\underline{\eta}_k$	index	s_k	$\underline{\eta}_k$
k			k			k		
1	1.0000	1.0000	23	0.0628	0.0637	45	0.0327	0.0329
2	0.5936	0.5936	24	0.0603	0.0611	46	0.0320	0.0322
3	0.4240	0.4240	25	0.0579	0.0587	47	0.0313	0.0315
4	0.3301	0.3301	26	0.0558	0.0565	48	0.0306	0.0309
5	0.2704	0.2704	27	0.0538	0.0544	49	0.0300	0.0303
6	0.2290	0.2290	28	0.0519	0.0525	50	0.0294	0.0297
7	0.1986	0.1986	29	0.0502	0.0507	51	0.0289	0.0291
8	0.1753	0.1753	30	0.0485	0.0491	52	0.0283	0.0285
9	0.1560	0.1570	31	0.0470	0.0475	53	0.0278	0.0280
10	0.1407	0.1421	32	0.0456	0.0461	54	0.0273	0.0275
11	0.1282	0.1298	33	0.0442	0.0447	55	0.0268	0.0270
12	0.1179	0.1194	34	0.0430	0.0434	56	0.0263	0.0265
13	0.1091	0.1106	35	0.0418	0.0422	57	0.0259	0.0261
14	0.1015	0.1030	36	0.0406	0.0410	58	0.0254	0.0256
15	0.0950	0.0964	37	0.0396	0.0399	59	0.0250	0.0252
16	0.0892	0.0906	38	0.0385	0.0389	60	0.0246	0.0248
17	0.0842	0.0854	39	0.0376	0.0379	61	0.0242	0.0244
18	0.0797	0.0808	40	0.0366	0.0370	62	0.0238	0.0240
19	0.0756	0.0767	41	0.0358	0.0361	63	0.0235	0.0236
20	0.0719	0.0730	42	0.0349	0.0352	64	0.0231	0.0232
21	0.0686	0.0696	43	0.0341	0.0344			
22	0.0656	0.0665	44	0.0334	0.0337			

since $s_n = \underline{\eta}_n$ (by the definition of s_k given in (3.3.31)). If $n \geq 9$, then

$$s_{n+1} \overset{c}{\leq} F\left(s_n\right) + 1 \overset{(3.3.35)\text{ and }b}{\leq} F\left(\underline{\eta}_n\right) + 1 \overset{a}{\leq} G\left(\underline{\eta}_n\right) + 1 = \underline{\eta}_{n+1},$$

i.e., (3.3.34) is true for $k = n$. Since n was arbitrary, we conclude that (3.3.34) is true for all $k \geq 1$. ■

Construction of F. The key to creating the function F with the three properties listed in Lemma III.8 is to use the fact that the Nitadori sequence is produced recursively through use of a monotonically increasing generator function G (3.3.33) and to choose F to be a monotonic increasing lower bound to G. Clearly, this choice will guarantee that F will possess the first two properties in Lemma III.8. By judiciously choosing F, we can also ensure that F has the third property in Lemma III.8.

Recall that $G = L \circ Z$ is the composition of two functions:

- $Z : (0, 1] \to \left(-\frac{1}{e}, -\frac{2}{e^2}\right]$ defined as

$$Z\left(w\right) \triangleq -\left(w+1\right) e^{-(w+1)} \tag{3.3.36}$$

  and note that Z is strictly increasing on $(0, 1]$ since

$$\frac{d}{dw} Z\left(w\right) = -e^{-(w+1)} + \left(w+1\right) e^{-(w+1)} = we^{-(w+1)} > 0.$$

- The principal branch of the Lambert W function $L : \left(-\frac{1}{e}, -\frac{2}{e^2}\right] \to (-1, 0]$ is defined such that for any $x \in \left(-\frac{1}{e}, -\frac{2}{e^2}\right]$, $L\left(x\right)$ is the unique value of $x \in (-1, 0]$ such that $we^w = x$. We note that L is strictly increasing on $\left(-\frac{1}{e}, -\frac{2}{e^2}\right]$.

Then $G : (0, 1] \to (-1, 0]$ is a strictly increasing function due to the combination of the monotonicity of both Z and L on their domains, and the Nitadori sequence $\underline{\eta}_k$ is generated by

$$\underline{\eta}_k = G\left(\underline{\eta}_{k-1}\right) + 1 = L \circ Z\left(\underline{\eta}_{k-1}\right) + 1 = L\left(-\left(\underline{\eta}_{k-1} + 1\right) e^{-\left(\underline{\eta}_{k-1}+1\right)}\right) + 1.$$

In order to create “F”, we will modify the individual functions that comprise G. First, we replace L by a lower bound function $L_{LB} : \left(-\frac{1}{e}, -\frac{2}{e^2}\right] \to (-1, 0]$ that is strictly increasing. To do this, we start with the composite representation for L given in [6]:

$$L = B \circ p, \tag{3.3.37}$$

where $p : \left(-\frac{1}{e}, -\frac{2}{e^2}\right] \to \left(0, \sqrt{2}\sqrt{1-\frac{2}{e}}\,\right]$ is

$$p(z) \overset{\triangle}{=} \sqrt{2}\sqrt{1+z\,e},$$

and

$$B(p) \overset{\triangle}{=} -1 + p - \frac{1}{3}p^2 + \frac{11}{72}p^3 - \frac{43}{540}p^4 + \frac{760}{17280}p^5 - \frac{221}{8505}p^6 + \ldots \tag{3.3.38}$$

is an infinite series that converges for $|p| < \sqrt{2}$, and is derived by inverting a power series expansion. (See [6] for details.) Focusing on the series B and truncating it to the fourth order, we define the polynomial $B_{LB} : \left(0, \sqrt{2}\sqrt{1-\frac{2}{e}}\,\right] \to (-1, 0]$ as

$$B_{LB} \overset{\triangle}{=} -1 + p - \frac{1}{3}p^2 + \frac{11}{72}p^3 - \frac{43}{540}p^4$$

which is a lower bound to B.[3] Replacing B with B_{LB} in (3.3.37) produces a lower bound to L: $L_{LB} \overset{\triangle}{=} B_{LB} \circ p(z)$ where

$$\begin{aligned} L_{LB}(z) &= B_{LB}(p(z)) \\ &= -1 + p(z) - \frac{1}{3}p^2(z) + \frac{11}{72}p^3(z) - \frac{43}{540}p^4(z)\,. \end{aligned}$$

The monotonicity of $L_{LB}(z)$, $z \in \left(-\frac{1}{e}, -\frac{2}{e^2}\right]$, follows from the monotonicity of p and B_{LB} which we now show:

1. Monotonicity of $p(z)$:

   $$\frac{d}{dz}p(z) = \frac{1}{\sqrt{2}}\frac{e}{\sqrt{1+z\,e}} > 0$$

   for all $z > -\frac{1}{e}$. Hence $p(z)$ is strictly increasing on $\left(-\frac{1}{e}, -\frac{2}{e^2}\right]$.

2. Monotonicity of $B_{LB}(p)$:

   $$\frac{d}{dp}B_{LB}(p) = 1 - \frac{2}{3}p + \frac{11}{24}p^2 - \frac{43}{135}p^3$$

[3] The proof of this statement is contained in Appendix C.

and when $p \in \left(0, \sqrt{2}\sqrt{1-\frac{2}{e}}\right]$, this derivative is positive, i.e., $B_{LB}(p)$ is strictly increasing on $\left(0, \sqrt{2}\sqrt{1-\frac{2}{e}}\right]$.

Now, consider the function

$$L_{LB} \circ Z(x) = B_{LB} \circ p \circ Z(x).$$

(See Figure 3.1 for a visualization.) While this function is both a lower bound to G and increasing, the form of this function is still not easy to work with. Thus, we will go one step further and create a lower bound to it that is also increasing. For this modification, we will lower bound the composition $p \circ Z$ using a lower bound to e^x that results from truncating a continued fraction expansion for e^x. Specifically, we will truncate the continued fraction expansion [1]

$$e^x = m_0 + \cfrac{x}{m_1 + \cfrac{x}{m_2 + \cfrac{x}{m_3 + \cfrac{x}{m_4 + \cdots}}}}, \tag{3.3.39}$$

where the convergents m_n are given by

$$m_{2n} = 2(-1)^n \text{ and } m_{2n+1} = (2n+1)^n, \quad n = 1, 2, \ldots,$$

with

$$m_0 = 1 \text{ and } m_1 = 1,$$

to a ratio of fourth order polynomials

$$e_{cf8}(x) \triangleq \frac{1680 + 180x^2 + 840x + 20x^3 + x^4}{1680 + 180x^2 - 840x - 20x^3 + x^4} \tag{3.3.40}$$

which was created by truncating to the eighth convergent of the expansion in (3.3.39). We remark that $e_{cf8}(-x) = (e_{cf8}(x))^{-1}$, just like $e^{-x} = (e^x)^{-1}$.

The main reasons behind using a continued fraction approximation to the exponential function are two-fold:

- Good accuracy is achieved using a small number of convergents (as opposed to

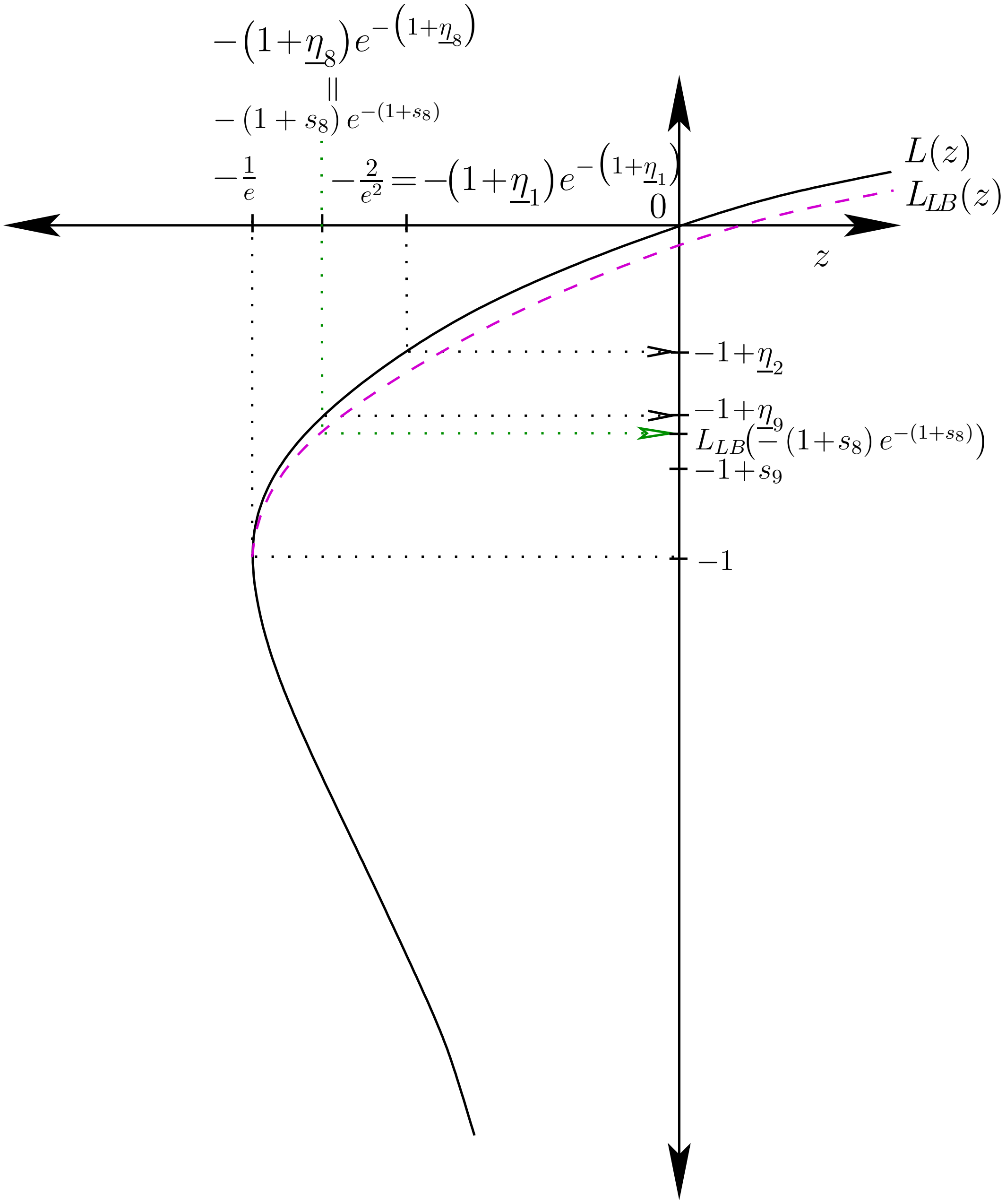

$-\left(1+\underline{\eta}_8\right)e^{-\left(1+\underline{\eta}_8\right)}$
$\|$
$-\left(1+s_8\right)e^{-(1+s_8)}$
$-\frac{1}{e}$
$-\frac{2}{e^2}=-\left(1+\underline{\eta}_1\right)e^{-\left(1+\underline{\eta}_1\right)}$
0
$L(z)$
$L_{LB}(z)$
z
$-1+\underline{\eta}_2$
$-1+\underline{\eta}_9$
$L_{LB}\left(-\left(1+s_8\right)e^{-(1+s_8)}\right)$
$-1+s_9$
-1

Figure 3.1: Illustration of $L(z) = B(p(z))$ and $L_{LB}(z) = B_{LB}(p(z))$, showing how $L_{LB}\left(-(1+s_8)e^{-(1+s_8)}\right) > -1 + s_9$, where $s_8 = \underline{\eta}_8$. Note that $L_{LB}(z)$ is a precursor to the final function $F(x)$ which is a lower bound to $G(x)$.

the number of terms required in a Taylor series expansion of the same function to achieve the same accuracy).

- The continued fraction expansion yields an approximation in a convenient polynomial form that is easy to work with.

Note:

$$\begin{aligned}
\frac{d}{dx}e_{cf8}\left(x\right) &= \frac{60x^2+4x^3+360x+840}{x^4-20x^3+180x^2-840x+1680} - \\
&\quad \frac{\left(x^4-20x^3+180x^2-840x+1680\right)\left(4x^3-60x^2+360x-840\right)}{\left(x^4-20x^3+180x^2-840x+1680\right)^2} \\
&= -\frac{40\left(x^6-54x^4+2520x^2-70560\right)}{\left(x^4-20x^3+180x^2-840x+1680\right)^2} \\
&> 0
\end{aligned}$$

when

$$|x| < \sqrt{2}\frac{\sqrt{\left(2304+147\sqrt{345}\right)^{\frac{2}{3}}-129+9\left(2304+147\sqrt{345}\right)^{\frac{1}{3}}}}{\left(2304+147\sqrt{345}\right)^{\frac{1}{6}}} \approx 6.101171636.$$

Lower bound to $p \circ Z$. Concentrating on the $p\left(Z\left(x\right)\right)$ term, we have

$$\begin{aligned}
p\left(Z\left(x\right)\right) &= \sqrt{2}\sqrt{1+Z\left(x\right)\cdot e} = \sqrt{2}\sqrt{1-\left(x+1\right)e^{-(x+1)}\cdot e} = \sqrt{2}\sqrt{1-\left(x+1\right)e^{-x}} \\
&\geq \sqrt{2}\sqrt{1-\left(x+1\right)e_{cf8}\left(-x\right)} \\
&= \sqrt{2}\sqrt{1-\left(x+1\right)\left(\frac{1680+180x^2+840x+20x^3+x^4}{1680+180x^2-840x-20x^3+x^4}\right)} \quad \text{(using (3.3.40))} \\
&= \sqrt{2}\cdot\sqrt{-\frac{\left(x^3-20x^2+140x-840\right)x^2}{x^4+20x^3+180x^2+840x+1680}} \\
&= \sqrt{2}x\cdot\sqrt{-\frac{x^3-20x^2+140x-840}{x^4+20x^3+180x^2+840x+1680}} \\
&= \sqrt{2}x\cdot\sqrt{\frac{840-140x+20x^2-x^3}{x^4+20x^3+180x^2+840x+1680}} \triangleq \left(p\circ Z\right)_{LB}\left(x\right). \qquad (3.3.41)
\end{aligned}$$

Remark on $(p \circ Z)_{LB}(x)$. The function $(p \circ Z)_{LB}(x)$ defined for $x \in [0,1]$ is strictly increasing since

$$\begin{aligned}\frac{d}{dx}(p \circ Z)_{LB}(x) &= \frac{\sqrt{2}}{2}\frac{\mathrm{num}_{pZLB}}{\mathrm{den}_{pZLB}} \\ &= -\frac{\sqrt{2}}{2}\frac{\left(x^7 + 40x^6 - 2160x^4 + 100800x^2 - 2822400\right)}{\sqrt{\frac{-(x^3-20x^2+140x-840)}{x^4+20x^3+180x^2+840x+1680}}\left(x^4 + 20x^3 + 180x^2 + 840x + 1680\right)^2}.\end{aligned}$$

This derivative is greater than zero when the polynomial term in the numerator is negative, and this polynomial was observed to be negative for $|x| < 5.8963$ when plotted.

Putting it all together to make F. We define

$$\begin{aligned}F(x) &\overset{\triangle}{=} B_{LB} \circ (p \circ Z)_{LB}(x) \\ &= -1 + (p \circ Z)_{LB}(x) - \frac{1}{3}(p \circ Z)^2_{LB}(x) + \frac{11}{72}(p \circ Z)^3_{LB}(x) - \frac{43}{540}(p \circ Z)^4_{LB}(x)\end{aligned} \tag{3.3.42}$$

and by construction, we know that F is both monotone increasing and a lower bound to G on $(0,1]$.

It remains to show the third property of Lemma III.8. To do this, we will find an upper bound to s_{k+1} and a lower bound to $F(s_k)+1$. Both of these will involve bounds to expressions of the form $(1+\alpha)^{\frac{1}{2}}$ that are obtained by truncating or tweaking the binomial expansions for such.

Binomial Theorem. From [1], for $|\alpha| < 1$,

$$\begin{aligned}(1+\alpha)^{\frac{1}{2}} &= 1 + \frac{\alpha}{2} + \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\alpha^2}{2!} + \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\left(\frac{1}{2}-2\right)\alpha^3}{3!} + \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\left(\frac{1}{2}-2\right)\left(\frac{1}{2}-3\right)\alpha^4}{4!} + \dots \\ &= 1 + \frac{1}{2}\alpha - \frac{1}{8}\alpha^2 + \frac{1}{16}\alpha^3 - \frac{5}{128}\alpha^4 + \dots\end{aligned}$$

and thus for $\alpha \in [0,1]$,

$$(1-\alpha)^{\frac{1}{2}} = 1 - \frac{1}{2}\alpha - \frac{1}{8}\alpha^2 - \frac{1}{16}\alpha^3 - \frac{5}{128}\alpha^4 - \dots. \tag{3.3.43}$$

Note that truncating the expansion in (3.3.43) at any point always results in an upper bound to $(1-\alpha)^{\frac{1}{2}}$.

Upper bound to s_{k+1}. To find an upper bound to s_{k+1}, we first write s_{k+1} from (3.3.31) (with $k = k+1$) as

$$
\begin{aligned}
s_{k+1} &= \frac{3}{2}\frac{1}{k+1}\left[1 - \frac{2}{k+1} + \frac{8}{(k+1)^2}\right]^{\frac{1}{2}} \\
&= \frac{3}{2}\frac{1}{k}\left[\frac{k^2}{(k+1)^2}\cdot\left(1 - \frac{2}{k+1} + \frac{8}{(k+1)^2}\right)\right]^{\frac{1}{2}} \\
&= \frac{3}{2}\frac{1}{k}\left[\frac{k^2}{(k+1)^2}\cdot\left(1 - \frac{2}{k+1} + \frac{8}{(k+1)^2}\right)\right]^{\frac{1}{2}} \times \frac{\Omega_k}{\Omega_k} \\
&= \frac{3}{2}\frac{1}{k} \times \Omega_k \left[\frac{k^2}{(k+1)^2}\cdot\left(\frac{1 - \frac{2}{k+1} + \frac{8}{(k+1)^2}}{\Omega_k^2}\right)\right]^{\frac{1}{2}} \\
&= \frac{3}{2}\frac{1}{k} \times \Omega_k \left[\frac{k^2}{(k+1)^2}\cdot\left(\frac{1 - \frac{2}{k+1} + \frac{8}{(k+1)^2}}{1 - \frac{2}{k} + \frac{8}{k^2}}\right)\right]^{\frac{1}{2}} \\
&= \frac{3}{2}\frac{1}{k} \times \Omega_k \left[\frac{k^4(k^2+7)}{(k+1)^4(k^2-2k+8)}\right]^{\frac{1}{2}} \\
&= \frac{3}{2}\frac{1}{k} \times \Omega_k \left[\frac{k^6+7k^4}{k^6+2k^5+6k^4+24k^3+41k^2+30k+8}\right]^{\frac{1}{2}}, \qquad (3.3.44)
\end{aligned}
$$

where we have used the definition of Ω_k given in (3.3.32).

Then using long division on the ratio of polynomials in (3.3.44), we have

$$
s_{k+1} = \frac{3}{2}\frac{1}{k} \times \Omega_k \left[1 - \frac{2}{k} + \frac{5}{k^2} - \frac{22}{k^3} + \frac{21}{k^4} + \frac{22}{k^5} + \frac{205}{k^6} + R_{7,s_{k+1}}\right]^{\frac{1}{2}}
$$

$$
\leq \frac{3}{2}\frac{1}{k} \times \Omega_k \left[1 - \frac{2}{k} + \frac{5}{k^2} - \frac{22}{k^3} + \frac{21}{k^4} + \frac{22}{k^5} + \frac{205}{k^6}\right]^{\frac{1}{2}}, \qquad (3.3.45)
$$

since

$$
R_{7,s_{k+1}} = -\frac{278 + \frac{1999}{k} + \frac{6276}{k^2} + \frac{9233}{k^3} + \frac{6326}{k^4} + \frac{1640}{k^5}}{k^6+2k^5+6k^4+24k^3+41k^2+30k+8} < 0
$$

for any $k \geq 1$. Thus by truncating $R_{7,s_{k+1}}$, we obtain the upper bound on the right hand side of (3.3.45). Since this upper bound is expressed as a square root, we will need to find an upper bound to it that is not expressed as a square root. Re-expressing (3.3.45) and using the binomial expansion in (3.3.43) with $\alpha = \frac{2}{k} - \frac{5}{k^2} + \frac{22}{k^3} - \frac{21}{k^4} - \frac{22}{k^5} - \frac{205}{k^6}$ and noting that $|\alpha| = \left|\frac{2}{k} - \frac{5}{k^2} + \frac{22}{k^3} - \frac{21}{k^4} - \frac{22}{k^5} - \frac{205}{k^6}\right| < 1$ when $k \geq 1$, we have for

$k \geq 5$[4]

$$
\begin{aligned}
(3.3.45)_{RHS} &= \frac{3}{2}\frac{1}{k} \times \Omega_k \left[1 - \left(\frac{2}{k} - \frac{5}{k^2} + \frac{22}{k^3} - \frac{21}{k^4} - \frac{22}{k^5} - \frac{205}{k^6}\right)\right]^{\frac{1}{2}} \\
&= \frac{3}{2}\frac{1}{k} \times \Omega_k \left[1 - \frac{1}{2}\left(\frac{2}{k} - \frac{5}{k^2} + \frac{22}{k^3} - \frac{21}{k^4} - \frac{22}{k^5} - \frac{205}{k^6}\right) - \frac{1}{8}\left(\frac{2}{k} - \frac{5}{k^2} + \right.\right. \\
&\quad \left.\frac{22}{k^3} - \frac{21}{k^4} - \frac{22}{k^5} - \frac{205}{k^6}\right)^2 - \frac{1}{16}\left(\frac{2}{k} - \frac{5}{k^2} + \frac{22}{k^3} - \frac{21}{k^4} - \frac{22}{k^5} - \frac{205}{k^6}\right)^3 - \\
&\quad \left.\frac{6}{128}\left(\frac{2}{k} - \frac{5}{k^2} + \frac{22}{k^3} - \frac{21}{k^4} - \frac{22}{k^5} - \frac{205}{k^6}\right)^4 - \ldots\right].
\end{aligned}
$$

To obtain an upper bound to s_{k+1}, we truncate this expansion after the fourth term[5] and expand

$$
\begin{aligned}
s_{k+1} \leq \frac{3}{2}\frac{1}{k} \times \Omega_k &\left[1 - \frac{1}{k} + \frac{2}{k^2} - \frac{9}{k^3} + \frac{1}{8k^4} + \frac{185}{8k^5} + \frac{2125}{16k^6} - \frac{1251}{8k^7} + \frac{12473}{16k^8} - \right. \\
&\frac{3951}{4k^9} + \frac{34175}{8k^{10}} - \frac{95603}{8k^{11}} + \frac{244177}{16k^{12}} - \frac{171987}{4k^{13}} + \frac{168381}{8k^{14}} - \frac{1097371}{8k^{15}} + \\
&\left.\frac{2945235}{16k^{16}} + \frac{1386825}{8k^{17}} + \frac{8615125}{16k^{18}}\right] \triangleq s_{k+1,UB} \qquad (3.3.46)
\end{aligned}
$$

Now that we have an upper bound to s_{k+1}, it is time to find a lower bound to $F(s_k)+1$.

Preparation for finding a lower bound to $F(s_k)+1$. $k \geq 3$. Since with $x = s_k$,

$$
F(s_k)+1 = (p \circ Z)_{LB}(s_k) - \frac{1}{3}(p \circ Z)^2_{LB}(s_k) + \frac{11}{72}(p \circ Z)^3_{LB}(s_k) - \frac{43}{540}(p \circ Z)^4_{LB}(s_k), \quad (3.3.47)
$$

we expect the square root terms in s_k (see (3.3.31), (3.3.32)) and $(p \circ Z)_{LB}(s_k)$ (see (3.3.41)) to appear throughout an expanded version of this function, and thus, finding a lower bound to $F(s_k)$ means swapping an upper bound for square root terms that are negative and using lower bounds to positive square root terms in s_k and $(p \circ Z)_{LB}(s_k)$.

[4] Note: Through out the remainder of this derivation, we have used the constraint $k \geq 5$ whenever we make decisions on truncating expansions and verifying bounds when, in reality, we really only needed to use the restriction that $k \geq 9$ (which comes from the definition of Ω_k). The resulting lower bound function F that is produced at the end of this discussion is, perhaps, slightly more complicated (and may be tighter) than necessary since it was derived to hold for $k \geq 5$ (which is overly restrictive) instead of $k \geq 9$. Nevertheless, the F function that we produce fulfills all of the requirements in Lemma III.8.

[5] Truncating after the third term (or any earlier term) did not yield an upper bound with sufficient accuracy to $(3.3.45)_{RHS}$ under the constraint that $k \geq 5$.

The two square root expressions that will be appear throughout the expression for $F(s_k) + 1$ are:

1. Ω_k which was introduced in (3.3.32)
2. $$\psi_k \triangleq \sqrt{\frac{k^4 - \frac{1}{4}\Omega_k k^3 + \frac{3}{56}\Omega_k^2 k^2 - \frac{9}{2240}\Omega_k^3 k}{k^4 + \frac{3}{4}\Omega_k k^3 + \frac{27}{112}\Omega_k^2 k^2 + \frac{9}{224}\Omega_k^3 k + \frac{27}{8960}\Omega_k^4}} \tag{3.3.48}$$

   which is an expression that will be seen later in $(p \circ Z)_{LB}(s_k)$.

Bounds for Ω_k. We define the following upper and lower bound to Ω_k:

$$\Omega_{k,UB} \triangleq 1 - \frac{1}{2}\left(\frac{2}{k} - \frac{8}{k^2}\right) = 1 - \frac{1}{k} + \frac{4}{k^2} \tag{3.3.49}$$

$$\Omega_{k,LB} \triangleq 1 - \frac{1}{k}. \tag{3.3.50}$$

It is clear that $\Omega_{k,UB}$ is an upper bound to Ω_k since

$$\begin{aligned}(\Omega_{k,UB})^2 - \Omega_k^2 &= \left(1 - \frac{1}{k} + \frac{4}{k^2}\right)^2 - \left(1 - \frac{2}{k} + \frac{8}{k^2}\right) \\ &= \left(1 - \frac{2}{k} + \frac{9}{k^2} - \frac{8}{k^3} + \frac{16}{k^4}\right) - \left(1 - \frac{2}{k} + \frac{8}{k^2}\right) \\ &= \frac{1}{k^2} - \frac{8}{k^3} + \frac{16}{k^4} = \frac{1}{k^4}\left(k^2 - 8k + 16\right) = \frac{1}{k^4}(k-4)^2 \\ &> 0\end{aligned}$$

for $k \geq 5$, and it is also clear that when $k \geq 5$, $\Omega_{k,LB}$ is a lower bound to Ω_k

$$\Omega_k^2 - (\Omega_{k,LB})^2 = \left(1 - \frac{2}{k} + \frac{8}{k^2}\right) - \left(1 - \frac{1}{k}\right)^2 = \left(1 - \frac{2}{k} + \frac{8}{k^2}\right) - \left(1 - \frac{2}{k} + \frac{1}{k^2}\right) = \frac{7}{k^2} > 0.$$

Summarizing: We have three functions, Ω_k defined for $k \geq 9$, and the bounds $\Omega_{k,UB}$ and $\Omega_{k,LB}$ to Ω_k that hold when $k \geq 9$.[6]

[6] Just to re-iterate, the bounds actually hold for $k \geq 5$.

A lower bound for ψ_k. First, we use long division to evaluate the ratio of polynomials underneath the square root in (3.3.48), where the ratio of polynomials is

$$\frac{k^4 - \frac{1}{4}\Omega_k k^3 + \frac{3}{56}\Omega_k^2 k^2 - \frac{9}{2240}\Omega_k^3 k}{k^4 + \frac{3}{4}\Omega_k k^3 + \frac{27}{112}\Omega_k^2 k^2 + \frac{9}{224}\Omega_k^3 k + \frac{27}{8960}\Omega_k^4}. \tag{3.3.51}$$

If we stop the long division after the fourth iteration[7], we have

$$\psi_k = \left[1 - \frac{\Omega_k}{k} + \frac{9}{16}\frac{\Omega_k^2}{k^2} - \frac{9}{40}\frac{\Omega_k^3}{k^3} + R_{4,(3.3.51)}\right]^{\frac{1}{2}}$$

where

$$R_{4,(3.3.51)} = \frac{\frac{9}{128}\Omega_k^4 + \frac{621}{17920}\frac{\Omega_k^5}{k} + \frac{1053}{143360}\frac{\Omega_k^6}{k^2} + \frac{243}{358400}\frac{\Omega_k^7}{k^3}}{k^4 + \frac{3}{4}\Omega_k k^3 + \frac{27}{112}\Omega_k^2 k^2 + \frac{9}{224}\Omega_k^3 k + \frac{27}{8960}\Omega_k^4}.$$

Since, for $k \geq 5$, the range where Ω_k is defined, $R_{4,(3.3.51)}$ is always positive, dropping it will produce a lower bound to ψ_k. With $\alpha = \left(\frac{\Omega_k}{k} - \frac{9}{16}\frac{\Omega_k^2}{k^2} + \frac{9}{40}\frac{\Omega_k^3}{k^3}\right)$ and noting that when $k \geq 5$, $|\alpha| < 1$, we can use the binomial expansion in (3.3.43) to express the newly created lower bound to ψ_k as an infinite series

$$\psi_k \geq \left[1 - \left(\frac{\Omega_k}{k} - \frac{9}{16}\frac{\Omega_k^2}{k^2} + \frac{9}{40}\frac{\Omega_k^3}{k^3}\right)\right]^{\frac{1}{2}} = 1 - \frac{1}{2}\left(\frac{\Omega_k}{k} - \frac{9}{16}\frac{\Omega_k^2}{k^2} + \frac{9}{40}\frac{\Omega_k^3}{k^3}\right) - \frac{1}{8}\left(\frac{\Omega_k}{k} - \frac{9}{16}\frac{\Omega_k^2}{k^2} + \frac{9}{40}\frac{\Omega_k^3}{k^3}\right)^2 - \frac{1}{16}\left(\frac{\Omega_k}{k} - \frac{9}{16}\frac{\Omega_k^2}{k^2} + \frac{9}{40}\frac{\Omega_k^3}{k^3}\right)^3 - \frac{5}{128}\left(\frac{\Omega_k}{k} - \frac{9}{16}\frac{\Omega_k^2}{k^2} + \frac{9}{40}\frac{\Omega_k^3}{k^3}\right)^4 - \dots. \tag{3.3.52}$$

To create a *useable* lower bound to ψ_k, we will have to truncate the infinite series in (3.3.52). Truncation of this binomial series, however, produces an upper bound and we require a lower bound. Thus, we first truncate the binomial expansion after four terms, which yields an upper bound to (3.3.51),

$$1 - \frac{1}{2}\left(\frac{\Omega_k}{k} - \frac{9}{16}\frac{\Omega_k^2}{k^2} + \frac{9}{40}\frac{\Omega_k^3}{k^3}\right) - \frac{1}{8}\left(\frac{\Omega_k}{k} - \frac{9}{16}\frac{\Omega_k^2}{k^2} + \frac{9}{40}\frac{\Omega_k^3}{k^3}\right)^2 - \frac{1}{16}\left(\frac{\Omega_k}{k} - \frac{9}{16}\frac{\Omega_k^2}{k^2} + \frac{9}{40}\frac{\Omega_k^3}{k^3}\right)^3,$$

and then, we "tweak" the resulting upper bound to obtain a lower bound: We drop all terms with powers of k less than or equal to -4. and increase the magnitude of the coefficient of the k^{-3} coefficient to get the following lower bound to (3.3.51) which

[7] While truncation after an even number of terms always yields a lower bound, we only obtain sufficiently tight lower bounds to ψ_k (under the constraint that $k \geq 5$) if we truncate after the fourth term. Moreover, truncating after the fourth term produces a lower bound to ψ_k with the simplest form (fewest number of terms).

is, in turn, also a lower bound to ψ_k:

$$\psi_k \geq 1 - \frac{1}{2}\frac{\Omega_k}{k} + \frac{5}{32}\frac{\Omega_k^2}{k^2} - \frac{12}{320}\frac{\Omega_k^3}{k^3} = 1 - \frac{1}{2}\frac{\Omega_k}{k} + \frac{5}{32}\frac{\Omega_k^2}{k^2} - \frac{3}{80}\frac{\Omega_k^3}{k^3} \triangleq \psi_{k,LB}. \tag{3.3.53}$$

To check that $\psi_{k,LB}$ in (3.3.53) is a lower bound to ψ_k:

$$\begin{aligned}
\psi_k^2 - \psi_{k,LB}^2 &= \left[\frac{k^4 - \frac{1}{4}\Omega_k k^3 + \frac{3}{56}\Omega_k^2 k^2 - \frac{9}{2240}\Omega_k^3 k}{k^4 + \frac{3}{4}\Omega_k k^3 + \frac{27}{112}\Omega_k^2 k^2 + \frac{9}{224}\Omega_k^3 k + \frac{27}{8960}\Omega_k^4}\right] - \left[\frac{1}{2}\frac{\Omega_k}{k} + \frac{5}{32}\frac{\Omega_k^2}{k^2} - \frac{3}{80}\frac{\Omega_k^3}{k^3}\right]^2 \\
&= \left[\frac{k^4 - \frac{1}{4}\Omega_k k^3 + \frac{3}{56}\Omega_k^2 k^2 - \frac{9}{2240}\Omega_k^3 k}{k^4 + \frac{3}{4}\Omega_k k^3 + \frac{27}{112}\Omega_k^2 k^2 + \frac{9}{224}\Omega_k^3 k + \frac{27}{8960}\Omega_k^4}\right] - \frac{1}{25600k^6}\Bigg[25600k^6 - \\
&\qquad 25600\Omega_k k^5 + 14400\Omega_k^2 k^4 - 5920\Omega_k^3 k^3 + 1585\Omega_k^4 k^2 - 300\Omega_k^5 k + 36\Omega_k^6\Bigg] \\
&= \frac{\text{num}_{\psi_k^2 - \psi_{k,LB}^2}}{\text{den}_{\psi_k^2 - \psi_{k,LB}^2}},
\end{aligned}$$

where

$$\begin{aligned}
\text{num}_{\psi_k^2 - \psi_{k,LB}^2} &\triangleq \Omega_k^3 \Big(1433600k^7 + 3001600\Omega_k k^6 + 331200\Omega_k^2 k^5 + 12240\Omega_k^3 k^4 - \\
&\qquad 4680\Omega_k^4 k^3 - 12555\Omega_k^5 k^2 - 4860\Omega_k^6 k - 972\Omega_k^7\Big) \qquad (3.3.54) \\
\text{den}_{\psi_k^2 - \psi_{k,LB}^2} &\triangleq 25600k^6 \Big(8960k^4 + 6720\Omega_k k^3 + 2160\Omega_k^2 k^2 + 360\Omega_k^3 k + 27\Omega_k^4\Big).
\end{aligned}$$

Since $\text{den}_{\psi_k^2 - \psi_{k,LB}^2} > 0$ when $k \geq 5$, then if $\text{num}_{\psi_k^2 - \psi_{k,LB}^2} > 0$, then $\psi_k^2 - \psi_{k,LB}^2 > 0$. We have used Maple to find that $\text{num}_{\psi_k^2 - \psi_{k,LB}^2}$ in (3.3.54) only has one real root with a value approximately equal to $.3051153731\Omega_k$. Since for each $k \geq 5$, k is also greater than or equal to the root of (3.3.54), and since the coefficient of the leading term of $\text{num}_{\psi_k^2 - \psi_{k,LB}^2}$ is also positive for every $k \geq 5$, we conclude that $\psi_k^2 - \psi_{k,LB}^2 > 0$ when $k \geq 5$.

A lower bound to $F(s_k) + 1$ when $k \geq 5$. Since $F(s_k) + 1$ is a fourth order polynomial in $(p \circ Z)_{LB}(s_k)$ (see (3.3.47)), it is time to evaluate $(p \circ Z)_{LB}(x)$. From

(3.3.41) with $x = s_k$

$$\begin{aligned}
&(p \circ Z)_{LB}(s_k) \\
&= \sqrt{2} s_k \cdot \sqrt{\frac{840 - 140 s_k + 20 s_k^2 - s_k^3}{s_k^4 + 20 s_k^3 + 180 s_k^2 + 840 s_k + 1680}} \\
&= \sqrt{2}\left(\frac{3}{2}\frac{1}{k}\Omega_k\right) \cdot \sqrt{\frac{840 - 140\left(\frac{3}{2}\frac{1}{k}\Omega_k\right) + 20\left(\frac{3}{2}\frac{1}{k}\Omega_k\right)^2 - \left(\frac{3}{2}\frac{1}{k}\Omega_k\right)^3}{\left(\frac{3}{2}\frac{1}{k}\Omega_k\right)^4 + 20\left(\frac{3}{2}\frac{1}{k}\Omega_k\right)^3 + 180\left(\frac{3}{2}\frac{1}{k}\Omega_k\right)^2 + 840\left(\frac{3}{2}\frac{1}{k}\Omega_k\right) + 1680}} \\
&= \frac{3}{2}\frac{1}{k} \times \Omega_k \times \sqrt{\frac{4k\left(2240k^3 - 560k^2\Omega_k + 120k\Omega_k^2 - 9\Omega_k^3\right)}{8960k^4 + 6720k^3\Omega_k + 2160k^2\Omega_k^2 + 360k\Omega_k^3 + 27\Omega_k^4}} \\
&= \frac{3}{2}\frac{1}{k} \times \Omega_k \times \sqrt{\frac{8960k^4 - 2240k^3\Omega_k + 480k^2\Omega_k^2 - 36k\Omega_k^3}{8960k^4 + 6720k^3\Omega_k + 2160k^2\Omega_k^2 + 360k\Omega_k^3 + 27\Omega_k^4}} \\
&= \frac{3}{2}\frac{1}{k} \times \Omega_k \sqrt{\frac{k^4 - \frac{1}{4}k^3\Omega_k + \frac{3}{56}k^2\Omega_k^2 - \frac{9}{2240}k\Omega_k^3}{k^4 + \frac{3}{4}k^3\Omega_k + \frac{27}{112}k^2\Omega_k^2 + \frac{9}{224}k\Omega_k^3 + \frac{27}{8960}\Omega_k^4}} \\
&= \frac{3}{2}\frac{1}{k} \times \Omega_k \psi_k. && (3.3.55)
\end{aligned}$$

Using (3.3.55) in (3.3.47), along with the definitions for Ω_k and ψ_k in (3.3.31) and (3.3.48), we have

$$F(s_k) + 1 = \frac{3}{2k} \times \Omega_k \times \frac{num_{(3.3.56)}}{den_{(3.3.56)}}, \tag{3.3.56}$$

where

$$\begin{aligned}
&\mathrm{num}_{(3.3.56)} \\
&= \psi_k k^{10} + \left(-\frac{1}{2}\Omega_k + \frac{3}{2}\psi_k\Omega_k\right)k^9 + \left(\frac{11}{32}\psi_k^3\Omega_k^2 - \frac{1}{4}\Omega_k^2 + \frac{117}{112}\psi_k\Omega_k^2\right)k^8 + \\
&\quad \left(\frac{33}{64}\psi_k^3\Omega_k^3 + \frac{99}{224}\psi_k\Omega_k^3 - \frac{361}{1120}\Omega_k^3\right)k^7 + \left(\frac{1287}{3584}\psi_k^3\Omega_k^4 + \frac{7803}{62720}\psi_k\Omega_k^4 + \frac{283}{2240}\Omega_k^4\right)k^6 + \\
&\quad \left(\frac{2997}{125440}\psi_k\Omega_k^5 - \frac{5899}{125440}\Omega_k^5 + \frac{1089}{7168}\psi_k^3\Omega_k^6\right)k^5 + \left(\frac{1539}{501760}\psi_k\Omega_k^6 + \frac{85833}{2007040}\psi_k^3\Omega_k^6 + \right. \\
&\quad \left. \frac{11469}{1254400}\Omega_k^6\right)k^4 + \left(-\frac{6579}{5017600}\Omega_k^7 + \frac{243}{1003520}\psi_k\Omega_k^7 + \frac{32967}{4014080}\psi_k^3\Omega_k^7\right)k^3 +
\end{aligned}$$

$$\begin{aligned}
&\left(\frac{4887}{40140800}\Omega_k^8+\frac{16929}{16056320}\psi_k^3\Omega_k^8+\frac{729}{80281600}\Omega_k^8\psi_k\right)k^2+\left(\frac{2673}{32112640}\psi_k^3\Omega_k^9-\right.\\
&\left.\frac{3483}{802816000}\Omega_k^9\right)k+\left(\frac{8019}{2569011200}\right)\Omega_k^{10}\psi_k^3\\
=\;&\psi_k k^{10}+\left(-\frac{1}{2}+\frac{3}{2}\psi_k\right)\Omega_k k^9+\left(\frac{11}{32}\psi_k^3-\frac{1}{4}+\frac{117}{112}\psi_k\right)\Omega_k^2k^8+\left(\frac{33}{64}\psi_k^3+\frac{99}{224}\psi_k-\right.\\
&\left.\frac{361}{1120}\right)\Omega_k^3k^7+\left(\frac{1287}{3584}\psi_k^3+\frac{7803}{62720}\psi_k+\frac{283}{2240}\right)\Omega_k^4k^6+\left(\frac{2997}{125440}\psi_k-\frac{5899}{125440}+\right.\\
&\left.\frac{1089}{7168}\psi_k^3\right)\Omega_k^5k^5+\left(\frac{1539}{501760}\psi_k+\frac{85833}{2007040}\psi_k^3+\frac{11469}{1254400}\right)\Omega_k^6k^4+\left(-\frac{6579}{5017600}+\right.\\
&\left.\frac{243}{1003520}\psi_k+\frac{32967}{4014080}\psi_k^3\right)\Omega_k^7k^3+\left(\frac{4887}{40140800}+\frac{16929}{16056320}\psi_k^3+\right.\\
&\left.\frac{729}{80281600}\psi_k\right)\Omega_k^8k^2+\left(\frac{2673}{32112640}\psi_k^3-\frac{3483}{802816000}\right)\Omega_k^9k+\left(\frac{8019}{2569011200}\right)\Omega_k^{10}\psi_k^3
\end{aligned}$$

$$\begin{aligned}
\mathrm{den}_{(3.3.56)}=k^{10}&+\frac{3}{2}\Omega_k k^9+\frac{117}{112}\Omega_k^2k^8+\frac{99}{224}\Omega_k^3k^7+\frac{7803}{62720}\Omega_k^4k^6+\frac{2997}{125440}\Omega_k^5k^5+\\
&\frac{1539}{501760}\Omega_k^6k^4+\frac{243}{1003520}\Omega_k^7k^3+\frac{729}{80281600}\Omega_k^8k^2. \qquad (3.3.57)
\end{aligned}$$

Using long division, one can express $\frac{num_{(3.3.56)}}{den_{(3.3.56)}}$ from (3.3.56) as the sum quotients

$$\frac{num_{(3.3.56)}}{den_{(3.3.56)}}=Q_1+Q_2+Q_3+Q_4+Q_5+\ldots, \qquad (3.3.58)$$

where Q_k is the quotient that results from the kth iteration of long division. We obtain a lower bound by stopping the long division process after the kth iteration when the remainder $R_k>0$ for $k\geq 5$ is produced. At the fourth iteration of long division on $\frac{num_{(3.3.56)}}{den_{(3.3.56)}}$, we have just such a remainder

$$\begin{aligned}
\frac{num_{(3.3.56)}}{den_{(3.3.56)}}&=Q_1+Q_2+Q_3+Q_4+R_{4,(3.3.59)}\\
&=\psi_k-\frac{1}{2}\frac{\Omega_k}{k}+\frac{1}{32}\left(11\psi_k^3+16\right)\frac{\Omega_k^2}{k^2}-\frac{11}{20}\frac{\Omega_k^3}{k^3}+R_{4,(3.3.59)}, \qquad (3.3.59)
\end{aligned}$$

where the remainder is

$$R_{4,(3.3.59)}=\frac{\mathrm{num}_{R_{4,(3.3.59)}}}{\mathrm{den}_{(3.3.56)}}$$

with

$$
\begin{aligned}
\mathrm{num}_{R_{4,(3.3.59)}} &= \frac{13}{20}\Omega_k^4 k^6 + \frac{59}{160}\Omega_k^5 k^5 + \frac{1131}{5600}\Omega_k^6 k^4 + \frac{10161}{179200}\Omega_k^7 k^3 + \frac{475659}{40140800}\Omega_k^8 k^2 + \\
&\quad \frac{628641}{401408000}\Omega_k^9 k + \frac{4131}{32112640}\Omega_k^{10} + \frac{8019}{1605632000}\frac{\Omega_k^{11}}{k} \\
&> 0
\end{aligned}
$$

for $k \geq 5$. Since $\mathrm{den}_{(3.3.56)}$ in (3.3.57) is positive when $k \geq 5$, if we drop the remainder $R_{4,(3.3.59)}$ from (3.3.59), we obtain a lower bound to $F(s_k) + 1$

$$
\frac{3}{2k} \times \Omega_k \times \left[\psi_k - \frac{1}{2}\frac{\Omega_k}{k} + \frac{1}{32}\left(11\psi_k^3 + 16\right)\frac{\Omega_k^2}{k^2} - \frac{11}{20}\frac{\Omega_k^3}{k^3}\right]. \tag{3.3.60}
$$

Substituting $\psi_{k,LB}$ from (3.3.53) for ψ in (3.3.60), we have

$$
\begin{aligned}
F(s_k) + 1 &\geq \frac{3}{2k} \times \Omega_k \times \left[\psi_{k,LB} - \frac{1}{2}\frac{\Omega_k}{k} + \frac{1}{32}\left(11\psi_{k,LB}^3 + 16\right)\frac{\Omega_k^2}{k^2} - \frac{11}{20}\frac{\Omega_k^3}{k^3}\right] \\
&= \frac{3}{2k} \times \Omega_k \times \left[\left(1 - \frac{1}{2}\frac{\Omega_k}{k} + \frac{5}{32}\frac{\Omega_k^2}{k^2} - \frac{3}{80}\frac{\Omega_k^3}{k^3}\right) - \frac{1}{2}\frac{\Omega_k}{k} + \frac{1}{32}\left(11\psi_k^3 + 16\right)\frac{\Omega_k^2}{k^2} - \frac{11}{20}\frac{\Omega_k^3}{k^3}\right] \\
&= \frac{3}{2k} \times \Omega_k \times \left[1 - \frac{\Omega_k}{k} + \frac{\Omega_k^2}{k^2} - \frac{353}{320}\frac{\Omega_k^3}{k^3} + \frac{429}{1024}\frac{\Omega_k^4}{k^4} - \frac{1243}{5120}\frac{\Omega_k^5}{k^5} + \frac{17061}{163840}\frac{\Omega_k^6}{k^6} - \right. \\
&\quad \left. \frac{11253}{327680}\frac{\Omega_k^7}{k^7} + \frac{230791}{26214400}\frac{\Omega_k^8}{k^8} - \frac{21879}{13107200}\frac{\Omega_k^9}{k^9} + \frac{297}{1310720}\frac{\Omega_k^{10}}{k^{10}} - \frac{297}{16384000}\frac{\Omega_k^{11}}{k^{11}}\right] \\
&\triangleq \left(F(s_k) + 1\right)_{LB,\Omega_k}.
\end{aligned} \tag{3.3.61}
$$

To obtain the final form of our lower bound to $F(s_k) + 1$, we must substitute the upper and lower bounds to Ω_k (appropriately) from (3.3.49) and (3.3.50) for Ω in (3.3.61), ignoring the Ω_k that has been factored out of the term in square brackets

$$
\begin{aligned}
\left(F(s_k) + 1\right)_{LB} &\triangleq \frac{3}{2k} \times \Omega_k \times \left[1 - \frac{\Omega_{k,UB}}{k} + \frac{\Omega_{k,LB}^2}{k^2} - \frac{353}{320}\frac{\Omega_{k,UB}^3}{k^3} + \frac{429}{1024}\frac{\Omega_{k,LB}^4}{k^4} - \right. \\
&\quad \frac{1243}{5120}\frac{\Omega_{k,UB}^5}{k^5} + \frac{17061}{163840}\frac{\Omega_{k,LB}^6}{k^6} - \frac{11253}{327680}\frac{\Omega_{k,UB}^7}{k^7} + \frac{230791}{26214400}\frac{\Omega_{k,LB}^8}{k^8} - \\
&\quad \left. \frac{21879}{13107200}\frac{\Omega_{k,UB}^9}{k^9} + \frac{297}{1310720}\frac{\Omega_{k,LB}^{10}}{k^{10}} - \frac{297}{16384000}\frac{\Omega_{k,UB}^{11}}{k^{11}}\right] \\
&= \frac{3}{2k} \times \Omega_k \times \left[1 - \frac{1}{k} + \frac{2}{k^2} - \frac{2273}{320k^3} + \frac{27537}{5120k^4} - \frac{110959}{5120k^5} + \right.
\end{aligned}
$$

$$\begin{aligned}
&\frac{6281797}{163840k^{6}}-\frac{31651633}{327680k^{7}}+\frac{3466650711}{26214400k^{8}}-\frac{4147238043}{13107200k^{9}}+\\
&\frac{1271653207}{2621440k^{10}}-\frac{16323232967}{16384000k^{11}}+\frac{85301099683}{65536000k^{12}}-\frac{60308757001}{32768000k^{13}}+\\
&\frac{5001219437}{3276800k^{14}}-\frac{88627207}{51200k^{15}}+\frac{174792316787}{131072000k^{16}}-\frac{128664779331}{65536000k^{17}}+\\
&\frac{147921825909}{65536000k^{18}}-\frac{1721884461}{655360k^{19}}+\frac{16140878259}{6553600k^{20}}-\frac{45267378783}{16384000k^{21}}+\\
&\frac{45132288993}{16384000k^{22}}-\frac{12881685663}{4096000k^{23}}+\frac{598051971}{204800k^{24}}-\frac{26108973}{10240k^{25}}+\\
&\frac{56723733}{32000k^{26}}-\frac{11214027}{8000k^{27}}+\frac{2028807}{2000k^{28}}-\frac{94743}{100k^{29}}+\frac{35937}{50k^{30}}-\frac{58806}{125k^{31}}+\\
&\left.\frac{26136}{125k^{32}}-\frac{9504}{125k^{33}}\right].
\end{aligned}\tag{3.3.62}$$

Finally, using the lower bound $\left(F\left(s_k\right)+1\right)_{LB}$ in (3.3.62) and the upper bound $s_{k+1,UB}$ in (3.3.46), we have

$$F\left(s_k\right)+1-s_{k+1}\geq\left(F\left(s_k\right)+1\right)_{LB}-s_{k+1,UB}$$

where the right hand side equals

$$\begin{aligned}
\frac{3}{2k}\times\Omega_k\times\Bigg[&\frac{607}{320k^{3}}+\frac{26897}{5120k^{4}}-\frac{229359}{5120k^{5}}-\frac{15478203}{163840k^{6}}+\frac{19589327}{327680k^{7}}-\frac{16969112489}{26214400k^{8}}+\\
&\frac{8799398757}{13107200k^{9}}-\frac{9926810793}{2621440k^{10}}+\frac{179471711033}{16384000k^{11}}-\frac{914847892317}{65536000k^{12}}+\\
&\frac{1348608746999}{32768000k^{13}}-\frac{63967638163}{3276800k^{14}}+\frac{6934547193}{51200k^{15}}-\frac{23952572803213}{131072000k^{16}}-\\
&\frac{11489535179331}{65536000k^{17}}-\frac{35139630174091}{65536000k^{18}}-\frac{1721884461}{655360k^{19}}+\frac{16140878259}{6553600k^{20}}-
\end{aligned}$$

$$\frac{45267378783}{16384000k^{21}}+\frac{45132288993}{16384000k^{22}}-\frac{12881685663}{4096000k^{23}}+\frac{598051971}{204800k^{24}}-\frac{26108973}{10240k^{25}}+$$
$$\frac{56723733}{32000k^{26}}-\frac{11214027}{8000k^{27}}+\frac{2028807}{2000k^{28}}-\frac{94743}{100k^{29}}+\frac{35937}{50k^{30}}-\frac{58806}{125k^{31}}+\frac{26136}{125k^{32}}-$$
$$\left.\frac{9504}{125k^{33}}\right]=\frac{\text{num}}{k^{33}}$$

and

$$\begin{aligned}\text{num} = {} & \frac{607}{320}k^{30}+\frac{26897}{5120}k^{29}-\frac{229359}{5120}k^{28}-\frac{15478203}{163840}k^{27}+\frac{19589327}{327680}k^{26}- \\ & \frac{16969112489}{26214400}k^{25}+\frac{8799398757}{13107200}k^{24}-\frac{9926810793}{2621440}k^{23}+\frac{179471711033}{16384000}k^{22}- \\ & \frac{914847892317}{65536000}k^{21}+\frac{1348608746999}{32768000}k^{20}-\frac{63967638163}{3276800}k^{19}+\frac{6934547193}{51200}k^{18}- \\ & \frac{23952572803213}{131072000}k^{17}-\frac{11489535179331}{65536000}k^{16}-\frac{35139630174091}{65536000}k^{15}- \\ & \frac{1721884461}{655360}k^{14}+\frac{16140878259}{6553600}k^{13}-\frac{45267378783}{16384000}k^{12}+\frac{45132288993}{16384000}k^{11}- \\ & \frac{12881685663}{4096000}k^{10}\frac{598051971}{204800}k^{9}-\frac{26108973}{10240}k^{8}+\frac{56723733}{32000}k^{7}-\frac{11214027}{8000}k^{6}+ \\ & \frac{2028807}{2000}k^{5}\frac{94743}{100}k^{4}+\frac{35937}{50}k^{3}-\frac{58806}{125}k^{2}+\frac{26136}{125}k-\frac{9504}{125}.\end{aligned} \tag{3.3.63}$$

Using Maple, we have discovered that num in (3.3.63) has only two real roots which are approximately equal to $-.6772223804$ and 4.740419301, and thus, we conclude that

$$F\left(s_k\right)+1-s_{k+1}>0$$

when $k\geq 5$.

We have now shown that for $k\geq 9$,[8] the function F defined in (3.3.42) satisfies Property c) from Lemma III.8. Since F also satisfies Properties a) and b) in Lemma III.8, and since for $k=1,2,\ldots,8$, $s_k=\underline{\eta}_k$, Lemma III.8 tells us that $s_k\leq\underline{\eta}_k$ for all $k\geq 1$. This conclusion completes Step 1 of the proof of the lower bound to $t_1^{(N)}$ in Theorem III.1.

[8] Actually, we have shown it to be true for $k\geq 5$.

Execution of Step 2. To show (3.3.30) is true, we evaluate the sum in (3.3.29) using the definition of s_k in (3.3.31) when $N > 9$:

$$\begin{aligned}
s_1 + s_N + 2\sum_{k=2}^{N-1} s_k &= s_1 + s_N + 2\sum_{m=2}^{8} s_m + 2\sum_{k=9}^{N-1} \frac{3}{2k}\Omega_k \\
&= \underline{\eta}_1 + s_N + 2\sum_{m=2}^{8} \underline{\eta}_m + 2\sum_{k=9}^{N-1} \frac{3}{2k}\Omega_k \\
&= 1 + s_N + 2\sum_{m=2}^{8} \underline{\eta}_m + 2\sum_{k=9}^{N-1} \frac{3}{2k}\Omega_k. \qquad (3.3.64)
\end{aligned}$$

Using the binomial expansion given in (3.3.43) with $\alpha = \left(\frac{2}{i} - \frac{8}{i^2}\right)$, we can express Ω_k as

$$\begin{aligned}
\Omega_k &= \left[1 - \left(\frac{2}{k} - \frac{8}{k^2}\right)\right]^{\frac{1}{2}} \\
&= 1 - \frac{1}{2}\left(\frac{2}{k} - \frac{8}{k^2}\right) - \frac{1}{8}\left(\frac{2}{k} - \frac{8}{k^2}\right)^2 - \frac{1}{16}\left(\frac{2}{k} - \frac{8}{k^2}\right)^3 - \frac{5}{128}\left(\frac{2}{k} - \frac{8}{k^2}\right)^4 - \frac{7}{256}\left(\frac{2}{k} - \frac{8}{k^2}\right)^5 - \ldots \\
&= 1 - \left(\frac{1}{k} - \frac{4}{k^2}\right) - \frac{1}{2}\left(\frac{1}{k^2} - \frac{8}{k^3} + \frac{16}{k^4}\right) - \frac{1}{16}\left(\frac{2}{k} - \frac{8}{k^2}\right)^3 - \frac{5}{128}\left(\frac{2}{k} - \frac{8}{k^2}\right)^4 - \frac{7}{256}\left(\frac{2}{k} - \frac{8}{k^2}\right)^5 - \ldots \\
&= 1 - \frac{1}{k} + \frac{4}{k^2} + \frac{1}{k^2}\left[-\frac{1}{2} + \frac{4}{k} - \frac{8}{k^2} - \frac{k^2}{16}\left(\frac{2}{k} - \frac{8}{k^2}\right)^3 - \frac{5k^2}{128}\left(\frac{2}{k} - \frac{8}{k^2}\right)^4 - \frac{7k^2}{256}\left(\frac{2}{k} - \frac{8}{k^2}\right)^5 - \ldots\right] \\
&= 1 - \frac{1}{k} + \frac{7}{2}\frac{1}{k^2} + \frac{1}{k^2}\left[\frac{4}{k} - \frac{8}{k^2} - \frac{k^2}{16}\left(\frac{2}{k} - \frac{8}{k^2}\right)^3 - \frac{5k^2}{128}\left(\frac{2}{k} - \frac{8}{k^2}\right)^4 - \frac{7k^2}{256}\left(\frac{2}{k} - \frac{8}{k^2}\right)^5 - \ldots\right] \\
&= 1 - \frac{1}{k} + \frac{7}{2}\frac{1}{k^2} + \epsilon_k,
\end{aligned}$$

where

$$\begin{aligned}
\epsilon_k &\triangleq \frac{1}{k^2}\left[\frac{4}{k} - \frac{8}{k^2} - \frac{k^2}{16}\left(\frac{2}{k} - \frac{8}{k^2}\right)^3 - \frac{5k^2}{128}\left(\frac{2}{k} - \frac{8}{k^2}\right)^4 - \frac{7k^2}{256}\left(\frac{2}{k} - \frac{8}{k^2}\right)^5 - \ldots\right] \\
&= \frac{1}{k^2}\left[\frac{4}{k} - \frac{8}{k^2} + \left(\Omega_k - \left[1 - \frac{1}{2}\left(\frac{2}{k} - \frac{8}{k^2}\right) - \frac{1}{8}\left(\frac{2}{k} - \frac{8}{k^2}\right)^2\right]\right)\right] \\
&= \frac{1}{2k^2}\left[2\left(\Omega_k - 1\right)k^2 + 2k - 7\right] \qquad (3.3.65)
\end{aligned}$$

is $o\left(k^{-2}\right)$. Using Maple, we find that the numerator of ϵ_k in (3.3.65) has three roots: $-.9151112632, .9004449069, 1.955242680$. Since the leading coefficient of the dominant term of this numerator is positive (because $\Omega_k > 1$ for $k \geq 4$), we know that

$\epsilon_k > 0$ when $k \geq 5$ since for these values of k is greater than the largest positive root of the numerator.

Then using (3.3.65) in (3.3.64),

$$\begin{aligned}
&1 + s_N + 2\sum_{m=2}^{8} \underline{\eta}_m + 2\sum_{k=9}^{N-1} \frac{3}{2k}\Omega_k \\
&= 1 + s_N + 2\sum_{m=2}^{8} \underline{\eta}_m + 3\sum_{k=9}^{N-1} \frac{1}{k}\left(1 - \frac{1}{k} + \frac{7}{2}\frac{1}{k^2} + \epsilon_k\right) \\
&= 1 + s_N + 2\sum_{m=2}^{8} \underline{\eta}_m + 3\sum_{k=9}^{N-1} \frac{1}{k} - \frac{1}{k^2} + \frac{7}{2}\frac{1}{k^3} + \frac{\epsilon_k}{k} \\
&= 1 + s_N + 2\sum_{m=2}^{8} \underline{\eta}_m + 3\left[\sum_{k=9}^{N-1} \frac{1}{k} - \sum_{k=9}^{N-1} \frac{1}{k^2} + \sum_{k=9}^{N-1} \frac{7}{2}\frac{1}{k^3} + \frac{\epsilon_k}{k}\right] \\
&= 1 + s_N + 2\sum_{m=2}^{8} \underline{\eta}_m + 3\left[\sum_{k=9}^{N-1} \frac{1}{k} - \sum_{k=9}^{N-1} \frac{1}{k^2} + \sum_{k=9}^{N-1} \frac{7}{2}\frac{1}{k^3} + \sum_{k=9}^{N-1} \frac{\epsilon_k}{k}\right]. \qquad (3.3.66)
\end{aligned}$$

Since Riemann upper and lower sums provide bounds to Riemann integrals, we have the following inequalities

$$\begin{aligned}
\sum_{i=9}^{N-1} \frac{1}{i} &\geq \int_{9}^{N} \frac{1}{x}dx = \log x\Big|_{9}^{N} = \log N - \log 9 \\
\sum_{i=9}^{N-1} \frac{1}{i^2} &\leq \int_{8}^{N-1} \frac{1}{x^2}dx = -\frac{1}{x}\Big|_{8}^{N-1} = \frac{1}{8} - \frac{1}{N-1} \\
\sum_{i=9}^{N-1} \frac{1}{i^3} &\geq \int_{9}^{N} \frac{1}{x^3}dx = -\frac{1}{2x^2}\Big|_{9}^{N} = \frac{1}{2\cdot 9^2} - \frac{1}{2N^2} = \frac{1}{162} - \frac{1}{2N^2},
\end{aligned}$$

and using these relationships in (3.3.66), we have

$$\begin{aligned}
(3.3.66)_{RHS} &\geq 1 + s_N + 2\sum_{m=2}^{8} \underline{\eta}_m + 3\left[\log N - \log 9 - \frac{1}{8} + \frac{1}{N-1} + \frac{7}{2}\left(\frac{1}{162} - \frac{1}{2N^2}\right)\right] + 3\sum_{k=9}^{N-1}\frac{\epsilon_k}{k} \\
&= 1 + s_N + 2\sum_{m=2}^{9} \underline{\eta}_m + 3\left[\log N + \frac{1}{N-1} - \frac{7}{4}\left(\frac{1}{N^2}\right)\right] + 3\left[\frac{7}{324} - \log 9 - \frac{1}{8}\right] + 3\sum_{k=9}^{N-1}\frac{\epsilon_k}{k}
\end{aligned}$$

$$\begin{aligned}
&= 3\left[\log N + \frac{1}{N-1} - \frac{7}{4}\frac{1}{N^2}\right] + \left[1 + s_N + 2\sum_{m=2}^{8}\underline{\eta}_m - 3\log 9 - \frac{3\cdot 67}{648}\right] + 3\sum_{k=9}^{N-1}\frac{\epsilon_k}{k} \\
&= \left[3\log N + \frac{3}{N-1} - \frac{7}{4}\frac{3}{N^2} + s_N\right] + \left[1 + 2\sum_{m=2}^{8}\underline{\eta}_m - 3\log 9 - \frac{67}{216}\right] + 3\sum_{k=9}^{N-1}\frac{\epsilon_k}{k} \\
&= \left[3\log N + \frac{3}{N-1} - \frac{21}{4}\frac{1}{N^2} + \frac{3}{2N}\left(1 - \frac{2}{N} + \frac{8}{N^2}\right)^{\frac{1}{2}}\right] + \left[1 + 2\sum_{m=2}^{8}\underline{\eta}_m - \right. \\
&\quad \left. 3\log 9 - \frac{67}{216}\right] + 3\sum_{k=9}^{N-1}\frac{\epsilon_k}{k} \\
&= \left[3\log N + \delta(N)\right] + \left[1 + 2\sum_{m=2}^{8}\underline{\eta}_m - 3\log 9 - \frac{67}{216}\right] + 3\sum_{k=9}^{N-1}\frac{\epsilon_k}{k}
\end{aligned}$$

$$\geq \left[3\log N + \delta(N)\right] + \left[1 + 2\sum_{m=2}^{8}\underline{\eta}_m - 3\log 9 - \frac{67}{216}\right] \tag{3.3.67}$$

$$> 3\log N + \delta\left(N\right) - 1.46004 \tag{3.3.68}$$

where

$$1 + 2\sum_{m=2}^{8}\underline{\eta}_m - 3\log 9 - \frac{67}{216} \approx -1.46003439868863$$

and

$$\delta\left(N\right) = \frac{3}{N-1} - \frac{21}{4}\frac{1}{N^2} + \frac{3}{2N}\left(1 - \frac{2}{N} + \frac{8}{N^2}\right)^{\frac{1}{2}}$$

is as defined in (3.2.3). Also, we note that in (3.3.67), we have dropped the term $3\sum_{k=9}^{N-1}\frac{\epsilon_k}{k}$ because it is strictly positive when $k \geq 5$.

Since (3.3.68) is the lower bound stated in Theorem III.1, this concludes **Step 2**.

Since, as already stated, (3.3.27) follows directly from the relationships obtained in **Step 1** and **Step 2**, we have completed the proof of Theorem III.1. ■

3.3.3 The Proof of Corollary III.2.

Now we prove that the rate at which $N^2 D\left(N\right)$ converges to $\frac{\beta}{12}\sigma^2$, which is the Panter-Dite constant scaled by the factor $\frac{\sigma^2}{12}$, is no greater than $\frac{9}{4}\sigma^2\left(\frac{2}{N} - \frac{8}{N^2}\right)$ and we remark that the extension to case where the variance of the source is given by σ^2 is not difficult since $\underline{\Delta}_N^{(N)} = \underline{\eta}_N\sigma$ for $N \geq 1$.

Proof of Corollary III.2. Again, just as in the proof of Theorem III.1, we prove this corollary for the unit variance exponential source.

To prove Corollary III.2, we use Nitadori's result and the bounds in Theorem III.1. To produce an upper bound to $D\left(N\right)$, we remember that the half steps of q_N^c were always larger than the corresponding half steps of q_N^*. Therefore, we have

$$\begin{aligned}
D\left(N\right) &= \left(\underline{\Delta}_N^{(N)}\right)^2 \quad \text{(Nitadori's distortion result)}\\
&= \left(\underline{\eta}_N\right)^2 \quad \text{(unit variance exponential source)}\\
&\leq \left(c_{N,c}\right)^2\\
&= \left(\underline{\Delta}_{N,c}^{(N)}\right)^2 \quad \text{(unit variance)}\\
&= \left[\Delta_l\left(-3\log\left(1-\frac{1}{N}\right)\right)\right]^2\\
&= \left[1-\frac{\Delta_{N,c}^{(N)}\left(1-\frac{1}{N}\right)^3}{1-\left(1-\frac{1}{N}\right)^3}\right]^2 \quad \text{(used (3.3.11))}\\
&= \left[1-\frac{-3\log\left(1-\frac{1}{N}\right)\left(1-\frac{1}{N}\right)^3}{1-\left(1-\frac{1}{N}\right)^3}\right]^2\\
&= \left[1-\frac{\left(3\left[\frac{1}{N}+\frac{1}{2}\frac{1}{N^2}+\frac{1}{3}\frac{1}{N^3}+\frac{1}{4}\frac{1}{N^4}+\ldots\right]\right)\left(1-\frac{1}{N}\right)^3}{1-\left(1-\frac{1}{N}\right)^3}\right]^2 \quad \text{(used (3.3.16))}\\
&= \left[1-\frac{\left(3\left[\frac{1}{N}+\frac{1}{2}\frac{1}{N^2}+\frac{1}{3}\frac{1}{N^3}+\frac{1}{4}\frac{1}{N^4}+\ldots\right]\right)\left(1-\frac{1}{N}\right)^3}{1-\left(1-\frac{3}{N}+\frac{3}{N^2}-\frac{1}{N^3}\right)}\right]^2\\
&= \left[\frac{\left(\frac{3}{N}-\frac{3}{N^2}+\frac{1}{N^3}\right)-\left(3\left[\frac{1}{N}+\frac{1}{2}\frac{1}{N^2}+\frac{1}{3}\frac{1}{N^3}+\frac{1}{4}\frac{1}{N^4}+\ldots\right]\right)\left(1-\frac{1}{N}\right)^3}{\frac{3}{N}-\frac{3}{N^2}+\frac{1}{N^3}}\right]^2\\
&= \left[\frac{\left(3-\frac{3}{N}+\frac{1}{N^2}\right)-\left(3\left[1+\frac{1}{2}\frac{1}{N}+\frac{1}{3}\frac{1}{N^2}+\frac{1}{4}\frac{1}{N^3}+\ldots\right]\right)\left(1-\frac{1}{N}\right)^3}{3-\frac{3}{N}+\frac{1}{N^2}}\right]^2\\
&= \left[\frac{\left(1-\frac{1}{N}+\frac{1}{3N^2}\right)-\left[1+\frac{1}{2}\frac{1}{N}+\frac{1}{3}\frac{1}{N^2}+\frac{1}{4}\frac{1}{N^3}+\ldots\right]\left(1-\frac{1}{N}\right)^3}{1-\frac{1}{N}+\frac{1}{3N^2}}\right]^2\\
&= \left[\frac{\left(1-\frac{1}{N}+\frac{1}{3N^2}\right)-\left[1+\frac{1}{2}\frac{1}{N}+\frac{1}{3}\frac{1}{N^2}+\frac{1}{4}\frac{1}{N^3}+\ldots\right]\left(1-\frac{3}{N}+\frac{3}{N^2}-\frac{1}{N^3}\right)}{1-\frac{1}{N}+\frac{1}{3N^2}}\right]^2\\
&= \frac{num}{1-\frac{1}{N}+\frac{1}{3N^2}}.
\end{aligned}$$

Concentrating on just the numerator, we have

$$
\begin{aligned}
num &= \left(1-\frac{1}{N}+\frac{1}{3N^2}\right)-\left[1+\frac{1}{2}\frac{1}{N}+\frac{1}{3}\frac{1}{N^2}+\frac{1}{4}\frac{1}{N^3}+\dots\right]\left(1-\frac{3}{N}+\frac{3}{N^2}-\frac{1}{N^3}\right)\\
&= \left(1-\frac{1}{N}+\frac{1}{3N^2}\right)-\left[1+\frac{1}{2N}+\frac{1}{3N^2}+\frac{1}{4N^3}+\dots\right]+\frac{3}{N}\left[1+\frac{1}{2N}+\frac{1}{3N^2}+\frac{1}{4N^3}+\dots\right]\\
&\quad -\frac{3}{N^2}\left[1+\frac{1}{2N}+\frac{1}{3N^2}+\frac{1}{4N^3}+\dots\right]+\frac{1}{N^3}\left[1+\frac{1}{2N}+\frac{1}{3N^2}+\frac{1}{4N^3}+\dots\right]\\
&= \left(-\frac{1}{N}+\frac{1}{3N^2}\right)-\left[\frac{1}{2N}+\frac{1}{3N^2}+\frac{1}{4N^3}+\dots\right]+\left[\frac{3}{N}+\frac{3}{2}\frac{1}{N^2}+\frac{1}{N^3}+\frac{3}{4}\frac{1}{N^4}+\dots\right]\\
&\quad -\left[\frac{3}{N^2}+\frac{3}{2}\frac{1}{N^3}+\frac{1}{N^4}+\frac{3}{4}\frac{1}{N^4}+\dots\right]+\left[\frac{1}{N^3}+\frac{1}{2}\frac{1}{N^4}+\frac{1}{3}\frac{1}{N^5}+\frac{1}{4}\frac{1}{N^6}+\dots\right]\\
&= \left(-\frac{1}{N}-\frac{1}{2N}+\frac{3}{N}\right)+\left(\frac{1}{3N^2}-\frac{1}{3N^2}+\frac{3}{2}\frac{1}{N^2}-\frac{3}{N^2}\right)+\left[-\frac{1}{4N^3}+\frac{1}{N^3}-\frac{3}{2}\frac{1}{N^3}+\frac{1}{N^3}\right]+\\
&\quad O\left(\frac{1}{N^4}\right)\\
&= \left(\frac{3}{2}\frac{1}{N}\right)+\left(-\frac{3}{2}\frac{1}{N^2}\right)+\left[-\frac{1}{4N^3}+\frac{2}{N^3}-\frac{3}{2}\frac{1}{N^3}\right]+O\left(\frac{1}{N^4}\right)\\
&= \left(\frac{3}{2}\frac{1}{N}\right)+\left(-\frac{3}{2}\frac{1}{N^2}\right)+\left[-\frac{1}{4N^3}+\frac{1}{2}\frac{1}{N^3}\right]+O\left(\frac{1}{N^4}\right)\\
&= \left(\frac{3}{2}\frac{1}{N}\right)+\left(-\frac{3}{2}\frac{1}{N^2}\right)+\left[\frac{1}{4N^3}\right]+O\left(\frac{1}{N^4}\right)\;=\;\frac{3}{2}\frac{1}{N}\left[1-\frac{1}{N}+\frac{1}{6}\frac{1}{N^2}+O\left(\frac{1}{N^3}\right)\right].
\end{aligned}
$$

Then

$$
\begin{aligned}
D\left(N\right) &\leq \left[\frac{3}{2}\frac{1}{N}\frac{1-\frac{1}{N}+\frac{1}{6}\frac{1}{N^2}+O\left(\frac{1}{N^3}\right)}{1-\frac{1}{N}+\frac{1}{3N^2}}\right]^2\\
&= \frac{9}{4}\frac{1}{N^2}\left[\frac{1-\frac{1}{N}+\frac{1}{6}\frac{1}{N^2}+O\left(\frac{1}{N^3}\right)}{1-\frac{1}{N}+\frac{1}{3}\frac{1}{N^2}}\right]^2\\
&= \frac{9}{4}\frac{1}{N^2}\left[1-\frac{1}{6}\frac{1}{N^2}+O\left(\frac{1}{N^3}\right)\right]^2\\
&= \frac{9}{4}\frac{1}{N^2}\left[1-\frac{1}{3}\frac{1}{N^2}+O\left(\frac{1}{N^3}\right)\right]. \qquad (3.3.69)
\end{aligned}
$$

Similarly for a lower bound to $D\left(N\right)$, we recall that the sequence s_k is a lower bound to the sequence of optimal half steps $\underline{\eta}_k$. Again, starting with Nitadori's result, we have

$$\begin{aligned} D\left(N\right) &= \left(\underline{\Delta}_N^{(N)}\right)^2 = \left(\underline{\eta}_N\right)^2 \\ &\geq \frac{1}{2}\left(s_N\right)^2 \\ &= \left[\frac{3}{2}\frac{1}{N}\left(1-\frac{2}{N}+\frac{8}{N^2}\right)^{\frac{1}{2}}\right]^2 \qquad \text{(from (3.3.31))} \\ &= \frac{9}{4}\frac{1}{N^2}\left(1-\frac{2}{N}+\frac{8}{N^2}\right). \end{aligned} \tag{3.3.70}$$

Noting that both bounds in (3.3.69) and (3.3.70) are less than $\frac{\beta}{12N^2} = \frac{9}{4N^2}$ (which means that $N^2 D\left(N\right)$ approaches $\frac{\beta}{12}$ from below), we have

$$\frac{9}{4}\left(\frac{1}{3}\frac{1}{N^2}+O\left(\frac{1}{N^3}\right)\right) < \frac{\beta}{12} - N^2 D\left(N\right) < \frac{9}{4}\left(\frac{2}{N}-\frac{8}{N^2}\right).$$

For our upper bound to the convergence rate to $\frac{\beta}{12}$, we choose $\frac{9}{2}\left(\frac{2}{N}-\frac{8}{N^2}\right)$, since it is the larger of the two bounds in the limit as N goes to infinity. ■

3.4 Discussion and Applications of Theorem III.1 and Its Corollaries.

This section is comprised of two discussions. The first one deals with support threshold estimation where we consider the functions given in Theorem III.1 as approximations to $t_1^{(N)}$. Following this, we look at support threshold estimation using the s_k sequence and compare the results we get against the bounds in Theorem III.1, as well as $t_1^{(N)}$. The second discussion focuses on quantizer design using the s_k sequence and on design using a variation on the s_k sequence. We compare the performance of these quantizers against the best performance achieved by optimal quantizers. Note that throughout these discussions, we will make our comments with regards to quantization of a unit variance exponential source, remarking that similar statements can be made for the quantization of an exponential source with arbitrary variance.

3.4.1 Accuracy of Bounds in Theorem III.1.

The upper and lower bounds

$$t_{1,c}^{(N)} = 3 \log N, \quad N \geq 1$$
$$t_{1,LB,Thm\ III.1}^{(N)} \triangleq 3 \log N + \delta(N) - 1.46004, \quad N > 9 \quad \text{(see (3.3.27))}$$

reported in Theorem III.1 can be used as estimators for the support threshold $t_1^{(N)}$ of optimal exponential quantizers. The suitability of these bounds as $t_1^{(N)}$ estimators is seen in how accurate they are in tracking the behavior of $t_1^{(N)}$ as the number of levels N increases. In Figure 3.2, we look at the quantities

$$t_{1,c}^{(N)} - t_1^{(N)}$$

and

$$t_{1,LB,Thm\ III.1}^{(N)} - t_1^{(N)}$$

at low numbers of levels ($N \leq 16$, the low rate case) and at high numbers of levels ($N \leq 4096$, the high rate case). From examination of the data shown, it appears that both $t_{1,c}^{(N)}$ and $t_{1,LB,Thm\ III.1}^{(N)}$ are able to track the rate of growth of $t_1^{(N)}$ as a function of the number of levels N since it appears that $t_{1,c}^{(N)} - t_1^{(N)}$ and $t_{1,LB,Thm\ III.1}^{(N)} - t_1^{(N)}$ are converging monotonically to constant values. It is also apparent that $t_{1,LB,Thm\ III.1}^{(N)}$ represents a better approximation to $t_1^{(N)}$ than $t_{1,c}^{(N)}$ when $N \geq 9$ since $\left|t_{1,LB,Thm\ III.1}^{(N)} - t_1^{(N)}\right| < 0.3$ as opposed to $\left|t_{1,c}^{(N)} - t_1^{(N)}\right| > 0.5$ when $N \geq 9$. We also remark that when $N \leq 8$, $t_{1,LB,Thm\ III.1}^{(N)}$ is greater than $t_1^{(N)}$ and thus $t_{1,LB,Thm\ III.1}^{(N)}$ is only a lower bound to $t_1^{(N)}$ when $N \geq 9$.

3.4.2 Tighter Support Threshold Estimation.

Since the sequence s_k is a lower bound to the Nitadori sequence $\underline{\eta}_k$ and because s_k, in contrast to $\underline{\eta}_k$, can be expressed in closed-form, it is natural to use s_k to estimate parameters that are used to design MMSE exponential quantizers. As a first application, we use s_k to estimate the key parameter or support threshold $t_1^{(N)}$ of an N-level optimal quantizer, which is an important initializing value used in the Lloyd-Max algorithm for the design of optimal quantizers [13], [15]. To construct an estimate for $t_1^{(N)}$, recall that $t_1^{(N)}$ is equal to a sum of Nitadori sequence terms

$$t_1^{(N)} = \underline{\eta}_1 + \underline{\eta}_N + 2\sum_{k=2}^{N-1} \underline{\eta}_k. \tag{3.4.71}$$

To create our support threshold estimate $t_{1,s}^{(N)}$, we replace each $\underline{\eta}_k$ for s_k for in (3.4.71) to get

$$\begin{aligned} t_{1,s}^{(N)} &\triangleq s_1 + s_N + 2\sum_{k=2}^{N-1} s_k \\ &= \begin{cases} t_1^{(N)} & \text{for } N \leq 8 \\ \begin{aligned} & t_1^{(8)} + \underline{\eta}_8 + s_N + 2\sum_{i=9}^{N-1} s_i \\ & \quad = t_1^{(8)} + \underline{\eta}_8 + \tfrac{3}{2N}\left(1 - \tfrac{2}{N} + \tfrac{8}{N^2}\right)^{\frac{1}{2}} + 2\textstyle\sum_{i=9}^{N-1} \tfrac{3}{2i}\left(1 - \tfrac{2}{i} + \tfrac{8}{i^2}\right)^{\frac{1}{2}} \end{aligned} & \text{for } N \geq 9 \end{cases} \end{aligned} \tag{3.4.72}$$

which is a lower bound to $t_1^{(N)}$ since $s_k \leq \underline{\eta}_k$ for all k.

For the unit variance exponential source, Figure 3.2 shows that the lower bound $t_{1,s}^{(N)}$ is a rather close approximation to $t_1^{(N)}$ since the absolute difference between $t_1^{(N)}$ and $t_{1,s}^{(N)}$ is less than 0.1, at least for all values of $N \leq 4096$. Thus for $N \leq 4096$, we have

$$t_1^{(N)} \geq t_{1,s}^{(N)} \geq t_1^{(N)} - 0.1$$

or equivalently,

$$t_{1,s}^{(N)} \leq t_1^{(N)} \leq t_{1,s}^{(N)} + 0.1. \tag{3.4.73}$$

Recall from Figure 3.2 that the absolute difference between $t_1^{(N)}$ and the lower bound $t_{1,LB,Thm\ III.1}^{(N)}$ given in Theorem III.1 is less than 0.15. We remark that while $t_{1,s}^{(N)}$ seems to be a better lower bound than $t_{1,LB,Thm\ III.1}^{(N)}$, $t_{1,LB,Thm\ III.1}^{(N)}$ can be expressed as a function of N in closed form and this is in contrast to $t_{1,s}^{(N)}$ which does not have a closed form expression in N but can be expressed as a the sum in (3.4.72). We remind ourselves that the ability to express the bound $t_{1,s}^{(N)}$ as the sum in (3.4.72) is (when $N \geq 9$), however, an improvement over the expression for $t_1^{(N)}$ in (3.4.71) which does not have a closed form expression even in terms of the individual $\underline{\eta}_k$ terms since $\underline{\eta}_k$ cannot be expressed in closed form.

Also recall from Figure 3.2 that the absolute difference between $t_1^{(N)}$ and the corresponding support threshold for the UTCC quantizer $t_{1,c}^{(N)} = 3\log N$ (see (3.3.8)) is less than 0.78 and that for $N = 128, 129, \ldots, 4096$, the difference appears to remain constant. Thus based on what has been seen, we remark that the upper bound $t_{1,c}^{(N)}$ to $t_1^{(N)}$ seems able to track the growth rate of $t_1^{(N)}$, but it is not nearly as tight a $t_1^{(N)}$ estimator as the lower bounds $t_{1,s}^{(N)}$ and $t_{1,LB,Thm\ III.1}^{(N)}$.

Overall, it appears that $t_{1,s}^{(N)}$ is the best approximation of $t_1^{(N)}$ of the estimates shown in Figure 3.2. Summarizing these observations, for $N \leq 4096$, we observe that

$$t_1^{(N)} \geq t_{1,LB,Thm\ III.1}^{(N)} \geq t_1^{(N)} - 0.2 \tag{3.4.74}$$

or

$$3\log N + \delta(N) - 1.46004 \leq t_1^{(N)} \leq 3\log N + \delta(N) - 1.26004$$

and

$$t_1^{(N)} \leq t_{1,c}^{(N)} = 3\log N \leq t_1^{(N)} + 0.8$$

or

$$3\log N \geq t_1^{(N)} \geq 3\log N - 0.8,$$

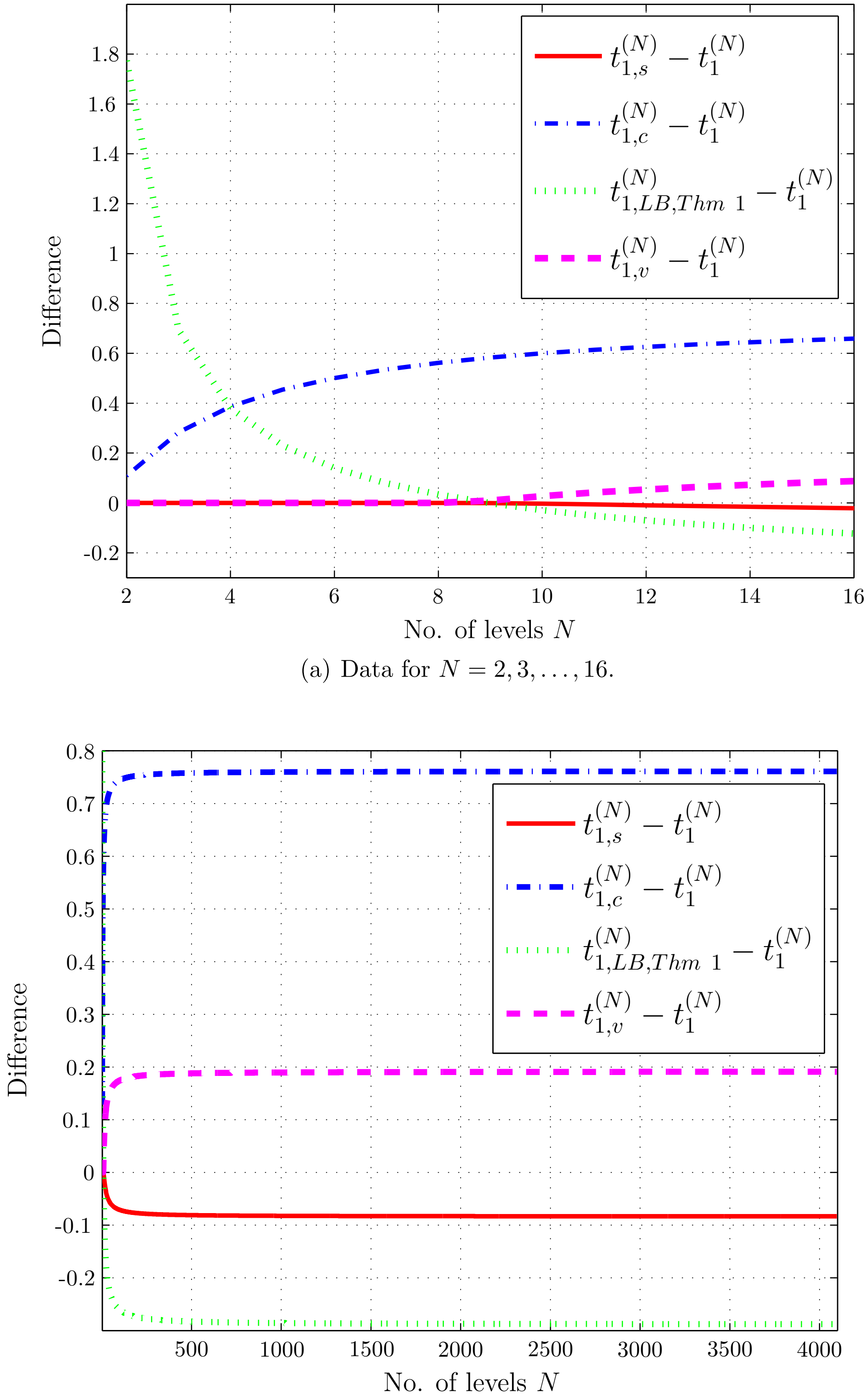
and

$$t_{1,s}^{(N)} \leq t_{1,c}^{(N)} - 0.9 = 3\log N - 0.9. \tag{3.4.75}$$

Combining the observations we have made from Figure 3.2 and using (3.4.73), (3.4.74),(3.4.75), we observe the following bounds to $t_{1,s}^{(N)}$ when $N \leq 4096$,

$$3\log N + \delta(N) - 1.36004 = t_{1,LB,Thm\ III.1}^{(N)} + 0.2 - 0.1 \leq t_1^{(N)} - 0.1 \leq t_{1,s}^{(N)} \leq 3\log N - 0.9.$$

As an additional remark, since $t_{1,s}^{(N)}$ appears to be a superior approximation to $t_1^{(N)}$ than is $t_{1,LB,Thm\ III.1}^{(N)}$, it may be possible to find a tighter theoretical lower bound to $t_1^{(N)}$ than the one reported in Theorem III.1. Indeed, the possibility of finding a tighter upper bound to $t_1^{(N)}$ also exists.



(a) Data for $N = 2, 3, \ldots, 16$.

(b) Data for $N = 2, 3, \ldots, 4096$.

Figure 3.2: Gauging the accuracy of the estimators $t_{1,s}^{(N)}, t_{1,v}^{(N)}$ against $t_1^{(N)}$. Also shown are the $t_1^{(N)}$ bounds from Theorem III.1, where $t_{1,c}^{(N)}$ is the upper bound.

3.4.3 Quantizer Design Using Half Step Sequences.

Recall from Chapter II that while quantizer design is typically expressed in terms of a set of quantization thresholds and reconstruction levels, optimal quantizer design can also be expressed succinctly as a single set of (lower) half step lengths. Indeed, for suboptimal quantizers that have been designed under the nearest neighbor constraint, this half step parameterization provides a complete and compact description of such quantizers. Motivated by this observation, in this section, we examine quantizer design using half steps under the nearest neighbor constraint using two different half step sequences.

3.4.3.1 Quantizer design using s_k.

Since the s_k sequence can be thought of as an approximation to the Nitadori sequence $\underline{\eta}_k$, it would be of interest to design suboptimal quantizers using s_k and then to measure the MSE performance as a function of the number of levels N in order to see how close to optimal these quantizers are. Since there is a closed form expression for s_k, this application of the s_k sequence provides a simpler, more practical way of designing quantizers with known performance.

An easy way to use s_k to design quantizers is to follow the method used to design optimal quantizers from the Nitadori sequence $\underline{\eta}_k$. First, we fix the number of levels N. Then we take the first N values of the s_k sequence, $s_1, s_2, \ldots, s_N$ and we assign the half steps of the quantizer to the values of s_k as $\underline{\Delta}_k = s_k$, for $k = 1, 2, \ldots, N$. The quantizer's quantization thresholds t_k and quantization levels l_k are determined as follows:

1. $t_{k,s}^{(N)} = \sum_{i=k+1}^{N} \Delta_{i,s}^{(N)} = \sum_{i=k+1}^{N} \underline{\Delta}_{i,s}^{(N)} + \underline{\Delta}_{i-1,x}^{(N)} = \sum_{i=k+1}^{N} s_i + s_{i-1}$, for $k = 1, 2, \ldots, N$ with $t_{N,s}^{(N)} \triangleq 0$ and $t_{0,s}^{(N)} \triangleq +\infty$.

2. $l_{k,s}^{(N)} \triangleq t_{k,s}^{(N)} + \underline{\Delta}_{k,s}^{(N)} = t_{k,s}^{(N)} + s_k$, for $k = 1, 2, \ldots, N$.

Using the general expression (2.2.1) from Chapter II, the MSE performance of such a quantizer with N levels is

$$D_{s_k}(N) \triangleq \sum_{k=1}^{N} \int\limits_{t_{k,s}^{(N)}}^{t_{k,s}^{(N)} + \Delta_{k,s}^{(N)}} \left(x - l_{k,s}^{(N)} \right)^2 e^{-x} dx. \tag{3.4.76}$$

Remarks on this particular method of quantizer design. As a consequence of the design algorithm described, since we have set the quantizer's half steps to values of the s_k sequence, we have produced a quantizer whose thresholds satisfy the nearest neighbor requirement of the optimality conditions. The centroid condition, however, has, in general, not been met by the quantizer's reconstruction levels. But even in spite of a lack of adherence to the centroid condition, quantizers produced in this manner, using s_k in place of $\underline{\eta}_k$, appear to have very good MSE performance. Evidence of this is seen in Figure 3.3 Row a where it also seems that such quantizers may be asymptotically optimal since the data appears to show that $\frac{D_{s_k}(N)}{D(N)}$ converge to 1 as N increases. Moreover, this convergence behavior of the ratio $\frac{D(N)}{\frac{\beta}{12}}$ can been seen even for small values of N, say $N \leq 128$ (see Figure 3.3 plot $(a-i)$) where the maximum value of this ratio (which occurs at around $N = 28$) is within 0.05% of 1.

Examining the performance of an N-level quantizer designed using s_k more closely by breaking down the expression for the MSE from (3.4.76), we have

$$D_{s_k}(N) \overset{IBP}{=} \sum_{k=1}^{N} -\left(x - l_{k,s}^{(N)}\right)^2 e^{-x}\Big|_{t_{k,s}^{(N)}}^{t_{k,s}^{(N)}+\Delta_{k,s}^{(N)}} + \int_{t_{k,s}^{(N)}}^{t_{k,s}^{(N)}+\Delta_{k,s}^{(N)}} 2\left(x - l_{k,s}^{(N)}\right) e^{-x} dx \tag{3.4.77}$$

$$= \rho_{s_k}(N) + d_{s_k}(N), \tag{3.4.78}$$

where we have used IBP

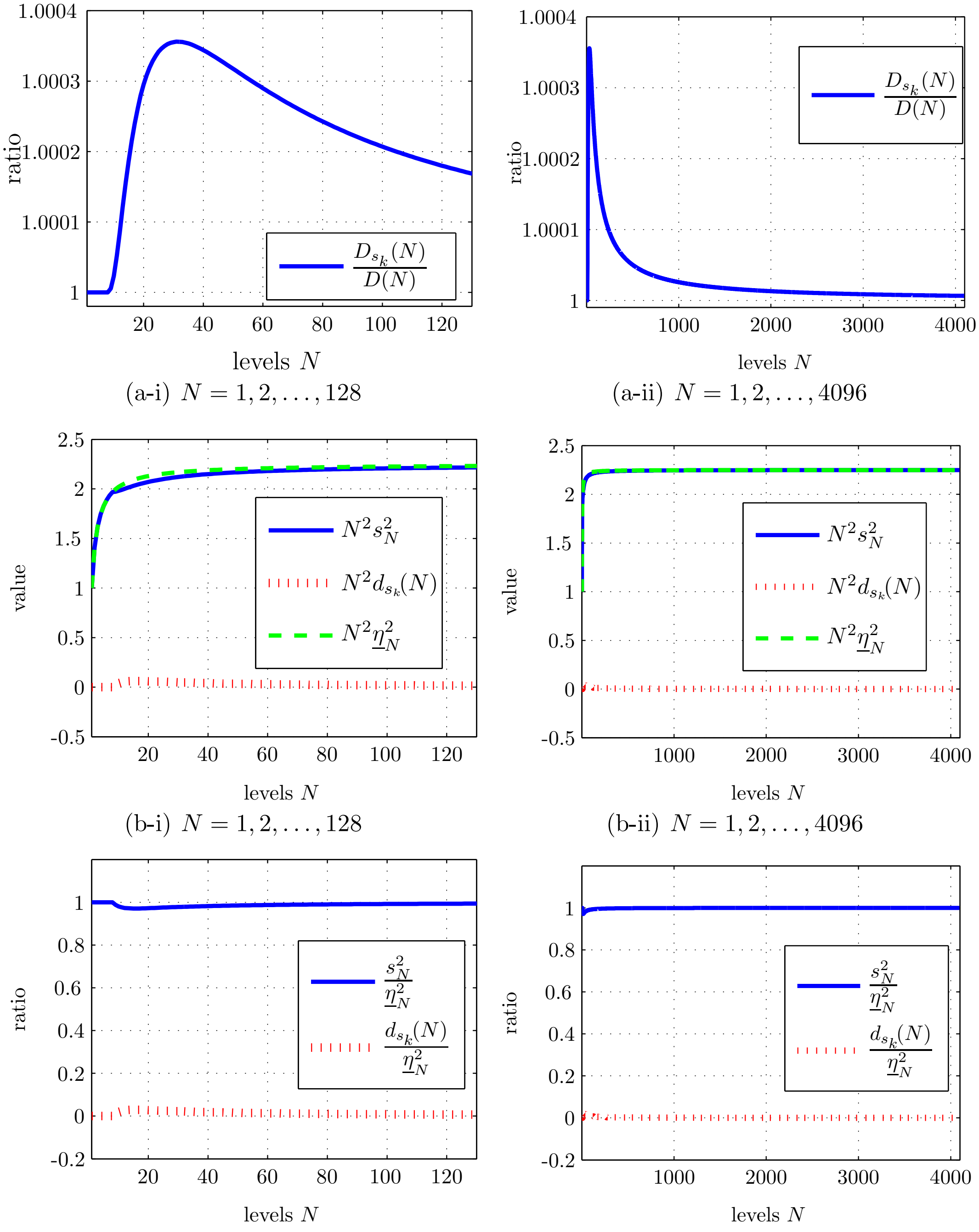
$$u = \left(x - l_{k,s}^{(N)}\right)^2 \qquad du = 2\left(x - l_{k,s}^{(N)}\right) dx$$

$$v = -e^{-x} \qquad dv = e^{-x} dx$$

to obtain (3.4.77) and we have defined

$$\rho_{s_k}(N) \triangleq \sum_{k=1}^{N} -\left(x - l_{k,s}^{(N)}\right)^2 e^{-x}\Big|_{t_{k,s}^{(N)}}^{t_{k,s}^{(N)}+\Delta_{k,s}^{(N)}}$$

$$d_{s_k}(N) \triangleq \sum_{k=1}^{N} \int_{t_{k,s}^{(N)}}^{t_{k,s}^{(N)}+\Delta_{k,s}^{(N)}} 2\left(x - l_{k,s}^{(N)}\right) e^{-x} dx.$$



(a-i) $N = 1, 2, \dots, 128$

(a-ii) $N = 1, 2, \dots, 4096$

(b-i) $N = 1, 2, \dots, 128$

(b-ii) $N = 1, 2, \dots, 4096$

(c-i) $N = 1, 2, \dots, 128$

(c-ii) $N = 1, 2, \dots, 4096$

Figure 3.3: The MSE performance of quantizers q_{s_k} designed using s_k compared against the performance of optimal quantizers. The data shown in plots (*-i) are for $N \leq 128$ and highlights the performance for quantizers designed for low levels, while the data shown in plots (*-ii) also depict asymptotic performance ($N \leq 4096$).

Evaluating just the first term in (3.4.78),

$$\begin{aligned}
\rho_{s_k}(N) &= \sum_{k=1}^{N} - \left[\left(t_{k,s}^{(N)} + \Delta_{k,s}^{(N)} - l_{k,s}^{(N)}\right)^2 e^{-\left(t_{k,s}^{(N)}+\Delta_{k,s}^{(N)}\right)} - \left(t_{k,s}^{(N)} - l_{k,s}^{(N)}\right)^2 e^{-t_{k,s}^{(N)}}\right] \\
&= \sum_{k=1}^{N} - \left[\underline{\Delta}_{k-1}^2 e^{-\left(t_{k,s}^{(N)}+\Delta_{k,s}^{(N)}\right)} - \underline{\Delta}_{k}^2 e^{-t_{k,s}^{(N)}}\right] \quad \text{(n.n. optimality and } \underline{\Delta}_0 = \infty\text{)} \\
&= \sum_{k=1}^{N} -\underline{\Delta}_{k-1}^2 e^{-t_{k-1,s}^{(N)}} + \underline{\Delta}_{k}^2 e^{-t_{k,s}^{(N)}} \\
&= \sum_{i=2}^{N} -\underline{\Delta}_{k-1}^2 e^{-t_{k-1,s}^{(N)}} + \underline{\Delta}_{k}^2 e^{-t_{k,s}^{(N)}} - \underline{\Delta}_{1}^2 e^{-t_{0,s}^{(N)}} \\
&= \sum_{k=2}^{N} -\underline{\Delta}_{k-1}^2 e^{-t_{k-1,s}^{(N)}} + \underline{\Delta}_{k}^2 e^{-t_{k,s}^{(N)}} - 0 \quad \text{(since } t_{0,s}^{(N)} = \infty\text{)} \\
&= \underline{\Delta}_{N}^2 e^{-t_{N,s}^{(N)}} = \underline{\Delta}_{N}^2 \quad \text{(since } t_{N,s}^{(N)} = 0\text{)} \\
&= s_N^2. \qquad (3.4.79)
\end{aligned}$$

The first remark we make is that, up to this point, we have not used any fact specific of the s_k sequence to arrive at (3.4.79) other than knowing that s_k is a sequence of half steps for the quantizer. Second, we can interpret the value of $\rho_{s_k}(N)$ as reflection of how well the nearest neighbor optimality condition is adhered to by scrutinizing how close $\rho_{s_k}(N)$ is to s_N^2. (In this case, they are equal since our quantizer satisfies the nearest neighbor condition. For an arbitrary quantizer q, $\rho_q(N)$ may not equal $\underline{\Delta}_N^{(N)^2}$.)

Examining the second term of (3.4.78), we can interpret $d_{s_k}(N)$ as indicating how well the centroid optimality condition is satisfied by our quantizer. Since the value of $d_{s_k}(N)$ is more obscure than $\rho_{s_k}(N)$, we will observed its behavior as a function of N by looking at numerical data.

Thus, using (3.4.79), the MSE of an N-level quantizer designed using the s_k sequence can be expressed as

$$D_{s_k}(N) = s_N^2 + d_{s_k}(N), \qquad (3.4.80)$$

where

$$d_{s_k}(N) = 2\sum_{k=1}^{N} \left(x - l_{k,s}^{(N)}\right) P_{\left[t_{k,s}^{(N)}, t_{k,s}^{(N)}+\Delta_{k,s}^{(N)}\right)} \qquad (3.4.81)$$

and, as already noted, can be considered *general* and applicable to any N-level, unit variance exponential quantizer that has been designed in accordance to a sequence of half steps since no fact specific to the s_k sequence was used to generate these two expressions. From (3.4.80), we study the contribution to distortion of each component expression by multiplying each component by N^2. Figure 3.3 Row b and Row c show evidence to support the conjecture that the first term s_N^2 in (3.4.80) is the dominant contributor to the MSE produced by such quantizers since it appears to converge to $\frac{\beta}{12}$ in Row b and $\frac{s_N^2}{\underline{\eta}_N^2}$ appears to converge to 1, while the second term in (3.4.80) appears to go to zero when multiplied by N^2 (in Row b) and the ratio of the second term over $\underline{\eta}_N^2$ appears to go to zero when N becomes large.

Since it is true that $N^2 s_N^2 \to \frac{\beta}{12}$ as $N \to \infty$ (see lemma below), if one could show analytically that the first term in (3.4.80) is dominant over the second term in (3.4.80), then designing quantizers using s_k in the manner we have described would yield asymptotically optimal quantizers. However, at present, this problem is still open for future work.

Lemma III.9. *$N^2 s_N^2 \to \frac{\beta}{12}$ as $N \to \infty$.*

Proof. Evaluating the limit, we have

$$\begin{aligned}\lim_{N\to\infty} N^2 D_{s_k}(N) &= \lim_{N\to\infty} N^2 s_N^2 = \lim_{N\to\infty} N^2 \left[\frac{3}{2N}\left(1-\frac{2}{N}+\frac{8}{N^2}\right)^{\frac{1}{2}}\right]^2 \\ &= \lim_{N\to\infty} \frac{9}{4}\left(1-\frac{2}{N}+\frac{8}{N^2}\right) = \frac{9}{4} = \frac{27}{12} = \frac{\beta}{12}\end{aligned}$$

since for the unit variance exponential source $\beta = 27$. ■

3.4.3.2 Simplified quantizer design.

Since it appears that designing quantizers using the s_k sequence yields quantizers with good MSE performance, it may be prudent to consider a further simplification of this design method. To this end, we use the exact same procedure to design quantizers but we swap s_k for the sequence $v_k \triangleq \frac{3}{2k}$ which is constructed by removing Ω_k from each term of the s_k sequence when $k \geq 9$, i.e., v_k is defined as

$$v_k = \underline{\eta}_k, \quad k = 1, 2, \ldots, 8 \qquad\qquad v_k = \frac{3}{2k}, \quad k \geq 9.$$

Thus, using (3.4.80), it is clear that the MSE of an N-level quantizer designed for the exponential source using the half step sequence v_k is

$$D_{v_k}(N) = v_N^2 + d_{v_k}(N),$$

where

$$d_{v_k}(N) = 2\sum_{k=1}^{N}\left(x - l_{k,v}^{(N)}\right) P_{\left[t_{k,v}^{(N)}, t_{k,v}^{(N)}+\Delta_{k,v}^{(N)}\right)}$$

which is similar to (3.4.81).

Figure 3.4(a) shows MSE performance data for such quantizers designed with v_k for the unit exponential source when $N \leq 128$, where the data has been presented as ratios: $\frac{D_{v_k}(N)}{D(N)}$ and $\frac{D_{s_k}(N)}{D(N)}$. Here, we observe that even for quantizers designed using the simplified sequence v_k, that $\frac{D_{v_k}(N)}{D(N)}$ lies within 0.3% of 1. In Figure 3.4(b), the contribution to MSE distortion given by each component in $N^2 D_{v_k}(N)$ is compared to $\frac{\beta}{12} = 2.25$, where we note that v_k overestimates $\underline{\eta}_k$ and consequently, the contribution of the second component to the overall distortion is negative. Thus it appears that the sequence v_k is an upper bound to the Nitadori sequence $\underline{\eta}_k$ which is in contrast to s_k which is a lower bound to $\underline{\eta}_k$. Summarizing, we know (from the proof of Part 2 of Theorem III.1)

$$s_k = v_k \cdot \Omega_k \leq \underline{\eta}_k$$

for all $k \geq 1$, and from the data, we observe the following trend

$$v_k \geq \underline{\eta}_k$$

when $N \leq 128$.

The support threshold for quantizers designed using v_k is

$$t_{1,v}^{(N)} = v_N + v_1 + 2\sum_{k=2}^{N-1} = \frac{3}{2N} + \underline{\eta}_1 + 2\sum_{k=2}^{8}\underline{\eta}_k + 2\sum_{k=9}^{N-1}\frac{3}{2k} = \frac{3}{2N} + \left(t_1^{(8)} - \underline{\eta}_8\right) + 3\sum_{k=9}^{N-1}\frac{1}{k}.$$

when $N > 9$. Using a Riemann integral, for an upper bound to the last sum

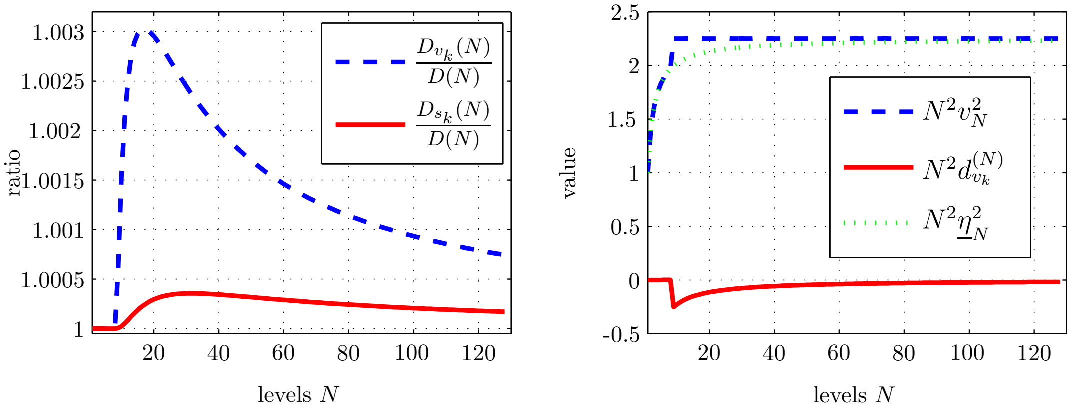
$$\sum_{k=9}^{N-1}\frac{1}{k} \leq \int_{8}^{N-1}\frac{1}{x}dx = \log x\big|_8^{N-1} = \log(N-1) - \log 8,$$

we obtain the following upper bound to $t_{1,v}^{(N)}$

$$\begin{aligned} t_{1,v}^{(N)} &\leq \frac{3}{2N} + \left(t_1^{(8)} - \underline{\eta}_8\right) + 3\log(N-1) - 3\log 8 \\ &< 3\log(N-1) + \frac{3}{2N} + -0.7965 = 3\log N + 3\log\left(1 - \frac{1}{N}\right) + \frac{3}{2N} - 0.7965 \quad (3.4.82) \end{aligned}$$

for $N \geq 9$. Since the support threshold estimate in (3.4.82) is greater than $t_1^{(N)}$ but it is less than $t_{1,c}^{(N)}$, we know the following:

$$t_{1,s}^{(N)} \leq t_1^{(N)} \leq t_{1,v}^{(N)} \leq t_{1,(3.4.82)}^{(N)} \leq t_{1,c}^{(N)}.$$



(a) Data for q_{v_k} and q_{s_k}.

(b) Data for q_{v_k} only.

Figure 3.4: The MSE performance of quantizers q_{v_k} designed using v_k compared against the performance of quantizers designed using v_k and optimal quantizers. The data shown in these plots is for $N \leq 128$. Plot (a) shows represents the distortion data in terms of the ratio of the MSE for quantizers designed with v_k over the minimum achievable MSE against the ratio of the MSE for quantizers designed with s_k over the minimum achievable MSE and Plot (b) highlights data for N^2 times the MSE contribution by each distortion component for quantizers designed with v_k.

Performance of quantizers designed using our companding method: UTCC (Uniform Threshold Compander with Centroid Reconstruction Levels) quantizers. For the sake of comparing the MSE performance of all quantizers discussed in this chapter against each other as well as against the performance of optimal quantizers, here, we briefly comment on the performance of quantizers designed using the companding method described in the proof of the upper bound to the support threshold $t_1^{(N)}$ in Theorem III.1. (For more remarks, see Appendix B.)

The MSE performance of an N-level UTCC quantizer can be expressed as the sum of two parts

$$
\begin{aligned}
D_{UTCC}(N) &= \sum_{k=1}^{N} \int_{t_{k,c}^{(N)}}^{t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}} \left(x - l_{k,c}^{(N)}\right)^2 f(x)\, dx \\
&\overset{IBP}{=} \left[\sum_{k=1}^{N} - \left(x - l_{k,c}^{(N)}\right)^2 e^{-x} \Big|_{t_{k,c}^{(N)}}^{t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}} \right] + d_{UTCC}(N), \qquad (3.4.83)
\end{aligned}
$$

where the sum in (3.4.83) (in square brackets) equals

$$
-\sum_{k=1}^{N}\left(t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}-l_{k,c}^{(N)}\right)^2 e^{-\left(t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}\right)} - \left(t_{k,c}^{(N)}-l_{k,c}^{(N)}\right)^2 e^{-t_{k,c}^{(N)}} = -\sum_{k=1}^{N} \overline{\Delta}_{k,c}^{2} e^{-\left(t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}\right)} - \underline{\Delta}_{k,c}^{2} e^{-t_{k,c}^{(N)}}
$$

and

$$
d_{UTCC}^{(N)} = 2\sum_{k=1}^{N} \left(\mu_k^{(N)} - l_k^{(N)}\right) P_{\left[t_{k,c}^{(N)},\, t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}\right)} = 0
$$

since UTCC quantizers have centroid reconstruction levels.

Evaluating (3.4.83) and using (3.3.8), we have

$$
\begin{aligned}
D_{UTCC}(N) &= \sum_{k=1}^{N} \underline{\Delta}_{k,c}^{2} e^{-t_{k,c}^{(N)}} - \overline{\Delta}_{k,c}^{2} e^{-\left(t_{k-1,c}^{(N)}\right)} \\
&= \sum_{k=1}^{N} \left[1 - \frac{-3\log\left(1-\frac{1}{k}\right)\cdot\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right]^2 \left(\frac{k}{N}\right)^3 - \left[\frac{-3\log\left(1-\frac{1}{k}\right)}{1-\left(1-\frac{1}{k}\right)^3} - 1\right]^2 \left(\frac{k-1}{N}\right)^3. \qquad (3.4.84)
\end{aligned}
$$

(See Appendix B for complete derivation.) While not compact, (3.4.84) gives an exact formula for computing the MSE of an N-level UTCC quantizer.

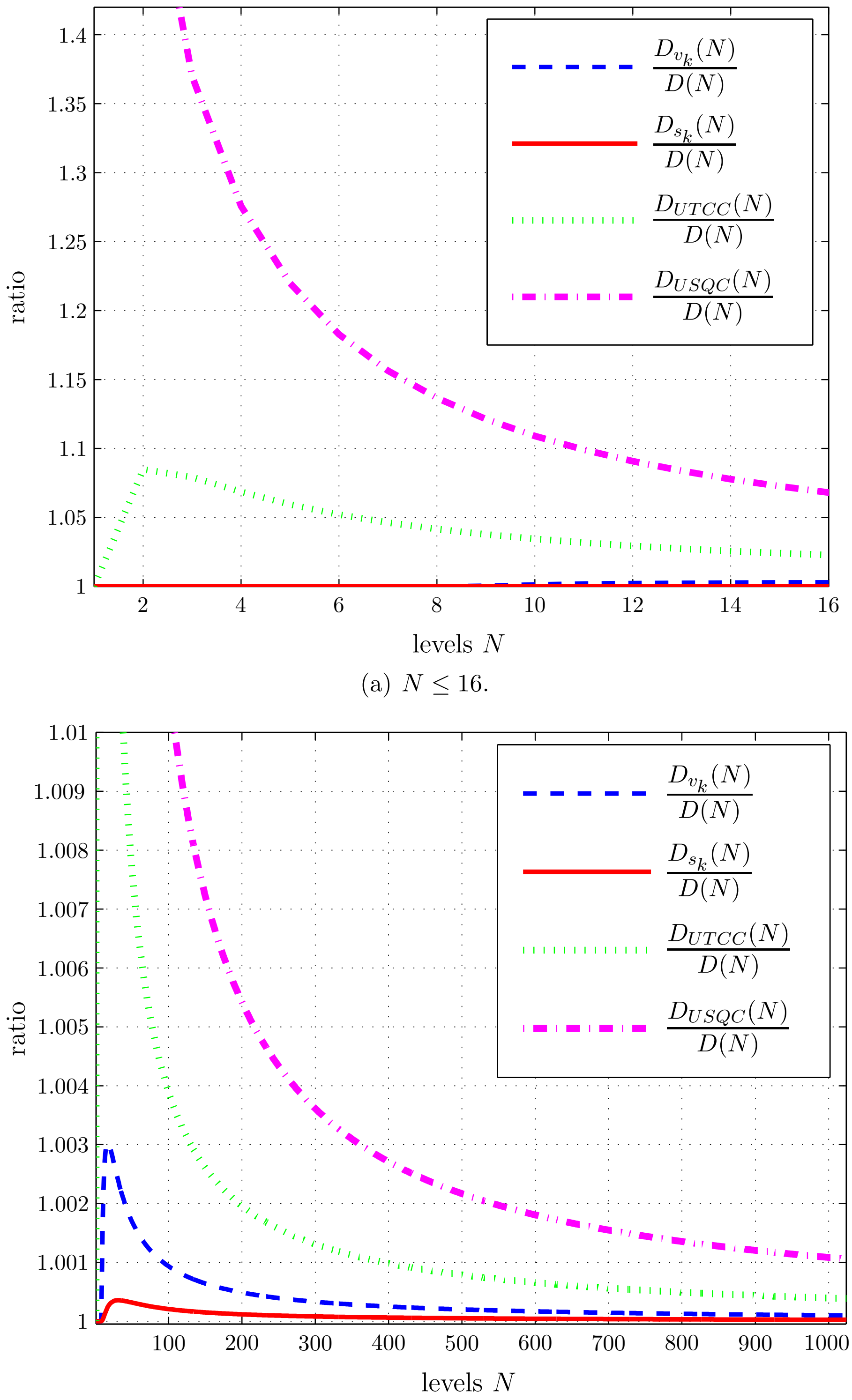
In Figure 3.5, we have plotted performance data for all of the quantizers that we have discussed in this chapter, including the performance of UTCC quantizers as indicated by (3.4.84). We see that for $N \leq 1024$, quantizers designed with s_k perform closest to optimal quantizers, followed by quantizers designed with v_k, then UTCC quantizers and finally, USQC quantizers. We remark that quantizers designed using s_k perform approximately and v_k perform significantly better than UTCC quantizers and USQC quantizers. More specifically, max $\frac{D_{c_k}(N)}{D(N)} = 1.08480647707816$ and occurs when $N = 2$, and for v_k, we have $\max \frac{D_{v_k}(N)}{D(N)} = 1.00302509854578$ when $N = 17$ and for s_k, we have $\max \frac{D_{s_k}(N)}{D(N)} = 1.00035593710511$ when $N = 31$.

3.4.4 Concluding Remarks.

In this section we have considered suboptimal quantizer design using the half step sequences s_k and v_k. To gauge how well these two design schemes are, we first compared their respective support thresholds $t_{1,s}^{(N)}$ and $t_{1,v}^{(N)}$, along with the support threshold $t_{1,c}^{(N)}$ of UTCC and USCQ quantizers, for use as estimates for the support threshold $t_1^{(N)}$ of optimal quantizers. As a second comparison, we looked at the MSE performance of all of these quantizers as well. From these comparisons, we observed that quantizers designed with s_k gave the closest approximation to $t_1^{(N)}$ and also had the smallest MSE of the quantizers discussed. Quantizers designed with v_k, while not as good as quantizers designed using s_k for support threshold estimation or for minimizing MSE, offer a much simpler, analytically tractable means of attaining quantizers with good performance and support threshold estimation. UTCC quantizers, while giving the least accurate estimates of $t_1^{(N)}$ and the worst MSE of the quantizers in the group (even though these quantizers are known to be asymptotically optimal), nevertheless yielded an sequence of support threshold estimates that as a function of N appears able to track to the growth rate of $t_1^{(N)}$ to within a constant term. Thus, from what we have observed, it appears that quantizers designed using either s_k or v_k are also asymptotically optimal.

During the course of studying quantizer design using half step sequences, we utilized a decomposition of the general MSE for an arbitrary quantizer that decomposes MSE into the sum of two expressions. The first expression is highly sensitive to whether the nearest neighbor condition has been satisfied and the second expression is highly sensitive to whether the centroid condition has been satisfied. Using this decomposition to compute the MSE of quantizers designed using half steps, we have been to make the following observation regarding quantizers designed using half step sequences:

If an N-level quantizer has been designed using a half step sequence h_k, $k \geq 1$, and if the resulting reconstruction levels are equal to the centroids of the quantization cells, then these quantizers are optimal. This conclusion is equivalent to the result by Fleischer [7] and Trushkin [23] which states that a quantizer designed for an exponential source whose thresholds and reconstruction levels satisfy the optimality conditions is optimal and unique. Elaborating further on this reasoning:



(a) $N \leq 16$.

(b) $N \leq 1024$.

Figure 3.5: Comparing the MSE performance of all quantizers discussed against the performance of optimal quantizers. Plot (a) is a close-up view with $N \leq 16$ and highlights trends for the low level case. Plot (b) shows when $N \leq 1024$ and illustrates the overall trend for the high level case.

1. Use of a half step design sequence means that the thresholds of any such quantizer satisfies the nearest neighbor condition for optimality, and also implies that the MSE contribution of the first term in the distortion sum is easily evaluated and is equal to $\rho_{h_k} = h_N^2$. (Refer to (3.4.79).)

2. The fact that the resulting reconstruction levels are centroids of the quantization cells means that the centroid condition for optimality is also satisfied by such quantizers and this implies that the second term's contribution to distortion is $d_{h_k} = 0$. (Refer to (3.4.81).)

3. Thus the overall MSE of an N-level quantizer designed using h_k is $D_{h_k}(N) = h_N^2$.

4. But we already know that quantizers that satisfy the two optimality conditions are both optimal and unique if the source is exponential ([7],[23]). Thus it is clear that $h_k = \underline{\eta}_k$, i.e., h_k is the Nitadori sequences, and so $D_{h_k}(N) = \underline{\eta}_N^2$ for all $N \geq 1$.

3.5 Future Work.

While examining the topics addressed in this chapter, several ideas arose that can be investigated or studied further in the future. The following list briefly describes these avenues for more work:

- **A tighter lower bound to $t_1^{(N)}$.** Recall that the support threshold lower bound $t_{1,LB,Thm\ III.1}^{(N)}$ (in Theorem III.1) was created as a lower bound to the support threshold function $t_{1,s}^{(N)}$. (See (3.3.29) and (3.3.30).) In the discussion section, based on what we observed in the data shown in Figure 3.2, there appears to be room to improve $t_{1,LB,Thm\ III.1}^{(N)}$ when compared to $t_{1,s}^{(N)}$. Thus, it may be possible to construct a new, tighter theoretical lower bound to $t_1^{(N)}$ for Theorem III.1.

- **Asymptotic optimality of quantizers designed using s_k and v_k.** From inspecting the performance data for quantizers designed using s_k and v_k in Figure 3.5, we observed that these quantizers produce MSE that is less than the MSE produced by q_{UTCC} and q_{USQC} quantizers. Since it is known that q_{USQC} quantizers are asymptotically optimal, we suspect that q_{s_k} and q_{v_k} are as well. Thus, it may be possible to prove theoretically that quantizers designed using s_k and v_k are asymptotically optimal.

- **A test for conformity to the optimality conditions.** Investigate the use of the MSE expression in (3.4.78) as a method to test an arbitrary quantizer q_N for conformity to the optimality conditions. Clearly, a first application would be to use (3.4.78) as a way to rule out quantizers that do not conform to these conditions by checking to see if

  $$\rho_{q_N}(N) = \left(\underline{\Delta}_{q_N}\right)^2 \tag{3.5.85}$$

  which is the square of the value of the half step of the cell containing the origin. If equality has not been achieved in (3.5.85), then clearly, the nearest neighbor rule has not been used in the design of q_N. Similarly, a check to see if

  $$d_{q_N}(N) = 0 \tag{3.5.86}$$

  would reveal if the centroid optimality condition has been followed. If equality has not been achieved in (3.5.86), then this optimality condition was not used in the design of q_N. Thus, a quantizer q_N which does not satisfy (3.5.85) and (3.5.86) is not optimal.

  The next step in the investigation would be to determine whether a quantizer could achieve (3.5.85) and (3.5.86) yet still not be optimal. This examination would require ascertaining the non-optimal design scenarios under which q_N could achieve equality in (3.5.85) and (3.5.86). In this way, a test for optimality may be constructed for exponential quantizers.

  Furthermore, since the MSE expression in (3.4.78) is applicable to any quantizer designed for a source with finite mean and variance, the next direction to take would be to try to construct tests for conformity to the optimality conditions and a (possible) test for optimality when for a quantizer that has been designed for a source that is not exponential, yet has finite first and second moments.

- **Investigate the importance of adhering to the nearest neighbor condition by studying its effect on MSE.** Begin by examining the performance of the family of quantizers which satisfy the nearest neighbor condition using the MSE expression in (3.4.78). The goal of this study would be to ascertain the influence of setting thresholds optimally (without regard to the placement of reconstruction levels) on MSE by constructing a performance bound on quantizers that satisfy the nearest neighbor condition and comparing it against the MSE achieved by optimal quantizers. Some questions to keep in mind after this comparison are:

1. Are these quantizers asymptotically optimal?
2. If not, what additional constraint would be necessary to achieve asymptotic optimality?

Next, relax the constraint that the nearest neighbor condition has been satisfied and replace it with the condition in (3.5.85). For this scenario, perform similar analysis, i.e., construct a performance bound under this scenario and compare it to the bound already created, check for asymptotic optimality and the design constraints that achieve asymptotic optimality.

Finally, perform this study for quantizers designed for sources with finite mean and variance.

CHAPTER IV

On the Asymptotic Behavior of Half Steps in General Exponential Optimal Scalar Quantizers

4.1 Introduction and Main Result.

Optimal scalar quantizer design is a non-trivial problem that involves specifying the exact position of quantization cell thresholds and reconstruction levels that minimize mean-squared error. For the *normalized* exponential source,[1] however, aided through implicit simplifications made possible by the source's memoryless property, i.e., $P\left(X \geq s+t\,|X \geq s\right) = P\left(X \geq t\right)$, for any $s, t \geq 0$, Nitadori solved the optimal scalar quantizer design problem by taking partial derivatives of the distortion function. He showed that for an optimal N-level quantizer, $\underline{\Delta}_k^{(N)}$ does not depend on N, and because of this, there is a sequence of numbers $\underline{\eta}_1, \underline{\eta}_2, \underline{\eta}_3, \ldots$, where $\underline{\eta}_k = \underline{\Delta}_k^{(N)}$, that provides a complete and unique[2] specification of an optimal N-level scalar quantizer designed for a normalized exponential source. The following relationships illustrate this fact:

1. The support threshold can be calculated by

$$t_1^{(N)} = \underline{\eta}_1 + \sum_{i=1}^{N-1} 2\underline{\eta}_i + \underline{\eta}_N \tag{4.1.1}$$

[1] We define the *normalized exponential source* to be the source with pdf of the form e^{-x}, $x \geq 0$, where $E\left[X\right] = 1$ and $\sigma^2 = 1$. Note that while Nitadori [19] actually considered optimal quantization of a two-sided, zero mean, unit variance exponential source, the resulting sequence that was derived by him also holds for optimal quantization of a (one-sided) normalized exponential source.

[2] Uniqueness for an exponential source was first proposed by Fleischer [7], but the argument was later corrected by Trushkin [23]. Fleischer [7] did prove uniqueness for strictly log convex sources, however.

and the reconstruction level in the outer region can be calculated as

$$\mu_1^{(N)} = t_1^{(N)} + \underline{\eta}_1. \tag{4.1.2}$$

2. For $i = 2, 3, \ldots, N$, the thresholds and reconstruction levels are determined using

$$t_{i+1}^{(N)} = t_i^{(N)} - \left(\underline{\eta}_i + \underline{\eta}_{i+1}\right) \tag{4.1.3}$$

and

$$\mu_{i+1} = t_i^{(N)} - \underline{\eta}_i. \tag{4.1.4}$$

3. $t_0^{(N)} = \infty$.

Nitadori's sequence $\underline{\eta}_k$ also provides a simple expression for the MSE performance [19] of each optimal, N-level, normalized exponential quantizer, namely,

$$\sum_{i=1}^{N} \int_{t_i^{(N)}}^{t_{i-1}^{(N)}} \left(u - \mu_i^{(N)}\right)^2 e^{-u}\, du = \underline{\eta}_N{}^2.$$

As already noted in Chapter II, knowing the set of N half steps belonging to an optimal quantizer is equivalent to knowing both the set of quantizer thresholds and levels that specify that quantizer. In general, the set of half steps belonging to an N-level quantizer depends on N. In the case of optimal quantization of exponential sources,[3] however, the set of half steps for an optimal N-level quantizer is independent of the number of levels N in that quantizer. For this case, quantizer design reduces down to truncating the Nitadori sequence after the Nth term with $\{\underline{\eta}_k\}_{k=1}^{N}$ providing a complete specification for an optimal N-level quantizer. While the fact that the Nitadori sequence provides a simple, exact specification for optimal quantizers of all levels is astonishing by itself, it is even more remarkable that this same sequence of values also provides the MSE performance of any optimal N-level quantizer.

[3] For the rest of this chapter, reference to an exponential source will refer to the normalized exponential source. We note, however, that for a zero mean, exponential source with variance σ^2, there exists a Nitadori sequence $\sigma\underline{\eta}_k$, $k = 1, 2, \ldots$, and that the MSE performance for an optimal N-level quantizer designed for this source is given by $\sigma^2\underline{\eta}_k{}^2$.

The utility of the Nitadori sequence $\underline{\eta}_k$ with regard to exponential MMSE quantizer design and performance naturally leads to a search for a similar result for MMSE quantization of other sources. Derivation of the Nitadori sequence required use of the exponential source's memoryless property, however, and since in general, an arbitrary source distribution does not possess this property, it does not appear possible to solve for a single sequence that specifies all N-level MMSE scalar quantizers for an arbitrary source. On closer inspection of how the memoryless property is used within the Nitadori sequence derivation, we find that only one *key effect* of the memoryless property is required:

$$f\left(x\right)=\int\limits_{x}^{\infty}f\left(u\right)du,$$

where $f\left(x\right)=e^{-x}$, $x\geq 0$. With this observation in mind and to improve the chances of success in deriving a result similar to Nitadori's with regard to MMSE quantizer design and performance, we limit the scope of our work in two ways:

1. **Large N.** Instead of considering N-level MMSE quantizer design for any value of N, we consider MMSE quantization when N is large.

2. **General exponential sources.** Instead of MMSE quantization design for an arbitrary source, we consider only sources from the general exponential (GE) family whose probability density functions have the form

$$f\left(x\right)=c_p\,e^{-\frac{x^p}{p}},\;x\geq 0, \tag{4.1.5}$$

   where $c_p>0$ is a proportionality constant and $p\geq 1$ is a parameter which indexes amongst members in this one-sided source family.[4] Members of this family include the one-sided exponential source ($p=1$) and the one-sided Gaussian source ($p=2$). The decision to focus on sources from this particular family is due to the fact that each pdf in the family possesses an asymptotic version of the key effect.[5]

Under this restricted, but still useful, quantization context, we will show that sequences that asymptotically describe N-level MMSE quantizers when truncated to N terms is possible. For each member of the GE distribution family, and for each

[4] In the literature, this family is sometimes referred to as generalized Gaussian.

[5] Also, since GE-sources have pdfs that are strictly convex, for each value of $N\geq 1$, MMSE quantizers designed for such a source are unique. ([7], [23])

quantization cell $k \geq 1$ of an MMSE quantizer designed for that particular source, we have discovered that the sequence

$$\underline{\alpha}_{k,p}^{(N)} \triangleq \underline{\Delta}_k^{(N)} \left(t_k^{(N)}\right)^{p-1} \tag{4.1.6}$$

asymptotically satisfies the generating relationship that produces the Nitadori sequence $\underline{\eta}_k$ as $N \rightarrow \infty$. This find is the main result of this chapter and a formal statement of this result is as follows:

Theorem IV.1. *Fix $k \geq 1$. For an optimal N-level quantizer designed for a GE-source with parameter $p \geq 1$, the kth quantization cell satisfies*

1. $$\lim_{N\rightarrow\infty} \underline{\alpha}_{k,p}^{(N)} = \underline{\eta}_k.$$

2. $$\lim_{N\rightarrow\infty} \frac{\underline{\Delta}_k^{(N)}}{\underline{\Delta}_{k+1}^{(N)}} = \frac{\underline{\eta}_k}{\underline{\eta}_{k+1}}.$$

With regard to asymptotic MMSE quantizer design, Theorem IV.1 indicates that the role of the Nitadori sequence $\underline{\eta}_k$ is broader than just design for the exponential source, as is the case in (non-asymptotic) MMSE quantization. While the Nitadori sequence $\underline{\eta}_k$ provides the exact half cell sizes for optimal N-level quantization of an exponential source, Theorem IV.1, Part 1 indicates that for a general exponential source with $p > 1$, the Nitadori sequence $\underline{\eta}_k$ provides a way to asymptotically approximate the optimal half cell sizes $\underline{\Delta}_k^{(N)}$ when N is large via

$$\underline{\Delta}_k^{(N)} \approx \frac{\underline{\eta}_k}{\left(t_k^{(N)}\right)^{p-1}} \tag{4.1.7}$$

and subsequently, to approximate the optimal set of thresholds and levels via (4.1.1)-(4.1.4). Moreover, for each $1 \leq k << N$, the approximation in (4.1.7) gives an indication of how the optimal half cell sizes $\underline{\Delta}_k^{(N)}$ vary across the GE-family of sources (via the p parameter) with respect to their optimal thresholds $t_k^{(N)}$.

Curiously and perhaps more useful than Part 1, Part 2 of Theorem IV.1 states that for every GE-source with $p > 1$, as N grows large, the ratio of optimal half steps $\frac{\underline{\Delta}_k^{(N)}}{\underline{\Delta}_{k+1}^{(N)}}$ belonging to adjacent quantization cells approaches a constant $\frac{\underline{\eta}_k}{\underline{\eta}_{k+1}}$ that, surprisingly, is independent of p. Since the limiting constant $\frac{\underline{\eta}_k}{\underline{\eta}_{k+1}}$ goes to 1 as the cell index $k \to \infty$, it follows from Theorem IV.1 that

$$\lim_{k\to\infty} \lim_{N\to\infty} \frac{\underline{\Delta}_k^{(N)}}{\underline{\Delta}_{k+1}^{(N)}} = 1. \tag{4.1.8}$$

For optimal quantizers that have cells which become smaller as N increases, it is expected that for any fixed $x > 0$, the half steps of cells near x are expected to be approximately the same since the conditional source density of these neighboring cells become asymptotically constant. Since for each $k \geq 1$, $t_k^{(N)} \to \infty$ when $N \to \infty$, (4.1.8), however, is a stronger statement than this because, for any fixed value of $k \geq 1$, it states that the ratio between the half steps on either side of $t_k^{(N)}$ converges to 1 as $t_k^{(N)}$ increases without bound:

Fix $k \geq 1$. For $\epsilon > 0$, there exists k_ϵ such that for all $k \geq k_\epsilon$,

$$\left| \frac{\underline{\Delta}_k^{(N)}}{\underline{\Delta}_{k+1}^{(N)}} - 1 \right| < \epsilon$$

when N is sufficiently large.

It is also interesting to note that the nature of Theorem IV.1 is complementary to the conventional sort of asymptotic quantization result. In particular, conventional asymptotic quantization theory, such as the Panter-Dite formula [20], deals with the performance of quantizers in the region where the pdf is large, and basically ignores the tail of the source pdf. In contrast, Theorem IV.1 is based exclusively on the tail of the source pdf. What is interesting is the degree of impact/influence that the tail behavior of the source has regarding MMSE quantization of cells in regions where the pdf is large.

The proof of Theorem IV.1 (to be given later) centers on showing that for each fixed value $k \geq 1$, $\lim_{N\to\infty} \underline{\alpha}_{k,p}^{(N)}$ exists and that as $N \to \infty$, $\underline{\alpha}_{k,p}^{(N)}$ asymptotically satisfies the Nitadori sequence generating equation, the recursive relationship that generates successive terms of the Nitadori sequence when initialized with $\underline{\eta}_1$. (More discussion on this generating relation is given later.) Our approach to the proof of Theorem IV.1 begins by studying a specific method of deriving the Nitadori sequence, or equivalently, of solving the N-level MMSE quantization design problem for an exponential source. This method uses only the aforementioned key effect of the exponential source's memoryless property along with the optimality conditions and makes it easy to point out when and where use of the key effect of the memoryless property is made. When presenting this derivation, we break with convention, which would dictate deriving the Nitadori generating equation first and then solving for the initial sequence value $\underline{\eta}_1$. Instead, we start by establishing the initial sequence value $\underline{\eta}_1$ since the fact $\underline{\eta}_1$ does not depend on the number of levels N is entirely due to the memoryless property and is the central fact that makes the existence of Nitadori sequence possible. After $\underline{\eta}_1$ is known, we proceed to derive the Nitadori generating equation, discovering, as in the case of the initial value $\underline{\eta}_1$, that successive terms of the Nitadori sequence are also independent of the number of levels N.

We present our work in this chapter by first giving a brief review of Nitadori's result using a derivation similar in approach and style to the one we use to prove our results regarding the sequence $\underline{\alpha}_{k,p}^{(N)}$. After the review, we state an asymptotic key effect that holds for the GE-source family and establish an asymptotic fact regarding the value of $\underline{\alpha}_{1,p}^{(N)}$, which is the foundation of Theorem IV.1 and is, in essence, analogous to the fact that $\underline{\eta}_1 = 1$.

Since Theorem IV.1 is a result concerning the behavior of $\underline{\alpha}_{k,p}^{(N)} = \underline{\Delta}_k^{(N)}(t_k^{(N)})^{p-1}$ in cells that lie in the tail region of the source pdf, our method of proof consists of investigating properties of quantizers that are optimal for tail regions of the source of the form $[\tau, \infty)$, asymptotically as τ goes to infinity. Referring to these quantizers as *conditionally optimal* because they have been optimized for the conditional distribution X given that $X \geq \tau$, our approach allows us to prove a result that is analogous to the statement in Theorem IV.1 but for conditionally optimal quantizers. We then use this intermediate result to prove the theorem.

Once the proof of Theorem IV.1 has been demonstrated, several corollaries to it are presented and proven true. The bulk of the rest of this chapter deals with applications of Theorem IV.1. The first application concerns half step estimation. The second application discusses support threshold estimation. Closing out the chapter, ideas for further study are proposed and briefly commented on.

While Nitadori's result assumed a two-sided source distribution, we will restrict the scope of our work to one-sided sources that have pdf support on the non-negative reals to improve the readability of the analysis, with the notion that extending the results to the two-sided case is easy to do.

4.2 Review of Nitadori's Result.

Instead of reviewing the exact method used by Nitadori in 1965, we will re-derive the sequence $\underline{\eta}_k$ for the one-sided exponential source in a manner similar to the method we will use to derive the sequence $\lim_{N\to\infty} \underline{\Delta}_k^{(N)}(t_k^{(N)})^{p-1}$ in Part 1 of Theorem IV.1.

The method of choice is a simple one, using only three relationships: the two optimality conditions and the memoryless property's key effect, along with a single tool, integration-by-parts (IBP). Since our goal will be to explicitly derive the Nitadori sequence, we will have obtained our goal when we have derived the sequence generator equation, a well-defined relationship between $\underline{\eta}_i$ and $\underline{\eta}_{i-1}$ for any $i \geq 2$, and the initial value $\underline{\eta}_1$ for this generator equation, so that when $\underline{\eta}_1$ is used with the generator equation, the terms of the Nitadori sequence $\underline{\eta}_k$ are generated in succession.

The re-derivation is organized as follows: We begin the derivation by considering the centroid of an arbitrary half open interval and the impact the key effect has on its expression, in a non-quantization setting. We use this general setting, not only to simplify the notation we use and to improve clarity of the discussion, but also to emphasize the fact that the resulting simplified expression for the centroid is a side effect of the key effect, which is a property of the exponential source and not due to any special circumstance/restrictions imposed by the MSE quantization scenario. Since the centroid optimality condition requires that all reconstruction levels of MMSE quantizers to be the centroids of the quantization cells they belong to, this is a reasonable way to begin the derivation.

Next, we consider the special case when the half open interval has infinite step size, and we switch our focus from the centroid to the half step of such an interval. Here, an important observation is made: The half step (or the distance between the centroid and the endpoint of the interval) of such an interval is independent of the endpoint of the interval. This is, perhaps, the most important observation to be made since without the independence of this particular half step from the threshold t, the Nitadori result would not be possible.

After those brief remarks, we will apply the simplified centroid expression to N-level, exponential MMSE quantization design to derive the initial value of the Nitadori sequence $\underline{\eta}_1$ and the Nitadori sequence generator equation, discovering that none of the half step sizes belonging to the cells of an optimal N-level quantizers depend on N.

4.2.1 Property of the Exponential Source: Impact of the Key Effect on the Conditional Mean.

Let

$$f(x) = e^{-x},\ x \geq 0,$$

be the one-sided exponential source pdf and let

$$Q(x) \overset{\triangle}{=} \int_{x}^{\infty} f(u)\,du$$

be the *tail function* for f. In this case, the exponential source's memoryless property ($P(X \geq s+t \,|\, X \geq s) = P(X \geq t)$, for any $s, t \geq 0$) produces the *key effect* for f, expressed as

$$Q(x) = f(x),\ x \geq 0.$$

Let $[t, t+\Delta)$ be a half open interval with lower threshold $t \geq 0$ and step size $\Delta > 0$. Consider now what happens to the conditional mean of X given $X \in [t, t+\Delta)$, or equivalently, the centroid

$$\mu_{[t,t+\Delta)} \overset{\triangle}{=} \frac{\int_{[s,t)} xf(x)\,dx}{P_{[s,t)}}$$

of this interval when the key effect is used to simplify it. Using integration by parts (IBP)

$$\begin{aligned} &u = x && du = dx \\ &v = -e^{-x} = -\int_{x}^{\infty} f(x)\,dx = -Q(x) \overset{k.e.}{=} -f(x) && dv = e^{-x}dx = f(x)\,dx \end{aligned} \tag{4.2.9}$$

we have

$$\mu_{[t,t+\Delta)} = \frac{1}{P_{[t,t+\Delta)}} \int_{t}^{t+\Delta} xf(x)\,dx = \frac{\int_t^{t+\Delta} xf(x)\,dx}{Q(t) - Q(t+\Delta)} \overset{IBP}{=} \frac{-xQ(x)|_t^{t+\Delta} + \int_t^{t+\Delta} Q(x)\,dx}{Q(t) - Q(t+\Delta)} \tag{4.2.10}$$

$$= \frac{tQ(t) - (t+\Delta)\,Q(t+\Delta)}{Q(t) - Q(t+\Delta)} + \frac{\int_t^{t+\Delta} Q(x)\,dx}{Q(t) - Q(t+\Delta)} \tag{4.2.11}$$

where IBP has been used to obtain the last equality in (4.2.10). Since no special property of the exponential source has been used yet, (4.2.11) is a general expression for the centroid $\mu_{[s,t)}$ for differentiable source pdfs, and it is completely expressed in terms of the thresholds $t,\ t+\Delta$ and the tail function $Q(x)$ evaluated at these thresholds.

Using the key effect on (4.2.11) simplifies the centroid expression to

$$\begin{aligned} \mu_{[t,t+\Delta)} &= \frac{tQ\left(t\right)-\left(t+\Delta\right)Q\left(t+\Delta\right)}{Q\left(t\right)-Q\left(t+\Delta\right)}+\frac{\int_t^{t+\Delta} f\left(x\right)dx}{Q\left(t\right)-Q\left(t+\Delta\right)} \\ &= \frac{tQ\left(t\right)-\left(t+\Delta\right)Q\left(t+\Delta\right)}{Q\left(t\right)-Q\left(t+\Delta\right)}+\frac{Q\left(t\right)-Q\left(t+\Delta\right)}{Q\left(t\right)-Q\left(t+\Delta\right)} \qquad (4.2.12) \\ &= \frac{tQ\left(t\right)-\left(t+\Delta\right)Q\left(t+\Delta\right)}{Q\left(t\right)-Q\left(t+\Delta\right)}+1, \qquad (4.2.13) \end{aligned}$$

where in (4.2.13), use of the key effect caused

$$\frac{\int_t^{t+\Delta} Q\left(x\right)dx}{Q\left(t\right)-Q\left(t+\Delta\right)}=1. \tag{4.2.14}$$

Special case $\Delta \to \infty$: Letting $\Delta \to \infty$, we now consider the half step length of the infinite half open interval with threshold t

$$\underline{\Delta}_{[t,\infty)} \overset{\triangle}{=} \mu_{[t,\infty)} - t.$$

From (4.2.13), we have

$$\mu_{[t,\infty)} = \lim_{\Delta\to\infty} \mu_{[t,t+\Delta)} = \frac{tQ\left(t\right)}{tQ\left(t\right)}+1=t+1$$

or equivalently,

$$\underline{\Delta}_{[t,\infty)} = 1. \tag{4.2.15}$$

In (4.2.15), we observe that the impact of the key effect on the conditional mean is to cause the half step of an infinite half open interval to be independent of the threshold t. Later more will be stated on this point. This independence from t or translation invariance is the most important consequence of the key effect and in effect, frees the half step solution to the N-level MMSE quantizer design problem from a dependence on the number of levels N.

4.2.2 Generating Nitadori's Sequence.

Fix $N \geq 1$ and consider an optimal N-level quantizer that has been designed for the exponential source. We begin with the half step for the outermost quantization cell $\underline{\Delta}_1^{(N)}$ and then we solve for the Nitadori generator equation that produces the remaining half steps $\underline{\Delta}_k^{(N)}$, $k = 2, 3, \ldots, N$, of the quantizer.

Initial value of the Nitadori sequence – The half step of the outermost cell: Since the outermost quantization cell ($k = 1$) of the optimal quantizer is $[t_1^{(N)}, \infty)$, using (4.2.15), we deduce that $\underline{\eta}_1 = \underline{\Delta}_1^{(N)} = \underline{\Delta}_{[t_1^{(N)},\infty)} = 1$ which is independent of $t_1^{(N)}$ and hence, also independent of N.

Finding the Nitadori recursion equation: The existence of the remaining terms of the Nitadori sequence $\underline{\eta}_k$, $k = 2, 3, \ldots, N$, is assured if it can be shown that $\underline{\Delta}_k^{(N)}$, for $k = 2, 3, \ldots, N$, does not depend on N. Demonstrating the independence of $\underline{\Delta}_k^{(N)}$ with respect to N is accomplished by finding a well-defined relationship between $\underline{\Delta}_k^{(N)}$ and $\underline{\Delta}_{k-1}^{(N)}$, for $k = 2, 3, \ldots, N-1$, that does not depend on N. This relationship emerges after applying both optimality conditions and the memoryless effect $Q(x) = f(x)$ to and then re-expressing the expression for the optimal reconstruction level for the kth quantization cell $[t_k^{(N)}, t_{k-1}^{(N)})$.

Fix $k \in \{2, 3, \ldots, N\}$. Let us start with (4.2.13) (which is the centroid optimality condition applied to the kth cell),

$$\begin{aligned}\mu_k^{(N)} &= \frac{1}{P_{\left[t_k^{(N)}, t_{k-1}^{(N)}\right)}} \int_{t_k^{(N)}}^{t_{k-1}^{(N)}} x f(x)\, dx \quad \text{(optimality cond.: cond. mean)} \\ &= \frac{t_k^{(N)} Q\left(t_k^{(N)}\right) - t_{k-1}^{(N)} Q\left(t_{k-1}^{(N)}\right)}{Q\left(t_k^{(N)}\right) - Q\left(t_{k-1}^{(N)}\right)} + 1\end{aligned}$$

or equivalently,

$$\left(Q\left(t_k^{(N)}\right) - Q\left(t_{k-1}^{(N)}\right)\right)\left(\mu_k^{(N)} - 1\right) = -t_{k-1}^{(N)} Q\left(t_{k-1}^{(N)}\right) + t_k^{(N)} Q\left(t_k^{(N)}\right)$$

or equivalently,

$$Q\left(t_k^{(N)}\right)\left(\mu_k^{(N)} - 1 - t_k^{(N)}\right) = Q\left(t_{k-1}^{(N)}\right)\left(\mu_k^{(N)} - 1 - t_{k-1}^{(N)}\right)$$

or equivalently with $\underline{\Delta}_k^{(N)} = \mu_k^{(N)} - t_k^{(N)}$ and $\underline{\Delta}_{k-1}^{(N)} = t_{k-1}^{(N)} - \mu_k^{(N)}$ (by the nearest neighbor condition),

$$Q\left(t_k^{(N)}\right)\left(\underline{\Delta}_k^{(N)} - 1\right) = Q\left(t_{k-1}^{(N)}\right)\left(-\underline{\Delta}_{k-1}^{(N)} - 1\right)$$

or equivalently with $Q\left(x\right) = f\left(x\right) = e^{-x}$ (key effect),

$$\begin{aligned} f\left(t_k^{(N)}\right)\left(\underline{\Delta}_k^{(N)} - 1\right) &= f\left(t_{k-1}^{(N)}\right)\left(-\underline{\Delta}_{k-1}^{(N)} - 1\right) \quad \text{(key effect)} \\ e^{-\left(\mu_k^{(N)} - \underline{\Delta}_k^{(N)}\right)}\left(\underline{\Delta}_k^{(N)} - 1\right) &= e^{-\left(\mu_k^{(N)} + \underline{\Delta}_{k-1}^{(N)}\right)}\left(-\underline{\Delta}_{k-1}^{(N)} - 1\right), \end{aligned} \tag{4.2.16}$$

where we have used $t_k^{(N)} = \mu_k^{(N)} - \underline{\Delta}_k^{(N)}$; $t_{k-1}^{(N)} = \mu_k^{(N)} + \underline{\Delta}_{k-1}^{(N)}$, or equivalently,

$$e^{\underline{\Delta}_k^{(N)}}\left(\underline{\Delta}_k^{(N)} - 1\right) = e^{-\underline{\Delta}_{k-1}^{(N)}}\left(-\underline{\Delta}_{k-1}^{(N)} - 1\right) \tag{4.2.17}$$

or finally,

$$e^{\underline{\Delta}_k^{(N)}}\left(1 - \underline{\Delta}_k^{(N)}\right) = e^{-\underline{\Delta}_{k-1}^{(N)}}\left(1 + \underline{\Delta}_{k-1}^{(N)}\right). \tag{4.2.18}$$

(4.2.18) is a relationship of the form

$$e^y\left(1 - y\right) = e^{-x}\left(1 + x\right)$$

that, when solved for y (more on this topic is below), produces $y = \underline{\Delta}_k^{(N)}$ if given $x = \underline{\Delta}_{k-1}^{(N)}$. It is clear that if $\underline{\Delta}_{k-1}^{(N)}$ is independent of N, then $\underline{\Delta}_k^{(N)}$ is also independent of N. Furthermore, if $\underline{\Delta}_{k-1}^{(N)}$ is independent of N, then the relationship in (4.2.18) is independent of N. Since $\underline{\eta}_1$ is independent of N and the generating relationship in (4.2.18) is independent of N, we conclude that the infinite sequence $\underline{\eta}_k$, $k = 1, 2, \ldots,$ is independent of N. Thus, the independence of $\underline{\Delta}_1^{(N)}$ from N causes all of the other half steps to be independent of N as well. This is a very important point.

It is now clear that (4.2.18) is the Nitadori sequence generator equation and the initial condition to be used with it is $\underline{\eta} = 1$. It is also clear that the solution to the N-level, exponential MMSE quantizer is found by truncating the Nitadori sequence to the first N terms and that these terms are equal to the set of N half steps that uniquely specify the quantizer.

Generating $\underline{\eta}_k$, $k = 2, 3, \ldots,$ using (4.2.17)**.** Since $\underline{\Delta}_l^{(N)} = \underline{\eta}_l$, $l = 1, 2, \ldots, N$, for each $2 \leq k \leq N$, solving for $\underline{\Delta}_k^{(N)}$ given $\underline{\Delta}_{k-1}^{(N)}$ in (4.2.17) is equivalent to solving for $\underline{\eta}_k$ given $\underline{\eta}_{k-1}$. To solve for $\underline{\eta}_k$, we multiply both sides of (4.2.17) by e^{-1} to get

$$e^{\underline{\eta}_k - 1}\left(\underline{\eta}_k - 1\right) = e^{-\underline{\eta}_{k-1} - 1}\left(-\underline{\eta}_{k-1} - 1\right). \tag{4.2.19}$$

Each side of (4.2.19) is of the form

$$z(W) \triangleq We^{W}, \tag{4.2.20}$$

where for the left hand side of (4.2.19), $W = \underline{\eta}_k - 1$ and for the right hand side of (4.2.19), $W = -\underline{\eta}_{k-1} - 1$. Since the exponential pdf is decreasing, intuitively, it is clear that $\underline{\eta}_k \leq \underline{\eta}_{k-1} \leq \underline{\eta}_1 = 1$ for all $k \geq 2$, and thus, $\underline{\eta}_k \in [0, 1]$ for all $k \geq 1$. Then for the left hand side of (4.2.19), $W \in [-1, 0]$ and for the right hand side of (4.2.19), $W \in [-2, -1]$. Given the value of $\underline{\eta}_{k-1}$, a unique solution for $\underline{\eta}_k$ on the left hand side of (4.2.19) is guaranteed since, for $W \in [-1, 0]$, the function $Z(W)$ is $1 - 1$. A plot of $z(W)$ is shown in Figure 4.1, along with the corresponding W values for $\underline{\eta}_1$, $\underline{\eta}_2$ and $\underline{\eta}_3$, as graphical examples of how successive values of the Nitadori sequence are generated once we know $\underline{\eta}_1$. Also, it can be shown (as seen in this figure) that the W's (as given by $-1 - \underline{\eta}_{k-1}$ and $-1 + \underline{\eta}_k$) are converging to -1 or, equivalently, that the $\underline{\eta}_k$'s are converging towards 0 as k increases.

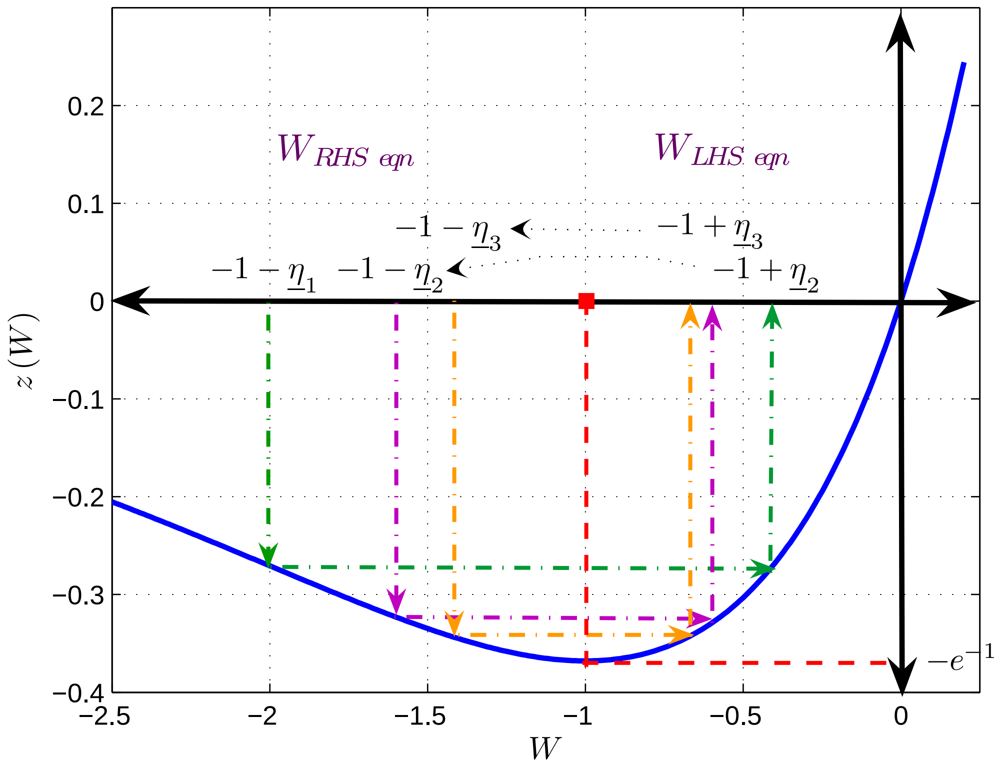


Figure 4.1: Illustration of the function $z(W) = We^{W}$. Also shown are the values of W corresponding to the values of $\underline{\eta}_1$, $\underline{\eta}_2$, $\underline{\eta}_3$.

Recap. To recap, in the case of the MMSE quantization of the exponential source, the initial value and the Nitadori recursion are given by

$$\underline{\eta}_1 = 1 \tag{4.2.21}$$

$$e^{\underline{\eta}_k}\left(1-\underline{\eta}_k\right) = e^{-\underline{\eta}_{k-1}}\left(1+\underline{\eta}_{k-1}\right) \tag{4.2.22}$$

for $k = 2, 3, \ldots$. (For a table of these values, refer to Table 3.1 in Chapter III. To see a plot of these values, go to Figure 2.6 in Chapter II.) Figure 4.2 is a visualization of the half steps of an optimal exponential quantizer where the half step values equal the Nitadori sequence values.

MMSE quantizer performance from the Nitadori sequence $\underline{\eta}_k$. Nitadori also showed that from the sequence $\underline{\eta}_k$, the exact mean-squared error distortion of any optimal N-level quantizer designed for the exponential source is given by $\underline{\eta}_N^2$. (See Chapter II, Section 2.7, (2.7.17).) His derivation (not given in Chapter II) is repeated below for reference.

$$\begin{aligned}
D(N) &= \sum_{i=1}^{N} \int_{t_i^{(N)}}^{t_i^{(N)}+\Delta_i^{(N)}} \left(x-\mu_i^{(N)}\right)^2 f(x)\,dx \\
&\overset{IBP}{=} \sum_{i=1}^{N} \left(x-\mu_i^{(N)}\right)^2 Q(x)\Big|_{t_i^{(N)}}^{t_i^{(N)}+\Delta_i^{(N)}} - \int_{t_i^{(N)}}^{t_i^{(N)}+\Delta_i^{(N)}} 2\left(x-\mu_i^{(N)}\right) Q(x)\,dx \\
&= \sum_{i=1}^{N} -\left(x-\mu_i^{(N)}\right)^2 f(x)\Big|_{t_i^{(N)}}^{t_i^{(N)}+\Delta_i^{(N)}} + \int_{t_i^{(N)}}^{t_i^{(N)}+\Delta_i^{(N)}} 2\left(x-\mu_i^{(N)}\right) f(x)\,dx \\
&= \sum_{i=1}^{N} -\left[\left(t_i^{(N)}+\Delta_i^{(N)}-\mu_i^{(N)}\right)^2 f\left(t_i^{(N)}+\Delta_i^{(N)}\right) - \left(t_i^{(N)}-\mu_i^{(N)}\right)^2 f\left(t_i^{(N)}\right)\right] + 0 \\
&= \sum_{i=1}^{N} -\left[\left(\underline{\underline{\Delta}}_{i-1}^{(N)}\right)^2 f\left(t_i^{(N)}+\Delta_i^{(N)}\right) - \left(\underline{\underline{\Delta}}_i^{(N)}\right)^2 f\left(t_i^{(N)}\right)\right] \quad \left(\text{n.n. cond.; } \underline{\underline{\Delta}}_0^{(N)} \overset{\triangle}{=} \infty\right) \\
&= \sum_{i=1}^{N} -\left(\underline{\underline{\Delta}}_{i-1}^{(N)}\right)^2 f\left(t_{i-1}^{(N)}\right) + \left(\underline{\underline{\Delta}}_i^{(N)}\right)^2 f\left(t_i^{(N)}\right) \\
&= \sum_{i=2}^{N} -\left(\underline{\underline{\Delta}}_{i-1}^{(N)}\right)^2 f\left(t_{i-1}^{(N)}\right) + \left(\underline{\underline{\Delta}}_i^{(N)}\right)^2 f\left(t_i^{(N)}\right) - \left(\underline{\underline{\Delta}}_1^{(N)}\right)^2 f\left(t_0^{(N)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=2}^{N} -\underline{\Delta}_{i-1}^{(N)^2} f\left(t_{i-1}^{(N)}\right) + \left(\underline{\Delta}_{i}^{(N)}\right)^2 f\left(t_{i}^{(N)}\right) - 0 \quad (\text{since } t_0^{(N)} = \infty) \\
&= \left(\underline{\Delta}_{N}^{(N)}\right)^2 f\left(t_{N}^{(N)}\right) = \left(\underline{\Delta}_{N}^{(N)}\right)^2 \quad (\text{since } t_N^{(N)} = 0) \\
&= \underline{\eta}_N^2.
\end{aligned}
$$

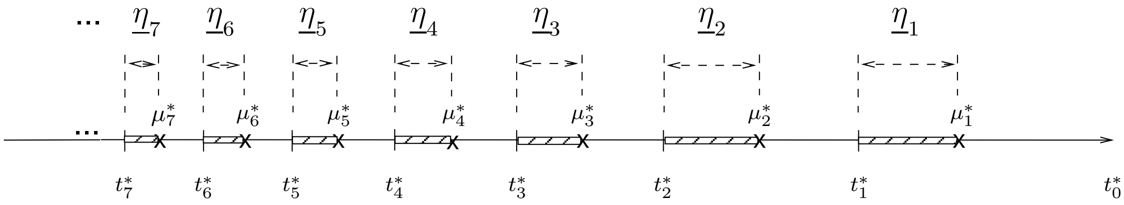


Figure 4.2: Visualization of the Nitadori sequence $\underline{\eta}_k$.

4.3 Rumination: Extending to General Exponential Sources.

Observations made from the Nitadori sequence derivation. It is clear from deriving the Nitadori sequence that the cornerstone to its existence lies in two facts:

1. **Fact #1:** The distance between $\mu_1^{(N)}$ and $t_1^{(N)}$ is fixed regardless of the value of $t_1^{(N)}$ as shown in (4.2.15), and thus $\underline{\Delta}_1^{(N)}$ is independent of N, the number of levels in the MMSE quantizer. This independence from N implies that dependence of $\mu_1^{(N)}$ and $t_1^{(N)}$ on N is exactly the same and by taking the difference between the two values, $\underline{\Delta}_1^{(N)}$ becomes independent of N.

2. **Fact #2:** The Nitadori generator equation is independent of N.

Uniqueness of the Nitadori sequence is assured since for each N, there is only one quantizer designed for an exponential source that minimizes MSE.

Does Fact #1 hold for any other source? Suppose we have an optimal N-level quantizer that has been designed for a source that, while not exponential, still has a differentiable pdf. Let us consider the expression for $\underline{\Delta}_{[t,\infty)}$ for this case, and compare it against the analogous expression shown in (4.2.15) that holds for the exponential case: Returning to (4.2.11), and knowing that $\Delta \to \infty$, we have

$$
\mu_{[t,\infty)} = \frac{tQ(t)}{Q(t)} + \frac{\int_t^\infty Q(x)\,dx}{Q(t)} = t + \frac{\int_t^\infty Q(x)\,dx}{Q(t)}
$$

or

$$\underline{\Delta}_{[t,\infty)} = \frac{\int_t^\infty Q\left(x\right)dx}{Q\left(t\right)}. \tag{4.3.23}$$

From (4.3.23), it is clear that for any fixed $t \geq 0$ and for any arbitrary source with a differentiable pdf, $\underline{\Delta}_{[t,\infty)} = \mu_{[t,\infty)} - t$ is not independent of the threshold t since the key effect (or any scaled version of the key effect) does not hold. Furthermore, suppose we allow $t \to \infty$. Using L'Hopital's rule, for a source with a differentiable pdf, we have

$$\lim_{t\to\infty} \underline{\Delta}_{[t,\infty)} = \lim_{t\to\infty} \frac{\int_t^\infty Q\left(x\right)dx}{\int_t^\infty f\left(u\right)du} \overset{L'H}{=} \lim_{t\to\infty} \frac{-Q\left(t\right)}{-f\left(t\right)} = \lim_{t\to\infty} \frac{Q\left(t\right)}{f\left(t\right)}. \tag{4.3.24}$$

In (4.3.24), we see that as t grows large, the value of $\underline{\Delta}_{[t,\infty)}$ converges to a non-zero constant if and only if asymptotically $Q\left(t\right)$ behaves like $f\left(t\right)$, i.e., there exists a constant $0 < c < \infty$ such that when t is large, $Q\left(t\right) \approx cf\left(t\right)$, or, in words, that a (scaled) version of the key effect holds for large t. Calling this phenomenon *tail exponentiality* (when in the tail region of the source, the pdf behaves like an exponential pdf), it is clear that tail exponentiality and not perfect exponentiality is required for (4.3.24) to equal a non-zero constant since the key effect holds only for exponential sources. In general, tail exponentiality is not property of a source and its pdf, for example, the Gaussian source. However, it is well-known that for this particular source, $\frac{Q(t)}{f(t)} \approx \frac{1}{t}$ when t is large, and this suggests looking for a limit to $\underline{\Delta}_{[t,\infty)}t$ as $t \to \infty$. By extension, these observations lead us to consider the behavior of $\underline{\Delta}_{[t,\infty)}t^{p-1}$ as $t \to \infty$ for sources belonging to the GE-family.

4.4 A Property of General Exponential Sources.

Recall that we have defined one-sided, zero mean general exponential (GE) source densities to have the form

$$f\left(x\right) = c_p\, e^{-\frac{x^p}{p}}, \text{ for all } x \geq 0,$$

where $c_p = \frac{1}{\int_0^\infty \exp(-\frac{x^p}{p})\,dx}$ and $p \geq 1$. In order to study $\lim_{t\to\infty} \underline{\Delta}_{[t,\infty)}t^{p-1}$, we start by looking at the expression for the centroid of an arbitrary interval when the source is from the GE-family.

Proposition IV.2. *Let $t > 0$ be any positive number and let $\Delta > 0$. For any GE-source X, the conditional mean of X given $X \in [t, t+\Delta)$ has the form*

$$\mu_{[t,t+\Delta)} = \frac{\frac{f(t)}{t^{p-2}} - \frac{f(t+\Delta)}{(t+\Delta)^{p-2}}}{Q(t) - Q(t+\Delta)} + \frac{(-p+2)\int_t^{t+\Delta} x^{-p+1} f(x)\,dx}{Q(t) - Q(t+\Delta)}.$$

Proof. Using

$$P_{[t,t+\Delta)} \triangleq \int_t^{t+\Delta} f(x)\,dx,$$

we start with a general expression for the centroid of $[t, t+\Delta)$

$$P_{[t,t+\Delta)}\mu_{[t,t+\Delta)} = \int_t^{t+\Delta} xf(x)\,dx = \int_t^{t+\Delta} \frac{x^{p-1}f(x)}{x^{p-2}}dx$$

and use integration by parts (IBP),

$$\begin{aligned} u &= x^{-p+2} & du &= (-p+2)\,x^{-p+1}dx \\ v &= -f(x) & dv &= x^{p-1}f(x)\,dx, \end{aligned}$$

to get

$$\begin{aligned} P_{[t,t+\Delta)}\mu_{[t,t+\Delta)} &= -\frac{f(x)}{x^{p-2}}\bigg|_t^{t+\Delta} + (-p+2)\int_t^{t+\Delta} x^{-p+1} f(x)\,dx \\ &= \frac{f(t)}{t^{p-2}} - \frac{f(t+\Delta)}{(t+\Delta)^{p-2}} + (-p+2)\int_t^{t+\Delta} x^{-p+1} f(x)\,dx. \end{aligned}$$

Since

$$P_{[t,t+\Delta)} = \int_t^{t+\Delta} f(x)\,dx = Q(t) - Q(t+\Delta),$$

when $y \geq 0$, we have

$$\mu_{[t,t+\Delta)} = \frac{\frac{f(t)}{t^{p-2}} - \frac{f(t+\Delta)}{(t+\Delta)^{p-2}}}{Q(t) - Q(t+\Delta)} + \frac{(-p+2)\int_t^{t+\Delta} x^{-p+1} f(x)\,dx}{Q(t) - Q(t+\Delta)}.$$

$\blacksquare$

Remarks: Highlighting the role of the key effect – Comparing Proposition IV.2 to (4.2.12)**.** The expression in Proposition IV.2, since it applies to any GE-source, is more general than the one in (4.2.12) which only holds for an exponential source and it is clear that when $p = 1$, Proposition IV.2 reduces exactly to (4.2.12). The case when $p > 1$ is more interesting. Comparing term-by-term using the expressions shown in Table 4.1, we see from the first term comparison, it appears that $\frac{f(t)}{t^{p-2}}$ from Proposition IV.2 corresponds to $Q(t)$ from (4.2.12).

Table 4.1: Comparing the terms from Proposition IV.2 to (4.2.11) and (4.2.12).

	Proposition IV.2 GE-sources	(4.2.11) Exp. source	(4.2.12) Exp. source key effect applied
First term	$\frac{\frac{f(t)}{t^{p-2}} - \frac{f(t+\Delta)}{(t+\Delta)^{p-2}}}{Q(t)-Q(t+\Delta)}$	$\frac{tQ(t)-(t+\Delta)Q(t+\Delta)}{Q(t)-Q(t+\Delta)}$	$\frac{tQ(t)-(t+\Delta)Q(t+\Delta)}{Q(t)-Q(t+\Delta)}$
Second term	$\frac{(-p+2)\int_t^{t+\Delta} \frac{f(x)}{x^{p-1}}dx}{Q(t)-Q(t+\Delta)}$	$\frac{\int_t^{t+\Delta} Q(x)dx}{Q(t)-Q(t+\Delta)}$	$\frac{Q(t)-Q(t+\Delta)}{Q(t)-Q(t+\Delta)} = 1$

When comparing second terms, the correspondence is more difficult to see. However, we will comment that use of the key effect on (4.2.11) simplified the second term to the value of 1 as seen in (4.2.12). Since the key effect holds only when $p = 1$, we will pay particular attention to what happens to the second term in Proposition IV.2 when we apply a substitute for the key effect that holds not only when $p = 1$, but when $p > 1$ as well.

4.5 Key Effect Substitute.

Several times now, we have pointed out that there appears to be a connection between $Q(x)$ and $\frac{f(x)}{x^{p-1}}$ when the source pdf belongs to the GE-family. It turns out that this connection can be expressed by a well-known asymptotic relationship between the tail function $Q(x)$ of a GE-source and its pdf $f(x)$. We state this approximation below.

Asymptotic tail function expression for general exponential distributions – The asymptotic key effect. For any $p \geq 1$, the tail function $Q(x)$ belonging to a GE-source can be expressed asymptotically [22] as: For large $x > 0$,

$$Q(x) = \frac{f(x)}{x^{p-1}} \times \left[1 + (-p+1)\left(\frac{1}{x^p} + (-2p+1)\frac{1}{x^{2p}} + O\left(\frac{1}{x^{3p}}\right)\right)\right].^{6} \qquad (4.5.25)$$

(See Appendix D for more details on this expansion.)

Equation (4.5.25) appears to be a good choice as an asymptotic form of the key effect of the memoryless property since for $x >> 0$, $Q(x) \approx \frac{f(x)}{x^{p-1}}$ when $p > 1$ and when $p = 1$, the source is exponential and (4.5.25) becomes the key effect $Q(x) = f(x)$ for large values of x. Since (4.5.25) is only a valid approximation of the tail function for general exponential sources (specifically when $p > 1$), the decision to use the approximation in (4.5.25) limits the scope of our work to asymptotic results that hold only when $x > 0$ is large and $p > 1$.

Impact of the key effect substitute on the conditional mean for GE-sources: Asymptotic expression for $\mu_{[t,\infty)} - t$ leading to the analogue to Nitadori initial sequence value. Just as in the Nitadori re-derivation, where the key effect was applied to $\mu_{[t,\infty)}$ for an exponential source, with $\Delta \to \infty$ to obtain $\mu_{[t,\infty)} - t = 1$ for all $t \geq 0$, we use the asymptotic tail function expression in (4.5.25) to simplify the expression in Proposition IV.2 to find an analogous asymptotic expression that holds for GE-source distributions.

Proposition IV.3. *Let $t > 0$. For any GE-source X, the conditional mean of X given $X \in [t, \infty)$ has the form*

$$\begin{aligned}\mu_{[t,\infty)} &= t\left[1 + t^{-p} + (-p+1)\left(2t^{-2p} + O\left(t^{-3p}\right)\right)\right] \\ &= t + \frac{1}{t^{p-1}} + \frac{2(-p+1)}{t^{2p-1}} + (-p+1)\,O\left(\frac{1}{t^{3p-1}}\right).\end{aligned}$$

[6] **Big O Notation conventions.** Let $f(x)$ and $g(x)$ be real-valued functions. When we say, "$f(x)$ is $O(g(x))$," we are using the conventional definition that states that there exists $x_0 > 0$ and $M > 0$ such that $|f(x)| \leq M|g(x)|$ when $x > x_0$. Suppose we now have two real functions $f_y(x)$ and $g_y(x)$ that have an implicit dependence on another real variable y. When we say, "$f_y(x)$ is $O_y(g_y(x))$," we mean that there exists $y_0 > 0$ and $M > 0$ such that $|f_y(x)| \leq M|g_y(x)|$ when $y > y_0$.

Proof. Letting $\Delta \to \infty$, Proposition IV.2 implies

$$\mu_{[t,\infty)} = \frac{\frac{f(t)}{t^{p-2}} + (-p+2)\int_t^\infty x^{-p+1} f(x)\,dx}{Q(t)}.$$

Since

$$\int_t^\infty x^{-p+1} f(x)\,dx = t^{-2p+2} f(t) + (-2p+2)\,t^{-3p+2} f(t) + (-2p+2)(-3p+2) \times$$
$$t^{-4p+2} f(t) + (-2p+2)(-3p+2)(-4p+2)\int_t^\infty x^{-4p+1} f(x)\,dx$$

and

$$Q(t) = \frac{f(t)}{t^{p-1}} \times \left[1 + (-p+1)\left(\frac{1}{t^p} + (-2p+1)\frac{1}{t^{2p}} + O\left(\frac{1}{t^{3p}}\right)\right)\right],$$

we have

$$\begin{aligned}
\mu_{[t,\infty)} &= f(t)\,t^{-p+2}\Bigg[1 + t^{-p} + (-2p+2)\,t^{-2p} + (-2p+2)(-3p+2)\,t^{-3p} + \\
&\quad (-2p+2)(-3p+2)(-4p+2)\frac{t^{p-2}}{f(t)}\int_t^\infty x^{-4p+1} f(x)\,dx\Bigg] \times \frac{t^{p-1}}{f(t)} \times \\
&\quad \left[1 + (-p+1)\left(\frac{1}{t^p} + (-2p+1)\frac{1}{t^{2p}} + O\left(\frac{1}{t^{3p}}\right)\right)\right]^{-1} \\
&= f(t)\,t^{-p+2}\Bigg[1 + t^{-p} + (-2p+2)\,t^{-2p} + (-2p+2)(-3p+2)\,t^{-3p} + \\
&\quad (-2p+2)(-3p+2)(-4p+2)\frac{t^{p-2}}{f(t)} \cdot f(t)\,t^{-p+2} O\left(t^{-4p}\right)\Bigg] \times \frac{t^{p-1}}{f(t)} \times \\
&\quad \left[1 + (-p+1)\left(\frac{1}{t^p} + (-2p+1)\frac{1}{t^{2p}} + O\left(\frac{1}{t^{3p}}\right)\right)\right]^{-1} \\
&= t\Bigg[1+t^{-p}+(-2p+2)t^{-2p}+(-2p+2)(-3p+2)t^{-3p}+(-2p+2)(-3p+2)\times \\
&\quad (-4p+2)\,O\left(t^{-4p}\right)\Bigg] \times \left[1 + (-p+1)\left(\frac{1}{t^p} + (-2p+1)\frac{1}{t^{2p}} + O\left(\frac{1}{t^{3p}}\right)\right)\right]^{-1}
\end{aligned}$$

since $\int_t^\infty x^{-4p+1} f(x)\,dx$ is $f(t)\,O\left(t^{-5p+2}\right)$ and is also $f(t)\,t^{-p+2} O\left(t^{-4p}\right)$ because

$$\int\limits_t^\infty x^{-4p+1} f(x)\,dx = \int\limits_t^\infty x^{-5p+2} \cdot x^{p-1} f(x)\,dx$$

$$\leq t^{-5p+2} \int\limits_t^\infty x^{p-1} f(x)\,dx = t^{-5p+2} \left. -f(x)\right|_t^\infty = t^{-5p+2} f(t)\,.$$

Using long division, we finally have

$$\mu_{[t,\infty)} = t\left[1 + t^{-p} + (-p+1)\left(2t^{-2p} + O\left(t^{-3p}\right)\right)\right].$$

■

Proposition IV.3 shows that the centroid $\mu_{[t,\infty)}$ of the half open interval $[t,\infty)$ can be expressed as $t + t^{p-1} + O\left(t^{2p-1}\right)$ which is function of both t and the source parameter p. Thus, it is clear that the distance between the centroid $\mu_{[t,\infty)}$ and t can be expressed as $t^{p-1} + O\left(t^{2p-1}\right)$. The next corollary is a direct consequence of this fact and establishes that the product $\underline{\Delta}_{[t,\infty)} t^{p-1} = \left(\mu_{[t,\infty)} - t\right) t^{p-1}$ converges to a limit as $t \to \infty$.

Corollary IV.4. $\lim_{t\to\infty} \left(\mu_{[t,\infty)} - t\right) t^{p-1} = 1$.
$\left(\mu_{[t,\infty)} - t\right) t^{p-1} = 1 + O\left(\frac{1}{t^p}\right)$.

Proof. From Proposition IV.3 and $t > 0$,

$$\begin{aligned}\left(\mu_{[t,\infty)} - t\right) t^{p-1} &= \left(\frac{1}{t^{p-1}} + \frac{2(-p+1)}{t^{2p-1}} + (-p+1)\,O\left(\frac{1}{t^{3p-1}}\right)\right) t^{p-1} \\ &= \left(1 + \frac{2(-p+1)}{t^p} + (-p+1)\,O\left(\frac{1}{t^{2p}}\right)\right),\end{aligned} \tag{4.5.26}$$

we observe that

$$\lim_{t\to\infty} \left(\mu_{[t,\infty)} - t\right) t^{p-1} = 1.$$

■

Corollary IV.4 presents us with a general, asymptotic version of the exponential source property $\underline{\Delta}_{[t,\infty)} = \mu_{[t,\infty)} - t = 1$, $t \geq 0$, a property which was shown using the key effect and the fact that the effect is valid for all values $t \geq 0$. Corollary IV.4, by comparison, relies on the GE-sources' asymptotic key effect (the asymptotic tail function approximation in (4.5.25)) which is only valid when $t > 0$ is large. The consequences of using this asymptotic key effect are easy to see in the expression shown in Corollary IV.4. When $p = 1$, Corollary IV.4 reduces to $\lim_{t\to\infty} \underline{\Delta}_{[t,\infty)} = 1$ which is trivially true since $\underline{\Delta}_{[t,\infty)} = 1$ for all $t \geq 0$ when the source is exponential. When $p > 1$, i.e., when we consider general exponential sources other than the exponential distribution, Corollary IV.4 shows us that another effect of using the asymptotic tail function in (4.5.25) is to append a polynomial multiplicative factor of t^{p-1} to $\underline{\Delta}_{[t,\infty)} = \mu_{[t,\infty)} - t$ so that $\left(\mu_{[t,\infty)} - t\right) t^{p-1} \approx 1$ only when t is large. Thus, we see that while the statement in Corollary IV.4 is more general (since it applies to sources other than exponential), it is a weaker statement regarding the behavior of $\underline{\Delta}_{[t,\infty)} t^{p-1}$ since it is a remark regarding limiting behavior in t rather than a fact that holds for arbitrary values of $t \geq 0$.

4.6 Initial Limiting Value of $\underline{\alpha}_{1,p}^{(N)}$ as $N \to \infty$ for GE-sources.

Here we discuss our asymptotic generalization to the initial value $\underline{\eta}_1 = \underline{\Delta}_1^{(N)}$ of the Nitadori sequence. Choose any GE-source by fixing $p \geq 1$. Consider now a sequence of MMSE quantizers (indexed by the number of levels N) that have all been designed for this GE-source. Recalling that the quantization interval containing the support threshold $t_1^{(N)}$ is $[t_1^{(N)}, \infty)$ and knowing that as N increases, the support threshold $t_1^{(N)}$ corresponding to each quantizer in the sequence also increases, Corollary IV.4 tells us that $\underline{\Delta}_1^{(N)} t_1^{(N)^{p-1}} = \left(\mu_{[t_1^{(N)},\infty)} - t_1^{(N)}\right) \left(t_1^{(N)}\right)^{p-1} \to 1$ as $N \to \infty$. Using the definition in (4.1.6), we have just demonstrated that for each $p \geq 1$,

$$\lim_{N\to\infty} \underline{\alpha}_{1,p}^{(N)} = \lim_{N\to\infty} \underline{\Delta}_1^{(N)} (t_1^{(N)})^{p-1} = 1$$

and this is the initial value of the limiting sequence in Part 1 of Theorem IV.1.

4.7 More Properties of GE-sources.

At this point in the discussion, if we were to strictly follow the Nitadori rederivation, we would begin considering asymptotic MMSE quantization of GE-sources. However, because our derivation becomes more complicated than the Nitadori derivation and to keep the notation as simple as possible for as long as possible, we now prove several more relationships in preparation for further discussion on asymptotic

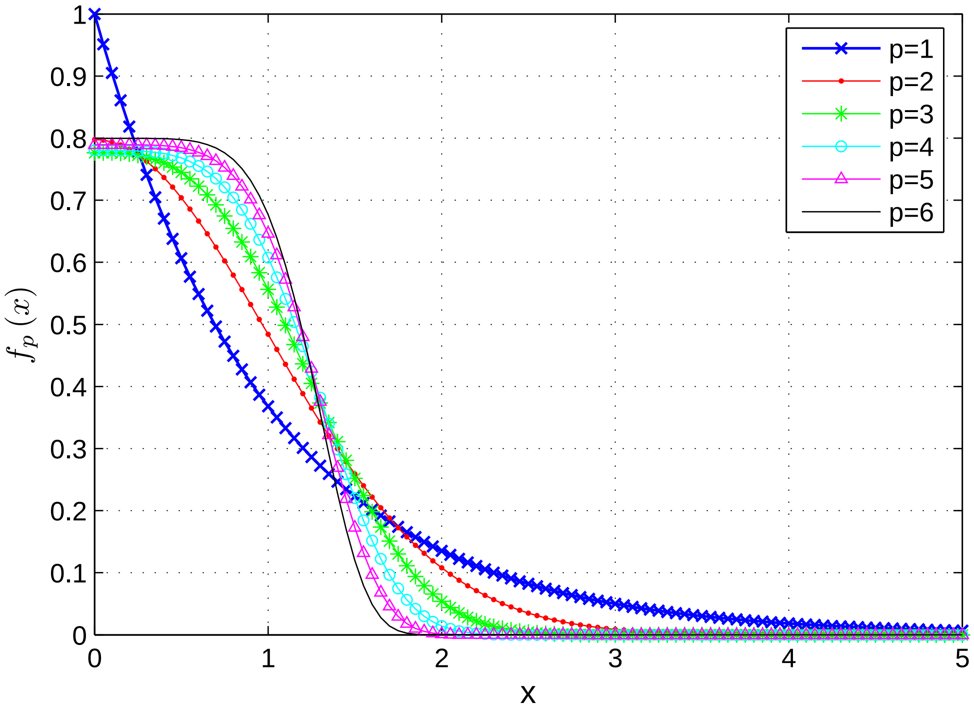


Figure 4.3: Plot of various GE-source pdfs with $p = 1, 2, 3, 4, 5, 6$.

optimal quantization of GE-sources, bearing in mind that the ultimate goal is, for each $p > 1$, to derive expressions for $\lim_{N\to\infty} \underline{\alpha}^{(\tau)}_{k|j,p}$ when $k > 1$. We note that these relationships are source properties only and hence, they do not rely on any construct related to quantization.

The lemma below states a relationship that we use to prove Theorem IV.1.

Lemma IV.5. *Let $t > s \geq 0$. Then for any GE-source pdf,*

1.

$$\frac{f(s)}{f(t)} = e^{-\frac{1}{p}(s^p - t^p)}$$

2.

$$\frac{f(s)}{f(t)} = \frac{\frac{1}{t^{p-2}} + (-p+2)\,\frac{Int(t)}{f(t)} - \frac{Q(t)}{f(t)}\mu_{[s,t)}}{\frac{1}{s^{p-2}} + (-p+2)\,\frac{Int(s)}{f(s)} - \frac{Q(s)}{f(s)}\mu_{[s,t)}},$$

where

$$Int(s) \triangleq \int\limits_{s}^{\infty} x^{-p+1} f(x)\,dx.$$

Proof.

Part 1. This proof is trivial.

$$\frac{f(s)}{f(t)} = \frac{c\,e^{-\frac{1}{p}s^p}}{c\,e^{-\frac{1}{p}t^p}} = \frac{e^{-\frac{1}{p}s^p}}{e^{-\frac{1}{p}t^p}} = e^{-\frac{1}{p}(s^p - t^p)}.$$

Part 2. To prove Lemma IV.5, part 2, we just re-express the relationship in Proposition IV.2 by bringing out the term $\frac{f(s)}{f(t)}$ to one side of the equality.

Let $k \geq 2$. Choose $t > s > 0$. From Proposition IV.2, we obtain

$$\begin{aligned}\mu_{[s,t)} &= \frac{\frac{f(s)}{s^{p-2}} - \frac{f(t)}{t^{p-2}} + (-p+2)\int_s^t x^{-p+1} f(x)\,dx}{Q(s) - Q(t)} \\ &= \frac{\frac{f(s)}{s^{p-2}} - \frac{f(t)}{t^{p-2}}}{Q(s) - Q(t)} + \frac{(-p+2)\left(\mathrm{Int}(s) - \mathrm{Int}(t)\right)}{Q(s) - Q(t)} \\ &= \frac{\left[\frac{f(s)}{s^{p-2}} + (-p+2)\,\mathrm{Int}(s)\right]}{Q(s) - Q(t)} - \frac{\left[\frac{f(t)}{t^{p-2}} + (-p+2)\,\mathrm{Int}(t)\right]}{Q(s) - Q(t)}. \qquad (4.7.27)\end{aligned}$$

Manipulating (4.7.27), we have

$$\mu_{[s,t)}\left(Q\left(s\right)-Q\left(t\right)\right)=\left[\frac{f\left(s\right)}{s^{p-2}}+\left(-p+2\right)\mathrm{Int}\left(s\right)\right]-\left[\frac{f\left(t\right)}{t^{p-2}}+\left(-p+2\right)\mathrm{Int}\left(t\right)\right]$$

which is equivalent to

$$\left[\frac{f\left(s\right)}{s^{p-2}}+\left(-p+2\right)\mathrm{Int}\left(s\right)-Q\left(s\right)\mu_{[s,t)}\right]=\left[\frac{f\left(t\right)}{t^{p-2}}+\left(-p+2\right)\mathrm{Int}\left(t\right)-Q\left(t\right)\mu_{[s,t)}\right]$$

which is also equivalent to

$$\frac{f\left(s\right)}{f\left(t\right)}=\frac{\frac{1}{t^{p-2}}+\left(-p+2\right)\frac{\mathrm{Int}(t)}{f(t)}-\frac{Q(t)}{f(t)}\mu_{[s,t)}}{\frac{1}{s^{p-2}}+\left(-p+2\right)\frac{\mathrm{Int}(s)}{f(s)}-\frac{Q(s)}{f(s)}\mu_{[s,t)}}.$$

■

The next lemma simply states that the centroid of a cell is always less than or equal to the midpoint of the cell for source pdfs that behave *nicely*.

Lemma IV.6. *Suppose f is a non-increasing, first order differentiable pdf with support on the non-negative reals. Then for any interval $[s,t)$ with $t>s\geq 0$, the centroid $\mu_{[s,t)}$ is less than or equal to the midpoint $m_{[s,t)}\overset{\triangle}{=}\frac{1}{2}\left(s+t\right)$.*

Proof. We show that $\mu_{[s,t)}-m_{[s,t)}\leq 0$. With $P_{[s,t)}=\int_s^t f\left(x\right)dx$,

$$\begin{aligned}P_{[s,t)}\mu_{[s,t)}&=\int_s^t xf\left(x\right)dx=\int_s^t m_{[s,t)}f\left(x\right)dx+\int_s^t\left(x-m_{[s,t)}\right)f\left(x\right)dx\\&=m_{[s,t)}\int_s^t f\left(x\right)dx+\int_s^t\left(x-m_{[s,t)}\right)\left(f\left(m_{[s,t)}\right)+R_0\left(x,m_{[s,t)}\right)\right)dx\\&=m_{[s,t)}P_{[s,t)}+\int_s^t\left(x-m_{[s,t)}\right)\left(f\left(m_{[s,t)}\right)+f'\left(\gamma_x\right)\left(x-m_{[s,t)}\right)\right)dx,\end{aligned}$$

where $R_0\left(x, m_{[s,t)}\right) = f'\left(\gamma_x\right)\left(x - m_{[s,t)}\right)$ is the 0th order Taylor's series remainder and $\gamma_x \in [s,t)$. Then

$$\begin{aligned}
P_{[s,t)}\left(\mu_{[s,t)} - m_{[s,t)}\right) &= \int\limits_s^t \left(x - m_{[s,t)}\right)\left(f\left(m_{[s,t)}\right) + f'\left(\gamma_x\right)\left(x - m_{[s,t)}\right)\right) dx \\
&= f\left(m_{[s,t)}\right) \int\limits_s^t \left(x - m_{[s,t)}\right) dx + \int\limits_s^t f'\left(\gamma_x\right)\left(x - m_{[s,t)}\right)^2 dx \\
&= 0 + \int\limits_s^t f'\left(\gamma_x\right)\left(x - m_{[s,t)}\right)^2 dx \\
&\leq 0
\end{aligned}$$

since f is non-increasing in $[s,t)$. ■

4.8 Foundation for the Recursion Derivation: Conditionally Optimal Quantizers.

Having established the initial value $\lim_{N\to\infty} \underline{\alpha}_{1,p}^{(N)} = 1$, it is time to concentrate on deriving an asymptotic recursion relationship between

$$\lim_{N\to\infty} \underline{\alpha}_{k,p}^{(N)} = \lim_{N\to\infty} \underline{\Delta}_k^{(N)} (t_k^{(N)})^{p-1}$$

and

$$\lim_{N\to\infty} \underline{\alpha}_{k-1,p}^{(N)} = \lim_{N\to\infty} \underline{\Delta}_{k-1}^{(N)} (t_{k-1}^{(N)})^{p-1}$$

when $k \geq 2$.

Fix $p \geq 1$. Recall that, for a fixed cell index $k \geq 1$, in order to ascertain the limiting behavior of $\lim_{N\to\infty} \underline{\alpha}_{k|j,p}^{(\tau)}$, we need only consider the behavior of the first k quantization cells of quantizers belonging to a sequence of optimal quantizers that is indexed by N, and because for any value of N, the first k quantization cells of an optimal quantizer are exactly the same as a k-cell optimal quantizer designed for the conditional distribution of X given $X \in [t_k^{(N)}, \infty)$, which we have already referred to, in the introduction, as a *conditionally optimal quantizer*, it makes sense to use conditionally optimal quantizers to study the behavior of $\underline{\Delta}_m^{(N)} (t_m^{(N)})^{p-1}$ for cells

$m = 1, 2, \ldots, k$.

Conditionally optimal τ, j quantizers and some of their properties. As illustrated in Figure 4.4, given a source X with pdf $f(x)$, $x \geq 0$, for each integer $j > 0$ and each $\tau \geq 0$, we define a conditionally optimal (or just optimal, for short) τ, j quantizer

$$q_{\tau,j} \overset{\triangle}{=} \left(t_{0|j}, t_{1|j}, t_{2|j}, \ldots, t_{j|j}; \mu_{1|j}, \mu_{2|j}, \ldots, \mu_{j|j}\right),$$

to be an optimal, j-level quantizer designed for a conditional source with the conditional distribution

$$f_{X|X\geq\tau}(x) \overset{\triangle}{=} \frac{1}{P_{[\tau,\infty)}} f(x),$$

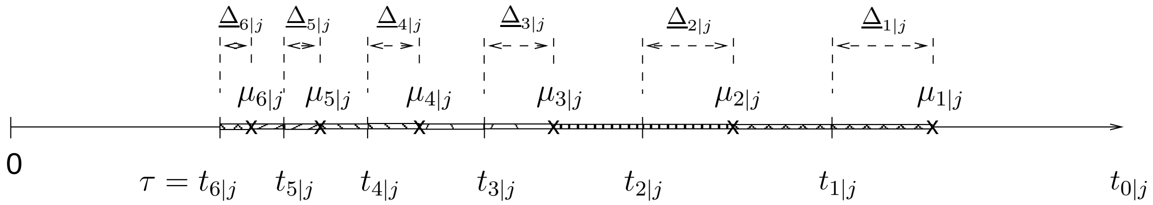
where $t_{i|j}$ and $\mu_{i|j}$ are the lower threshold and reconstruction level belonging to the ith quantization cell of the τ, j quantizer with $\tau = t_{j|j}$ and $t_{0|j} = +\infty$. (Note that while $t_{k|j}$ and $\mu_{k|j}$ depend on the source X (as indicated by the value of p) and on the value of τ, we have omitted this dependence in the notation to facilitate readability.) Hence, the vector of thresholds and reconstruction levels $q_{\tau,j}$ satisfy the optimality criteria for MMSE quantization and, if the source X is GE, then $q_{\tau,j}$ is unique.

As with MMSE quantizers, we make the following definitions for the kth quantization cell of an optimal τ, j quantizer:

$$\begin{aligned}
\underline{\Delta}_{k|j} &\overset{\triangle}{=} \mu_{k|j} - t_{k|j} \quad \text{(the half step)} \\
\Delta_{k|j} &\overset{\triangle}{=} t_{k-1|j} - t_{k|j} \quad \text{(the step size)} \\
\underline{\alpha}^{(\tau)}_{k|j,p} &\overset{\triangle}{=} \underline{\Delta}_{k|j} t_{k|j}{}^{p-1}, \;\; 1 \leq k \leq j \\
\alpha^{(\tau)}_{k|j,p} &\overset{\triangle}{=} \Delta_{k|j} (t_{k|j})^{p-1}, \;\; 1 \leq k \leq j \qquad (4.8.28) \\
\underline{r}^{(\tau)}_{k|j,p} &\overset{\triangle}{=} \frac{\underline{\Delta}_{k-1|j}}{\underline{\Delta}_{k|j}}, \;\; 2 \leq k \leq j.
\end{aligned}$$

We remark that, for any fixed $k \geq 1$, any MMSE scalar quantizer with thresholds $\{t_i^{(N)}\}_{i=0}^{N}$, reconstruction levels $\{\mu_i^{(N)}\}_{i=1}^{N}$, and $N \geq k$ quantization levels "contains" exactly N optimal τ, j quantizers. Let $\tau = t_j^{(N)}$. Then, we have on the right of (and including) $t_j^{(N)}$ an optimal τ, j quantizer, i.e., for each $1 \leq j \leq N$,

$$q_{\tau,j} = (t_0^{(N)}, t_1^{(N)}, t_2^{(N)}, \ldots, t_j^{(N)}; \mu_1^{(N)}, \mu_2^{(N)}, \ldots, \mu_j^{(N)})$$



$\underline{\Delta}_{6|j}$
$\underline{\Delta}_{5|j}$
$\underline{\Delta}_{4|j}$
$\underline{\Delta}_{3|j}$
$\underline{\Delta}_{2|j}$
$\underline{\Delta}_{1|j}$
$\mu_{6|j}$
$\mu_{5|j}$
$\mu_{4|j}$
$\mu_{3|j}$
$\mu_{2|j}$
$\mu_{1|j}$
0
$\tau = t_{6|j}$
$t_{5|j}$
$t_{4|j}$
$t_{3|j}$
$t_{2|j}$
$t_{1|j}$
$t_{0|j}$

Figure 4.4: Example of an optimal $\tau, 6$ quantizer. Note: The hash marks in the picture are intended to illustrate that adjacent upper and lower half steps belonging to neighboring reconstruction cells are equal in size.

is an optimal τ, j quantizer with j-levels.

Working with optimal τ, j quantizers (indexed in τ) instead of optimal quantizers (indexed in N), we can recast Theorem IV.1 as:

Theorem IV.7. *Choose $p \geq 1$ and for each quantization cell k of an optimal τ, j quantizer, define*

$$\underline{\nu}_k \stackrel{\triangle}{=} \lim_{\tau \to \infty} \underline{\alpha}^{(\tau)}_{k|j,p} = \lim_{\tau \to \infty} \underline{\Delta}_{k|j,p} (t_{k|j})^{p-1}$$

if the limit exists.

1. *(a) $\underline{\nu}_1$ exists and is equal to $\underline{\eta}_1$.*

   (b) If $\underline{\nu}_{k-1}$ exists, then $\underline{\nu}_k$ exists and satisfies

$$e^{\underline{\nu}_k} \left(1 - \underline{\nu}_k\right) = e^{-\underline{\nu}_{k-1}} \left(1 + \underline{\nu}_{k-1}\right). \tag{4.8.29}$$

   (c) $\underline{\nu}_k = \underline{\eta}_k$, $k \geq 1$.

2. *For $1 \leq k \leq j-1$, $\underline{r}_k \stackrel{\triangle}{=} \lim_{\tau \to \infty} \underline{r}^{(\tau)}_{k|j,p}$ exists and*

$$\underline{r}_k = \frac{\underline{\eta}_k}{\underline{\eta}_{k+1}}. \tag{4.8.30}$$

Before proving Theorem IV.7, we present a corollary along with its proof. This corollary finishes the proof of Theorem IV.1 by providing the connection between the statement in Theorem IV.7, which is a fact regarding optimal τ, j quantization of GE-sources, and the statement in Theorem IV.1, which is a fact regarding MMSE quantization of GE-sources.

Corollary IV.8. *Choose $p \geq 1$and fix $k \geq 1$. For any sequence of optimal scalar quantizers designed for the GE-source indicated by the value of p, where the sequence is indexed by the number of levels N, the following two facts are true:*

1. $\underline{\alpha}_{k,p}^{(N)} \to \underline{\eta}_k$ *as $N \to \infty$. (Theorem IV.1, Part 1.)*
2. $\frac{\underline{\Delta}_k^{(N)}}{\underline{\Delta}_{k+1}^{(N)}} \to \frac{\underline{\eta}_k}{\underline{\eta}_{k+1}}$ *as $N \to \infty$. (Theorem IV.1, Part 2.)*

Proof. Fix $p \geq 1$. Consider any sequence q_N of optimal, N-level quantizers designed for this GE-source, where the sequence is indexed by N.

Part 1. Choose any $k \geq 1$. Then for all $N \geq k$, q_N contains an optimal τ, k quantizer where $\tau = t_{k|k} = t_k^{(N)}$. Then

$$\underline{\alpha}_{k,p}^{(N)} = \underline{\Delta}_k^{(N)}(t_k^{(N)})^{p-1} = \underline{\Delta}_{k|k}(t_{k|k})^{p-1} = \alpha_{k|k,p}^{(\tau)}$$

when $N \geq k$.

Since letting $N \to \infty$ (for the optimal quantizers) implies that $\tau \to \infty$ (for the *embedded* sequence of optimal τ, j quantizers), using Theorem IV.7, Part 1, we have

$$\lim_{N\to\infty} \underline{\alpha}_{k,p}^{(N)} = \lim_{\tau\to\infty} \alpha_{k|k,p}^{(\tau)} = \underline{\nu}_k \overset{Thm\ IV.7,P.1}{=} \underline{\eta}_k.$$

■

Part 2. Fix $k \geq 1$ and choose $j = k+1$. Since for any $N > k+1$, q_N contains an optimal $\tau, k+1$ quantizer, we have a sequence of optimal $\tau, k+1$ quantizers, indexed by N, when $N \geq k+1$. With

$$\underline{\Delta}_k^{(N)} = \underline{\Delta}_{k|j} \quad \text{and} \quad \underline{\Delta}_{k+1}^{(N)} = \underline{\Delta}_{k+1|j},$$

and the fact that as $N \to \infty$, it is clear that $\tau \to \infty$ since $\tau = t_{k+1}^{(N)}$ and $t_{k+1}^{(N)} \to \infty$ if $N \to \infty$, we have

$$\lim_{N\to\infty} \frac{\underline{\Delta}_k^{(N)}}{\underline{\Delta}_{k+1}^{(N)}} = \lim_{\tau\to\infty} \frac{\underline{\Delta}_{k|j}}{\underline{\Delta}_{k+1|j}} = \lim_{\tau\to\infty} \underline{r}_{k|j,p}^{(\tau)} \overset{Thm\ IV.7,P.2}{=} \underline{r}_k.$$

■

A property of optimal τ, j quantizers designed for any source. Returning to the business of proving Theorem IV.7, the following fact is quite general in that it holds for any source and depends only on the MMSE optimality conditions.

Lemma IV.9. *Suppose we have an optimal τ, j quantizer where $j \geq 2$. Then for any $1 \leq k \leq j$,*

$$\underline{\Delta}_{k|j} = \sum_{n=2}^{k} (-1)^{k-n} \Delta_{n|j} + (-1)^{k-1} \underline{\Delta}_{1|j}.$$

Proof. Since for optimal τ, j quantizers, $\underline{\Delta}_{i|j} = \overline{\Delta}_{i+1|j}$ for all $i = 1, 2, \ldots, j-1$, then for any $1 \leq k \leq j$,

$$\begin{aligned}
\underline{\Delta}_{k|j} &= \underline{\Delta}_{k|j} + \overline{\Delta}_{k|j} - \underline{\Delta}_{k-1|j} - \overline{\Delta}_{k-1|j} + \underline{\Delta}_{k-2|j} + \overline{\Delta}_{k-2|j} - \cdots + (-1)^{k-2} \underline{\Delta}_{2|j} + \\
&\quad (-1)^{k-2} \overline{\Delta}_{2|j} + (-1)^{k-1} \underline{\Delta}_{1|j} \\
&= \left(\underline{\Delta}_{k|j} + \overline{\Delta}_{k|j}\right) - \left(\underline{\Delta}_{k-1|j} + \overline{\Delta}_{k-1|j}\right) + \left(\underline{\Delta}_{k-2|j} + \overline{\Delta}_{k-2|j}\right) - \cdots + (-1)^{k-2} \times \\
&\quad \left(\underline{\Delta}_{2|j} + \overline{\Delta}_{2|j}\right) + (-1)^{k-1} \underline{\Delta}_{1|j} \\
&= \Delta_{k|j} - \Delta_{k-1|j} + \Delta_{k-2|j} - \cdots + (-1)^{k-2} \Delta_{2|j} + (-1)^{k-1} \underline{\Delta}_{1|j} \\
&= \sum_{n=2}^{k} (-1)^{k-n} \Delta_{n|j} + (-1)^{k-1} \underline{\Delta}_{1|j}.
\end{aligned}$$

■

Note that while we have stated Lemma IV.9 as a property of optimal τ, j quantizers, the relationship actually holds for any quantizer whose thresholds and reconstruction levels satisfy the optimality conditions.

A property of optimal τ, j quantizers designed for non-increasing, first order differentiable pdf sources. The next fact states that the step sizes $\underline{\Delta}_{k|j}$ of an optimal τ, j quantizer are non-increasing in the cell index k when the source pdf is non-increasing and first order differentiable. Proving this fact requires just a simple application of Lemma IV.6: Since the reconstruction levels and thresholds for cells in an optimal τ, j quantizer, $j \geq 2$, satisfy the optimality conditions for MMSE quantization, if f is non-increasing and differentiable, then Lemma IV.6 tells us that $\underline{\Delta}_{k|j} \leq \overline{\Delta}_{k|j}$, and thus, the step sizes $\left\{\Delta_{k|j}, k \leq j\right\}$ are non-increasing. Lemma IV.10 is the formal statement of this observation and we have stated it without proof.

Lemma IV.10. *Suppose we have an optimal τ, j quantizer, $j \geq 2$, designed for a non-increasing pdf. Let $\left[t_{m|j}, t_{m-1|j}\right)$ and $\left[t_{n|j}, t_{n-1|j}\right)$, $n > m$, be any two quantization cells of this quantizer with $m, n \leq j$. Then $\Delta_{m|j} \geq \Delta_{n|j}$.*

Asymptotic properties of optimal τ, j quantizers designed for GE-sources. With Lemma IV.9 and Lemma IV.10 in hand, we begin our study of the asymptotic behavior of $\underline{\Delta}_{k|j}(t_{k|j})^{p-1}$ in optimal τ, j quantizers designed for general exponential sources by establishing three asymptotic quantizer properties that will allow us to formulate a generator equation for the limits of $\underline{\Delta}_{k|j}(t_{k|j})^{p-1}$ as $\tau \to \infty$.

The next lemma expresses the reconstruction level of a quantization cell in an asymptotic form that depends on the values of the step sizes and thresholds of the cells to its left.

Lemma IV.11. *Suppose we have an optimal τ, j quantizer where $j \geq 2$. If the quantizer was designed for a GE-source, then for $1 \leq k \leq j$,*

$$\mu_{k|j} = t_{k|j} + \sum_{n=2}^{k} (-1)^{k-n}\Delta_{n|j} + (-1)^{k-1}\left[\frac{1}{{t_{1|j}}^{p-1}} + (-p+1)\left(\frac{2}{{t_{1|j}}^{2p-1}} + O_\tau\left(\frac{1}{{t_{1|j}}^{3p-1}}\right)\right)\right].$$

Proof. Using Lemma IV.9, for any $1 \leq k \leq j$,

$$\begin{aligned}\mu_{k|j} = \mu_{k|j} \pm t_{k|j} &= t_{k|j} + \left(\mu_{k|j} - t_{k|j}\right) = t_{k|j} + \underline{\Delta}_{k|j} \\ &= t_{k|j} + \sum_{n=2}^{k} (-1)^{k-n}\, \Delta_{n|j} + (-1)^{k-1}\, \underline{\Delta}_{1|j}.\end{aligned}$$

Since $t_{k|j} \geq \tau >> 0$ and $\underline{\Delta}_{1|j} = \mu_{1|j} - t_{1|j}$, we use Proposition IV.3 to get

$$\mu_{k|j} = t_{k|j} + \sum_{n=2}^{k} (-1)^{k-n}\Delta_{n|j} + (-1)^{k-1}\left[\frac{1}{{t_{1|j}}^{p-1}} + (-p+1)\left(\frac{2}{{t_{1|j}}^{2p-1}} + O_\tau\left(\frac{1}{{t_{1|j}}^{3p-1}}\right)\right)\right].$$

■

The following lemma provides two simple, but useful, properties of optimal τ, j quantizers that are designed for GE-sources. These properties reveal the relative behavior of quantizer thresholds (to each other) as $\tau \to \infty$.

Lemma IV.12. *Consider an optimal τ, j quantizer designed for a GE-source.*

1. *If $j \geq 1$, then $\frac{t_{1|j} - t_{j|j}}{\tau} \to 0$ as $\tau \to \infty$.*
2. *Suppose $j \geq 2$. Let $t_{m|j}$, $t_{n|j}$, $m, n \leq j$, be any thresholds of the quantizer. Then the ratio $\frac{t_{m|j}}{t_{n|j}} \to 1$ as $\tau \to \infty$.*

Proof.

Part 1. Suppose $j \geq 1$. If $j = 1$, then $\frac{t_{1|j} - t_{j|j}}{\tau} = 0$ for all τ so $\lim_{\tau \to \infty} \frac{t_{1|j} - t_{j|j}}{\tau} = 0$. Now let $j > 1$. Since $\frac{t_{1|j} - t_{j|j}}{\tau} > 0$ and

$$\frac{t_{1|j} - t_{j|j}}{\tau} = \frac{\sum_{i=2}^{j} \Delta_{i|j}}{\tau} \leq j \cdot \frac{\Delta_{2|j}}{\tau},$$

then

$$\lim_{\tau \to \infty} \frac{t_{1|j} - t_{j|j}}{\tau} \leq \lim_{\tau \to \infty} j \cdot \frac{\Delta_{2|j}}{\tau} = j \cdot \lim_{\tau \to \infty} \frac{\Delta_{2|j}}{\tau} = j \cdot \lim_{t_{j|j} \to \infty} \frac{\Delta_{2|j}}{t_{j|j}} = j \cdot 0 = 0$$

since for the second to last equality, $\Delta_{2|j} = \underline{\Delta}_{2|j} + \underline{\Delta}_{1|j} \leq 2\underline{\Delta}_{1|j}$ (Lemma IV.6) and there exists $0 < c < \infty$ such that $\underline{\Delta}_{1|j} < c$ (Corollary IV.4).

Part 2. Since for any $1 \leq m \leq j$,

$$0 \leq \frac{t_{m|j} - t_{j|j}}{\tau} \leq \frac{t_{1|j} - t_{j|j}}{\tau},$$

from Lemma IV.12, Part 1, we know that $\lim_{\tau \to \infty} \frac{t_{m|j} - t_{j|j}}{\tau} = 0$ and thus

$$\begin{aligned}
\lim_{\tau \to \infty} \frac{t_{m|j}}{t_{n|j}} &= \lim_{\tau \to \infty} \frac{\frac{t_{m|j}}{\tau}}{\frac{t_{n|j}}{\tau}} = \lim_{\tau \to \infty} \frac{\frac{\tau + \left(t_{m|j} - \tau\right)}{\tau}}{\frac{\tau + \left(t_{n|j} - \tau\right)}{\tau}} = \lim_{\tau \to \infty} \frac{1 - \frac{\left(\tau - t_{m|j}\right)}{\tau}}{1\frac{\left(\tau - t_{n|j}\right)}{\tau}} = \frac{1 - \lim_{\tau \to \infty} \frac{\tau - t_{m|j}}{\tau}}{1 - \lim_{\tau \to \infty} \frac{\tau - t_{n|j}}{\tau}} \\
&= \frac{1 - \lim_{\tau \to \infty} \frac{t_{j|j} - t_{m|j}}{\tau}}{1 - \lim_{\tau \to \infty} \frac{t_{j|j} - t_{n|j}}{\tau}} = 1.
\end{aligned}$$

■

Using the definition of $\alpha_{k|j,p}^{(\tau)}$ in (4.8.28), since

$$\begin{aligned}
\alpha_{k|j,p}^{(\tau)} &= \left(\underline{\Delta}_{k|j} + \overline{\Delta}_{k|j}\right) (t_{k|j})^{p-1} = \left(\underline{\Delta}_{k|j} + \underline{\Delta}_{k-1|j}\right) (t_{k|j})^{p-1} \\
&= \underline{\Delta}_{k|j} (t_{k|j})^{p-1} + \underline{\Delta}_{k-1|j} (t_{k-1|j})^{p-1} \cdot \left(\frac{t_{k|j}}{t_{k-1|j}}\right)^{p-1} \\
&= \underline{\alpha}_{k|j,p}^{(\tau)} + \underline{\alpha}_{k-1|j,p}^{(\tau)} \cdot \left(\frac{t_{k|j}}{t_{k-1|j}}\right)^{p-1},
\end{aligned} \tag{4.8.31}$$

we see that from Lemma IV.12, Part 2 and (4.8.31), we now know that

$$\lim_{\tau\to\infty} \alpha^{(\tau)}_{k|j,p} = \lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k|j,p} + \lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k-1|j,p}$$

if these limits exist. Existence of these limits and the fact that these limits asymptotically satisfy the Nitadori generating recursion is established in the proof of Theorem IV.7 for optimal τ, j quantizers and holds for MMSE quantizers as well.

Applying Lemma IV.5 to optimal τ, j quantization. We have one more lemma to prove, and it is an important one since it provides the means by which we can generalize the method used in the Nitadori derivation to derive an asymptotic recursion relation between successive terms of $\underline{\alpha}^{(\tau)}_{k|j,p}$. Recall that Lemma IV.5 was derived for any arbitrary interval $[s, t)$, $t \geq s > 0$, and without any sort of quantization scheme in mind. We are now going to translate the result in Lemma IV.5 to optimal τ, j quantizers when τ is large since this is the scenario we are working in. For an optimal τ, j quantizer with quantization cell index $2 \leq k \leq j$, we can re-express Lemma IV.5, Part 1 as

$$\frac{f\left(t_{k|j}\right)}{f\left(t_{k-1|j}\right)} = e^{\beta^{(\tau)}_{k|j,p}}, \tag{4.8.32}$$

where $\beta^{(\tau)}_{k|j,p} \stackrel{\triangle}{=} -\frac{1}{p}\left({t_{k|j}}^p - {t_{k-1|j}}^p\right)$ is, in some sense, asymptotically the same as $\underline{\alpha}^{(\tau)}_{k|j,p}$ when $t_{k|j} >> 0$, as will be made clear in a later discussion.

We are now ready to apply Part 2 of Lemma IV.5 to optimal τ, j quantizers with large $\tau > 0$.

Lemma IV.13. *Suppose we have an optimal τ, j quantizer designed for a GE-source with $j \geq 2$. For $2 \leq k \leq j$,*

$$\begin{aligned} \frac{f\left(t_{k|j}\right)}{f\left(t_{k-1|j}\right)} &= \left(1 + o_\tau\left(\frac{1}{t_{k|j}}\right)\right) \times \frac{num_k}{den_k} \\ &= \left(1 - \frac{\underline{\Delta}_{k|j}}{t_{k-1|j}}\right)^{2p-2} \times \frac{num_k}{den_k} \\ &= \left(\frac{t_{k|j}}{t_{k-1|j}}\right)^{2p-2} \times \frac{num_k}{den_k}, \end{aligned}$$

where

$$\begin{aligned}
num_k &= 1 + \underline{\Delta}_{k-1|j} {t_{k-1|j}}^{p-1} + (-p+1)\, O_\tau\!\left(\frac{1}{t_{k-1|j}}\right) \\
&= 1 + \underline{\alpha}^{(\tau)}_{k-1|j,p} + (-p+1)\, O_\tau\!\left(\frac{1}{t_{k-1|j}}\right) \\
den_k &= 1 - \underline{\Delta}_{k|j} {t_{k|j}}^{p-1} - (-p+1)\, O_\tau\!\left(\frac{1}{t_{k|j}}\right) \\
&= 1 - \underline{\alpha}^{(\tau)}_{k|j,p} - (-p+1)\, O_\tau\!\left(\frac{1}{t_{k|j}}\right).
\end{aligned}$$

Proof. To convert the result in Lemma IV.5 into an asymptotic form, we substitute asymptotic expressions for the tail function terms $Q\,(s)\,,\, Q\,(t)$ and the integral function terms $\mathrm{Int}\,(s)\,,\, \mathrm{Int}\,(t)$. To do this, we use the asymptotic key effect from (4.5.25) as a direct substitute for the $Q\,(s)\,,\, Q\,(t)$ terms, and in an indirect way when formulating an asymptotic expression for the integral $\mathrm{Int}\,(s)$ with $s > 0$.

Let $k \geq 2$ and let $\tau >> 0$. Since $t_{k|j} \geq \tau >> 0$,

$$\begin{aligned}
\mathrm{Int}\,\left(t_{k|j}\right) &= \int_{t_{k|j}}^{\infty} x^{-p+1} f\,(x)\, dx \\
&= f\,\left(t_{k|j}\right) \left({t_{k|j}}^{-2p+2} + (-2p+2)\, {t_{k|j}}^{-3p+2} + (-p+1)\, O_\tau\!\left({t_{k|j}}^{-4p+2}\right)\right) \\
&\quad \text{(General-exp-type-conditional-mean-vby.tex, Aside section.)} \\
&= f\,\left(t_{k|j}\right) {t_{k|j}}^{-2p+2} \left(1 + (-2p+2)\, {t_{k|j}}^{-p} + (-p+1)\, O_\tau\!\left({t_{k|j}}^{-2p}\right)\right) \\
&= f\,\left(t_{k|j}\right) {t_{k|j}}^{-2p+2} \left(1 + (-p+1) \left(2{t_{k|j}}^{-p} + O_\tau\!\left({t_{k|j}}^{-2p}\right)\right)\right) \\
&= f\,\left(t_{k|j}\right) {t_{k|j}}^{-2p+2} \left(1 + (-p+1)\, O_\tau\!\left({t_{k|j}}^{-p}\right)\right). \qquad (4.8.33)
\end{aligned}$$

Substituting the expressions for $\mathrm{Int}\,\left(t_{k-1|j}\right)\,,\ \mathrm{Int}\,\left(t_{k|j}\right)$ into the expression in Part 2 of Lemma IV.5, we have

$$\frac{f\,\left(t_{k|j}\right)}{f\,\left(t_{k-1|j}\right)} = \frac{\frac{1}{{t_{k-1|j}}^{p-2}} + (-p+2)\, \frac{f\left(t_{k-1|j}\right)\left(t_{k-1|j}\right)^{-2p+2}\left(1+(-p+1)O_\tau\left({t_{k-1|j}}^{-p}\right)\right)}{f\left(t_{k-1|j}\right)} - \frac{Q\left(t_{k-1|j}\right)}{f\left(t_{k-1|j}\right)}\mu_{k|j}}{\frac{1}{{t_{k|j}}^{p-2}} + (-p+2)\, \frac{f\left(t_{k|j}\right){t_{k|j}}^{-2p+2}\left(1+(-p+1)O_\tau\left({t_{k|j}}^{-p}\right)\right)}{f\left(t_{k|j}\right)} - \frac{Q\left(t_{k|j}\right)}{f\left(t_{k|j}\right)}\mu_{k|j}}$$

$$
\begin{aligned}
&= \frac{\frac{1}{{t_{k-1|j}}^{p-2}} + (-p+2)\left(t_{k-1|j}\right)^{-2p+2}\left(1+(-p+1)\,O_\tau\!\left({t_{k-1|j}}^{-p}\right)\right) - \frac{Q\left(t_{k-1|j}\right)}{f\left(t_{k-1|j}\right)}\mu_{k|j}}{\frac{1}{{t_{k|j}}^{p-2}} + (-p+2)\,{t_{k|j}}^{-2p+2}\left(1+(-p+1)\,O_\tau\!\left({t_{k|j}}^{-p}\right)\right) - \frac{Q\left(t_{k|j}\right)}{f\left(t_{k|j}\right)}\mu_{k|j}} \\
&= \left(\frac{{t_{k|j}}^{p-2}}{{t_{k-1|j}}^{p-2}}\right) \times \\
&\left(\frac{1+(-p+2)\,{t_{k-1|j}}^{-p}\left(1+(-p+1)\,O_\tau\!\left({t_{k-1|j}}^{-p}\right)\right) - \frac{{t_{k-1|j}}^{p-2}Q\left(t_{k-1|j}\right)}{f\left(t_{k-1|j}\right)}\mu_{k|j}}{1+(-p+2)\,{t_{k|j}}^{-p}\left(1+(-p+1)\,O_\tau\!\left({t_{k|j}}^{-p}\right)\right) - \frac{{t_{k|j}}^{p-2}Q\left(t_{k|j}\right)}{f\left(t_{k|j}\right)}\mu_{k|j}}\right) \\
&= \left(\frac{t_{k|j}}{t_{k-1|j}}\right)^{p-2} \times \\
&\left(\frac{1+(-p+2)\,{t_{k-1|j}}^{-p}\left(1+(-p+1)\,O_\tau\!\left({t_{k-1|j}}^{-p}\right)\right) - {t_{k-1|j}}^{p-2}\mu_{k|j}\frac{Q\left(t_{k-1|j}\right)}{f\left(t_{k-1|j}\right)}}{1+(-p+2)\,{t_{k|j}}^{-p}\left(1+(-p+1)\,O_\tau\!\left({t_{k|j}}^{-p}\right)\right) - {t_{k|j}}^{p-2}\mu_{k|j}\frac{Q\left(t_{k|j}\right)}{f\left(t_{k|j}\right)}}\right). \qquad (4.8.34)
\end{aligned}
$$

Define

$$
num_k \triangleq 1+(-p+2)\,{t_{k-1|j}}^{-p}\left(1+(-p+1)\,O_\tau\!\left({t_{k-1|j}}^{-p}\right)\right) - {t_{k-1|j}}^{p-2}\mu_{k|j}\frac{Q\left(t_{k-1|j}\right)}{f\left(t_{k-1|j}\right)}
$$

and

$$
den_k \triangleq 1+(-p+2)\,{t_{k|j}}^{-p}\left(1+(-p+1)\,O_\tau\!\left({t_{k|j}}^{-p}\right)\right) - {t_{k|j}}^{p-2}\mu_{k|j}\frac{Q\left(t_{k|j}\right)}{f\left(t_{k|j}\right)}.
$$

Evaluating and simplifying num_k by using the fact that when $y >> 1$,

$$
\frac{Q(y)}{f(y)} = \frac{1}{y^{p-1}} \times \left[1+(-p+1)\left(\frac{1}{y^p} + (-2p+1)\frac{1}{y^{2p}} + O\left(\frac{1}{y^{3p}}\right)\right)\right], \qquad (4.8.35)
$$

we have

$$
\begin{aligned}
num_k &= 1+(-p+2)\,{t_{k-1|j}}^{-p}\left(1+(-p+1)\,O_\tau\!\left({t_{k-1|j}}^{-p}\right)\right) - {t_{k-1|j}}^{p-2}\mu_{k|j}\frac{Q\left(t_{k-1|j}\right)}{f\left(t_{k-1|j}\right)} \\
&= 1+(-p+2)\,{t_{k-1|j}}^{-p}\left(1+(-p+1)\,O_\tau\!\left({t_{k-1|j}}^{-p}\right)\right) - {t_{k-1|j}}^{p-2}\mu_{k|j} \times \\
&\quad \frac{1}{{t_{k-1|j}}^{p-1}} \times \left[1+(-p+1)\left(\frac{1}{{t_{k-1|j}}^{p}} + (-2p+1)\frac{1}{{t_{k-1|j}}^{2p}} + O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{3p}}\right)\right)\right] \\
&= 1+(-p+2)\,{t_{k-1|j}}^{-p}\left(1+(-p+1)\,O_\tau\!\left({t_{k-1|j}}^{-p}\right)\right) - {t_{k-1|j}}^{-1}\mu_{k|j} \times \\
&\quad \left[1+(-p+1)\left(\frac{1}{{t_{k-1|j}}^{p}} + (-2p+1)\frac{1}{{t_{k-1|j}}^{2p}} + O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{3p}}\right)\right)\right]. \qquad (4.8.36)
\end{aligned}
$$

Using the fact that

$$\mu_{k|j} = t_{k-1|j} - \underline{\Delta}_{k-1|j},$$

then

$${t_{k-1|j}}^{-1}\mu_{k|j} = {t_{k-1|j}}^{-1}\left(t_{k-1|j} - \underline{\Delta}_{k-1|j}\right) = 1 - \frac{\underline{\Delta}_{k-1|j}}{t_{k-1|j}},$$

and then $\mu_{k|j}$ in (4.8.36) becomes

$$\begin{aligned}
num_k &= 1 + \frac{(-p+2)}{{t_{k-1|j}}^p}\left(1 + (-p+1)\,O_\tau\!\left(\frac{1}{{t_{k-1|j}}^p}\right)\right) - \left(1 - \frac{\underline{\Delta}_{k-1|j}}{t_{k-1|j}}\right) \times \Bigg[1+ \\
&\quad (-p+1)\left(\frac{1}{{t_{k-1|j}}^p} + (-2p+1)\,\frac{1}{{t_{k-1|j}}^{2p}} + O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{3p}}\right)\right)\Bigg] \\
&= 1 + \frac{(-p+2)}{{t_{k-1|j}}^p}\left(1 + (-p+1)\,O_\tau\!\left(\frac{1}{{t_{k-1|j}}^p}\right)\right) - \Bigg[1 + (-p+1)\Bigg(\frac{1}{{t_{k-1|j}}^p} + \\
&\quad (-2p+1)\,\frac{1}{{t_{k-1|j}}^{2p}} + O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{3p}}\right)\Bigg)\Bigg] + \frac{\underline{\Delta}_{k-1|j}}{t_{k-1|j}} \times \Bigg[1 + (-p+1)\Bigg(\frac{1}{{t_{k-1|j}}^p} + \\
&\quad (-2p+1)\,\frac{1}{{t_{k-1|j}}^{2p}} + O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{3p}}\right)\Bigg)\Bigg] \\
&= \frac{(-p+2)}{{t_{k-1|j}}^p}\left(1 + (-p+1)\,O_\tau\!\left(\frac{1}{{t_{k-1|j}}^p}\right)\right) - (-p+1)\Bigg(\frac{1}{{t_{k-1|j}}^p} + (-2p+1)\times \\
&\quad \frac{1}{{t_{k-1|j}}^{2p}} + O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{3p}}\right)\Bigg) + \frac{\underline{\Delta}_{k-1|j}}{t_{k-1|j}} \times \Bigg[1 + (-p+1)\Bigg(\frac{1}{{t_{k-1|j}}^p} + (-2p+1)\times \\
&\quad \frac{1}{{t_{k-1|j}}^{2p}} + O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{3p}}\right)\Bigg)\Bigg] \\
&= \frac{1}{{t_{k-1|j}}^p} \times \Bigg\{(-p+2)\left(1 + (-p+1)\,O_\tau\!\left(\frac{1}{{t_{k-1|j}}^p}\right)\right) - (-p+1)\times \\
&\quad \left(1 + (-2p+1)\,\frac{1}{{t_{k-1|j}}^p} + O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{2p}}\right)\right) + \underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1} \times \\
&\quad \left[1 + (-p+1)\left(\frac{1}{{t_{k-1|j}}^p} + (-2p+1)\,\frac{1}{{t_{k-1|j}}^{2p}} + O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{3p}}\right)\right)\right]\Bigg\} \\
&= \frac{1}{{t_{k-1|j}}^p} \times \Bigg\{\left[(-p+2) - (-p+1) + \underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1}\right] + \Bigg[-\frac{(-p+1)(-2p+1)}{{t_{k-1|j}}^p} + \\
&\quad \underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1}\frac{(-p+1)}{{t_{k-1|j}}^p} + (-p+2)\,(-p+1)\,O_\tau\!\left(\frac{1}{{t_{k-1|j}}^p}\right)\Bigg] + \Bigg[-(-p+1)\times
\end{aligned}$$

$$
\begin{aligned}
&O_\tau\Bigg(\frac{1}{{t_{k-1|j}}^{2p}}\Bigg)+\underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1}\frac{(-2p+1)}{{t_{k-1|j}}^{2p}}\Bigg]+\underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1}O_\tau\Bigg(\frac{1}{{t_{k-1|j}}^{3p}}\Bigg)\Bigg\} \\
&=\frac{1}{{t_{k-1|j}}^{p}}\times\Bigg\{\Bigg[1+\underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1}\Bigg]+\frac{(-p+1)}{{t_{k-1|j}}^{p}}\Bigg[-(-2p+1)+\underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1}+ \\
&(-p+2)O_{t_{k-1|j}}(1)\Bigg]+\frac{1}{{t_{k-1|j}}^{2p}}\Bigg[-(-p+1)\,O_{t_{k-1|j}}(1)+\underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1}(-2p+1)\Bigg]+ \\
&\underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1}O_\tau\Bigg(\frac{1}{{t_{k-1|j}}^{3p}}\Bigg)\Bigg\} \\
&=\frac{1}{{t_{k-1|j}}^{p}}\times\Bigg\{\Bigg[1+\underline{\Delta}_{k-1|j}{t_{k-1|j}}^{p-1}\Bigg]+(-p+1)\,O_\tau\Bigg(\frac{1}{{t_{k-1|j}}^{p}}\Bigg)\Bigg\}.
\end{aligned}
$$

Similarly using (4.8.35), we have

$$
\begin{aligned}
den_k &= 1+\frac{(-p+2)}{{t_{k|j}}^{p}}\left(1+(-p+1)\,O_\tau\Bigg(\frac{1}{{t_{k|j}}^{p}}\Bigg)\right)-{t_{k|j}}^{p-2}\mu_{k|j}\frac{Q\left(t_{k|j}\right)}{f\left(t_{k|j}\right)} \\
&= 1+\frac{(-p+2)}{{t_{k|j}}^{p}}\left(1+(-p+1)\,O_\tau\Bigg(\frac{1}{{t_{k|j}}^{p}}\Bigg)\right)-{t_{k|j}}^{p-2}\mu_{k|j}\times\frac{1}{{t_{k|j}}^{p-1}}\times \\
&\quad\Bigg[1+(-p+1)\left(\frac{1}{{t_{k|j}}^{p}}+(-2p+1)\,\frac{1}{{t_{k|j}}^{2p}}+O_\tau\Bigg(\frac{1}{{t_{k|j}}^{3p}}\Bigg)\right)\Bigg] \\
&= 1+\frac{(-p+2)}{{t_{k|j}}^{p}}\left(1+(-p+1)\,O_\tau\Bigg(\frac{1}{{t_{k|j}}^{p}}\Bigg)\right)-{t_{k|j}}^{-1}\mu_{k|j}\Bigg[1+(-p+1)\times \\
&\quad\left(\frac{1}{{t_{k|j}}^{p}}+(-2p+1)\,\frac{1}{{t_{k|j}}^{2p}}+O_\tau\Bigg(\frac{1}{{t_{k|j}}^{3p}}\Bigg)\right)\Bigg].
\end{aligned}
$$

Since

$$
\mu_{k|j}=t_{k|j}+\underline{\Delta}_{k|j}
$$

then

$$
{t_{k|j}}^{-1}\mu_{k|j}=1+\frac{\underline{\Delta}_{k|j}}{t_{k|j}},
$$

and we have

$$
\begin{aligned}
den_k = 1+\frac{(-p+2)}{{t_{k|j}}^{p}}\left(1+(-p+1)\,O_\tau\Bigg(\frac{1}{{t_{k|j}}^{p}}\Bigg)\right)-\left(1+\frac{\underline{\Delta}_{k|j}}{t_{k|j}}\right)\times\Bigg[1+(-p+1)\times \\
\left(\frac{1}{{t_{k|j}}^{p}}+\frac{(-2p+1)}{{t_{k|j}}^{2p}}+O_\tau\Bigg(\frac{1}{{t_{k|j}}^{3p}}\Bigg)\right)\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= 1+\frac{(-p+2)}{{t_{k|j}}^{p}}\Bigg(1+(-p+1)\,O_\tau\bigg(\frac{1}{{t_{k|j}}^{p}}\bigg)\Bigg)-\bigg[1+(-p+1)\bigg(\frac{1}{{t_{k|j}}^{p}}+\frac{(-2p+1)}{{t_{k|j}}^{2p}}+\\
&\quad O_\tau\bigg(\frac{1}{{t_{k|j}}^{3p}}\bigg)\bigg)\bigg]-\frac{\underline{\Delta}_{k|j}}{t_{k|j}}\times\Bigg[1+(-p+1)\bigg(\frac{1}{{t_{k|j}}^{p}}+\frac{(-2p+1)}{{t_{k|j}}^{2p}}+O_\tau\bigg(\frac{1}{{t_{k|j}}^{3p}}\bigg)\bigg)\Bigg]\\
&= \frac{(-p+2)}{{t_{k|j}}^{p}}\Bigg(1+(-p+1)\,O_\tau\bigg(\frac{1}{{t_{k|j}}^{p}}\bigg)\Bigg)-(-p+1)\bigg(\frac{1}{{t_{k|j}}^{p}}+\frac{(-2p+1)}{{t_{k|j}}^{2p}}+\\
&\quad O_\tau\bigg(\frac{1}{{t_{k|j}}^{3p}}\bigg)\bigg)-\frac{\underline{\Delta}_{k|j}}{t_{k|j}}\times\Bigg[1+\bigg(\frac{(-p+1)}{{t_{k|j}}^{p}}+\frac{(-2p+1)}{{t_{k|j}}^{2p}}+O_\tau\bigg(\frac{1}{{t_{k|j}}^{3p}}\bigg)\bigg)\Bigg]\\
&= \frac{1}{{t_{k|j}}^{p}}\times\Bigg\{(-p+2)\Bigg(1+(-p+1)\,O_\tau\bigg(\frac{1}{{t_{k|j}}^{p}}\bigg)\Bigg)-(-p+1)\bigg(1+\frac{(-2p+1)}{{t_{k|j}}^{p}}+\\
&\quad O_\tau\bigg(\frac{1}{{t_{k|j}}^{2p}}\bigg)\bigg)-\underline{\Delta}_{k|j}{t_{k|j}}^{p-1}\times\Bigg[1+(-p+1)\bigg(\frac{1}{{t_{k|j}}^{p}}+\frac{(-2p+1)}{{t_{k|j}}^{2p}}+O_\tau\bigg(\frac{1}{{t_{k|j}}^{3p}}\bigg)\bigg)\Bigg]\Bigg\}\\
&= \frac{1}{{t_{k|j}}^{p}}\times\Bigg\{\bigg[(-p+2)-(-p+1)-\underline{\Delta}_{k|j}{t_{k|j}}^{p-1}\bigg]+\bigg[-\frac{(-p+1)(-2p+1)}{{t_{k|j}}^{p}}-\\
&\quad \underline{\Delta}_{k|j}{t_{k|j}}^{p-1}\frac{(-p+1)}{{t_{k|j}}^{p}}+(-p+2)(-p+1)\,O_\tau\bigg(\frac{1}{{t_{k|j}}^{p}}\bigg)\bigg]+\bigg[-\underline{\Delta}_{k|j}{t_{k|j}}^{p-1}\times\\
&\quad \frac{(-p+1)(-2p+1)}{{t_{k|j}}^{2p}}-(-p+1)\,O_\tau\bigg(\frac{1}{{t_{k|j}}^{2p}}\bigg)\bigg]-\underline{\Delta}_{k|j}{t_{k|j}}^{p-1}(-p+1)\,O_\tau\bigg(\frac{1}{{t_{k|j}}^{3p}}\bigg)\Bigg\}\\
&= \frac{1}{{t_{k|j}}^{p}}\times\Bigg\{\bigg[1-\underline{\Delta}_{k|j}{t_{k|j}}^{p-1}\bigg]+\frac{(-p+1)}{{t_{k|j}}^{p}}\bigg[-(-2p+1)-\underline{\Delta}_{k|j}{t_{k|j}}^{p-1}+\\
&\quad (-p+2)\,O_{t_{k|j}}(1)\bigg]+\frac{(-p+1)}{{t_{k|j}}^{2p}}\bigg[-\underline{\Delta}_{k|j}{t_{k|j}}^{p-1}(-2p+1)-O_{t_{k|j}}(1)\bigg]-\\
&\quad \underline{\Delta}_{k|j}{t_{k|j}}^{p-1}(-p+1)\,O_\tau\bigg(\frac{1}{{t_{k|j}}^{3p}}\bigg)\Bigg\}\\
&= \frac{1}{{t_{k|j}}^{p}}\times\Bigg\{\Big[1-\underline{\Delta}_{k|j}{t_{k|j}}^{p-1}\Big]+(-p+1)\,O_\tau\bigg(\frac{1}{{t_{k|j}}^{p}}\bigg)\Bigg\}.
\end{aligned}
$$

Since $\left(\frac{t_{k|j}}{t_{k-1|j}}\right)\to 1$ as $\tau\to\infty$ (from Lemma IV.12, Part 2), (4.8.34) now becomes

$$
\frac{f\left(t_{k|j}\right)}{f\left(t_{k-1|j}\right)}=\left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)\times\frac{num_k}{den_k},
$$

where

$$num_k = \left[1 + \underline{\Delta}_{k-1|j} {t_{k-1|j}}^{p-1}\right] + (-p+1)\, O_\tau\!\left(\frac{1}{{t_{k-1|j}}^{p}}\right)$$
$$den_k = \left[1 - \underline{\Delta}_{k|j} {t_{k|j}}^{p-1}\right] + (-p+1)\, O_\tau\!\left(\frac{1}{{t_{k|j}}^{p}}\right).$$

■

Lemma IV.13 is an important step in the process of deriving an asymptotic version of the Nitadori recursion for optimal τ, j quantizers designed for general exponential sources. As can be seen Table 4.2, it is clear when comparing left hand terms and comparing right hand terms that Lemma IV.13 is an asymptotic generalization of (4.2.16) since when $p = 1$, the expressions are the same, except for multiplication by -1 on both the left hand side and the right hand side of the expressions taken from Lemma IV.13. We remark that it is crucial that we are able to make this generalization for (4.2.16) since this equation is the direct result of applying the key effect in the Nitadori derivation.

Table 4.2: Comparing relations in Lemma IV.13 to (4.2.16) in the Nitadori derivation.

	Lemma IV.13 relation (GE-sources) (optimal τ, j quantizers)	Nitadori (4.2.16) (Exp. source) (MMSE quantizers)
LHS	$f\left(t_{k\|j}\right)\left(1-\frac{\underline{\Delta}_{k\|j}}{t_{k-1\|j}}\right)^{2p-2} \times \left(1-\underline{\Delta}_{k\|j} {t_{k\|j}}^{p-1} - (-p+1)\, O_\tau\!\left(\frac{1}{t_{k\|j}}\right)\right)$	$f\left(t_k^{(N)}\right)\left(\underline{\Delta}_k^{(N)} - 1\right)$
RHS	$f\left(t_{k-1\|j}\right) \times \left(1+\underline{\Delta}_{k-1\|j} {t_{k-1\|j}}^{p-1} + (-p+1)\, O_\tau\!\left(\frac{1}{t_{k-1\|j}}\right)\right)$	$f\left(t_{k-1}^{(N)}\right)\left(-\underline{\Delta}_{k-1}^{(N)} - 1\right)$

To review, we have been able to follow the outline of the steps taken in the Nitadori sequence derivation up to and including the point at which the key effect was applied. We are now ready to derive an analogous asymptotic form of the Nitadori sequence generator equation in (4.2.19), and this derivation will shown in the next section with the proof of Theorem IV.7.

4.9 Proof of Theorem IV.7.

Fix $p \geq 1$ and fix $j \geq 2$. We start by assuming we have an optimal τ, j quantizer that has been designed for the GE-source indicated by the value of p.

Part 1(*a*). *Show $\underline{\nu}_1$ exists and equals $\underline{\eta}_1$.*

We remark that we have already shown this to be true for MMSE quantizers, but, for completeness, we briefly go over it again for optimal τ, j quantizers.

Since optimal τ, j quantizers are optimal for the source with conditional distribution $f_{X|X\geq\tau}(x)$, we can apply Corollary IV.4 to the first ($k = 1$) quantization cell $[t_{1|j}, \infty)$ to get

$$\lim_{\tau\to\infty} \underline{\Delta}_{1|j} {t_{1|j}}^{p-1} = \lim_{\tau\to\infty} \left(\mu_{\left[t_{1|j},\infty\right)} - t_{1|j}\right) {t_{1|j}}^{p-1} \overset{\text{(Cor. IV.4)}}{=} 1,$$

since $t_{1|j} \geq t_{j|j} = \tau$ and so $t_{1|j} \to \infty$. Thus $\underline{\nu}_1 = \lim_{\tau\to\infty} \underline{\Delta}_{1|j} {t_{1|j}}^{p-1}$ exists and since $\underline{\eta}_1 = 1$, we have $\underline{\nu}_1 = \underline{\eta}_1$. ■

Part 1(*b*). *Show if $\underline{\nu}_{k-1}$ exists, then $\underline{\nu}_k$ exists and satisfies* (4.8.29)*:*

$$e^{\underline{\nu}_k} \left(1 - \underline{\nu}_k\right) = e^{-\underline{\nu}_{k-1}} \left(1 + \underline{\nu}_{k-1}\right).$$

This proof has five steps. Choose $k \in \{2, 3, \ldots, j\}$.

Step 1. Show

$$e^{\underline{\alpha}^{(\tau)}_{k|j,p}} \left(1 - \underline{\alpha}^{(\tau)}_{k|j,p}\right) = e^{-\underline{\alpha}^{(\tau)}_{k-1|j,p} \times \left(1 + o_\tau\left(\frac{1}{t_{k|j}}\right)\right)} \times \left[1 + \underline{\alpha}^{(\tau)}_{k-1|j,p}\right] \times \left(1 + o_\tau\left(\frac{1}{t_{k|j}}\right)\right) \times e^{-\left(\beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p}\right)}$$
$$\times \left[1 + \frac{(-p+1)\, O_\tau\left(\frac{1}{t_{k-1|j}}\right)}{1 + \underline{\alpha}^{(\tau)}_{k-1|j,p}}\right] \times \left[1 - \frac{(-p+1)\, O_\tau\left(\frac{1}{t_{k|j}}\right)}{1 - \underline{\alpha}^{(\tau)}_{k|j,p}}\right]^{-1} \tag{4.9.37}$$

which is an approximation to the recursive equation in (4.2.19).

To show (4.9.37), we re-express the equation in Lemma IV.13 by moving quantities from either side of the equal sign and by substituting in $\underline{\alpha}^{(\tau)}_{k-1|j,p}$, $\underline{\alpha}^{(\tau)}_{k|j,p}$, $\beta^{(\tau)}_{k|j,p}$ where possible.

With $t_{k|j} \geq \tau >> 0$, we begin by equating (4.8.32) to the expression in Lemma IV.13

to obtain

$$e^{\beta^{(\tau)}_{k|j,p}} = \left(1 + o_\tau\left(\frac{1}{t_{k|j}}\right)\right) \times \frac{1 + \underline{\Delta}_{k-1|j} {t_{k-1|j}}^{p-1} + (-p+1)\, O_\tau\!\left(\frac{1}{t_{k-1|j}}\right)}{1 - \underline{\Delta}_{k|j} {t_{k|j}}^{p-1} - (-p+1)\, O_\tau\!\left(\frac{1}{t_{k|j}}\right)}$$
$$= \left(1 + o_\tau\left(\frac{1}{t_{k|j}}\right)\right) \times \frac{1 + \underline{\alpha}^{(\tau)}_{k-1|j,p} + (-p+1)\, O_\tau\!\left(\frac{1}{t_{k-1|j}}\right)}{1 - \underline{\alpha}^{(\tau)}_{k|j,p} - (-p+1)\, O_\tau\!\left(\frac{1}{t_{k|j}}\right)}. \tag{4.9.38}$$

By multiplying both sides of (4.9.38) by the denominator of the right side of (4.9.38), we have

$$e^{\beta^{(\tau)}_{k|j,p}} \times \left[1 - \underline{\alpha}^{(\tau)}_{k|j,p} - (-p+1)\, O_\tau\!\left(\frac{1}{t_{k|j}}\right)\right]$$
$$= \left(1 + o_\tau\left(\frac{1}{t_{k|j}}\right)\right) \times \left[1 + \underline{\alpha}^{(\tau)}_{k-1|j,p} + (-p+1)\, O_\tau\!\left(\frac{1}{t_{k-1|j}}\right)\right]. \tag{4.9.39}$$

Concentrating on just the left-hand side of (4.9.39), we multiply the left side by

$$\frac{e^{\alpha^{(\tau)}_{k|j,p}}\left(1 - \underline{\alpha}^{(\tau)}_{k|j,p}\right)}{e^{\alpha^{(\tau)}_{k|j,p}}\left(1 - \underline{\alpha}^{(\tau)}_{k|j,p}\right)}$$

to get

$$(4.9.39)_{LHS} = \frac{e^{\alpha^{(\tau)}_{k|j,p}}\left(1 - \underline{\alpha}^{(\tau)}_{k|j,p}\right)}{e^{\alpha^{(\tau)}_{k|j,p}}\left(1 - \underline{\alpha}^{(\tau)}_{k|j,p}\right)} \times e^{\beta^{(\tau)}_{k|j,p}} \times \left[1 - \underline{\alpha}^{(\tau)}_{k|j,p} - (-p+1)\, O_\tau\!\left(\frac{1}{t_{k|j}}\right)\right]$$
$$= e^{\alpha^{(\tau)}_{k|j,p}}\left(1 - \underline{\alpha}^{(\tau)}_{k|j,p}\right) \times \frac{e^{\beta^{(\tau)}_{k|j,p}}}{e^{\alpha^{(\tau)}_{k|j,p}}} \times \frac{1 - \underline{\alpha}^{(\tau)}_{k|j,p} - (-p+1)\, O_\tau\!\left(\frac{1}{t_{k|j}}\right)}{1 - \underline{\alpha}^{(\tau)}_{k|j,p}}$$
$$= e^{\alpha^{(\tau)}_{k|j,p}}\left(1 - \underline{\alpha}^{(\tau)}_{k|j,p}\right) \times e^{\beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p}} \times \left[1 - \frac{(-p+1)\, O_\tau\!\left(\frac{1}{t_{k|j}}\right)}{1 - \underline{\alpha}^{(\tau)}_{k|j,p}}\right].$$

Since

$$\alpha^{(\tau)}_{k|j,p} = \Delta_{k|j} {t_{k|j}}^{p-1} = \left(\underline{\Delta}_{k|j} + \underline{\Delta}_{k-1|j}\right) {t_{k|j}}^{p-1} = \underline{\Delta}_{k|j} {t_{k|j}}^{p-1} + \underline{\Delta}_{k-1|j} {t_{k|j}}^{p-1}$$
$$= \underline{\alpha}^{(\tau)}_{k|j,p} + \underline{\Delta}_{k-1|j}\left(t_{k-1|j} - \Delta_{k|j}\right)^{p-1} = \underline{\alpha}^{(\tau)}_{k|j,p} + \underline{\Delta}_{k-1|j} {t_{k-1|j}}^{p-1}\left(1 - \frac{\Delta_{k|j}}{t_{k-1|j}}\right)^{p-1}$$

$$
\begin{aligned}
&= \underline{\alpha}^{(\tau)}_{k|j,p} + \underline{\alpha}^{(\tau)}_{k-1|j,p}\left(1-\frac{\Delta_{k|j}}{t_{k-1|j}}\right)^{p-1} = \underline{\alpha}^{(\tau)}_{k|j,p} + \underline{\alpha}^{(\tau)}_{k-1|j,p}\left(\frac{t_{k-1|j}-\Delta_{k|j}}{t_{k-1|j}}\right)^{p-1}\\
&= \underline{\alpha}^{(\tau)}_{k|j,p} + \underline{\alpha}^{(\tau)}_{k-1|j,p}\left(\frac{t_{k|j}}{t_{k-1|j}}\right)^{p-1}\\
&= \underline{\alpha}^{(\tau)}_{k|j,p} + \underline{\alpha}^{(\tau)}_{k-1|j,p}\left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right) \quad \text{(Lemma IV.12, Part 2),}
\end{aligned}
$$

we find that

$$
\begin{aligned}
&(4.9.39)_{LHS}\\
&\quad = e^{\underline{\alpha}^{(\tau)}_{k|j,p}+\underline{\alpha}^{(\tau)}_{k-1|j,p}\left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)}\left(1-\underline{\alpha}^{(\tau)}_{k|j,p}\right)\times e^{\beta^{(\tau)}_{k|j,p}-\alpha^{(\tau)}_{k|j,p}}\times\left[1-\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{k|j}}\right)}{1-\underline{\alpha}^{(\tau)}_{k|j,p}}\right].
\end{aligned}
$$

Concentrating on the right-hand side of (4.9.39), we have

$$
\begin{aligned}
(4.9.39)_{RHS} &= \left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)\times\left[1+\underline{\alpha}^{(\tau)}_{k-1|j,p}+(-p+1)\,O_\tau\!\left(\frac{1}{t_{k-1|j}}\right)\right]\times\frac{1+\underline{\alpha}^{(\tau)}_{k-1|j,p}}{1+\underline{\alpha}^{(\tau)}_{k-1|j,p}}\\
&= \left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)\times\left[\frac{1+\underline{\alpha}^{(\tau)}_{k-1|j,p}+(-p+1)\,O_\tau\!\left(\frac{1}{t_{k-1|j}}\right)}{1+\underline{\alpha}^{(\tau)}_{k-1|j,p}}\right]\times\left(1+\underline{\alpha}^{(\tau)}_{k-1|j,p}\right)\\
&= \left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)\times\left[1+\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{k-1|j}}\right)}{1+\underline{\alpha}^{(\tau)}_{k-1|j,p}}\right]\times\left(1+\underline{\alpha}^{(\tau)}_{k-1|j,p}\right).
\end{aligned}
$$

Multiplying both sides of (4.9.39) by

$$
e^{-\underline{\alpha}^{(\tau)}_{k-1|j,p}\left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)}\times e^{\alpha^{(\tau)}_{k|j,p}-\beta^{(\tau)}_{k|j,p}}\times\left[1-\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{k|j}}\right)}{1-\underline{\alpha}^{(\tau)}_{k|j,p}}\right]^{-1},
$$

we have

$$
\begin{aligned}
&e^{\underline{\alpha}^{(\tau)}_{k|j,p}}\left(1-\underline{\alpha}^{(\tau)}_{k|j,p}\right)\\
&= \left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)\times\left[1+\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{k-1|j}}\right)}{1+\underline{\alpha}^{(\tau)}_{k-1|j,p}}\right]\times\left(1+\underline{\alpha}^{(\tau)}_{k-1|j,p}\right)\times e^{-\underline{\alpha}^{(\tau)}_{k-1|j,p}\left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)}\\
&\quad\times\ e^{\alpha^{(\tau)}_{k|j,p}-\beta^{(\tau)}_{k|j,p}}\left[1-\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{k|j}}\right)}{1-\underline{\alpha}^{(\tau)}_{k|j,p}}\right]^{-1}
\end{aligned}
$$

$$= e^{-\underline{\alpha}_{k-1|j,p}^{(\tau)}\left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)} \times \left(1+\underline{\alpha}_{k-1|j,p}^{(\tau)}\right)$$

$$\times \left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right) \times e^{\alpha_{k|j,p}^{(\tau)}-\beta_{k|j,p}^{(\tau)}} \times \left[1+\frac{(-p+1)\, O_\tau\left(\frac{1}{t_{k-1|j}}\right)}{1+\underline{\alpha}_{k-1|j,p}^{(\tau)}}\right] \times \left[1-\frac{(-p+1)\, O_\tau\left(\frac{1}{t_{k|j}}\right)}{1-\underline{\alpha}_{k|j,p}^{(\tau)}}\right]^{-1},$$

which is (4.9.37).

Step 2. Show $\limsup_{\tau\to\infty} \underline{\alpha}_{k|j,p}^{(\tau)} < 1$.

The purpose of this step is to prove that the term $\left[1-\frac{(-p+1)O_\tau\left(\frac{1}{t_{k|j}}\right)}{1-\underline{\alpha}_{k|j,p}^{(\tau)}}\right]$ in (4.9.37) is bounded away from zero so that the term $\left[1-\frac{(-p+1)O_\tau\left(\frac{1}{t_{k|j}}\right)}{1-\underline{\alpha}_{k|j,p}^{(\tau)}}\right]^{-1}$ does not blow up as $\tau \to \infty$.

This is proof by contradiction. Since $\underline{\alpha}_{k|j,p}^{(\tau)} \leq \underline{\alpha}_{2|j,p}^{(\tau)}$ and $t_{k|j} \geq \tau$ for all $2 \leq k \leq j$, it suffices to show that $\limsup_{t_{2|j}\to\infty} \underline{\alpha}_{2|j,p}^{(\tau)} < 1$. With

$$\underline{\alpha}_{2|j,p}^{(\tau)} = \underline{\Delta}_{2|j} {t_{2|j}}^{p-1} \leq \underline{\Delta}_{1|j} {t_{2|j}}^{p-1} \leq \underline{\Delta}_{1|j} {t_{1|j}}^{p-1} = \underline{\alpha}_{1|j,p}^{(\tau)},$$

we know that

$$\limsup_{\tau\to\infty} \underline{\alpha}_{2|j,p}^{(\tau)} \leq \limsup_{\tau\to\infty} \underline{\alpha}_{1|j,p}^{(\tau)} = \lim_{\tau\to\infty} \underline{\alpha}_{1|j,p}^{(\tau)} = 1. \quad \text{(Corollary IV.4)}$$

Now suppose $\limsup_{\tau\to\infty} \underline{\alpha}_{2|j,p}^{(\tau)} = 1$. Then

$$\begin{aligned} 0 &= \left(1 - \limsup_{\tau\to\infty} \underline{\alpha}_{2|j,p}^{(\tau)}\right) \\ &= \left(1 - \limsup_{\tau\to\infty} \underline{\alpha}_{2|j,p}^{(\tau)} - \lim_{\tau\to\infty} (-p+1)\, O_\tau\left(\frac{1}{t_{2|j}}\right)\right) \\ &= \liminf_{\tau\to\infty} \left(1 - \underline{\alpha}_{2|j,p}^{(\tau)} - (-p+1)\, O_\tau\left(\frac{1}{t_{2|j}}\right)\right). \end{aligned} \tag{4.9.40}$$

Working to find an equivalent expression to the right hand side of (4.9.40), we can re-write (4.9.37) from Step 1 with $k = 2$, to obtain

$$e^{\underline{\alpha}^{(\tau)}_{2|j,p}} \left(1-\underline{\alpha}^{(\tau)}_{2|j,p}\right)\times\left[1-\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{2|j}}\right)}{1-\underline{\alpha}^{(\tau)}_{2|j,p}}\right]$$
$$= e^{-\underline{\alpha}^{(\tau)}_{1|j,p}\times\left(\frac{t_{2|j}}{t_{1|j}}\right)^{p-1}} \times\left[1+\underline{\alpha}^{(\tau)}_{1|j,p}\right]\times\left(\frac{t_{2|j}}{t_{1|j}}\right)^{2p-2}\times e^{-\left(\beta^{(\tau)}_{2|j,p}-\alpha^{(\tau)}_{2|j,p}\right)}\times\left[1+\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{1|j}}\right)}{1+\underline{\alpha}^{(\tau)}_{1|j,p}}\right]$$

or equivalently,

$$e^{\underline{\alpha}^{(\tau)}_{2|j,p}} \left(1-\underline{\alpha}^{(\tau)}_{2|j,p}\right)\times\left[\frac{1-\underline{\alpha}^{(\tau)}_{2|j,p}-(-p+1)\,O_\tau\!\left(\frac{1}{t_{2|j}}\right)}{1-\underline{\alpha}^{(\tau)}_{2|j,p}}\right]$$
$$= e^{-\underline{\alpha}^{(\tau)}_{1|j,p}\times\left(\frac{t_{2|j}}{t_{1|j}}\right)^{p-1}} \times\left[1+\underline{\alpha}^{(\tau)}_{1|j,p}\right]\times\left(\frac{t_{2|j}}{t_{1|j}}\right)^{2p-2}\times e^{-\left(\beta^{(\tau)}_{2|j,p}-\alpha^{(\tau)}_{2|j,p}\right)}\times\left[1+\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{1|j}}\right)}{1+\underline{\alpha}^{(\tau)}_{1|j,p}}\right]$$

or equivalently,

$$\left(1-\underline{\alpha}^{(\tau)}_{2|j,p}\right)\times\left[\frac{1-\underline{\alpha}^{(\tau)}_{2|j,p}-(-p+1)\,O_\tau\!\left(\frac{1}{t_{2|j}}\right)}{1-\underline{\alpha}^{(\tau)}_{2|j,p}}\right]=$$
$$e^{-\underline{\alpha}^{(\tau)}_{2|j,p}}\times e^{-\underline{\alpha}^{(\tau)}_{1|j,p}\times\left(\frac{t_{2|j}}{t_{1|j}}\right)^{p-1}} \times\left[1+\underline{\alpha}^{(\tau)}_{1|j,p}\right]\times\left(\frac{t_{2|j}}{t_{1|j}}\right)^{2p-2}\times e^{-\left(\beta^{(\tau)}_{2|j,p}-\alpha^{(\tau)}_{2|j,p}\right)}\times$$
$$\left[1+\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{1|j}}\right)}{1+\underline{\alpha}^{(\tau)}_{1|j,p}}\right]$$

or equivalently,

$$\left[1-\underline{\alpha}^{(\tau)}_{2|j,p}-(-p+1)\,O_\tau\!\left(\frac{1}{t_{2|j}}\right)\right]$$
$$= e^{-\underline{\alpha}^{(\tau)}_{1|j,p}\times\left(\frac{t_{2|j}}{t_{1|j}}\right)^{p-1}} \times\left[1+\underline{\alpha}^{(\tau)}_{1|j,p}\right]\times\left(\frac{t_{2|j}}{t_{1|j}}\right)^{2p-2}\times e^{-\underline{\alpha}^{(\tau)}_{2|j,p}}\times e^{-\left(\beta^{(\tau)}_{2|j,p}-\alpha^{(\tau)}_{2|j,p}\right)}\times$$
$$\left[1+\frac{(-p+1)\,O_\tau\!\left(\frac{1}{t_{1|j}}\right)}{1+\underline{\alpha}^{(\tau)}_{1|j,p}}\right]$$

$$= e^{-\underline{\alpha}^{(\tau)}_{1|j,p} \times \left(\frac{t_{2|j}}{t_{1|j}}\right)^{p-1}} \times \left[1 + \underline{\alpha}^{(\tau)}_{1|j,p}\right] \times \left(\frac{t_{2|j}}{t_{1|j}}\right)^{2p-2} \times e^{-\beta^{(\tau)}_{2|j,p}} \times \left[1 + \frac{(-p+1)\, O_\tau\!\left(\frac{1}{t_{1|j}}\right)}{1 + \underline{\alpha}^{(\tau)}_{1|j,p}}\right] \tag{4.9.41}$$

We see that the left hand side of (4.9.41) is almost the same as the right hand side of (4.9.40). Now taking the lim inf as $\tau \to \infty$ on both sides of (4.9.41), we obtain an equivalent expression to the right hand side of (4.9.40)

$$\begin{aligned}
&\liminf_{\tau\to\infty} \left[1 - \underline{\alpha}^{(\tau)}_{2|j,p} - (-p+1)\, O_\tau\!\left(\frac{1}{t_{2|j}}\right)\right] \\
&= \liminf_{\tau\to\infty} e^{-\underline{\alpha}^{(\tau)}_{1|j,p} \times \left(\frac{t_{2|j}}{t_{1|j}}\right)^{p-1}} \times \left[1 + \underline{\alpha}^{(\tau)}_{1|j,p}\right] \times \left(\frac{t_{2|j}}{t_{1|j}}\right)^{2p-2} \times e^{-\beta^{(\tau)}_{2|j,p}} \times \left[1 + \frac{(-p+1)\, O_\tau\!\left(\frac{1}{t_{1|j}}\right)}{1 + \underline{\alpha}^{(\tau)}_{1|j,p}}\right] \\
&= e^{-\lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{1|j,p} \times \lim_{\tau\to\infty} \left(\frac{t_{2|j}}{t_{1|j}}\right)^{p-1}} \times \lim_{\tau\to\infty} \left[1 + \underline{\alpha}^{(\tau)}_{1|j,p}\right] \times \lim_{\tau\to\infty} \left(\frac{t_{2|j}}{t_{1|j}}\right)^{2p-2} \times \liminf_{\tau\to\infty} e^{-\beta^{(\tau)}_{2|j,p}} \\
&\quad \times \lim_{\tau\to\infty} \left[1 + \frac{(-p+1)\, O_\tau\!\left(\frac{1}{t_{1|j}}\right)}{1 + \underline{\alpha}^{(\tau)}_{1|j,p}}\right] \\
&= e^{-1} \times 2 \times 1 \times \liminf_{\tau\to\infty} e^{-\beta^{(\tau)}_{2|j,p}} \times 1 \quad \text{(Part } 1(a)) \\
&= 2e^{-1} \times e^{-\limsup_{\tau\to\infty} \beta^{(\tau)}_{2|j,p}}
\end{aligned} \tag{4.9.42}$$

Substituting (4.9.42) for the right hand side of (4.9.40), we have

$$0 = 2e^{-1} \times e^{-\limsup_{\tau\to\infty} \beta^{(\tau)}_{2|j,p}}.$$

Thus if we can prove that $\beta^{(\tau)}_{2|j,p}$ is bounded for all τ, we will have our contradiction. Consider expanding the expression for $\beta^{(\tau)}_{2|j,p}$

$$\begin{aligned}
\beta^{(\tau)}_{2|j,p} &= -\frac{1}{p}\left({t_{2|j}}^p - {t_{1|j}}^p\right) = -\frac{1}{p}\left({t_{2|j}}^p - \left(t_{2|j} + \Delta_{2|j}\right)^p\right) = -\frac{1}{p}\left({t_{2|j}}^p - \sum_{n=0}^{p} \binom{p}{n} {\Delta_{2|j}}^n {t_{2|j}}^{p-n}\right) \\
&= -\frac{1}{p}\left({t_{2|j}}^p - {t_{2|j}}^p - \sum_{n=1}^{p} \binom{p}{n} {\Delta_{2|j}}^n {t_{2|j}}^{p-n}\right) = \frac{1}{p}\left(\sum_{n=1}^{p} \binom{p}{n} {\Delta_{2|j}}^n {t_{2|j}}^{p-n}\right) \\
&\le \frac{1}{p}\left(\sum_{n=1}^{p} \binom{p}{n} \Delta_{2|j} {t_{2|j}}^{p-1}\right) \quad \text{(since } \Delta_{2|j} {t_{2|j}}^{p-1} \text{ is the dominant term)} \\
&= \frac{1}{p}\left(2^p - 1\right) \Delta_{2|j} {t_{2|j}}^{p-1} \le \frac{1}{p}\left(2^p - 1\right) 2\underline{\Delta}_{1|j} {t_{1|j}}^{p-1}
\end{aligned}$$

where the last inequality is due to the fact that $t_{2|j} \leq t_{1|j}$, and from Lemma IV.6, $\Delta_{2|j} = \underline{\Delta}_{2|j} + \overline{\Delta}_{2|j} \leq 2\overline{\Delta}_{2|j} = 2\underline{\Delta}_{1|j}$.

Since $\limsup_{\tau\to\infty} \beta^{(\tau)}_{2|j,p} \leq \limsup_{\tau\to\infty} \frac{1}{p}\left(2^p - 1\right) 2\underline{\Delta}_{1|j} {t_{1|j}}^{p-1} = \frac{2}{p}\left(2^p - 1\right)$, $\beta^{(\tau)}_{2|j,p}$ is bounded for all τ and (4.9.43) becomes

$$0 \leq 2e^{-1} \cdot e^{-\frac{2}{p}(2^p-1)} > 0.$$

Contradiction. Therefore, we conclude that $2 \leq k \leq j$, $\limsup_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k|j,p} < 1$.

We now know that the term $\left[1 - \frac{(-p+1)O_\tau\left(\frac{1}{t_{k|j}}\right)}{1-\underline{\alpha}^{(\tau)}_{k|j,p}}\right]$ in (4.9.37) is bounded away from zero and thus the term $\left[1 - \frac{(-p+1)O_\tau\left(\frac{1}{t_{k|j}}\right)}{1-\underline{\alpha}^{(\tau)}_{k|j,p}}\right]^{-1}$ does not blow up as $\tau \to \infty$.

Step 3. Show $\lim_{\tau\to\infty} \beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p} = 0$.

We are examining another term in (4.9.37), the term $e^{-\left(\beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p}\right)}$, whose behavior as $\tau \to \infty$ we want to understand.

Since $t_{k|j} \geq \tau$, it suffices to show that $\lim_{t_{k|j}\to\infty} \beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p} = 0$. With $t_{k-1|j} = t_{k|j} + \Delta_{k|j}$, we have

$$\begin{aligned}
\beta^{(\tau)}_{k|j,p} &= -\frac{1}{p}\left({t_{k|j}}^p - {t_{k-1|j}}^p\right) = -\frac{1}{p}\left({t_{k|j}}^p - \left(t_{k|j} + \Delta_{k|j}\right)^p\right) = -\frac{1}{p}\left({t_{k|j}}^p - \sum_{n=0}^{p} \binom{p}{n} {\Delta_{k|j}}^n {t_{k|j}}^{p-n}\right) \\
&= -\frac{1}{p}\left({t_{k|j}}^p - {t_{k|j}}^p - p\Delta_{k|j}{t_{k|j}}^{p-1} - \sum_{n=2}^{p} \binom{p}{n} {\Delta_{k|j}}^n {t_{k|j}}^{p-n}\right) \\
&= \left(\Delta_{k|j}{t_{k|j}}^{p-1} + \frac{1}{p}\sum_{n=2}^{p} \binom{p}{n} {\Delta_{k|j}}^n {t_{k|j}}^{p-n}\right) = \left(\alpha^{(\tau)}_{k|j,p} + \frac{1}{p}\sum_{n=2}^{p} \binom{p}{n} {\Delta_{k|j}}^n {t_{k|j}}^{p-n}\right). \quad (4.9.43)
\end{aligned}$$

Then

$$\beta^{(\tau)}_{k|j,p} \geq \alpha^{(\tau)}_{k|j,p}$$

and

$$\liminf_{\tau\to\infty} \beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p} \geq 0.$$

Also, since

$$t_{k|j}{}^{p-2}\Delta_{k|j}{}^{2} = \max\left\{t_{k|j}{}^{p-2}\Delta_{k|j}{}^{2}, t_{k|j}{}^{p-3}\Delta_{k|j}{}^{3}, \ldots, t_{k|j}{}^{2}\Delta_{k|j}{}^{p-2}, t_{k|j}\Delta_{k|j}{}^{p-1}, \Delta_{k|j}{}^{p}\right\}$$

when τ is large in (4.9.43), we know that

$$\beta^{(\tau)}_{k|j,p} \leq \alpha^{(\tau)}_{k|j,p} + (p-1)\, O_\tau\!\left(t_{k|j}{}^{p-2}\Delta_{k|j}{}^{2}\right)$$

or

$$\beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p} \leq (p-1)\, O_\tau\!\left(t_{k|j}{}^{p-1}\Delta_{k|j} \cdot \left(\frac{\Delta_{k|j}}{t_{k|j}}\right)\right).$$

Since from Step 2 and Corollary IV.4,

$$\limsup_{\tau\to\infty} \Delta_{k|j} t_{k|j}{}^{p-1} \leq \limsup_{\tau\to\infty} 2\underline{\Delta}_{k-1|j} t_{k-1|j}{}^{p-1} \leq 2$$

for all $k \geq 2$, and $\Delta_{k|j}$ is bounded for $k \geq 2$, we know that

$$\limsup_{\tau\to\infty} \beta^{(\tau)}_{k|j,p} \leq \limsup_{\tau\to\infty} (p-1)\, O_\tau\!\left(t_{k|j}{}^{p-1}\Delta_{k|j} \cdot \left(\frac{\Delta_{k|j}}{t_{k|j}}\right)\right) = 0$$

and thus,

$$\limsup_{\tau\to\infty} \beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p} \leq 0.$$

Therefore since $\liminf_{\tau\to\infty} \beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p} = \limsup_{\tau\to\infty} \beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p}$,

$$\lim_{\tau\to\infty} \beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p} = 0.$$

Step 4. Confirm the following statements are true:

$$\begin{aligned}
\text{term}_1 &= \lim_{\tau\to\infty} e^{-\left(\beta^{(\tau)}_{k|j,p} - \alpha^{(\tau)}_{k|j,p}\right)} = 1 \quad \text{(from Step 3)}\\
\text{term}_2 &= \lim_{\tau\to\infty} \left[1 + \frac{(-p+1)\, O_\tau\!\left(\frac{1}{t_{k-1|j}}\right)}{1 + \underline{\alpha}^{(\tau)}_{k-1|j,p}}\right] = 1, \text{ if } \lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k-1|j,p} \text{ exists.}\\
\text{term}_3 &= \lim_{\tau\to\infty} \left[1 - \frac{(-p+1)\, O_\tau\!\left(\frac{1}{t_{k|j}}\right)}{1 - \underline{\alpha}^{(\tau)}_{k|j,p}}\right]^{-1} = 1. \quad \text{(from Step 2)}
\end{aligned}$$

We want to determine the asymptotic behavior of the complete expression in (4.9.37) as $\tau \to \infty$. We do this by first pinpointing the asymptotic behavior of each individual term in (4.9.37) from Step 1.

For term$_1$, since e^x is a strictly monotone function, we apply Step 3 to obtain the result for any $2 \leq k \leq j$.

For term$_2$, if $\lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k-1|j,p}$ exists, then it is clear that since $\underline{\alpha}^{(\tau)}_{k|j,p} \geq 0$, that $\lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k-1|j,p} \geq 0$ and thus term$_2 = 1$.

For term$_3$, we know from Step 2 that $\limsup_{\tau\to\infty} \alpha^{(\tau)}_{k|j,p} < 1$. Then $1-\underline{\alpha}^{(\tau)}_{k|j,p}$ is bounded away from zero for all $\tau > 0$. Thus term$_3 = 1$.

Step 5. Show if $\underline{\nu}_{k-1}$ exists, then $\underline{\nu}_k$ exists and satisfies

$$e^{\underline{\nu}_k} \left(1 - \underline{\nu}_k\right) = e^{-\underline{\nu}_{k-1}} \left[1 + \underline{\nu}_{k-1}\right].$$

This is the final step of **Part 1(b).**

Assume $\underline{\nu}_{k-1} = \lim_{\tau\to\infty} \underline{\Delta}_{k-1|j} t_{k-1|j}{}^{p-1}$ exists. From Step 2, we know that $\underline{\nu}_{k-1} < 1$ since $k \geq 2$. Then taking the limit as $\tau \to \infty$ to each side of (4.9.37) and starting on the right-hand side of (4.9.37), we have

$$\lim_{\tau\to\infty} (4.9.37)_{RHS} = \lim_{\tau\to\infty} e^{-\underline{\alpha}^{(\tau)}_{k-1|j,p} \times \left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right)} \times \left[1+\underline{\alpha}^{(\tau)}_{k-1|j,p}\right] \times \left(1+o_\tau\left(\frac{1}{t_{k|j}}\right)\right) \times$$
$$e^{-\left(\beta^{(\tau)}_{k|j,p}-\alpha^{(\tau)}_{k|j,p}\right)} \times \left[1+\frac{(-p+1)\, O_\tau\left(\frac{1}{t_{k-1|j}}\right)}{1+\underline{\alpha}^{(\tau)}_{k-1|j,p}}\right] \times \left[1-\frac{(-p+1)\, O_\tau\left(\frac{1}{t_{k|j}}\right)}{1-\underline{\alpha}^{(\tau)}_{k|j,p}}\right]^{-1}.$$

Since $\underline{\nu}_{k-1}$ exists, we can use all of the results from Step 4 to evaluate this limit as $\tau \to \infty$:

$$\begin{aligned}
\lim_{\tau\to\infty} (4.9.37)_{RHS} &= \lim_{\tau\to\infty} e^{-\underline{\alpha}^{(\tau)}_{k-1|j,p}} \times \lim_{\tau\to\infty} \left(1 + \underline{\alpha}^{(\tau)}_{k-1|j,p}\right) \quad \text{(Step 4, terms 1, 2, 3, 4)} \\
&= e^{-\lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k-1|j,p}} \times \left(1 + \lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k-1|j,p}\right) \\
&= e^{-\underline{\nu}_{k-1}} \left[1 + \underline{\nu}_{k-1}\right],
\end{aligned}$$

where, in the first equality, we have used the fact $\lim_{\tau\to\infty} \left(1 + o_\tau\left(\frac{1}{t_{k|j}}\right)\right) = 1$, along with the observation that since $\tau = t_{j|j}$ and $t_{j|j} \leq t_{k|j}$, as $\tau \to \infty$, we know that

$t_{k|j} \to \infty$. Taking the limit as $\tau \to \infty$ on the left-hand side of (4.9.37), we have

$$\lim_{\tau\to\infty} (4.9.37)_{LHS} = \lim_{\tau\to\infty} e^{\underline{\alpha}^{(\tau)}_{k|j,p}} \left(1 - \underline{\alpha}^{(\tau)}_{k|j,p}\right),$$

and thus

$$\begin{aligned} \lim_{\tau\to\infty} e^{\underline{\alpha}^{(\tau)}_{k|j,p}} \left(1 - \underline{\alpha}^{(\tau)}_{k|j,p}\right) &= \lim_{\tau\to\infty} (4.9.37)_{LHS} = \lim_{\tau\to\infty} (4.9.37)_{RHS} \\ &= e^{-\underline{\nu}_{k-1}} \left[1 + \underline{\nu}_{k-1}\right]. \end{aligned} \tag{4.9.44}$$

Next, we want to show that $\lim_{\tau\to\infty} \alpha^{(\tau)}_{k|j,p}$ exists. We do this by examining the function in $(4.9.37)_{LHS}$ and use proof by contradiction: Consider the function $h(x) \overset{\triangle}{=} e^{x}(1-x)$ defined over $[0,1]$. Then h is 1-1, onto and differentiable on $[0,1]$, and it is clear that h is strictly decreasing over the same closed interval. Suppose that $\lim_{\tau\to\infty} \alpha^{(\tau)}_{k|j,p}$ does not exist. Then there are two values $a_1, a_2 \in [0,1]$ (Step 2), such that $a_1 \neq a_2$, and two increasing subsequences τ_n, τ_m such that $\lim_{m\to\infty} \tau_m = \infty$, $\lim_{n\to\infty} \tau_n = \infty$ and $\lim_{m\to\infty} \underline{\alpha}_{k|j,p}{}^{(\tau_m)} = a_1 \neq a_2 = \lim_{n\to\infty} \underline{\alpha}_{k|j,p}{}^{(\tau_n)}$. Since $h(x)$ is continuous,

$$h\left(\lim_{m\to\infty} \underline{\alpha}_{k|j,p}{}^{(\tau_m)}\right) = \lim_{m\to\infty} h\left(\underline{\alpha}_{k|j,p}{}^{(\tau_m)}\right) = \lim_{\tau\to\infty} h\left(\underline{\alpha}^{(\tau)}_{k|j,p}\right) \overset{(4.9.44)}{=} e^{-\underline{\nu}_{k-1}}\left(1 + \underline{\nu}_{k-1}\right)$$

and

$$h\left(\lim_{n\to\infty} \underline{\alpha}_{k|j,p}{}^{(\tau_n)}\right) = \lim_{n\to\infty} h\left(\underline{\alpha}_{k|j,p}{}^{(\tau_n)}\right) = \lim_{\tau\to\infty} h\left(\underline{\alpha}^{(\tau)}_{k|j,p}\right) \overset{(4.9.44)}{=} e^{-\underline{\nu}_{k-1}}\left(1 + \underline{\nu}_{k-1}\right).$$

But $h(x)$ is also $1-1$ so $h\left(\lim_{m\to\infty} \underline{\alpha}_{k|j,p}{}^{(\tau_m)}\right) \neq h\left(\lim_{n\to\infty} \underline{\alpha}_{k|j,p}{}^{(\tau_n)}\right)$. Contradiction. Thus $a_1 = a_2$ and $\lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k|j,p}$ exists and we now know that

$$e^{\underline{\nu}_k}\left(1 - \underline{\nu}_k\right) = e^{-\underline{\nu}_{k-1}}\left[1 + \underline{\nu}_{k-1}\right].$$

■

Proof of Part 1(c). *Show* $\underline{\nu}_k = \underline{\eta}_k$, $k \geq 1$.

This is proof by induction. From Theorem IV.7, Part 1 (a), we already know that $\underline{\nu}_1 = 1 = \underline{\eta}_1$. Now assume that $\underline{\nu}_{k-1} = \underline{\eta}_{k-1}$. Using Theorem IV.7, Part 1 (b), we know that $\underline{\nu}_k$ satisfies

$$e^x\,(1-x) = e^{-\underline{\nu}_{k-1}}\left[1+\underline{\nu}_{k-1}\right] = e^{-\underline{\eta}_{k-1}}\left[1+\underline{\eta}_{k-1}\right]. \tag{4.9.45}$$

Since $\underline{\eta}_{k-1} \in [0,1]$, there is only one real solution to (4.9.45). Since (4.9.45) is the Nitadori recursion relation from (4.2.19), we conclude that $\underline{\nu}_k = \underline{\eta}_k$. ■

(This is the end of the first part of Theorem IV.7.)

Part 2. *Show that for an optimal τ, j quantizer with $j \geq 2$, for each $1 \leq k \leq j-1$, with*

$$r^{(\tau)}_{k|j,p} = \frac{\underline{\Delta}_{k|j}}{\underline{\Delta}_{k+1|j}},$$

that $r_k = \lim_{\tau\to\infty} r^{(\tau)}_{k|j,p}$ exists and is given by

$$r_k = \frac{\underline{\eta}_k}{\underline{\eta}_{k+1}}.$$

Assuming that we already know that $\underline{\nu}_k = \lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k|j,p}$ exists and that $\underline{\nu}_k = \underline{\eta}_k$ (from Theorem IV.7, Part 1(c)), for any $1 \leq k \leq j-1$,

$$r^{(\tau)}_{k|j,p} = \frac{\underline{\Delta}_{k|j}}{\underline{\Delta}_{k+1|j}} = \frac{\underline{\alpha}^{(\tau)}_{k|j,p}}{\underline{\alpha}^{(\tau)}_{k+1|j,p}} \cdot \left(\frac{t_{k+1|j}}{t_{k|j}}\right)^{p-1}.$$

Then using Lemma IV.12, Part 2, we know that

$$r_k = \lim_{\tau\to\infty} r^{(\tau)}_{k|j,p} = \frac{\underline{\eta}_k}{\underline{\eta}_{k+1}}.$$

(This is the end of the proof of the second part of Theorem IV.7.)
The proof of Theorem IV.7 is now complete. ■

4.10 Corollaries to Theorem IV.7.

In this section we present several corollaries to Theorem IV.7 which hold true for optimal τ, j quantizers designed for GE-sources, remarking that these corollaries also hold true for optimal scalar quantizers designed for the same sources.

The first corollary to Theorem IV.7 shows that when considering limits to $\underline{\Delta}_{k|j}(t_{l|j})^{p-1}$, $k \neq l$, as $\tau \to \infty$, this limit exists and that it is the index k of the half step, not the index l of the threshold that determines what this limit will be.

Corollary IV.14. *Suppose we have an optimal τ, j quantizer, $j \geq 2$. Suppose for some $1 \leq k \leq j$, $\lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k|j,p}$ exists, then for any $1 \leq l \leq j$, $\lim_{\tau\to\infty} \underline{\Delta}_{k|j}(t_{l|j})^{p-1} = \underline{\eta}_k$.*

Proof. First, let $1 \leq k \leq j$ and suppose that $\lim_{\tau\to\infty} \underline{\alpha}^{(\tau)}_{k-1|j,p}$ exists. Now let $1 \leq l \leq j$. If $t_{k|j} > 0$, then from Lemma IV.12, Part 2, we know that

$$\lim_{\tau\to\infty} \left(\frac{t_{l|j}}{t_{k|j}}\right)^{p-1} = 1,$$

and thus

$$\begin{aligned}\lim_{\tau\to\infty} \underline{\Delta}_{k|j}(t_{l|j})^{p-1} &= \lim_{\tau\to\infty} \underline{\Delta}_{k|j}\left(t_{k|j} \cdot \frac{t_{l|j}}{t_{k|j}}\right)^{p-1} = \lim_{\tau\to\infty} \underline{\Delta}_{k|j}(t_{k|j})^{p-1} \times \left(\frac{t_{l|j}}{t_{k|j}}\right)^{p-1} \\ &= \lim_{\tau\to\infty} \underline{\Delta}_{k|j}(t_{k|j})^{p-1} \times \lim_{\tau\to\infty} \left(\frac{t_{l|j}}{t_{k|j}}\right)^{p-1} = \underline{\eta}_k.\end{aligned}$$

■

Similar to the definition of $\underline{\eta}_k$ with respect to the half step $\underline{\Delta}_{k|j}$, we define $\eta_k \overset{\triangle}{=} \lim_{\tau\to\infty} \Delta_{k|j}(t_{k|j})^{p-1}$, a limit that focuses on the interaction between the whole step $\Delta_{k|j}$ and the cell threshold $t_{k|j}$. The next corollary shows that the relationship between η_k, $\underline{\eta}_k$, and $\underline{\eta}_{k-1}$ is much like the relationship between the size of an optimal τ, j quantization cell $\Delta_{k|j}$ and its two half cell widths $\underline{\Delta}_{k|j}$ and $\underline{\Delta}_{k-1|j}$. Not only does this corollary provide a quick way to generate the sequence η_k from the sequence $\underline{\eta}_k$, but the fact that η_k, $\underline{\eta}_k$, and $\underline{\eta}_{k-1}$ "behave" like $\Delta_{k|j}$, $\underline{\Delta}_{k|j}$, and $\underline{\Delta}_{k-1|j}$ provides some insight on how we might use η_k and $\underline{\eta}_k$ for asymptotic cell size estimation, a topic that will be discussed later.

Corollary IV.15. *Suppose we have an optimal τ, j quantizer designed for a GE-source with $j \geq 2$. Define $\eta_k \overset{\triangle}{=} \lim_{\tau\to\infty} \alpha^{(\tau)}_{k|j,p}$. Then for $2 \leq k \leq j$, η_k exists and*

$$\eta_k = \underline{\eta}_k + \underline{\eta}_{k-1}.$$

Proof. Let $2 \leq k \leq j$, $j \geq 2$. Since $\Delta_{k|j}(t_{k|j})^{p-1} = \underline{\Delta}_{k|j}(t_{k|j})^{p-1} + \underline{\Delta}_{k-1|j}(t_{k|j})^{p-1}$, using Corollary IV.14, we have

$$\begin{aligned}\lim_{\tau\to\infty} \Delta_{k|j}(t_{k|j})^{p-1} &= \lim_{\tau\to\infty} \underline{\Delta}_{k|j}(t_{k|j})^{p-1} + \underline{\Delta}_{k-1|j}(t_{k|j})^{p-1} \\ &= \lim_{\tau\to\infty} \underline{\Delta}_{k|j}(t_{k|j})^{p-1} + \lim_{\tau\to\infty} \underline{\Delta}_{k-1|j}(t_{k|j})^{p-1} \\ &= \underline{\eta}_k + \lim_{\tau\to\infty} \underline{\Delta}_{k-1|j}(t_{k|j})^{p-1} \\ &= \underline{\eta}_k + \underline{\eta}_{k-1}.\end{aligned}$$

Thus $\eta_k = \underline{\eta}_k + \underline{\eta}_{k-1}$. ■

The following corollary is an extension of the result for $\underline{r}_k$ in Theorem IV.7. Just as η_k focuses on the whole step $\Delta_{k|j}$ as compared to $\underline{\eta}_k$ and its focus on $\underline{\Delta}_{k|j}$, we define r_k as a ratio of whole steps as opposed to $\underline{r}_k$ which is a ratio of half steps. The corollary also shows how to generate the extension of r_k from the sequence η_k.

Corollary IV.16. *Suppose we have an optimal τ, j quantizer designed for a GE-source with $j \geq 3$. Then for $2 \leq k \leq j-1$, define*

$$r^{(\tau)}_{k|j,p} \triangleq \frac{\Delta_{k|j}}{\Delta_{k+1|j}}.$$

Then

$$r_k \triangleq \lim_{\tau\to\infty} r^{(\tau)}_{k|j,p}$$

exists and

$$r_k = \frac{\eta_k}{\eta_{k+1}}.$$

Proof. Let $2 \leq k \leq j-1$, $j \geq 3$. Since

$$\Delta_{k+1|j} = \underline{\Delta}_{k|j} + \underline{\Delta}_{k+1|j}$$

and if $t_{k|j} > 0$, we have

$$\lim_{\tau\to\infty} r^{(\tau)}_{k|j,p} = \lim_{\tau\to\infty} \frac{\Delta_{k|j}(t_{k|j})^{p-1}}{\left(\underline{\Delta}_{k|j} + \underline{\Delta}_{k+1|j}\right)(t_{k|j})^{p-1}}$$

$$
\begin{aligned}
&= \frac{\lim_{\tau\to\infty} \Delta_{k|j}(t_{k|j})^{p-1}}{\lim_{\tau\to\infty} \underline{\Delta}_{k|j}(t_{k|j})^{p-1} + \lim_{\tau\to\infty} \underline{\Delta}_{k+1|j}(t_{k|j})^{p-1}} \\
&= \frac{\underline{\eta}_k + \underline{\eta}_{k-1}}{\underline{\eta}_k + \underline{\eta}_{k+1}} = \frac{\eta_k}{\eta_{k+1}},
\end{aligned}
$$

where to get the second equality we have used the fact that limit in the numerator exists by Corollary IV.15 and that the limit in the denominator exists by Corollary IV.14 and Theorem IV.7. The last equality comes from Corollary IV.15. Thus $r_k = \frac{\eta_k}{\eta_{k+1}}$. ■

Intuitively, since we've shown in Corollary IV.15 that η_k "behaves" like a step size, we see that the ratio r_k can be "thought of" as the ratio of consecutive cell sizes in an optimal τ, j quantizer when $\tau >> 0$.

The following corollary will be used later during the discussion on applications of Theorem IV.1.

Corollary IV.17. *Suppose we have an optimal τ, j quantizer designed for a GE-source with $p > 1$ and $j \geq 2$. Fix $2 \leq k \leq j$. Then $\lim_{\tau\to\infty} \frac{\underline{\Delta}_{k|j}}{\underline{\Delta}_{1|j}} = \frac{1}{\underline{\rho}_1 \cdot \underline{\rho}_2 \cdots \underline{\rho}_{k-1}}$, where $\underline{\rho}_k \overset{\triangle}{=} \frac{\underline{\eta}_k}{\underline{\eta}_{k+1}}$.*

Proof. With $j \geq 2$, for any $2 \leq k \leq j$,

$$
\frac{\underline{\Delta}_{k|j}}{\underline{\Delta}_{1|j}} = \frac{\underline{\Delta}_{k|j}}{\underline{\Delta}_{k-1|j}} \cdot \frac{\underline{\Delta}_{k-1|j}}{\underline{\Delta}_{k-2|j}} \cdots \frac{\underline{\Delta}_{2|j}}{\underline{\Delta}_{1|j}} = \frac{1}{\underline{r}^{(\tau)}_{k-1|j,p}} \cdot \frac{1}{\underline{r}^{(\tau)}_{k-2|j,p}} \cdots \frac{1}{\underline{r}^{(\tau)}_{1|j,p}},
$$

where $\underline{r}^{(\tau)}_{k|j,p} \overset{\triangle}{=} \frac{\underline{\Delta}_{k|j}}{\underline{\Delta}_{k+1|j}}$. Then

$$
\begin{aligned}
\lim_{\tau\to\infty} \frac{\underline{\Delta}_{k|j}}{\underline{\Delta}_{1|j}} &= \lim_{\tau\to\infty} \frac{1}{\underline{r}^{(\tau)}_{k-1|j,p} \cdot \underline{r}^{(\tau)}_{k-2|j,p} \cdots \underline{r}^{(\tau)}_{1|j,p}} \\
&= \frac{1}{\lim_{\tau\to\infty} \underline{r}^{(\tau)}_{k-1|j,p} \cdot \lim_{\tau\to\infty} \underline{r}^{(\tau)}_{k-2|j,p} \cdots \lim_{\tau\to\infty} \underline{r}^{(\tau)}_{1|j,p}} \\
&= \frac{1}{\underline{\rho}_{k-1} \cdot \underline{\rho}_{k-2} \cdots \underline{\rho}_1}.
\end{aligned}
$$

■

There are obvious, analogous corollaries to Theorem IV.1. Rather than making formal statements for optimal scalar quantization of GE-sources, Table 4.3 summarizes these three corollaries to Theorem IV.1 and Theorem IV.7.

Table 4.4 shows some values from the sequences $\underline{\eta}_k$, η_k, $\underline{r}_k$, and r_k.

Table 4.3: Table of Additional Corollaries that Hold for Optimal Scalar Quantization of GE-sources.

Corollary	Optimal τ, j quantization	Optimal scalar quantization
IV.14	$\underline{\Delta}_{k\|j}(t_{l\|j})^{p-1} \xrightarrow{\tau\to\infty} \underline{\eta}_k$	$\underline{\Delta}_k^{(N)}(t_l^{(N)})^{p-1} \xrightarrow{N\to\infty} \underline{\eta}_k$
IV.15	$\Delta_{k\|j}(t_{k\|j})^{p-1} \xrightarrow{\tau\to\infty} \underline{\eta}_k + \underline{\eta}_{k-1}$	$\Delta_k^{(N)}(t_k^{(N)})^{p-1} \xrightarrow{N\to\infty} \underline{\eta}_k + \underline{\eta}_{k-1}$
IV.16	$\frac{\Delta_{k\|j}}{\Delta_{k+1\|j}} \xrightarrow{\tau\to\infty} \frac{\eta_k}{\eta_{k+1}}$	$\frac{\Delta_k^{(N)}}{\Delta_{k+1}^{(N)}} \xrightarrow{N\to\infty} \frac{\eta_k}{\eta_{k+1}}$
IV.17	$\frac{\underline{\Delta}_{k\|j}}{\underline{\Delta}_{1\|j}} \xrightarrow{\tau\to\infty} \frac{1}{\underline{\rho}_{k-1}\cdot\underline{\rho}_{k-2}\cdots\underline{\rho}_1}$	$\frac{\underline{\Delta}_k^{(N)}}{\underline{\Delta}_1^{(N)}} \xrightarrow{N\to\infty} \frac{1}{\underline{\rho}_{k-1}\cdot\underline{\rho}_{k-2}\cdots\underline{\rho}_1}$

4.11 Applications.

4.11.1 Asymptotic Approximations to $\underline{\Delta}_k^{(N)}$.

Since the facts stated in Theorem IV.1 describe limiting relationships between the Nitadori sequence and 1) the combined behavior of the half step and quantization threshold of a quantization cell, and 2) the behavior of the half steps of neighboring quantization cells in optimal GE-quantizers, these facts naturally lend themselves to the creation of a hamster estimator which uses the values of the Nitadori sequence. We define, for each $p > 1$ and $N \geq 2$, our half step estimator to be

$$\widehat{\underline{\Delta}}_k^{(N)} \triangleq \frac{\underline{\eta}_k}{(t_k^{(N)})^{p-1}}. \tag{4.11.46}$$

Given this estimator, a natural question to ask is: For a given GE-source with parameter p, how large must N be in order for (4.11.46) to yield good approximations to $\underline{\Delta}_k^{(N)}$ (and by association, $\frac{\underline{\Delta}_k^{(N)}}{\underline{\Delta}_{k+1}^{(N)}}$)? Another related question is: How does changing the value of p affect the accuracy of the approximations given by Theorem IV.1? To investigate these questions, we have computed the threshold and half step data of optimal scalar quantizers for GE-sources with $p = 1.5, 2$ (Gaussian case), 10 for values of $N = 32, 64, 128, 256, 512, 1024$. (For reference, see Figure 4.5 for an illustration of the pdfs of the GE-sources with $p = 1.5, 2, 10$, along with the pdf for the one-sided exponential source ($p = 1$).) For each N-level, p-determined, optimal scalar quantizer, we have used this data to compute $\widehat{\underline{\Delta}}_k^{(N)}$ for each quantization cell k in the N-level quantizer, and we compare it to the actual, corresponding value for $\underline{\Delta}_k^{(N)}$.

Table 4.4: Table of Selected $\underline{\eta}_k$, η_k, $\underline{r}_k$, and r_k Sequence Values.

k	$\underline{\eta}_k$	η_k	$\underline{r}_k$	r_k
1	1.0000	Not defined	1.6846	Not defined
2	0.5936	1.5936	1.4002	1.5661
3	0.4240	1.0176	1.2844	1.3495
4	0.3301	0.7540	1.2209	1.2558
5	0.2704	0.6004	1.1807	1.2025
6	0.2290	0.4993	1.1530	1.1678
7	0.1986	0.4276	1.1326	1.1434
8	0.1753	0.3739	1.1170	1.1252
16	0.0906	0.1870	1.0604	1.0624
32	0.0461	0.0936	1.0307	1.0312
64	0.0232	0.0468	1.0155	1.0156
128	0.0117	0.0234	1.0078	1.0078
256	0.0058	0.0117	1.0039	1.0039
512	0.0029	0.0059	1.0020	1.0020
1024	0.0015	0.0029	1.0010	1.0010
2048	$7.3222e-004$	0.0015	1.0005	1.0005

Does the data show $\widehat{\underline{\Delta}}_k^{(N)}$ is a good approximation for $\underline{\Delta}_k^{(N)}$? When is $\widehat{\underline{\Delta}}_k^{(N)}$ useful? In Figure 4.6, we have plotted data for the ratio $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}} = \frac{\underline{\eta}_k}{\underline{\Delta}_k^{(N)}(\underline{t}_k^{(N)})^{p-1}}$ versus cell index k. On initial inspection of the curves shown in this figure, it is easy to see the macro trends influenced by the facts of Theorem IV.1 as well as trends that are not explained or predicted by Theorem IV.1. The trends resulting from basing our estimator $\widehat{\underline{\Delta}}_k^{(N)}$ on Theorem IV.1 are:

- The numerical results confirm our theoretical expectations, namely that, for all values of p and for all values of N, the ratio $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}} = \frac{\underline{\eta}_k}{\underline{\Delta}_k^{(N)}(\underline{t}_k^{(N)})^{p-1}}$ is close to one for small k (which corresponds to quantization cells far from the origin, residing in the pdf tail region) and this is true even for moderate values of k.
- In general, the closer p is to 1, the better $\widehat{\underline{\Delta}}_k^{(N)}$ is at estimating $\underline{\Delta}_k^{(N)}$.
- As N increases, the number of quantization cells for which $\widehat{\underline{\Delta}}_k^{(N)}$ is a reliable estimator increases. Again, this is due to the fact that as N increases, the number of quantization cells that lies in the pdf tail region increases.

The trends that we observe that are outside the scope of Theorem IV.1 are:

- For quantization cells that are close to the origin (i.e., k close to N), $\widehat{\underline{\Delta}}_k^{(N)}$ loses its ability to accurately estimate $\underline{\Delta}_k^{(N)}$. This observation is not unexpected since:

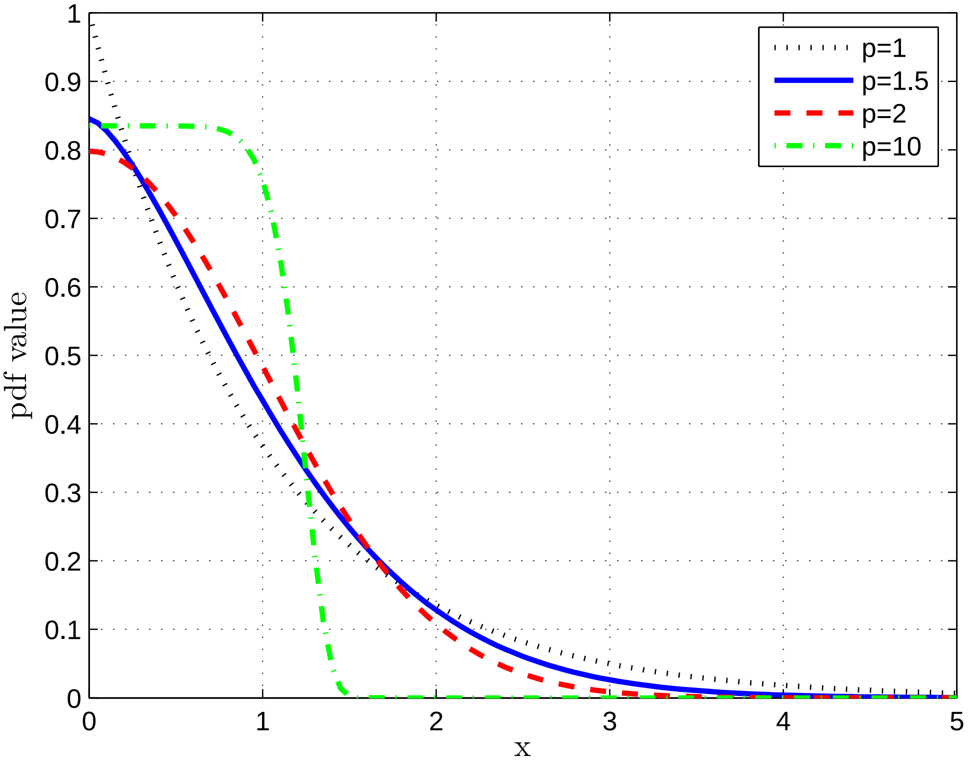


Figure 4.5: GE-sources considered for data comparison with $p = 1, 1.5, 2, 10$.

1) Theorem IV.1 does not contain information regarding the behavior of $\widehat{\underline{\Delta}}_k^{(N)}$ as N increases for cells near the origin, and 2) close to the origin, $t_k^{(N)} << 1$ and is on the same order as $\underline{\Delta}_k^{(N)}$ and thus taking $t_k^{(N)}$ to the power $(p-1)$ causes $\widehat{\underline{\Delta}}_k^{(N)}$ to blow up.

- For quantization cells lying somewhere between the origin and the pdf tail region, $\widehat{\underline{\Delta}}_k^{(N)}$ seems to have some ability to track the size of $\underline{\Delta}_k^{(N)}$ (most likely due to the fact that $t_k^{(N)}$ is bounded away from zero in this region) but this ability declines as k increases to N (since $t_k^{(N)}$ decreases towards zero). Theorem IV.1 also makes no statement about how $\widehat{\underline{\Delta}}_k^{(N)}$ behaves in this part of the pdf support region so this observation is somewhat interesting.

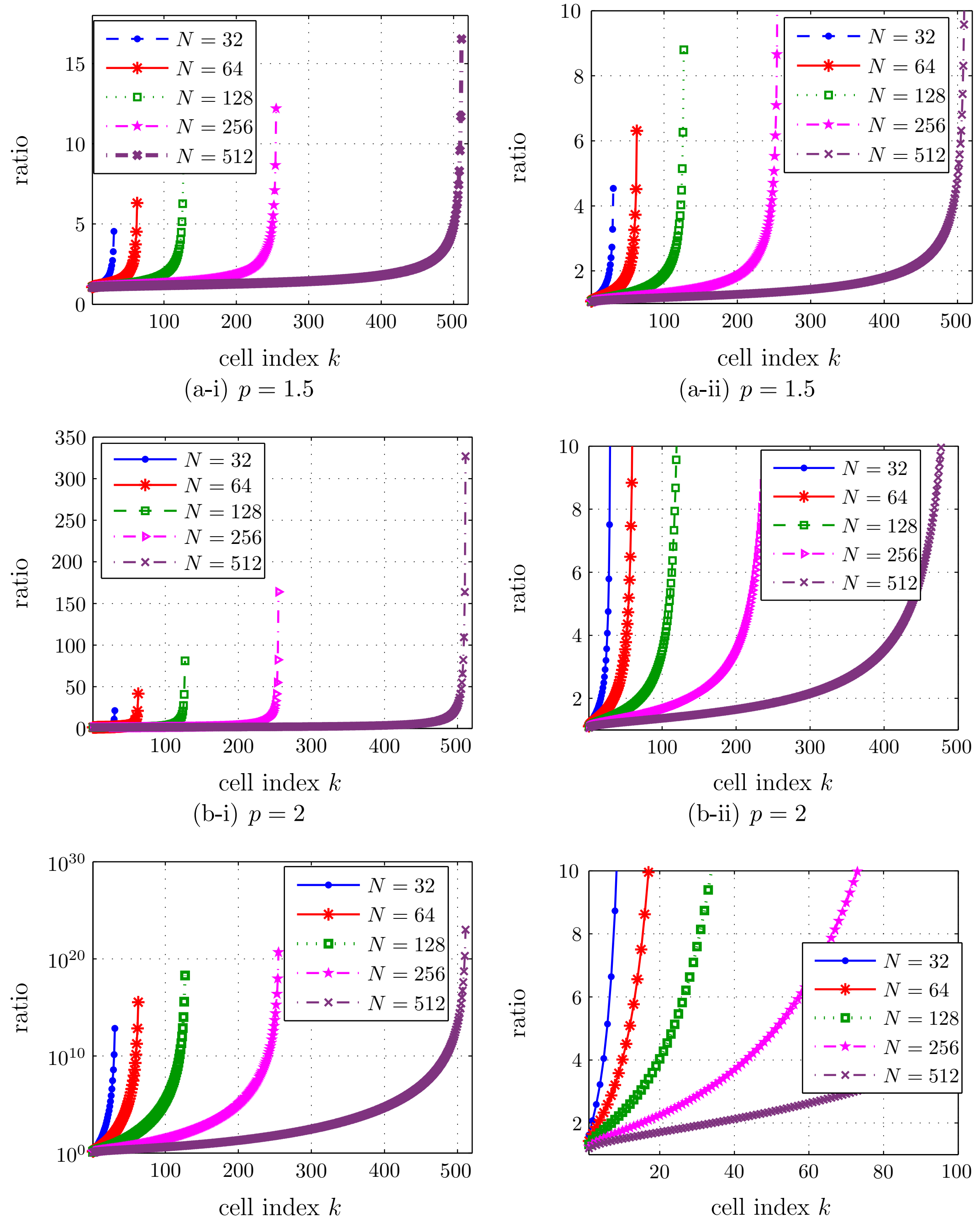
In Figure 4.7, we have re-plotted data for the ratio $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ versus relative cell index $\frac{k}{N}$. Curiously, we observe that, for each value of p, the ratio curves for $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ appear to be superimposed on top of each other, indicating that the behavior of $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ as a function of $\frac{k}{N}$ is relatively insensitive to the value of N. To investigate this apparent insensitivity to N (and bearing in mind that the scales used to plot the data may be skewing our perception of insensitivity), we look at how changing the value of N affects the accuracy achieved by $\widehat{\underline{\Delta}}_k^{(N)}$. To do this, for each p and each N, we fix a value $T \geq 0$ and initially define the *maximum cell index for tolerance* $T\%$

$$k_{T,N,p} \triangleq \left\{ k \in \{1, 2, \ldots, N\} : \ \forall i \leq k, \ \left| \frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}} - 1 \right| \leq \frac{T}{100} \right\}, \tag{4.11.47}$$

i.e., $k_{T,N,p}$ is the largest cell index for which $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ is within $T\%$ of the value 1. Using this definition, we define the *maximum relative cell index for tolerance* $T\%$ as

$$\frac{k_{T,N,p}}{N}.$$

In Figure 4.8, we have plotted $\frac{k_{T,N,p}}{N}$ for tolerances $T = 10\%, 20\%, 30\%, 50\%, 100\%, 200\%$, and we note that across all values of p shown, the curves appear to be converging as N increases, with $\frac{k_{T,N,p}}{N}$ increasing as N increases. This observation concurs with the observation made from viewing Figure 4.7 where we saw that for each source, as indicated by p, the ratio curves seem to be superimposed on top of each other. If we use the fact that for GE-sources, the point densities converge to optimal point densities as N increases, we can postulate a limiting connection (in N) between $\frac{k_{T,N,p}}{N}$ and the tail function of the optimal cumulative point distribution $1 - \Lambda_p(x)$, $x \geq 0$, to hypothesize that for each $T\%$, there is a limiting value x for which $1 - \Lambda_p(x) = \lim_{N \to \infty} \frac{k_{T,N,p}}{N}$ such that for any $y \geq x$, $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ is within $T\%$ of 1. Finally, we note again that as p increases, the estimator $\widehat{\underline{\Delta}}_k^{(N)}$ produces less accurate estimates for $\underline{\Delta}_k^{(N)}$ and this is easily seen in our data by the fact that for fixed T and fixed N, we see that the values for $\frac{k_{T,N,1.5}}{N} \geq \frac{k_{T,N,2}}{N} \geq \frac{k_{T,N,10}}{N}$.



(a-i) $p = 1.5$

(a-ii) $p = 1.5$

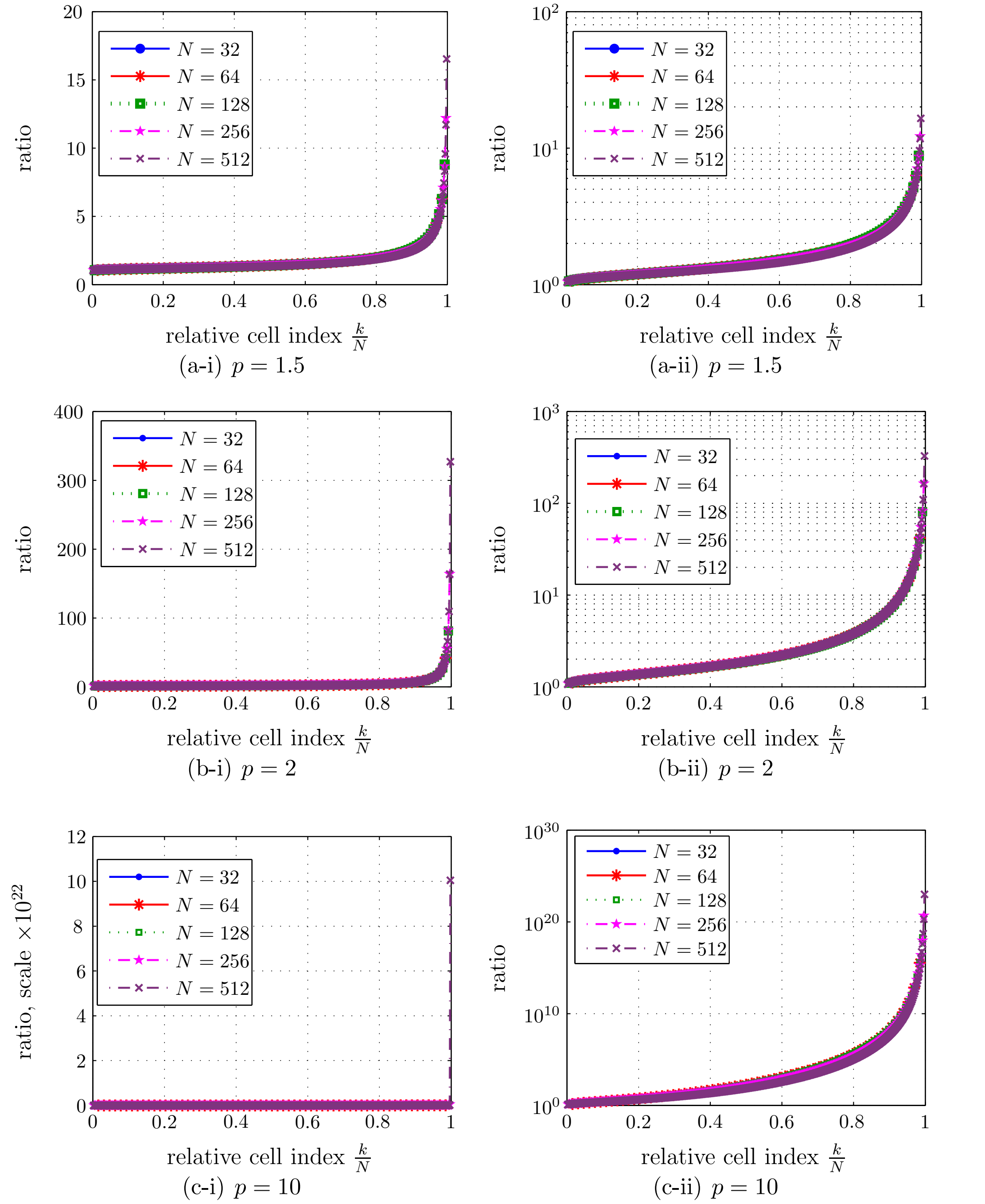
(b-i) $p = 2$

(b-ii) $p = 2$

(c-i) $p = 10$

(c-ii) $p = 10$

Figure 4.6: The ratio $\frac{\widehat{\Delta}_k^{(N)}}{\Delta_k^{(N)}}$ plotted versus cell index k for various N-level GE-source optimal scalar quantizers in two different scales: (a) $p = 1.5$, (b) $p = 2$, (c) $p = 10$. Note: The larger the cell index k is, the closer the quantization cell is to the origin.



(a-i) $p = 1.5$

(a-ii) $p = 1.5$

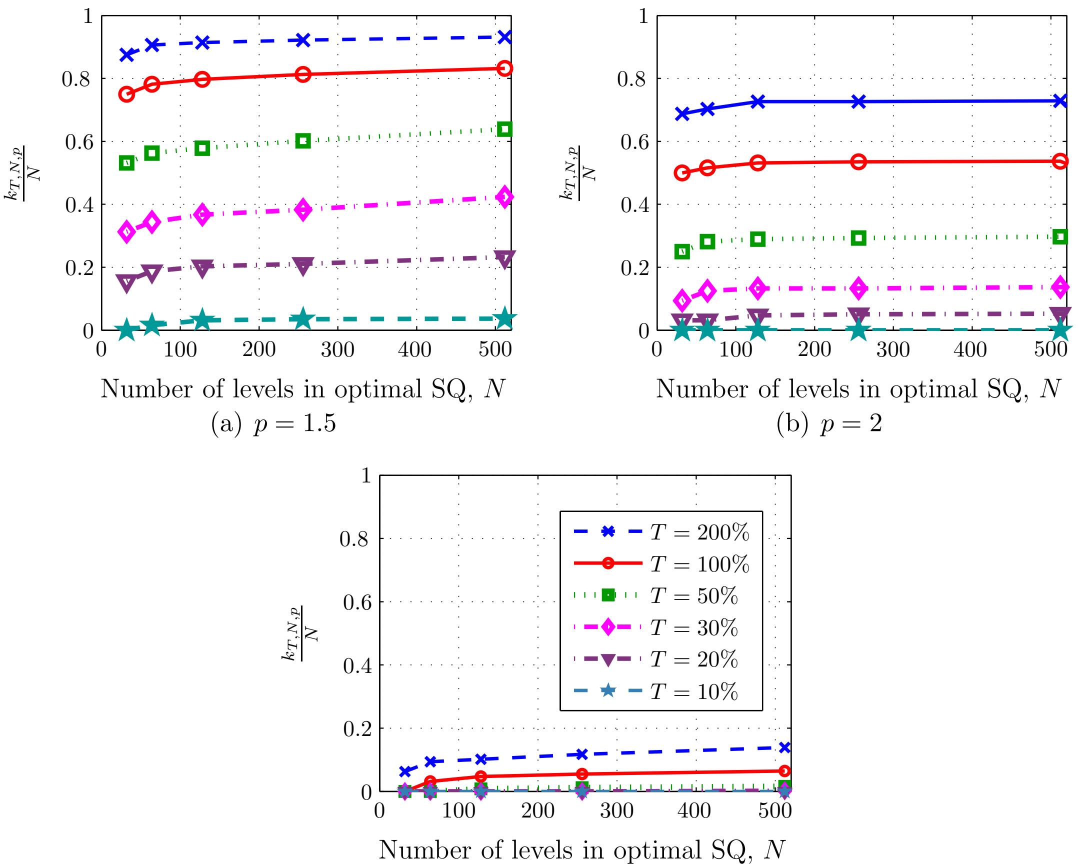
(b-i) $p = 2$

(b-ii) $p = 2$

(c-i) $p = 10$

(c-ii) $p = 10$

Figure 4.7: The ratio $\frac{\widehat{\Delta}_k^{(N)}}{\Delta_k^{(N)}}$ plotted against the relative cell index $\frac{k}{N}$ for various N-level GE-source optimal scalar quantizers: (a) $p = 1.5$, (b) $p = 2$, (c) $p = 10$. Note: (*-i) plots are linear scale and (*-ii) are semilog scale. Also, note that the larger the relative cell index $\frac{k}{N}$ is, the closer the quantization cell is to the origin.



(a) $p = 1.5$

(b) $p = 2$

(c) $p = 10$

Figure 4.8: Maximum relative index for tolerance $T\%$ (given by $\frac{k_{T,N,p}}{N}$) vs. N for various N-level GE-MMSE quantizers with $N = 32, 64, 128, 256, 512$ and $p = 1.5, 2, 10$.

The flip-side of looking at when our estimator can be used reliably is to consider when our estimator is not useful at all. Re-examining the plots shown in Figure 4.7, we notice that for relative cell indices around $0.85 - 0.95$, every ratio curve (regardless of p or N) shows a sudden increase in slope, indicating a dramatic *breakdown* in the estimator's ability to approximate $\underline{\Delta}_k^{(N)}$. To be more concrete, we somewhat arbitrarily define the *knee* of the $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ curve to be the point at which the slope of each of the semilog plots in Figure 4.7 is 45° (relative to the scale shown in those plots). The corresponding relative index will be called the *relative breakdown index* $b_{knee,N,p}$ so that, in this way, we have tied relative indices to the knee of the ratio curves shown. Related to $b_{knee,N}$, we also define the breakdown threshold $t_{knee,N,p} \triangleq t^{(N)}_{\text{round}(N \times b_{knee,N,p})}$. Table 4.5 lists the actual slope values at the knee of the curve and it also lists the approximate values for $b_{knee,N,p}$ and $t_{knee,N,p}$ according to p and N, where the value for $b_{knee,N}$ corresponds to the $x-$coordinate of the data point that is closest to the point of tangency between the $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ curve and a 45° line.

For each p, the values of $b_{knee,N,p}$ do not appear to vary much with N. However, we note that as p increases, for each N, $b_{knee,N,p}$ appears to decrease. (The average values are $\overline{b}_{knee,1.5} = 0.92658, \overline{b}_{knee,2} = 0.89804, \overline{b}_{knee,10} = 0.87656$.) We remark that the corresponding breakdown thresholds are all much less than 1, indicating that they all reside near the origin. Since for these sources, it is known that the point densities of N-level optimal quantizers converge to the optimal point density for each source [2] and because we have already seen that changing N does not appear to have much of an effect on $b_{knee,N,p}$, the fact that, for each p, the values for $t_{knee,N,p}$ appear to cluster is not unexpected. It is interesting to note, however, that $t_{knee,N,2}$ seems to be larger than either $t_{knee,N,1.5}$ and $t_{knee,N,10}$.

Table 4.5: Table of slope values and breakdown thresholds t_{knee} at the knee of the curve b_{knee}.

p	actual slope	breakdown threshold t_{knee}, $N = 32, 64, 128, 256, 512$
1.5	18.52	$0.1523, 0.1537, 0.1744, 0.1765, 0.1816$
2	35.71	$0.1338, 0.1348, 0.1534, 0.1525, 0.1527$
10	3.33×10^{11}	$0.0824, 0.0825, 0.0919, 0.0989, 0.1078$

Thus, summarizing what we have seen in our data, we make the following observations regarding the usefulness of the $\widehat{\underline{\Delta}}_k^{(N)}$ estimator:

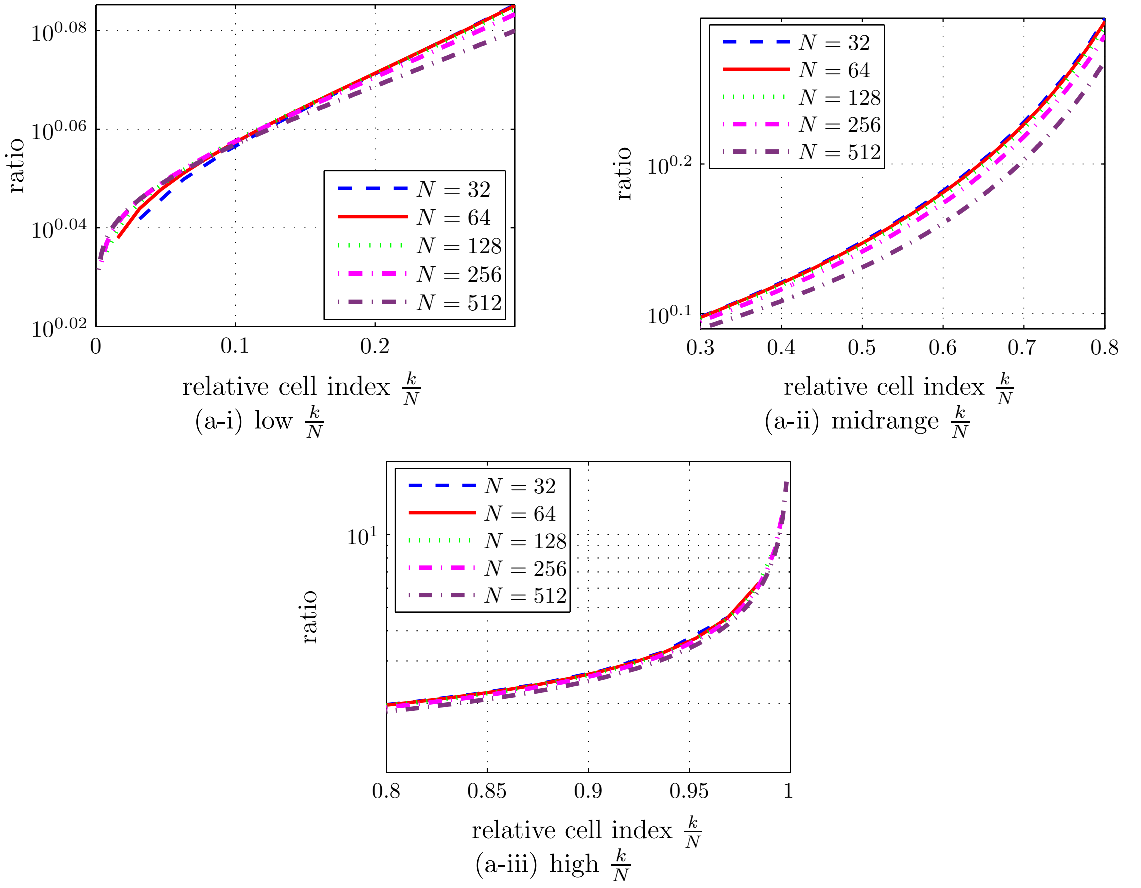
1. The overall performance of the estimator is best when p is close to 1. As p increases away from 1, the estimator becomes less accurate.

2. For each p, the estimator is very good for cells in the pdf tail region (when $k << N$ and $\frac{k}{N} << 1$), and surprisingly, it is moderately good (appears to be bounded) for $.2 < \frac{k}{N} < .87$ which we will refer to as the mid-range of relative indices.

3. To find the quantization cells for which the estimator has good performance, we can use $\frac{k_{T,N,p}}{N}$.

4. For relative cell indices greater than 0.87, the estimator fails. Alternate methods should be used for the quantization cells that corresponding to relative cell indices in this region.

5. There appears to be rapid convergence (in N) for the function $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ with respect to relative index $\frac{k}{N}$. For each p, there may exist a limiting function which gives the best performance that $\widehat{\underline{\Delta}}_k^{(N)}$ can attain (as measured by the ratio $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$) for a given value of $\frac{k}{N}$.

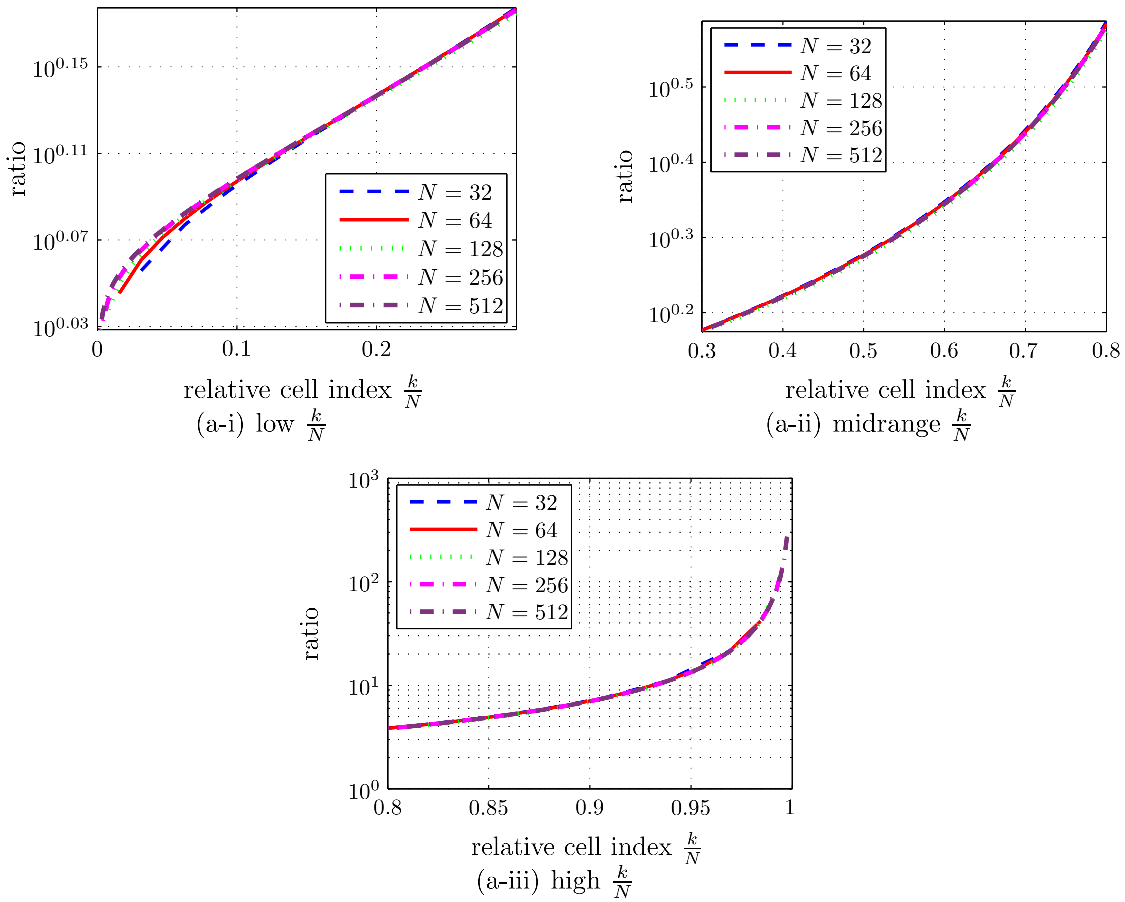


(a-i) low $\frac{k}{N}$

(a-ii) midrange $\frac{k}{N}$

(a-iii) high $\frac{k}{N}$

Figure 4.9: Case $p = 1.5$: Plotting $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ according to relative cell index range. Three regions for $\frac{k}{N}$ that span $[0, 1]$ are shown.

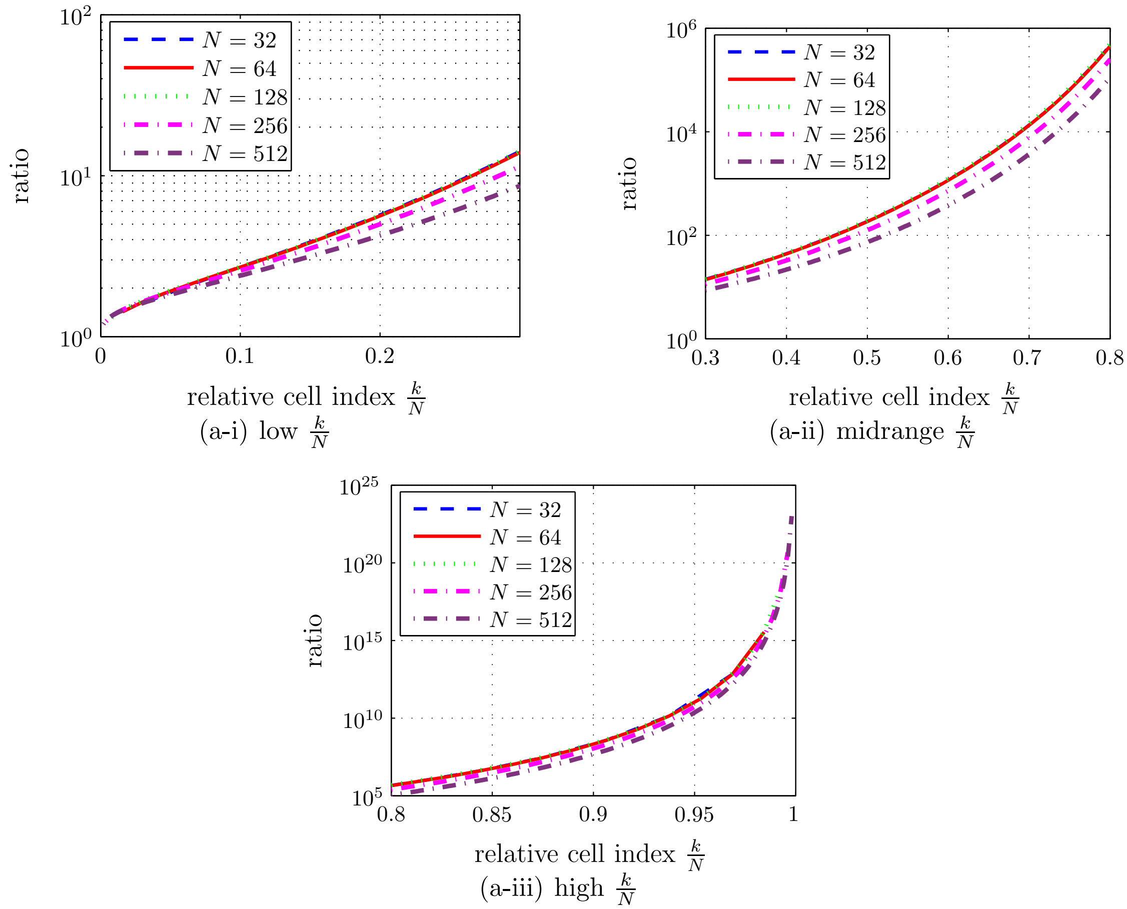


(a-i) low $\frac{k}{N}$

(a-ii) midrange $\frac{k}{N}$

(a-iii) high $\frac{k}{N}$

Figure 4.9: Case $p = 2$: Plotting $\frac{\widehat{\Delta}_k^{(N)}}{\Delta_k^{(N)}}$ according to relative cell index range. Three regions for $\frac{k}{N}$ that span $[0, 1]$ are shown.

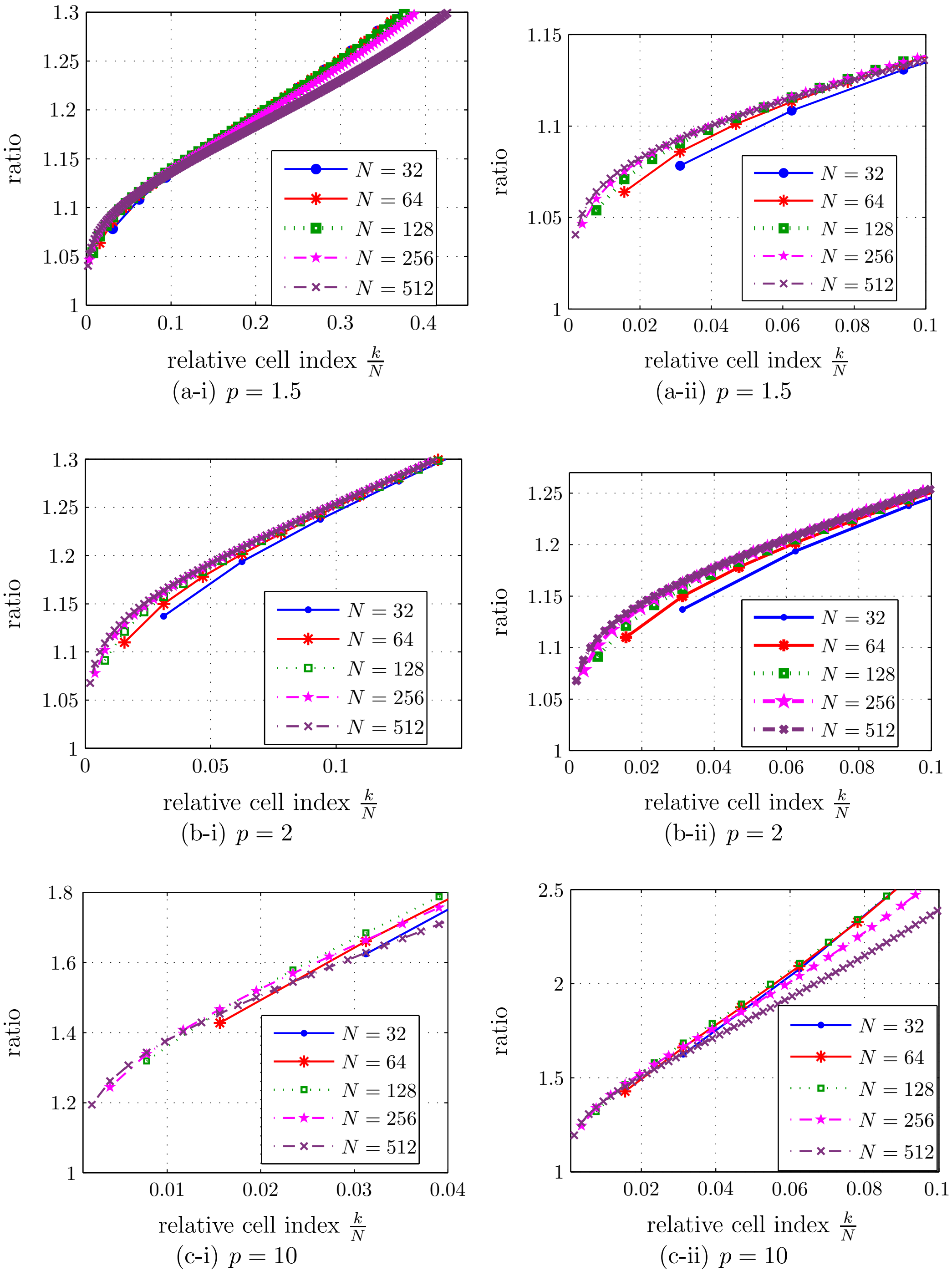


(a-i) low $\frac{k}{N}$

(a-ii) midrange $\frac{k}{N}$

(a-iii) high $\frac{k}{N}$

Figure 4.9: Case $p = 10$: Plotting $\frac{\widehat{\Delta}_k^{(N)}}{\Delta_k^{(N)}}$ according to relative cell index range. Three regions for $\frac{k}{N}$ that span $[0, 1]$ are shown.



(a-i) $p = 1.5$

(a-ii) $p = 1.5$

(b-i) $p = 2$

(b-ii) $p = 2$

(c-i) $p = 10$

(c-ii) $p = 10$

Figure 4.10: Close-up view: The ratio $\frac{\widehat{\Delta}_k^{(N)}}{\Delta_k^{(N)}}$ re-plotted against the relative cell index $\frac{k}{N}$ for various N-level GE-source optimal scalar quantizers in linear scale, but at a close-up view of the small relative cell index region (outer support region): (a) $p = 1.5$, (b) $p = 2$, (c) $p = 10$. Note: (*-i) plots and (*-ii) plots are at different scales.

Table 4.6: Table of $(\frac{k_{T,N,p}}{N})_{LB}$ values for $T = 5\%, 10\%, 20\%, 50\%$, for GE-sources with p=1.5, 2, 10.

$T\%$	$p = 1.5$	$p = 2$	$p = 10$
	$N = 32, \ldots, 512$	$N = 32, \ldots, 2048$	$N = 32, \ldots, 1024$
50%	0.5313 for $N \geq 32$	0.2500 for $N \geq 32$	0.0078 for $N \geq 128$
30%	0.3125 for $N \geq 32$	0.0938 for $N \geq 32$	0.0020 for $N \geq 512$
20%	0.1563 for $N \geq 32$	0.0313 for $N \geq 32$	0 for $N \leq 1024$
10%	0.0156 for $N \geq 32$	0 for $N \leq 2048$	0 for $N \leq 1024$
5%	0 for $N \leq 512$	0 for $N \leq 2048$	0 for $N \leq 1024$

Table 4.6 lists some value of $(\frac{k_{T,N,p}}{N})_{LB}$ for various N and $p = 1.5, 2, 10$.

4.11.2 Support Threshold Estimation.

Since the Nitadori sequence $\underline{\eta}_N$ provides an exact description of any N-level MMSE quantizer designed for an exponential source, the question naturally arises of how the Nitadori sequence $\underline{\eta}_N$ can be used to aid in the design of general exponential MMSE quantizers. Since for GE-sources with $p > 1$ and $N >> 0$, for each fixed k, Theorem IV.1 yields the relationship $\underline{\Delta}_k^{(N)} \approx \frac{\eta_k}{(t_k^{(N)})^{p-1}}$ and this relationship does not lead to a direct design method for MMSE quantizers, since the approximation for $\underline{\Delta}_k^{(N)}$ also requires knowledge of each $t_k^{(N)}$, we focus on estimating design parameters that assist the MMSE design process.

Key Parameter Estimation for GE-MMSE quantizers. As stated in the brief review in Chapter II, the *key parameter* or support threshold $t_1^{(N)}$ of an N-level optimal scalar quantization is historically important because it is used as an initializing value for the Lloyd Max algorithm ([13], [15]). In the case of exponential MMSE quantizers, computing the exact support threshold $t_1^{(N)}$ using the Nitadori sequence is a simple matter of summing up the first N terms of the Nitadori sequence since $\underline{\eta}_k = \underline{\Delta}_k^{(N)}$, irrespective of N: Starting with the general expression for the support threshold,

$$
\begin{aligned}
t_1^{(N)} &= \sum_{i=2}^{N} \Delta_i^{(N)} = \sum_{i=2}^{N} \underline{\Delta}_i^{(N)} + \overline{\Delta}_i^{(N)} = \sum_{i=2}^{N} \underline{\Delta}_i^{(N)} + \underline{\Delta}_{i-1}^{(N)} \\
&= \underline{\Delta}_1^{(N)} + \underline{\Delta}_N^{(N)} + 2\sum_{i=2}^{N-1} \underline{\Delta}_i^{(N)} \qquad (4.11.48) \\
&\overset{\text{exp. case}}{=} \underline{\eta}_1 + \underline{\eta}_N + 2\sum_{i=2}^{N-1} \underline{\eta}_i = t_1^{(N)}\text{exp. src.}\cdot \qquad (4.11.49)
\end{aligned}
$$

Now, consider the expression for the support threshold when we have an N-level MMSE quantizer designed for a GE-source with $p > 1$ and $N >> 0$. In this case, Theorem IV.1 and Lemma IV.12, Part 2 tell us that for fixed $k \geq 1$, where k does not depend on N,

$$
\underline{\Delta}_k^{(N)} \overset{\textit{Thm. IV.1}}{\approx} \frac{\underline{\eta}_k}{(t_k^{(N)})^{p-1}} \triangleq \widehat{\underline{\Delta}}_k^{(N)} \overset{\textit{Lemma IV.12, P. 2}}{\approx} \frac{\underline{\eta}_k}{(t_1^{(N)})^{p-1}}. \qquad (4.11.50)
$$

Using the expression for $t_1^{(N)}$ in (4.11.48) and keeping in mind the restrictions required in using Theorem IV.1 and Lemma IV.12, Part 2, we have

$$
\begin{aligned}
t_1^{(N)} &= \underline{\Delta}_1^{(N)} + \underline{\Delta}_N^{(N)} + 2\sum_{i=2}^{N-1} \underline{\Delta}_i^{(N)} \\
&\overset{?}{\approx} \frac{\underline{\eta}_1}{(t_1^{(N)})^{p-1}} + \frac{\underline{\eta}_N}{(t_N^{(N)})^{p-1}} + 2\sum_{i=2}^{N-1} \frac{\underline{\eta}_i}{(t_i^{(N)})^{p-1}} \triangleq t_{1,\,(4.11.51)}^{(N)}, \quad ((4.11.50),\ \text{Thm. IV.1}) \\
&\qquad (4.11.51)
\end{aligned}
$$

where we point out that in order to get $t_{1,\,(4.11.51)}^{(N)}$, we have applied Theorem IV.1 in a non-rigorous way to *each* half step $\underline{\Delta}_k^{(N)}$ in the expression for $t_1^{(N)}$, and we have used the symbol $\overset{?}{\approx}$ to highlight this fact. We also remark that in order to get the expression on the right-hand side of (4.11.51), we are ignoring the fact that $t_N^{(N)} = 0$ in the second term. (We make a further note on this remark by stating that we could drop this term from the expression in (4.11.51) since this term is supposed to represent $\underline{\Delta}_N^{(N)}$, and $\underline{\Delta}_N^{(N)}$, being the smallest half step in the support, is asymptotically insignificant to compared to the value of $t_1^{(N)}$.)

Continuing with this speculative line of thought, we have

$$\begin{aligned} t_{1,\ (4.11.51)}^{(N)} &= \frac{\eta_1}{(t_1^{(N)})^{p-1}} \left(\frac{t_1^{(N)}}{t_1^{(N)}}\right)^{p-1} + \frac{\eta_N}{(t_1^{(N)})^{p-1}} \left(\frac{t_1^{(N)}}{t_N^{(N)}}\right)^{p-1} + 2\sum_{i=2}^{N-1} \frac{\eta_i}{(t_i^{(N)})^{p-1}} \left(\frac{t_1^{(N)}}{t_i^{(N)}}\right)^{p-1} \\ &\overset{?}{\approx} \frac{\eta_1}{(t_1^{(N)})^{p-1}} + \frac{\eta_N}{(t_1^{(N)})^{p-1}} + 2\sum_{i=2}^{N-1} \frac{\eta_i}{(t_1^{(N)})^{p-1}} = \frac{t_1^{(N)} \text{exp. src.}}{(t_1^{(N)})^{p-1}}, \end{aligned}$$

where at $\overset{?}{\approx}$, similar to what was done in to create $t_{1,\ (4.11.51)}^{(N)}$, we have non-rigorously applied Lemma IV.12, Part 2 (see also (4.11.50)) because, strictly speaking, we cannot say that

$$\lim_{N\to\infty} \frac{t_1^{(N)}}{t_k^{(N)}}$$

exists when the cell index k is a function of N.

Summarizing, we have

$$t_1^{(N)} \overset{?\ \text{Thm. IV.1}\ ?}{\approx} t_{1,\ (4.11.51)}^{(N)} \overset{?\ \text{L IV.12, P.2}\ ?}{\approx} \frac{t_1^{(N)} \text{exp. src.}}{(t_1^{(N)})^{p-1}},$$

or equivalently,

$$(t_1^{(N)})^p \overset{?\ \text{Thm. IV.1}\ ?}{\approx} (t_1^{(N)})^{p-1} \cdot t_{1,\ (4.11.51)}^{(N)} \overset{?\ \text{L IV.12, P.2}\ ?}{\approx} t_1^{(N)} \text{exp. src.}$$

or equivalently,

$$t_1^{(N)} \overset{?\ \text{Thm. IV.1}\ ?}{\approx} \left((t_1^{(N)})^{p-1} \cdot t_{1,\ (4.11.51)}^{(N)}\right)^{\frac{1}{p}} \overset{?\ \text{L IV.12, P.2}\ ?}{\approx} (t_1^{(N)} \text{exp. src.})^{\frac{1}{p}}.$$

Define

$$t_{1,\,\text{est}}^{(N)} \triangleq (t_1^{(N)} \text{exp. src.})^{\frac{1}{p}}. \tag{4.11.52}$$

While it is obvious from the discussion that the relationship between $t_1^{(N)}$ and $t_{1,\,\text{est}}^{(N)}$ is ambiguous since the construction of $t_{1,\,\text{est}}^{(N)}$ was based on using asymptotic relationships that are only valid when quantization cells reside in the tail region of the pdf support, nevertheless, we will use $t_{1,\,\text{est}}^{(N)}$ to estimate $t_1^{(N)}$ in spite of these issues, noting that it will be interesting to see how accurate $t_{1,\,\text{est}}^{(N)}$ is. (As will be seen later, $t_{1,\,\text{est}}^{(N)}$ turns out to be surprisingly good.)

Observation on support and support length estimation for optimal τ, j quantizers. Having constructed a support threshold estimator for MMSE quantizers when $N >> 0$ using Theorem IV.1, we turn our attention to the support length $L_{\tau,j,p} \overset{\triangle}{=} t_{1|j} - t_{j|j}$ of optimal τ, j quantizers and estimation based on the statements in Theorem IV.7.

Consider an optimal τ, j quantizer with $j \geq 2$ fixed. First, we will remark that it is clear that $L_{\tau,j,p} \to 0$ as $\tau \to \infty$, since $L_{\tau,j,p}$ equals a fixed sum of half steps $\underline{\Delta}_{k|j}$ and each of these half steps $\underline{\Delta}_{k|j}$ are decreasing to zero as τ increases. To discover what more Theorem IV.7 can tell us about $L_{\tau,j,p}$, we will start by using an approach similar to one used in the previous discussion, but this time we can be more rigorous. Using Theorem IV.7 and Lemma IV.12, Part 2 (applied to optimal τ, j quantizers), we have

$$L_{\tau,j,p} = t_{1|j} - t_{j|j} = \underline{\Delta}_{1|j} + \underline{\Delta}_{j|j} + 2\sum_{i=2}^{j-1} \underline{\Delta}_{i|j} \tag{4.11.53}$$

$$\overset{\tau>>0}{\approx} \frac{\underline{\eta}_1}{(t_{1|j})^{p-1}} + \frac{\underline{\eta}_j}{(t_{1|j})^{p-1}} + 2\sum_{i=2}^{j-1} \frac{\underline{\eta}_i}{(t_{i|j})^{p-1}}, \tag{4.11.54}$$

where (4.11.53) is a general expression for the support length of a τ, j quantizer and is analogous to (4.11.48) which equals the support threshold for a j-level MMSE quantizer. Then multiplying both sides by $(t_{1|j})^{p-1}$, we have

$$L_{\tau,j,p} \cdot (t_{1|j})^{p-1} \approx \underline{\eta}_1 + \underline{\eta}_j + 2\sum_{i=2}^{j-1} \underline{\eta}_i \overset{(4.11.49)}{=} t_1^{(j)} \text{exp. src.}$$

or when $\tau >> 0$,

$$L_{\tau,j,p} \approx \frac{t_1^{(j)} \text{exp. src.}}{(t_{1|j})^{p-1}},$$

or more rigorously,

$$\lim_{\tau \to \infty} L_{\tau,j,p} \cdot (t_{1|j})^{p-1} = t_1^{(j)} \text{exp. src.} \cdot \tag{4.11.55}$$

Note that since for optimal τ, j quantizers, the smallest threshold $t_{j|j}$ grows away from the origin as τ increases, we do not have validity issues as is the case with $t_1^{(N)}$ estimation of MMSE quantizers. Thus (4.11.55) is a valid asymptotic result regarding the support length of optimal τ, j quantizers.

GE-MMSE quantization: How does the estimate $t^{(N)}_{1,\,\mathbf{est}}$ compare to the actual values for $t^{(N)}_1$? Before looking at the data shown in Figure 4.14, we will ruminate over what we might expect to see and then check our thoughts against the actual data. Our approach in this discussion will be to run-through the iterations we took to create $t^{(N)}_{1,\,\mathrm{est}}$ (where each step culminates in the creation of a new intermediate estimator), pausing after each step taken, to see if the intermediate support threshold estimator just created is either too big or too small relative to $t^{(N)}_1$.

Fix $N >> 0$. Beginning with the general expression for $t^{(N)}_1$ from (4.11.48)

$$t^{(N)}_1 = \underline{\Delta}^{(N)}_1 + \underline{\Delta}^{(N)}_N + 2\sum_{i=2}^{N-1} \underline{\Delta}^{(N)}_i,$$

the first change we make is to drop $\underline{\Delta}^{(N)}_N$ since its contribution to $t^{(N)}_1$ is negligible when $N >> 0$, so that we have

$$t^{(N)}_{1,as} \triangleq \underline{\Delta}^{(N)}_1 + 2\sum_{i=2}^{N-1} \underline{\Delta}^{(N)}_i, \tag{4.11.56}$$

where it is clear that $\lim_{N\to\infty} |t^{(N)}_1 - t^{(N)}_{1,as}| = 0$.

To create the first support threshold estimator $\hat{t}$, we substitute $\frac{\eta_1}{t^{p-1}}$ for $\underline{\Delta}^{(N)}_1$ in (4.11.56) to get an equation that $\hat{t}$ should satisfy:

$$t = \frac{\eta_1}{t^{p-1}} + 2\sum_{i=2}^{N-1} \underline{\Delta}^{(N)}_i$$

or

$$t - c_1 = \frac{\eta_1}{t^{p-1}} \tag{4.11.57}$$

where

$$c_1 = 2\sum_{i=2}^{N-1} \underline{\Delta}^{(N)}_i.$$

As seen in Figure 4.12, $\hat{t}$ is the unique real solution to (4.11.57). What is the relationship between $\hat{t}$ and $t^{(N)}_{1,as}$? Since

$$t^{(N)}_1 < \underline{\Delta}^{(N)}_1 + c_1$$

and since $\underline{\Delta}_1^{(N)} < \frac{\underline{\eta}_1}{(t_1^{(N)})^{p-1}}$, then

$$t_1^{(N)} < \frac{\underline{\eta}_1}{(t_1^{(N)})^{p-1}} + c_1$$

thus the unique real solution $\hat{t}$ to (4.11.57) satisfies $\hat{t} > t_{1,as}^{(N)}$.

For a second estimator $\hat{\hat{t}}$, we make another substitution by replacing c_1 for the constant

$$c_2 = 2 \sum_{i=2}^{N-1} \frac{\underline{\eta}_i}{(t_i^{(N)})^{p-1}},$$

which is created by swapping each $\underline{\Delta}_i^{(N)}$ in the sum by $\frac{\underline{\eta}_i}{(t_i^{(N)})^{p-1}}$, where we have assumed that $t_2^{(N)}, t_3^{(N)}, \ldots, t_{N-1}^{(N)}$ are known values, and we define $\hat{\hat{t}}$ to be the unique positive, real value that satisfies

$$t - c_2 = \frac{\underline{\eta}_1}{t^{p-1}}. \tag{4.11.58}$$

Using our empirical observation that $\underline{\Delta}_k^{(N)} < \frac{\underline{\eta}_k}{(t_k^{(N)})^{p-1}}$, it is clear that $c_2 > c_1$, and thus we know that $\hat{\hat{t}} > \hat{t} > t_{1,as}^{(N)}$. (This last relationship is shown pictorially in Figure 4.12.)

To create our final support threshold estimator $t_{1,\,\text{est}}^{(N)}$, we replace c_2 in (4.11.58) for the expression

$$2 \sum_{i=2}^{N-1} \frac{\underline{\eta}_i}{t^{p-1}}$$

which is not a constant but a function in t. This leads to $t_{1,\,\text{est}}^{(N)}$ as the unique real-valued solution to

$$\begin{aligned} t &= \frac{\underline{\eta}_1}{t^{p-1}} + 2 \sum_{i=2}^{N-1} \frac{\underline{\eta}_i}{t^{p-1}} \\ &= \frac{1}{t^{p-1}} \cdot \left(\underline{\eta}_1 + 2 \sum_{i=2}^{N-1} \underline{\eta}_i \right) \\ &= \frac{1}{t^{p-1}} \cdot t_1^{(N)}\text{exp. src.} \cdot \end{aligned} \tag{4.11.59}$$

What is the relationship between $t_{1,\,\text{est}}^{(N)}$ and $\hat{\hat{t}}$? Between $t_{1,\,\text{est}}^{(N)}$ and $\hat{t}$? Between $t_{1,\,\text{est}}^{(N)}$

and $t_{1,as}^{(N)}$?

To answer the first question, consider again the expression for which $\hat{\hat{t}}$ satisfies

$$\hat{\hat{t}} = \frac{\eta_1}{\hat{\hat{t}}^{p-1}} + c_2.$$

If we replace each $t_k^{(N)}$, $k = 2, 3, \ldots, N-1$ for $\hat{\hat{t}} > t_2^{(N)}$ (since $\hat{\hat{t}} > t_{1,as}^{(N)}$) in this equation, then

$$\hat{\hat{t}} > \frac{\eta_1}{\hat{\hat{t}}^{p-1}} + 2\sum_{i=2}^{N-1} \frac{\eta_i}{\hat{\hat{t}}^{p-1}}. \tag{4.11.60}$$

In order to achieve equality in (4.11.60), $t_{1,\text{ est}}^{(N)}$ must be smaller than $\hat{\hat{t}}$.

To see what relationship $t_{1,\text{ est}}^{(N)}$ has with respect to $\hat{t}$, we re-examine the relationship that $\hat{t}$ satisfies

$$\hat{t} = \frac{\eta_1}{\hat{t}^{p-1}} + c_1.$$

Since empirically we know that $\underline{\Delta}_k^{(N)} < \frac{\eta_k}{(t_k^{(N)})^{p-1}}$, then

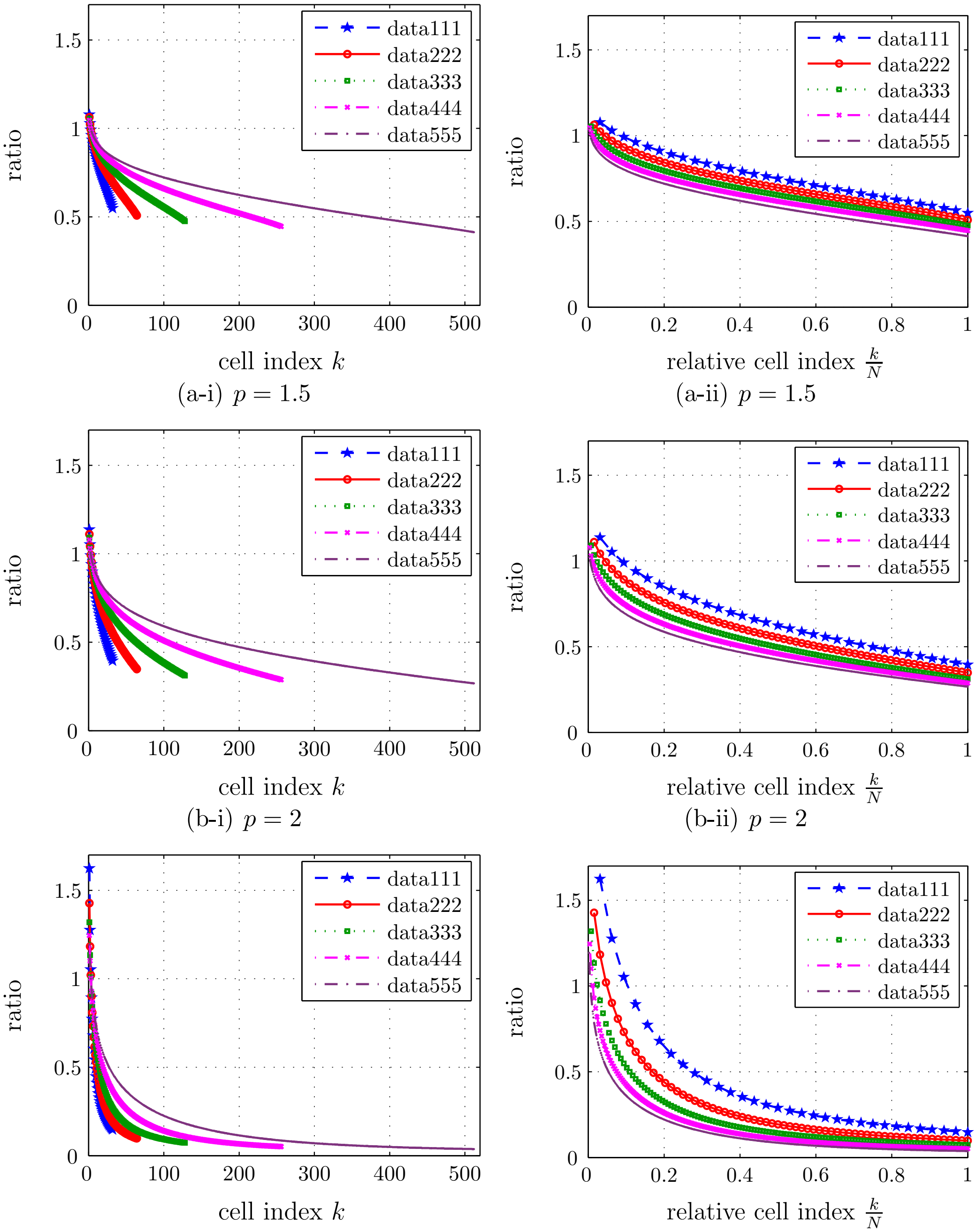
$$\hat{t} > \frac{\eta_1}{\hat{t}^{p-1}} + 2\sum_{i=2}^{N-1} \frac{\eta_i}{\hat{t}^{p-1}}$$

since $\hat{t} > t_{1,as}^{(N)} \geq t_2^{(N)}$. To achieve equality in this expression, $\hat{t}$ must be decreased so that we have $t_{1,\text{ est}}^{(N)} < \hat{t}$. Since $\hat{t} > t_{1,as}^{(N)}$, the relationship between $t_{1,\text{ est}}^{(N)}$ and $t_{1,as}^{(N)}$ (and hence $t_1^{(N)}$) is still unclear even though we know that $\hat{\hat{t}} > \hat{t} > t_{1,\text{ est}}^{(N)}$ and $\hat{\hat{t}} > \hat{t} > t_{1,as}^{(N)}$. Figure 4.13 illustrates the relative locations of $\hat{t}, \hat{\hat{t}}$, and $t_{1,\text{ est}}^{(N)}$.

Now, let us see what the data shows. We have computed data for GE-sources with $p = 1.5, 2, 10$. As seen in plots on the left-side of Figure 4.14, the actual values for $t_1^{(N)}$ and the corresponding estimate $t_{1,\text{ est}}^{(N)}$ for N-level MMSE quantizers, the estimator $t_{1,\text{ est}}^{(N)}$ appears to be tracking the behavior of $t_1^{(N)}$ as a function of N quite well, especially in light of the nature of $t_{1,\text{ est}}^{(N)}$'s construction. On closer inspection, $t_{1,\text{ est}}^{(N)}$ appears to be diverging from each other as N increases, with $t_{1,\text{ est}}^{(N)}$ underestimating $t_1^{(N)}$. This observation seems to support the notion that $t_1^{(N)}$ is much too large when using it in $\frac{\eta_k}{(t_1^{(N)})^{p-1}}$ to approximate $\underline{\Delta}_k^{(N)}$. The rate of divergence, however, appears to be decreasing as N increases. This decrease in rate is more easily seen in the plots on the right-hand side of Figure 4.14, where the ratio $\frac{t_1^{(N)}}{t_{1,\,\mathrm{est}}^{(N)}}$ seems to be flattening out as N increases.

Also from Figure 4.14, it appears that, as p increases, the rate at which the curves on the right-hand side are flattening increases as N increases. In other words, it appears that, if $\frac{t_1^{(N)}}{t_{1,\,\mathrm{est}}^{(N)}}$ is converging to a limit in N, then the larger p is, the faster $\frac{t_1^{(N)}}{t_{1,\,\mathrm{est}}^{(N)}}$ converges in N. For a reason that would support this observation, see the shape of the pdfs for the GE-sources with $p = 1.5, 2, 10$ in Figure 4.5. As p increases, the pdf shape becomes more and more like that of the uniform distribution on $[0, 1]$.[8] Thus, as the number of levels in an N-level MMSE quantizer increases, we would expect that $t_1^{(N)}$ would grow much more slowly for an optimal quantizer designed for a GE-source with high p over that of an optimal quantizer designed for a GE-source with low p.

[8] It can be easily shown that as $p \to \infty$, the GE-pdf converges to the uniform pdf on $[0, 1]$.



(a-i) $p = 1.5$

(a-ii) $p = 1.5$

(b-i) $p = 2$

(b-ii) $p = 2$

(c-i) $p = 10$

(c-ii) $p = 10$

Figure 4.11: Weaker half step estimator: (a) The ratio $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ plotted against cell index k, and (b) the ratio $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ plotted against relative cell index $\frac{k}{N}$, for various N-level GE-source optimal scalar quantizers.

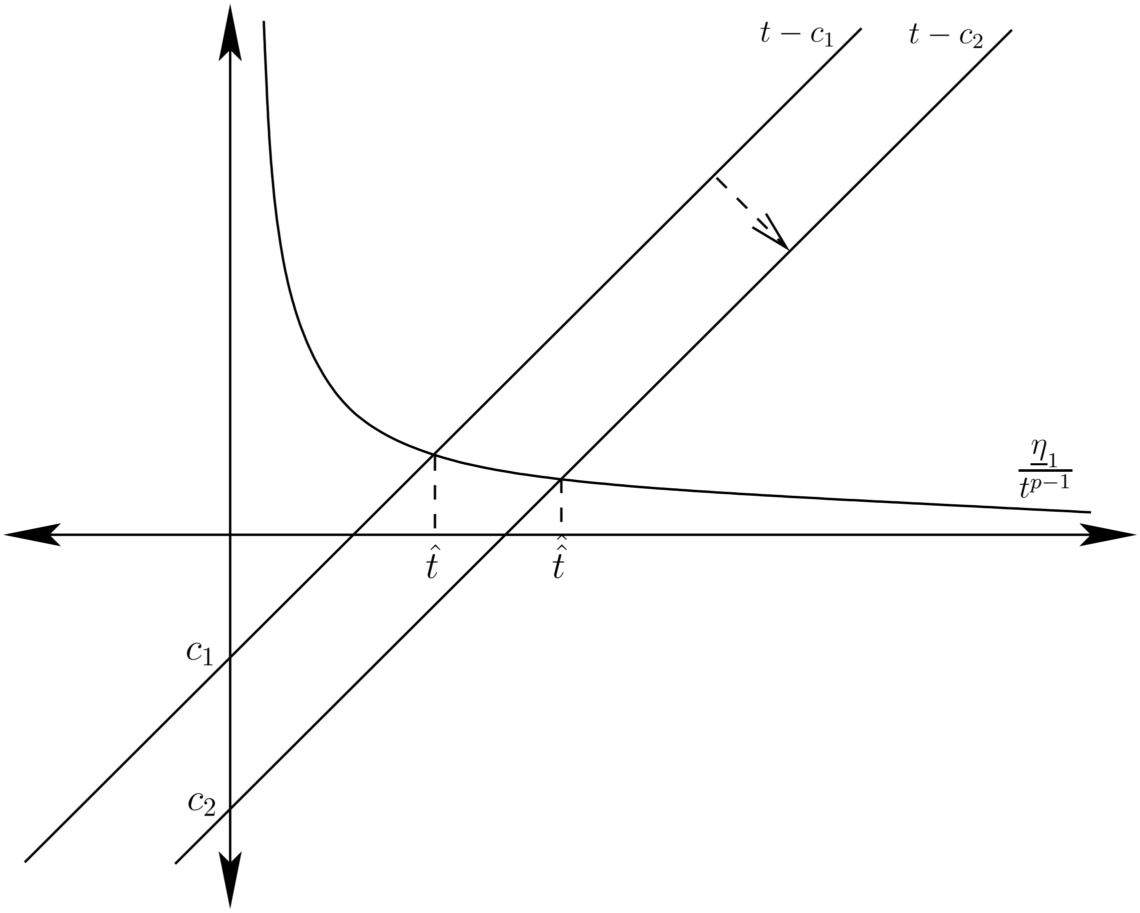


Figure 4.12: Illustration of $\hat{t}$ and $\hat{\hat{t}}$ as the intersection point of two sets of functions: $t - c_1$ and $\frac{\eta_1}{t^{p-1}}$, and $t - c_2$ and $\frac{\eta_1}{t^{p-1}}$.

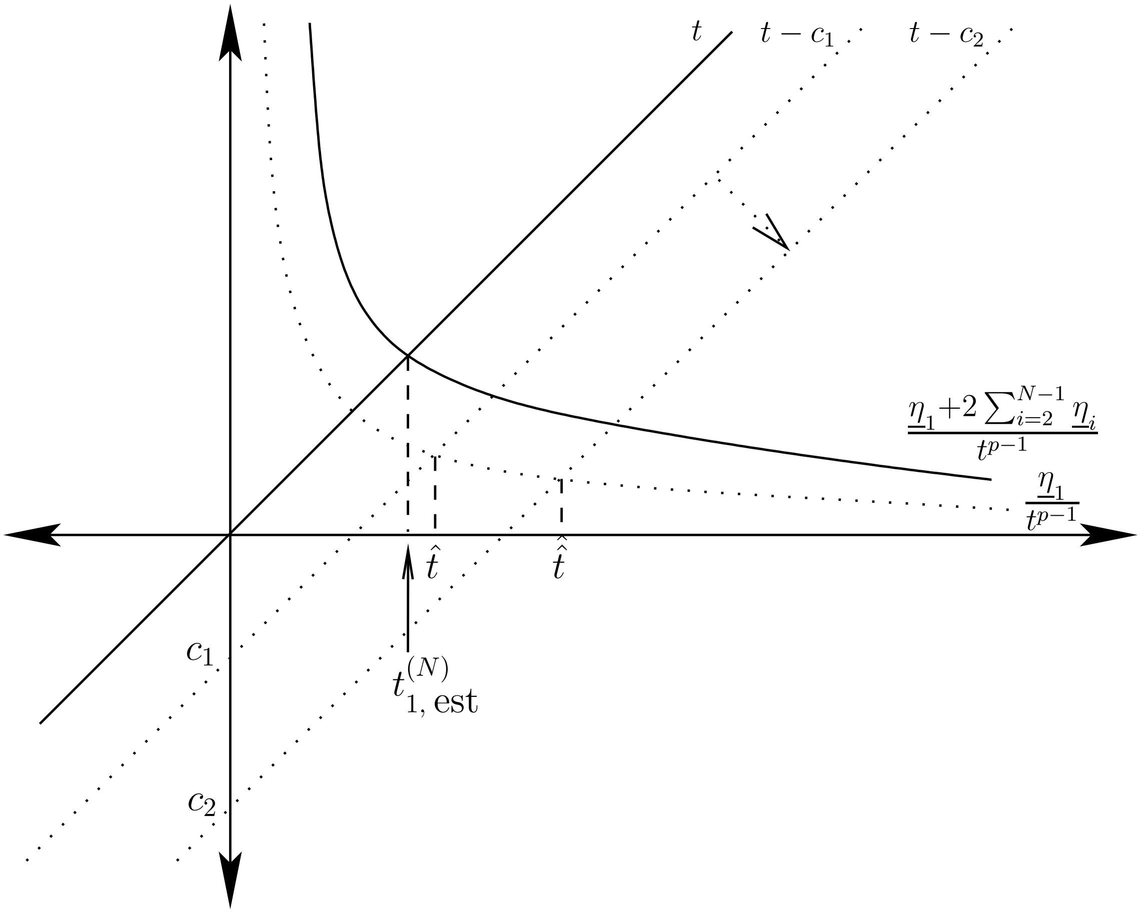

t
$t - c_1$
$t - c_2$
$\frac{\underline{\eta}_1 + 2\sum_{i=2}^{N-1} \underline{\eta}_i}{t^{p-1}}$
$\frac{\underline{\eta}_1}{t^{p-1}}$
$\hat{t}$
$\hat{\hat{t}}$
c_1
c_2
$t_{1,\,\mathrm{est}}^{(N)}$

Figure 4.13: Illustration of the relative locations of $t_{1,\,\mathrm{est}}^{(N)}$, $\hat{t}$ and $\hat{\hat{t}}$ when $N >> 0$.

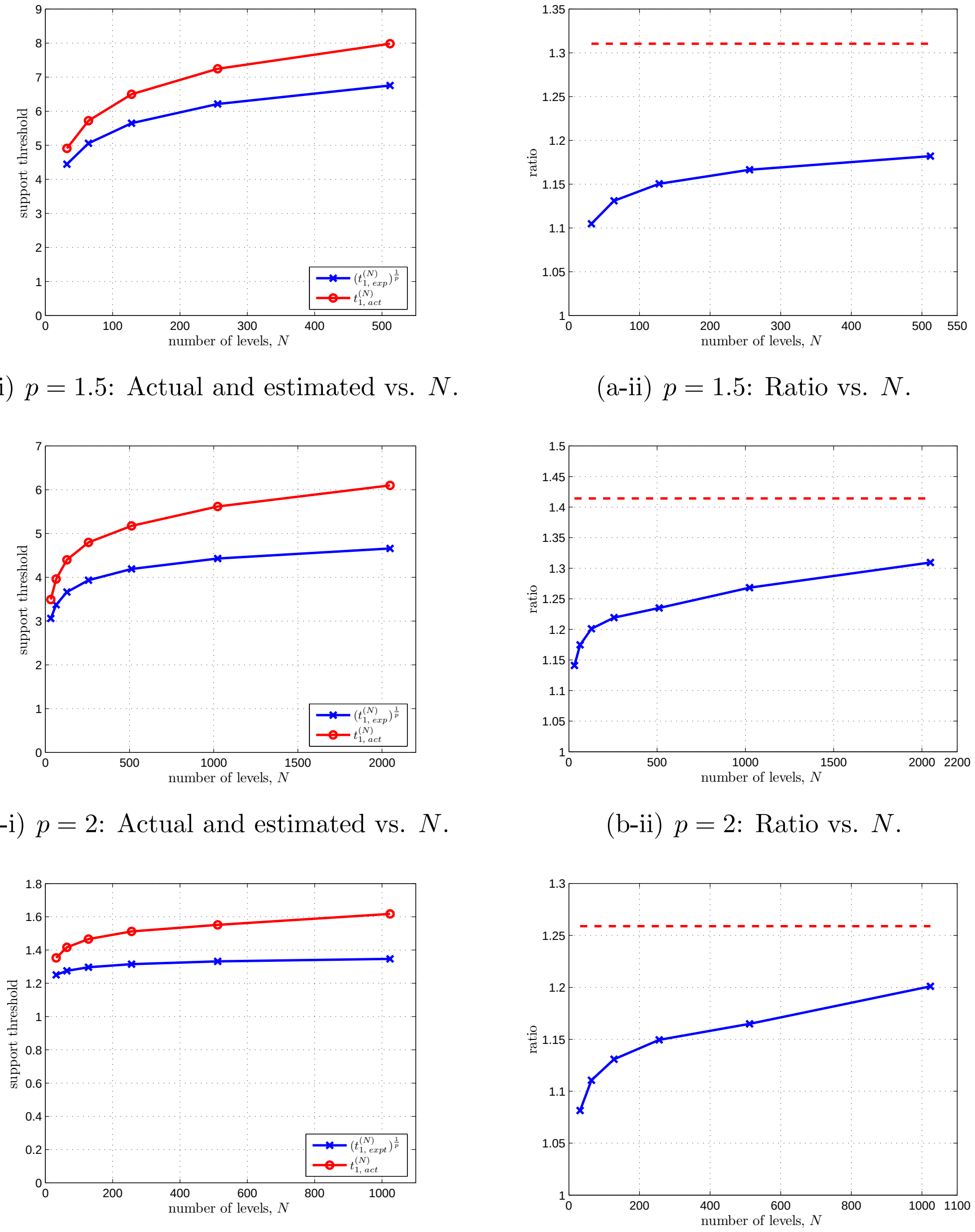
For clues as to why $t_{1,\,\text{est}}^{(N)}$ is performing so well as an estimator to $t_1^{(N)}$, see Figure 4.15, where we have plotted data for the ratio $\frac{\frac{\eta_k}{(t_1^{(N)})^{p-1}}}{\underline{\Delta}_k^{(N)}}$ as a function of both the cell index k and the relative cell index $\frac{k}{N}$ for various values of N and p. From what is shown, $\frac{\eta_k}{(t_1^{(N)})^{p-1}} < \underline{\Delta}_k^{(N)}$, for $k = 1, 2, \ldots, N$, and thus from (4.1.1), we conclude

$$t_1^{(N)} < t_{1,\,\text{est}}^{(N)}.$$

We also note that in contrast to the behavior of $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ as a function of relative cell index $\frac{k}{N}$ does not exhibit the same phenomenon of lying on top of each other across all values of N. Rather, the curves in Figure 4.15 appear to be converging towards the function

$$l\,(x) = \begin{cases} 1 & \frac{k}{N} = 0 \\ 0 & \text{else} \end{cases} \tag{4.11.61}$$

which might seem contradictory to the fact that $t_{1,\,\text{est}}^{(N)}$ grows with N, but in actuality, it is not because as N increases, while the proportion of $\frac{k}{N}$ having $\frac{\eta_k}{(t_1^{(N)})^{p-1}}$ approaching zero increases, as N increases, the value of $\frac{\eta_k}{(t_1^{(N)})^{p-1}}$ is always greater than zero, and thus for each value of N, when summing over all values of k, $t_{1,\,\text{est}}^{(N)}$ still grows with N.



(a-i) $p = 1.5$: Actual and estimated vs. N.

(a-ii) $p = 1.5$: Ratio vs. N.

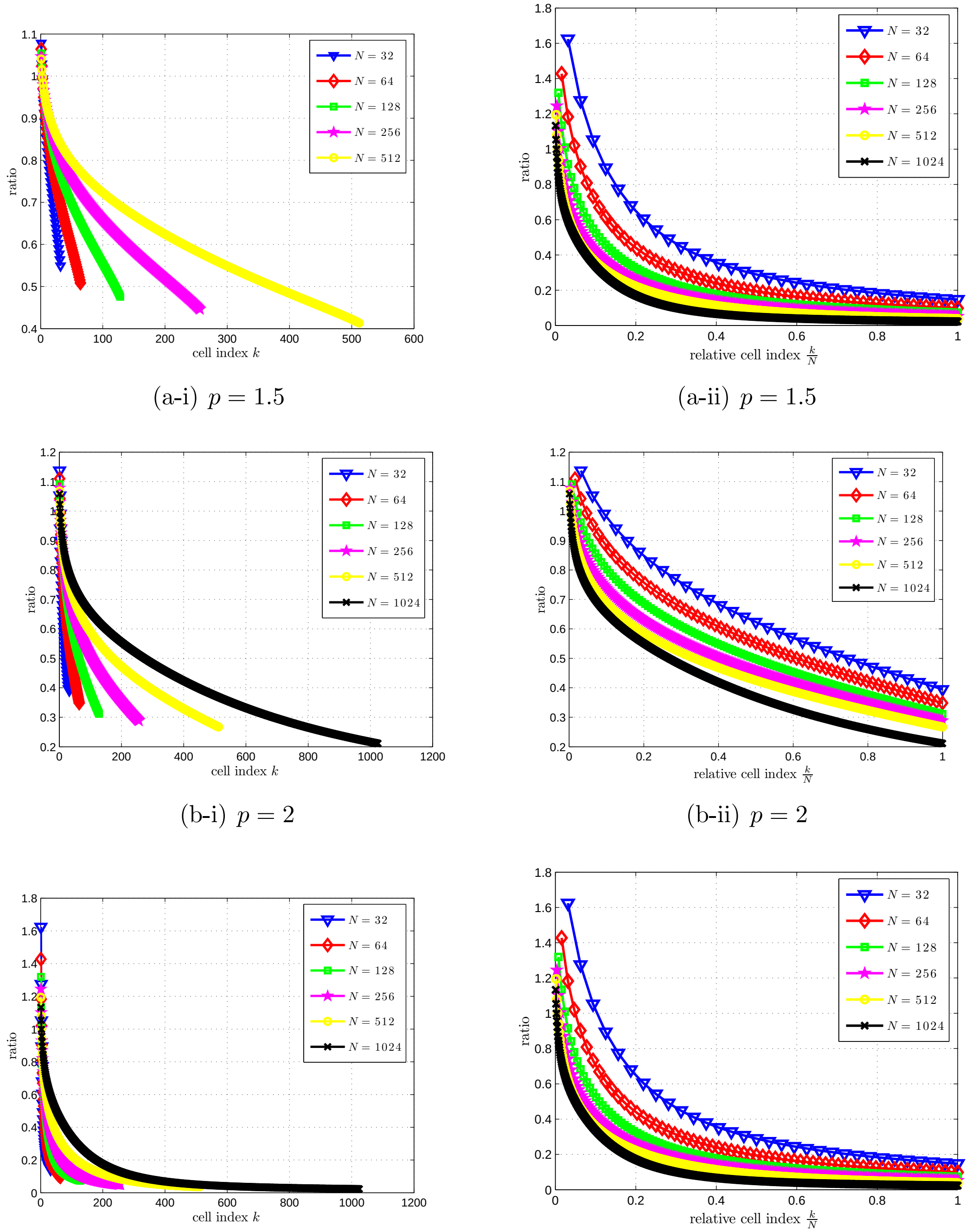
(b-i) $p = 2$: Actual and estimated vs. N.

(b-ii) $p = 2$: Ratio vs. N.

(c-i) $p = 10$: Actual and estimated vs. N.

(c-ii) $p = 10$: Ratio vs. N.

Figure 4.14: Comparing support threshold estimates created using the Nitadori sequence against the actual $t_1^{(N)}$ vs. N for various N-level MMSE quantizers designed for GE-sources with $p = 1.5, 2, 10$. The plots on the left-hand side show actual $t_1^{(N)}$ and estimated support threshold $t^{(N)}_{1,\,\text{est}}$ vs. N. The plots on the right-hand side show the ratio $\frac{t_1^{(N)}}{t^{(N)}_{1,\,\text{est}}}$ vs. N. Note: The dashed lines in the plots on the right-hand side represent the asymptotic limit $p^{\frac{1}{p}}$ in (4.11.63).



(a-i) $p = 1.5$

(a-ii) $p = 1.5$

(b-i) $p = 2$

(b-ii) $p = 2$

(c-i) $p = 10$

(c-ii) $p = 10$

Figure 4.15: The ratio $\frac{\frac{\eta_k}{(t_1^{(N)})^{p-1}}}{\Delta_k^{(N)}}$ plotted versus cell index k (*-i) and versus relative cell index $\frac{k}{N}$ (*-ii) for various optimal N-level GE-quantizers: (a) $p = 1.5$, (b) $p = 2$, (c) $p = 10$. Note: The larger the cell index k is, the closer the quantization cell is to the origin. Also, all plots are linear scale.

More discussion of the support threshold estimator $t_{1,\,\mathbf{est}}^{(N)}$ for MMSE quantization of GE-sources when $p > 1$. Ideally, in order to see how good the asymptotic usefulness of the estimator in (4.11.52) is, we would compare it against a closed-form analytic expression for $t_1^{(N)}$ for each value of p. Since closed-form expressions for $t_1^{(N)}$ are not available for GE-sources (if there were, it would obviate the need for support threshold estimation of GE-MMSE quantizers), we will compare our estimator to known closed-form functions that also estimate $t_1^{(N)}$. We choose the functions described in [18], which were derived through informal arguments, were compared against empirical support threshold data, and were concluded to be accurate and correct. Specifically from [18], the functions that we compare against, which we will refer to as the support threshold benchmark functions $(t_1^{(N)})_{(bm)}$, have the form

$$(t_1^{(N)})_{(bm)} = (3 \cdot p \ln N)^{\frac{1}{p}} \left(1 + o\left(\ln N\right)\right).$$

With $t_{1,\,\mathrm{est}}^{(N)} \triangleq (t_1^{(N)}{}_{\mathrm{exp.\ src.}})^{\frac{1}{p}}$, which is the estimator in (4.11.52), we have

$$\frac{(t_1^{(N)})_{(bm)}}{t_{1,\,\mathrm{est}}^{(N)}} = \frac{p^{\frac{1}{p}} \cdot (3 \ln N)^{\frac{1}{p}}}{(t_1^{(N)}{}_{\mathrm{exp.\ src.}})^{\frac{1}{p}}} \left(1 + o\left(\ln N\right)\right).$$

As was discussed in the previous chapter, the support threshold for an N-level, exponential MMSE quantizer is

$$t_1^{(N)}{}_{\mathrm{exp.\ src.}} = (3 \ln N) \left(1 + o\left(\ln N\right)\right).$$

Then

$$\begin{aligned} \frac{(t_1^{(N)})_{(bm)}}{t_{1,\,\mathrm{est}}^{(N)}} &= \frac{p^{\frac{1}{p}} \cdot (3 \ln N)^{\frac{1}{p}} \left(1 + o\left(\ln N\right)\right)}{(3 \ln N)^{\frac{1}{p}} \left(1 + o\left(\ln N\right)\right)^{\frac{1}{p}}} \\ &= p^{\frac{1}{p}} \frac{\left(1 + o\left(\ln N\right)\right)}{\left(1 + o\left(\ln N\right)\right)^{\frac{1}{p}}} \end{aligned} \tag{4.11.62}$$

or equivalently,

$$\lim_{N \to \infty} \frac{(t_1^{(N)})_{(bm)}}{t_{1,\,\mathrm{est}}^{(N)}} = p^{\frac{1}{p}}. \tag{4.11.63}$$

From (4.11.63), we now have evidence to support the conjecture that $\frac{t_1^{(N)}}{t_{1,\,\mathrm{est}}^{(N)}}$ is converging. Again, returning to Figure 4.14 (plots on the right-hand side), we can now remark that for each value of p shown, convergence of $\frac{t_1^{(N)}}{t_{1,\,\mathrm{est}}^{(N)}}$ towards $p^{\frac{1}{p}}$, which is the dashed line shown in the figure, is seen.

Using (4.11.63) and taking the limit as $p \to \infty$, we have

$$\lim_{p\to\infty} \lim_{N\to\infty} \frac{(t_1^{(N)})_{(bm)}}{t_{1,\,\mathrm{est}}^{(N)}} = 1 \tag{4.11.64}$$

since $\lim_{p\to\infty} p^{\frac{1}{p}} = 1$. The implication of (4.11.64) is that for large values of N, the estimator function $t_{1,\,\mathrm{est}}^{(N)}$ becomes a better approximation for the benchmark function $(t_1^{(N)})_{(bm)}$ as p is increased, which in turn, implies that $t_{1,\,\mathrm{est}}^{(N)}$ becomes a better approximation for $t_1^{(N)}$. Evidence of this trend as a function of p can also be seen in Figure 4.14 (right-hand side).

Thus, from what we've shown and as a final remark, we can now propose an improved support threshold estimator

$$\widetilde{t_1^{(N)}} \triangleq p^{\frac{1}{p}} \cdot \left(t_1^{(N)}{}_{\text{exp. src.}}\right)^{\frac{1}{p}},$$

which is just $t_{1,\,\mathrm{est}}^{(N)}$ multiplied by the factor $p^{\frac{1}{p}}$, that will be asymptotically correct in N assuming $(t_1^{(N)})_{(bm)}$ from [18] is correct, i.e.,

$$\lim_{N\to\infty} \frac{t_1^{(N)}}{\widetilde{t_1^{(N)}}} = 1,$$

if $(t_1^{(N)})_{(bm)}$ from [18] is correct. Moreover, we note that as p increases, $\widetilde{t_1^{(N)}}$ becomes closer and closer to our original support threshold estimator $t_{1,\,\mathrm{est}}^{(N)}$, i.e.,

$$\lim_{p\to\infty} \frac{\widetilde{t_1^{(N)}}}{t_{1,\,\mathrm{est}}^{(N)}} = 1,$$

and so for large values of p, our original support threshold estimator $t_{1,\,\mathrm{est}}^{(N)}$ is expected to perform nearly as well as our improved estimator $\widetilde{t_1^{(N)}}$ when N is large.

4.12 Future Work.

Several topics for future work arose during our study on the role of the Nitadori sequence and MMSE quantization of GE-sources. We briefly list them below:

- **Extending Theorem IV.1 to GE-sources with** $0 < p < 1$**.** It may be possible to extend the results of Theorem IV.1 to include GE-sources with $p \in (0,1)$. For such sources, the pdf tail is quite heavy and thus, we expect, for example, that both $t_1^{(N)}$ and $\underline{\Delta}_1^{(N)}$ will grow as N increases. Such behavior, especially in $\underline{\Delta}_1^{(N)}$ is, in stark contrast to the case when $p > 1$ (pdf's with lighter tails), where $\underline{\Delta}_k^{(N)}$ decreases (while $t_1^{(N)}$ increases) as N grows. Furthermore, if Theorem IV.1 can be shown to be true for any GE-source with $p > 0$, it would interesting to see how quickly, $\underline{\alpha}_{k,p}^{(N)} = \underline{\Delta}_k^{(N)} \left(t_k^{(N)}\right)^{p-1}$ converges to $\underline{\eta}_k$, when $p < 1$ as opposed to when $p \geq 1$.

- **Characterizing the behavior of** $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ **as a function of** $\frac{k}{N}$**.** In the discussion dealing with half step approximation and support threshold estimation, we witnessed an interesting phenomenon in the data shown in Figure 4.7, where for each value of N shown, the data curves for $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ as a function of relative cell index $\frac{k}{N}$ appear, not only to have the same kind of shape, but also to be lying on top of each other. This observation could be suggesting that as a function of N, $\frac{\widehat{\underline{\Delta}}_k^{(N)}}{\underline{\Delta}_k^{(N)}}$ in terms of relative cell index $\frac{k}{N}$ may be converging to a limiting function. It would be interesting to see not only if such a limiting function exists, but also to know what it is.

  Tie-in with the asymptotic theory of optimal point densities. Since, due to our indexing scheme, $\frac{k}{N}$ indicates the number of quantization levels to the right of the kth quantization threshold, if such a limiting function does indeed exist, we can use the asymptotic theory of optimal point densities (see Chapter II, Section 2.6) to suggest that the accuracy of $\widehat{\underline{\Delta}}_k^{(N)}$ with respect to $\underline{\Delta}_k^{(N)}$ as N increases (with k increasing proportionally to N) is tied to a specific location x along the real axis.

- **Issues related to rigorous support threshold estimation.** Recall that $t_{1,\,\mathrm{est}}^{(N)}$ was created using the asymptotic facts in Theorem IV.1 and Lemma IV.12 to estimate *every* half step in an N-level optimal quantizer (not just the ones in the pdf's tail region). This approach produced a surprisingly good approximation to $t_1^{(N)}$, even though it lacked a completely sound theoretical basis. A search for a more formally strict way to apply these results to support threshold estimation is a topic for further study.

Another topic for investigation that also pertains to $t_{1,\,\mathrm{est}}^{(N)}$ concerns that fact that we could not, at this point, rigorously evaluate $t_{1,\,\mathrm{est}}^{(N)}$'s effectiveness in approximating $t_1^{(N)}$. Perhaps more progress can be made if a different approach were used to think about the relationship between $t_{1,\,\mathrm{est}}^{(N)}$ and $t_1^{(N)}$.

In the very last section on support threshold estimation, the factor $p^{\frac{1}{p}}$ turned out to be the proportionality constant that our initial support threshold estimator $t_{1,\,\mathrm{est}}^{(N)}$ lacked. With more work, it may be possible to theoretically substantiate the existence and need for this factor when using the Nitadori sequence to estimate the support threshold of an optimal GE-quantizers with large N.

- **Generalizing Nitadori's MSE result.** Finally, another topic, not yet addressed, has its roots in the fact that the Nitadori sequence also gives the exact MSE performance of optimal scalar quantizers designed for an exponential source (Nitadori's second result, Chapter II, Section 2.7). Since GE-sources have pdfs that are a natural extension of the exponential source's pdf and since we now know that the Nitadori sequence provides an asymptotic connection between the product $\underline{\Delta}_k^{(N)} \left(t_k^{(N)}\right)^{p-1}$ for GE-MMSE quantizers and the half steps of exponential MMSE quantizers, we feel that there may be an analogous extension of Nitadori's second result for optimal exponential quantizers (repeated here, from (2.7.17) in Chapter II)

$$D\,(N) = \left(\underline{\eta}_N\right)^2$$

to optimal GE-quantizers as well.

APPENDICES

APPENDIX A

The Upper Bound to $t_1^{(N)}$ of Optimal, M-level Laplacian Quantizers: M Odd Case

This appendix gives a brief comment on the upper bound to $t_1^{(N)}$ stated in (3.2.4) of Corollary III.3, Part 1a), when the number of levels M is odd. While the case when M is even is not explicitly discussed because it is a straightforward application of Theorem III.1, we will mention specific results from it that are used in the case when M is odd. Furthermore, we will only consider the case when $\sigma^2 = 1$ since extending this result to arbitrary $\sigma^2 > 0$ is a simple matter of multiplying everything by σ.

Notation. In the discussion below, we will be referring to the quantization parameters of both exponential (one-sided, unit variance) quantizers and Laplacian (two-sided, unit variance) quantizers. To avoid confusion, we will use the following conventions for this appendix only:

- For N-level exponential quantizers, the half step of the kth quantization cell is $\underline{\underline{\Delta}}_k^{(N)}$ and the support threshold is $t_1^{(N)}$, and for UTCC exponential quantizers, the half step is $\underline{\underline{\Delta}}_{k,c}^{(N)}$ and the support threshold is $t_{1,c}^{(N)}$. (This is the same notation used in Chapter III.)

- For M-level Laplacian quantizers, the half step of the kth quantization cell is $\overset{\star}{\underline{\underline{\Delta}}}{}_k^{(M)}$ and the support threshold is $\overset{\star}{t}{}_1^{(M)}$, and for UTCC Laplacian quantizers, the half step is $\overset{\star}{\underline{\underline{\Delta}}}{}_{k,c}^{(M)}$ and the support threshold is $\overset{\star}{t}{}_{1,c}^{(M)}$.

Since a Laplacian source with variance σ^2 is the same as a one-sided exponential source with variance $= \frac{1}{2}\sigma^2$ defined on both the negative reals and on the non-negative reals, the following relationships hold between the quantization parameters of Laplacian (unit variance) quantizers and exponential (unit variance) quantizers when $M = 2N$ (even): For an optimal Laplacian quantizer,

$$\overset{\star}{\underline{\Delta}}{}_k^{(M)} = \frac{1}{\sqrt{2}}\underline{\Delta}_k^{(N)} \tag{A.1}$$

$$\overset{\star}{t}{}_1^{(M)} = \frac{1}{\sqrt{2}}t_1^{(N)} \tag{A.2}$$

when $k = 1, 2, \ldots, N$, and for a UTCC Laplacian quantizer,

$$\overset{\star}{\underline{\Delta}}{}_{k,c}^{(M)} = \frac{1}{\sqrt{2}}\underline{\Delta}_{k,c}^{(N)} \tag{A.3}$$

$$\overset{\star}{t}{}_{1,c}^{(M)} = \frac{1}{\sqrt{2}}t_{1,c}^{(N)} \tag{A.4}$$

when $k = 1, 2, \ldots, N$.

M **odd.** Let $M = 2N + 1$. (M is odd.) We show that if $M \geq 7$ (or $N \geq 3$),

$$\begin{aligned}
\overset{\star}{t}{}_1^{(M)} &= \overset{\star}{t}{}_1^{(2N+1)} = \overset{\star}{t}{}_1^{(2N)} + \overset{\star}{\underline{\Delta}}{}_N^{(M)} \overset{(A.1)}{=} \overset{\star}{t}{}_1^{(2N)} + \frac{1}{\sqrt{2}}\underline{\Delta}_N^{(N)} \overset{(A.2)}{=} \frac{1}{\sqrt{2}}\big(t_1^{(N)} + \underline{\Delta}_N^{(N)} \pm t_{1,c}^{(N)}\big) \\
&= \frac{1}{\sqrt{2}}t_{1,c}^{(N)} + \frac{1}{\sqrt{2}}\underline{\Delta}_N^{(N)} - \frac{1}{\sqrt{2}}\big(t_{1,c}^{(N)} - t_1^{(N)}\big) \\
&< \frac{1}{\sqrt{2}}t_{1,c}^{(N)} + \frac{1}{\sqrt{2}}\big(\underline{\Delta}_N^{(N)} - \underline{\Delta}_N^{(N)}\big) \qquad (A.5) \\
&= \frac{1}{\sqrt{2}}t_{1,c}^{(N)} \overset{(A.4)}{=} \overset{\star}{t}{}_{1,c}^{(2N)}
\end{aligned}$$

which is the upper bound stated in Corollary III.3.

To show (A.5), it must be true that

$$t_{1,c}^{(N)} - t_1^{(N)} > \underline{\Delta}_N^{(N)} \tag{A.6}$$

when $N \geq 3$. To show (A.6), recall that when M is even, using Lemma III.4, Lemma III.5, Part 2, followed by Lemma III.5, Part 1, that the half steps of optimal quantizers and UTCC quantizers both designed for the one-sided, unit variance exponential source satisfy $\underline{\Delta}_k^{(N)} < \underline{\Delta}_{k,c}^{(N)}$ for $k \geq 2$, (i.e., the lower half step of an optimal quantizer are always less than the corresponding lower half step of a UTCC quantizer when $k \geq 2$). Since $t_1^{(N)}$ and $t_{1,c}^{(N)}$ is equal to the sum of the half steps for an optimal quantizer and likewise for a UTCC quantizer, it is clear that as N

increases, the difference between $t_1^{(N)}$ and $t_{1,c}^{(N)}$ increases. Since $\underline{\Delta}_N^{(N)}$ and $\underline{\Delta}_{N,c}^{(N)}$ are both (strictly) decreasing as N increases, there exists an N_0 such that for all $N \geq N_0$, the difference between $t_{1,c}^{(N)}$ and $t_1^{(N)}$ is greater than $\underline{\Delta}_N^{(N)}$, i.e., there exists N_0 such that for all $N \geq N_0$,

$$\begin{aligned}
\underline{\Delta}_N^{(N)} &< \big(\underline{\Delta}_{N,c}^{(N)} - \underline{\Delta}_N^{(N)}\big) + 2\Big(\sum_{i=2}^{N-1} \underline{\Delta}_{i,c}^{(N)} - \underline{\Delta}_i^{(N)}\Big) + \big(\underline{\Delta}_{1,c}^{(N)} - \underline{\Delta}_1^{(N)}\big) \\
&= \Big(\underline{\Delta}_{N,c}^{(N)} + 2\sum_{i=2}^{N-1} \underline{\Delta}_{i,c}^{(N)} + \underline{\Delta}_{1,c}^{(N)}\Big) - \Big(\underline{\Delta}_N^{(N)} + 2\sum_{i=2}^{N-1} \underline{\Delta}_i^{(N)} + \underline{\Delta}_1^{(N)}\Big) \\
&< t_{1,c}^{(N)} - t_1^{(N)}. \qquad \text{(A.7)}
\end{aligned}$$

By trial and error, the smallest N_0 for which (A.7) holds is when $N_0 = 3$ (which is equivalent to $M_0 = 6$.

Thus, when $M \geq 7$, (A.5) is true and consequently, the statement in Corollary III.3, Part 1a) for the case when M is odd is true.

APPENDIX B

Facts About Exponential *UTCC* and *USQC* Quantizers

The following appendix contains both facts and remarks regarding the MSE performance of UTCC and USQC quantization systems designed for a one-sided exponential source with unit variance, including the derivation for the MSE expressions of these quantizers. The MSE derivations (along with necessary facts) are presented first, followed by remarks regarding performance. Note that the terminology used in this appendix is the same as that used in Chapter III and it is necessary to be familiar with the material in that chapter to understand what is presented here.

Recall that the only difference between an N-level UTCC quantization system and an N-level USQC quantization system is the fact that the UTCC quantizer uses centroid reconstruction levels while the USQC quantizers uses reconstruction levels that are determined by mapping the midpoints of an N-level USQ defined over $[0, 1]$ with $(C^*)^{-1}$, the inverse to the asymptotically optimal compressing function C^*. In order to formulate the MSE performance expressions for UTCC and USQC quantizers, we repeat, from Chapter III, the specifications for the thresholds, step sizes and reconstruction levels of both types of quantizers for easy reference. From

(3.3.8), (3.3.9), (3.3.10):

$$\begin{array}{lrll}
\text{UTCC and USQC} & \Delta_{k,c}^{(N)} = -3\log\left(1-\dfrac{1}{k}\right), & k=1,2,\cdots,N & \text{(B.1)} \\
\text{UTCC and USQC} & t_{k,c}^{(N)} = -3\log\left(\dfrac{k}{N}\right), & k=0,1,\cdots,N & \text{(B.2)} \\
\text{UTCC reconstruction levels} & \mu_{k,c}^{(N)} = t_{k,c}^{(N)} + \underline{\Delta}_{k,c}^{(N)}, & k=1,2,\cdots,N & \\
\text{USQC reconstruction levels} & l_{k,USQC}^{(N)} = -3\log\left(\dfrac{2k-1}{2N}\right), & k=1,2,\cdots,N, & \text{(B.3)}
\end{array}$$

where (B.3), included above, but not given in Chapter III, is derived in the last section of this appendix.

Also for reference are the expressions for the half steps of a N-level UTCC quantizers:

$$\underline{\Delta}_{k,c}^{(N)} = 1 - \frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3} \tag{B.4}$$

and

$$\overline{\Delta}_{k,c}^{(N)} = \frac{\Delta_{k,c}^{(N)}}{1-\left(1-\frac{1}{k}\right)^3} - 1. \tag{B.5}$$

MSE for UTCC quantizers. The MSE performance of an N-level UTCC quantizer can be expressed as the sum of two parts

$$\begin{aligned}
D_{UTCC}(N) &= \sum_{k=1}^{N} \int_{t_{k,c}^{(N)}}^{t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}} \left(x-\mu_{k,c}^{(N)}\right)^2 f(x)\,dx \\
&\overset{IBP}{=} \left[\sum_{k=1}^{N} -\left(x-\mu_{k,c}^{(N)}\right)^2 e^{-x}\Bigg|_{t_{k,c}^{(N)}}^{t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}}\right] + d_{UTCC}(N).
\end{aligned}$$

Before evaluating the first expression, we see that the second expression $d_{c,k}$ can be simplified since UTCC quantizers have centroid reconstruction levels

$$d_{UTCC}^{(N)} = 2\sum_{k=1}^{N}\left(\mu_k^{(N)}-\mu_{k,c}^{(N)}\right) P_{\left[t_{k,c}^{(N)},\,t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}\right)} = 0.$$

Thus the first expression is equal to D_{c_k}. The first expression, however, is not so easily to reduce (which is in contrast to the case for optimal quantizers and for quantizers designed using the sequences s_k and v_k). The reason for this is because UTCC quantizers do not satisfy the nearest neighbor optimality condition, i.e., $\overline{\Delta}_k \neq \underline{\Delta}_{k-1}$ for any $k \in \{1, 2, \ldots, N\}$ (as was shown in Lemma III.7. Working on the first term, we have

$$\begin{aligned}
&\sum_{k=1}^{N} -\left(x-\mu_{k,c}^{(N)}\right)^2 e^{-x}\Bigg|_{t_{k,c}^{(N)}}^{t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}} \\
&= -\left[\sum_{k=1}^{N}\left(t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}-\mu_{k,c}^{(N)}\right)^2 e^{-\left(t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}\right)} - \left(t_{k,c}^{(N)}-\mu_{k,c}^{(N)}\right)^2 e^{-t_{k,c}^{(N)}}\right] \\
&= -\left[\sum_{k=1}^{N}\overline{\Delta}_{k,c}^2 e^{-\left(t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}\right)} - \underline{\Delta}_{k,c}^2 e^{-t_{k,c}^{(N)}}\right],
\end{aligned}$$

where we define $\overline{\Delta}_{1,c} = \infty$. Since $t_{k,c}^{(N)} + \Delta_{k,c}^{(N)} = t_{k-1,c}^{(N)}$ and because we have established that the first expression is equal to D_{UTCC}, we have

$$D_{UTCC}(N) = \sum_{k=1}^{N} \underline{\Delta}_{k,c}^2 e^{-t_{k,c}^{(N)}} - \overline{\Delta}_{k,c}^2 e^{-\left(t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}\right)} = \sum_{k=1}^{N} \underline{\Delta}_{k,c}^2 e^{-t_{k,c}^{(N)}} - \overline{\Delta}_{k,c}^2 e^{-\left(t_{k-1,c}^{(N)}\right)}.$$

Finally, using (B.2), (B.4) and (B.5), we can express the MSE of an N-level UTCC quantizer as

$$\begin{aligned}
D_{UTCC}(N) &= \sum_{k=1}^{N} \underline{\Delta}_{k,c}^2 e^{-t_{k,c}^{(N)}} - \overline{\Delta}_{k,c}^2 e^{-\left(t_{k-1,c}^{(N)}\right)} \\
&= \sum_{k=1}^{N}\left[1-\frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right]^2 e^{3\log\left(\frac{k}{N}\right)} - \left[\frac{\Delta_{k,c}^{(N)}}{1-\left(1-\frac{1}{k}\right)^3}-1\right]^2 e^{3\log\left(\frac{k-1}{N}\right)} \\
&= \sum_{k=1}^{N}\left[1-\frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right]^2 \left(\frac{k}{N}\right)^3 - \left[\frac{\Delta_{k,c}^{(N)}}{1-\left(1-\frac{1}{k}\right)^3}-1\right]^2 \left(\frac{k-1}{N}\right)^3 \\
&= \sum_{k=1}^{N}\left[1-\frac{-3\log\left(1-\frac{1}{k}\right)\cdot\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right]^2 \left(\frac{k}{N}\right)^3 - \left[\frac{-3\log\left(1-\frac{1}{k}\right)}{1-\left(1-\frac{1}{k}\right)^3}-1\right]^2 \left(\frac{k-1}{N}\right)^3, \quad \text{(B.6)}
\end{aligned}$$

where in the last equality we have used (B.1). While not compact, (B.6) gives an exact, closed-form formula for computing the MSE of an N-level UTCC quantizer.

MSE for USQC quantizers. The MSE of an N-level USQC quantizer can also be decomposed as the sum of two expressions

$$\begin{aligned} D_{USQC}\left(N\right) &= \sum_{k=1}^{N} \int\limits_{t_{k,c}^{(N)}}^{t_{k,c}^{(N)}+\Delta_{k,c}^{(N)}} \left(x - l_{k,USQC}^{(N)}\right)^2 f\left(x\right) dx \\ &\overset{IBP}{=} \left[\sum_{k=1}^{N} - \left(x - l_{k,USQC}^{(N)}\right)^2 e^{-x} \Big|_{t_{k,c}^{(N)}}^{t_{k,c}^{(N)}+\Delta_{k}^{(N)}} \right] + d_{USQC}\left(N\right), \qquad \text{(B.7)} \end{aligned}$$

where we note that the thresholds (and consequently the step sizes) of USQC quantizers are the same as for the corresponding UTCC quantizer.

Using (B.2) and (B.3),we evaluate the first expression in (B.7)

$$\sum_{k=1}^{N} - \left(x - l_{k,USQC}^{(N)}\right)^2 e^{-x} \Big|_{t_{k,c}^{(N)}}^{t_{k-1,c}^{(N)}}$$

$$\begin{aligned} &= \sum_{k=1}^{N} - \left(t_{k-1,c}^{(N)} - l_{k,USQC}^{(N)}\right)^2 e^{-t_{k-1,c}^{(N)}} + \left(t_{k,c}^{(N)} - l_{k,USQC}^{(N)}\right)^2 e^{-t_{k,c}^{(N)}} \\ &= \sum_{k=1}^{N} - \left(-3\log\left(\frac{k-1}{N}\right) - \left(-3\log\left(\frac{2k-1}{2N}\right)\right)\right)^2 e^{--3\log\left(\frac{k-1}{N}\right)} + \left(-3\log\left(\frac{k}{N}\right) - \right. \\ &\left. \left(-3\log\left(\frac{2k-1}{2N}\right)\right)\right)^2 e^{--3\log\left(\frac{k}{N}\right)} \\ &= \sum_{k=1}^{N} - \left(-3\log\left(\frac{k-1}{N} \times \frac{2N}{2k-1}\right)\right)^2 e^{3\log\left(\frac{k-1}{N}\right)} + \left(-3\log\left(\frac{k}{N} \times \frac{2N}{2k-1}\right)\right)^2 e^{3\log\left(\frac{k}{N}\right)} \\ &= \sum_{k=1}^{N} - \left(-3\log\left(\frac{2k-2}{2k-1}\right)\right)^2 \left(\frac{k-1}{N}\right)^3 + \left(-3\log\left(\frac{2k}{2k-1}\right)\right)^2 \left(\frac{k}{N}\right)^3 \\ &= \sum_{k=1}^{N} - \left(3\log\left(\frac{2k-1}{2k-2}\right)\right)^2 \left(\frac{k-1}{N}\right)^3 + \left(3\log\left(\frac{2k-1}{2k}\right)\right)^2 \left(\frac{k}{N}\right)^3 \\ &= \sum_{k=1}^{N} - \left(3\log\left(1 + \frac{1}{2k-2}\right)\right)^2 \left(\frac{k-1}{N}\right)^3 + \left(3\log\left(1 - \frac{1}{2k}\right)\right)^2 \left(\frac{k}{N}\right)^3. \qquad \text{(B.8)} \end{aligned}$$

Working on the second expression,

$$
\begin{aligned}
&d_{USQC}\left(N\right)\\
&= 2\sum_{k=1}^{N}\left(\mu_k^{(N)} - l_{k,USQC}^{(N)}\right) P_{\left[t_{k,c}^{(N)},t_{k,c}^{(N)}+\Delta_k^{(N)}\right)}\\
&= 2\sum_{k=1}^{N}\left(t_{k,c}^{(N)} + \underline{\Delta}_{k,c}^{(N)} - l_{k,USQC}^{(N)}\right) P_{\left[t_{k,c}^{(N)},t_{k,c}^{(N)}+\Delta_k^{(N)}\right)}\\
&= 2\sum_{k=1}^{N}\left(-3\log\left(\frac{k}{N}\right)+\left(1-\frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right)-\left(-3\log\left(\frac{k}{N}-\frac{1}{2N}\right)\right)\right) P_{\left[t_{k,c}^{(N)},t_{k,c}^{(N)}+\Delta_k^{(N)}\right)}\\
&= 2\sum_{k=1}^{N}\left(-3\log\left(\frac{k}{N}\right)+3\log\left(\frac{2k-1}{2N}\right)+\left(1-\frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right)\right) P_{\left[t_{k,c}^{(N)},t_{k,c}^{(N)}+\Delta_k^{(N)}\right)}\\
&= 2\sum_{k=1}^{N}\left(-3\log\left(\frac{k}{N}\times\frac{2N}{2k-1}\right)+\left(1-\frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right)\right) P_{\left[t_{k,c}^{(N)},t_{k,c}^{(N)}+\Delta_k^{(N)}\right)}\\
&= 2\sum_{k=1}^{N}\left(-3\log\left(\frac{2k}{2k-1}\right)+\left(1-\frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right)\right) P_{\left[t_{k,c}^{(N)},t_{k,c}^{(N)}+\Delta_k^{(N)}\right)}\\
&= 2\sum_{k=1}^{N}\left(3\log\left(\frac{2k-1}{2k}\right)+\left(1-\frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right)\right) P_{\left[t_{k,c}^{(N)},t_{k,c}^{(N)}+\Delta_k^{(N)}\right)}\\
&= 2\sum_{k=1}^{N}\left(3\log\left(1-\frac{1}{2k}\right)+\left(1-\frac{\Delta_{k,c}^{(N)}\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right)\right) P_{\left[t_{k,c}^{(N)},t_{k,c}^{(N)}+\Delta_k^{(N)}\right)}\\
&= 2\sum_{k=1}^{N}\left(3\log\left(1-\frac{1}{2k}\right)+\left(1-\frac{-3\log\left(1-\frac{1}{k}\right)\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right)\right) P_{\left[t_{k,c}^{(N)},t_{k-1,c}^{(N)}\right)}\\
&= 2\sum_{k=1}^{N}\left(3\log\left(1-\frac{1}{2k}\right)+\left(1-\frac{-3\log\left(1-\frac{1}{k}\right)\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right)\right)\left[\left(\frac{k}{N}\right)^3-\left(\frac{k-1}{N}\right)^3\right],
\end{aligned}
\tag{B.9}
$$

where again, we have used the expression for reconstruction levels in (B.3), half steps in (B.4), step sizes in (B.1), and

$$
\begin{aligned}
P_{\left[t_{k,c}^{(N)},t_{k-1,c}^{(N)}\right)} &= \int_{t_{k,c}^{(N)}}^{t_{k-1,c}^{(N)}} e^{-x}dx = -e^{-x}\Big|_{t_{k,c}^{(N)}}^{t_{k-1,c}^{(N)}} = e^{t_{k,c}^{(N)}} - e^{t_{k-1,c}^{(N)}}\\
&= e^{--3\log\left(\frac{k}{N}\right)} - e^{--3\log\left(\frac{k-1}{N}\right)} = \left(\frac{k}{N}\right)^3-\left(\frac{k-1}{N}\right)^3.
\end{aligned}
$$

Thus, with (B.8) and (B.9), we have

$$D_{USQC}\left(N\right) = \sum_{k=1}^{N} - \left(3\log\left(1+\frac{1}{2k-2}\right)\right)^2\left(\frac{k-1}{N}\right)^3 + \left(3\log\left(1-\frac{1}{2k}\right)\right)^2\left(\frac{k}{N}\right)^3 + \\ 2\left(3\log\left(1-\frac{1}{2k}\right)+\left(1-\frac{-3\log\left(1-\frac{1}{k}\right)\left(1-\frac{1}{k}\right)^3}{1-\left(1-\frac{1}{k}\right)^3}\right)\right)\left[\left(\frac{k}{N}\right)^3-\left(\frac{k-1}{N}\right)^3\right].$$

Comments. Here are some remarks on the MSE performance of UTCC quantizers and USQC quantizers:

- We expect the performance of UTCC quantizers to be better than that of USQC quantizers with step sizes $\Delta_{k,c}^{(N)}$ (as defined in (3.3.8)) which are quantizers that are well-known and well-studied.
- Improved performance is due to the fact that for each N, while UTCC quantizers and USQC quantizers share the same quantization cells, they differ in their placement of the reconstruction levels: Specifically, UTCC's have centroid reconstruction levels whereas USQC's do not.
- Asymptotic optimality: Since it is known that a sequence in N of USQC quantizers is asymptotically optimal [5], we can conclude that a sequence in N of UTCC quantizers is also asymptotically optimal.

Derivation of (B.3)**: Reconstruction levels of an N-level USQC quantizer.** Since the reconstruction levels of an N-level USQC quantizer come from the midpoints of the quantization cells of the USQ with support on $[0, 1]$ and step size $\Delta = \frac{1}{N}$ as

$$y_k \triangleq 1 - \frac{k}{N} + \frac{1}{2}\frac{1}{N} = 1 - \frac{2k-1}{2N},$$

we can express the reconstruction levels as

$$l_{k,USQC}^{(N)} \triangleq C^{*-1}\left(y_k\right) = C^{*-1}\left(1 - \frac{2k-1}{2N}\right)$$

(where C^{*-1} is the inverse to the optimal compressor function for the one-sided exponential source) or equivalently, using the definition of C^* from (3.3.7) in Chapter III

$$C^*\left(l_{k,USQC}^{(N)}\right) = 1 - e^{-\frac{l_{k,USQC}^{(N)}}{3}} = 1 - \frac{2k-1}{2N}$$

or equivalently,

$$e^{-\frac{l_{k,USQC}^{(N)}}{3}} = \frac{2k-1}{2N}$$

or equivalently, we get (B.3)

$$l_{k,USQC}^{(N)} = -3\log\left(\frac{2k-1}{2N}\right).$$

Since $t_{k,c}^{(N)} = -3\log\left(\frac{k}{N}\right)$ and since the k-th (lower) half step is defined to be the distance between the reconstruction level and the (lower) threshold of the kth quantization cell, the (induced) half step of the kth cell of an N-level USQC quantizer is

$$\underline{\Delta}_{k,USQC}^{(N)} \stackrel{\triangle}{=} l_{k,USQC}^{(N)} - t_{k,c}^{(N)} = -3\log\left(1-\frac{1}{2k}\right).$$

APPENDIX C

The B_{LB} Lower Bound Proof

This appendix contains the proof that $B_{LB}\left(p\right)$ is a lower bound to the function $B\left(p\right)$ for $p \in \left[0, \sqrt{2\left(1-\frac{2}{e}\right)}\right]$, and is self-contained in that all of the necessary facts and definitions previously established in the chapters of this thesis have been re-stated here. Note that there are function definitions used here that are slightly different than those used in the chapters and that this was done to facilitate the discussion as much as possible. We also make several new definitions that further clarify the proof.

Facts, definitions and terminology.

1. Recall that from [6], a portion of the principal branch of the Lambert W function $W_0\left(p\right)$ can be expressed as $W_0\left(p\right) = B \circ p\left(z\right)$, where

$$B(p)=\sum_{n=0}^{\infty}\xi_n p^n = -1+p-\frac{1}{3}p^2+\frac{11}{72}p^3-\frac{43}{540}p^4+\frac{760}{17280}p^5-\frac{221}{8505}p^6+\ldots \quad \text{(C.1)}$$

is a power series in p with a region of convergence defined as $ROC_p \triangleq \left[0, \sqrt{2}\right)$[1] and $p\left(z\right) = \sqrt{2}\sqrt{1+z\,e}$ with $z \in \left[-\frac{1}{e}, 0\right]$. As reported in [6], the series coeffi-

[1] In [6], ROC_p is actually defined to be $\left(-\sqrt{2}, \sqrt{2}\right)$. If $p\left(z\right)$ is alternatively defined by multiplying the current definition by -1, other branches of the Lambert W function may be approximated using the composition $B \circ p\left(z\right)$. However, since we are not concerned with these other branches, we have defined ROC_p to be non-negative.

cients ξ_n are obtained via a series inversion and can be generated via

$$\xi_n = \frac{n-1}{n+2}\left(\frac{\xi_{n-2}}{2} + \frac{\gamma_{n-2}}{4}\right) - \frac{\gamma_n}{2} - \frac{\xi_{n-1}}{n+1}$$
$$\gamma_n = \sum_{m=2}^{n-1} \xi_m \, \xi_{n+1-m}$$

with $\gamma_0 = 2,\ \gamma_1 = -1,\ \xi_0 = -1,\ \xi_1 = 1$.

2. **Establishing the regions of interest.** Since we are ultimately concerned with approximating a generating function (which utilizes the principal branch of the Lambert W function $W_0(z)$) for the terms of the Nitadori sequence $\underline{\eta}_k$, we are only interested in specific subsets of the domains of $W_0(z)$ and $p(z)$. To make the discussion easier to read, we will refer to each the following half open intervals as a *region of interest* (ROI): Let the region of interest with respect to z be defined as $ROI_z \triangleq \left(-\frac{1}{e}, -\frac{2}{e^2}\right]$. Also, since $p(z) = \sqrt{2(1+ze)}$ is an increasing function over ROI_z, and hence bijective on ROI_z, we define the region of interest with respect to p as $ROI_p \triangleq p(ROI_z) = \left(0, \sqrt{2\left(1-\frac{2}{e}\right)}\right]$. Finally, since the principal branch of the Lambert W function is also increasing over ROI_z, we define the region of interest with respect to w as $ROI_w \triangleq L(ROI_z) = \left(-1, L\left(-\frac{2}{e^2}\right)\right]$, where $L\left(-\frac{2}{e^2}\right) \approx -.4063757398$.

   Additionally, we will refer to the *closure of a region interest* as union of an ROI with its lower endpoint: The closure of ROI_p will be given by $\overline{ROI}_p \triangleq \left[0, \sqrt{2\left(1-\frac{2}{e}\right)}\right]$, the closure of ROI_z will be given by $\overline{ROI}_z \triangleq \left[-\frac{1}{e}, -\frac{2}{e^2}\right]$ and the closure of ROI_w will be given by $\overline{ROI}_w \triangleq \left[-1, LambertW\left(-\frac{2}{e^2}\right)\right]$.

3. Definition of $L(z), B_{LB}(p)$ and construction of $L_{LB}(z)$. Since we are only interested in a $W_0(z)$ defined over the restricted domain $ROI_z \subset \left[-\frac{1}{e}, 0\right]$ and because we want to be able to distinguish this function from the principal branch of the Lambert W function $W_0(z)$ defined over the domain $\left[-\frac{1}{e}, 0\right]$ for which $W_0 = B \circ p(z)$, we define the function

$$L(z) \triangleq W(z) = B \circ p(z) \tag{C.2}$$

   when $z \in \overline{ROI}_z$. Based on this definition, we also define

$$L_{LB}(z) \triangleq B_{LB} \circ p(z) \tag{C.3}$$

when $z \in \overline{ROI}_z$, where

$$B_{LB}\left(p\right) \triangleq -1 + p - \frac{1}{3}p^2 + \frac{11}{72}p^3 - \frac{43}{540}p^4 \tag{C.4}$$

is the partial sum of $B\left(p\right)$ that is truncated to the fourth order and defined over $\overline{ROI}_p$.

Lemma C.1. *$B_{LB}\left(p\right) \leq B\left(p\right)$ for $p \in \overline{ROI}_p$.*

Before proceeding to the proof of Lemma C.1, we need a few facts that we will prove in the following two lemmas.

Lemma C.2. *Define $Z\left(w\right) \triangleq we^w$ for $w \in \overline{ROI}_w$. With this definition, the following are true:*

1. *$\frac{dL}{dz}\left(z\right)$ over ROI_z is well-defined and can be determined from:*

$$\frac{dL}{dz}\left(z\right) = \frac{1}{\left(w+1\right)e^w} \tag{C.5}$$

where $z = Z\left(w\right)$ and $w \in ROI_w$.

2. *$\frac{dB}{dp}\left(p\right)$ over $\overline{ROI}_p$ exists, and for $p \in ROI_p$, it can be determined from:*

$$\frac{dB}{dp}\left(p\right) = \left.\frac{dB}{dp}\right|_{p=p\left(Z\left(w\right)\right)} = \frac{\sqrt{2}\sqrt{1+we^{w+1}}}{\left(w+1\right)e^{w+1}} \tag{C.6}$$

where $p\left(Z\left(w\right)\right) = \sqrt{2\left(1+we^{w+1}\right)}$ and $w \in ROI_w$.

3. *$\frac{dB_{LB}}{dp}\left(p\right)$ over $\overline{ROI}_p$ is well-defined and has the form*

$$\frac{dB_{LB}}{dp}\left(p\right) = 1 - \frac{2}{3}p + \frac{11}{24}p^2 - \frac{43}{135}p^3 \tag{C.7}$$

which can also be determined from

$$\frac{dB_{LB}}{dp}\left(p\right) = \left.\frac{dB_{LB}}{dp}\right|_{p=p\left(z\right)} = \left(1 + \frac{11}{12}\left(1+we^{w+1}\right)\right) - \left(\frac{2}{3} + \frac{86}{135}\left(1+we^{w+1}\right)\right) \times \sqrt{2}\sqrt{1+we^{w+1}} \tag{C.8}$$

where $p = \sqrt{2\left(1+we^w\right)}$ and $w \in \overline{ROI}_w$.

Proof. We prove each item stated in Lemma C.2.

1. With the inverse to $L(z)$ defined as

$$Z(w) \stackrel{\triangle}{=} we^{w} \tag{C.9}$$

for $w \in \overline{ROI}_w$, (C.5) is established by using the Inverse Function Theorem [12]: L is differentiable on ROI_z because $Z(w)$ is differentiable on ROI_w and $\frac{dZ}{dw} > 0$ over ROI_w:

$$\frac{dL}{dz}(z) = \frac{1}{\frac{dZ}{dw}} = \frac{1}{(w+1)e^{w}}, \tag{C.10}$$

where $z = Z(w)$, $w \in ROI_w$.

2. Since $B(p)$ is a power series and $\overline{ROI}_p \subset ROC_p$, we know that $B(p)$ is (infinitely) differentiable on $\overline{ROI}_p$. However, given the form of $B(p)$ in (C.1), it is difficult to directly determine the expression for $\frac{dB}{dp}(p)$. So proceeding in an indirect manner, using the definition for $L(z)$ in (C.2), we know that

$$\frac{dL}{dz} = \left.\frac{dB}{dp}\right|_{p=p(z)} \times \frac{dp}{dz} \tag{C.11}$$

for $z \in ROI_z$. From (C.11), we can use the fact that

$$\frac{dp}{dz} = \frac{e}{p(z)} > 0 \tag{C.12}$$

when $z \in ROI_z$ along with the expression in (C.5) from Part 1 of this lemma, to obtain

$$\left.\frac{dB}{dp}\right|_{p=p(z)} = \frac{dL}{dz} \times \frac{1}{\frac{dp}{dz}} = \frac{1}{(w+1)e^{w}} \times \frac{p(z)}{e}. \tag{C.13}$$

With $z = Z(w) = we^{w}$ for $w \in ROI_w$, we use (C.13) to get

$$\left.\frac{dB}{dp}\right|_{p=p(Z(w))} = \frac{p(z)}{(w+1)e^{w+1}} = \frac{\sqrt{2}\sqrt{1+ze}}{(w+1)e^{w+1}} = \frac{\sqrt{2}\sqrt{1+we^{w+1}}}{(w+1)e^{w+1}} \tag{C.14}$$

when $w \in ROI_w$. Since $p(Z(w))$ is bijective over ROI_w (onto ROI_p), it is clear that for $w \in ROI_w$,

$$\frac{dB}{dp}(p) = \left.\frac{dB}{dp}\right|_{p=p(z)}.$$

Thus $\frac{dB}{dp}(p)$ for $p \in ROI_p$ can be determined from (C.14).

3. Since $B_{LB}(p)$ is a polynomial (see (C.4)),

$$\frac{dB_{LB}}{dp} = 1 - \frac{2}{3}p + \frac{11}{24}p^2 - \frac{43}{135}p^3 \tag{C.15}$$

and it is defined over $\overline{ROI}_p$.

If we make the substitutions $z = Z(w)$ and $p(z) = p(Z(w)) = \sqrt{2}\sqrt{1+we^{w+1}}$, we have

$$\begin{aligned}\left.\frac{dB_{LB}}{dp}\right|_{p=p(Z(w))} &= 1 - \frac{2}{3}p(Z(w)) + \frac{11}{24}p(Z(w))^2 - \frac{43}{135}p(Z(w))^3 \\ &= 1 - \frac{2}{3}\left(\sqrt{2}\sqrt{1+we^{w+1}}\right) + \frac{11}{24}\left(\sqrt{2}\sqrt{1+we^{w+1}}\right)^2 - \\ &\quad \frac{43}{135}\left(\sqrt{2}\sqrt{1+we^{w+1}}\right)^3 \\ &= 1 - \frac{2}{3}\sqrt{2}\sqrt{1+we^{w+1}} + \frac{11}{12}\left(1+we^{w+1}\right) - \frac{86}{135}\sqrt{2}\times \\ &\quad \left(1+we^{w+1}\right)\sqrt{1+we^{w+1}} \\ &= \left(1 + \frac{11}{12}\left(1+we^{w+1}\right)\right) - \left(\frac{2}{3} + \frac{86}{135}\left(1+we^{w+1}\right)\right) \times \\ &\quad \sqrt{2}\sqrt{1+we^{w+1}},\end{aligned} \tag{C.16}$$

where $z = Z(w)$ and $w \in \overline{ROI}_w$. Since $p(Z(w))$ is the composition of two bijective functions, $p(z)$ and $Z(w)$, it is clear that $\frac{dB_{LB}}{dp}(p)$ over $\overline{ROI}_p$ can be determined from (C.16).

■

Lemma C.3. *$\frac{dB}{dp}(p) \geq \frac{dB_{LB}}{dp}(p)$ over $\overline{ROI}_p$.*

Proof. There are two parts to this proof. First, we consider the value of $\frac{dB_{LB}}{dp}(p)$ and $\frac{dB}{dp}(p)$ at $p = 0$ and then we consider the relationship between $\frac{dB_{LB}}{dp}(p)$ and $\frac{dB}{dp}(p)$ for $p \in ROI_p$.

For the first part, by taking the first derivative with respect to p of (C.1) and using (C.15), we have at $p = 0$

$$\frac{dB}{dp}(p)\bigg|_{p=0} = 1 = \frac{dB_{LB}}{dp}(p)\bigg|_{p=0}.$$

Since $\frac{dB}{dp}(p) = \frac{dB_{LB}}{dp}(p)$ when p equals the lower end point of $\overline{ROI}_p$, it remains to show that

$$\frac{dB}{dp}(p) \geq \frac{dB_{LB}}{dp}(p) \tag{C.17}$$

for $p \in ROI_p$.

Using the relationships in Lemma C.2 from (C.6) and (C.8), we know that (C.17) is true for $p \in ROI_p$ if and only if

$$\frac{\sqrt{2}\sqrt{1+we^{w+1}}}{(w+1)\,e^{w+1}} \geq \left(1+\frac{11}{12}\left(1+we^{w+1}\right)\right) - \left(\frac{2}{3}+\frac{86}{135}\left(1+we^{w+1}\right)\right)\sqrt{2}\sqrt{1+we^{w+1}} \tag{C.18}$$

is true for $w \in ROI_w$.

With some algebraic manipulation (C.18) becomes

$$\sqrt{2}\sqrt{1+we^{w+1}} \geq \left[\left(1+\frac{11}{12}\left(1+we^{w+1}\right)\right) - \left(\frac{2}{3}+\frac{86}{135}\left(1+we^{w+1}\right)\right)\sqrt{2}\sqrt{1+we^{w+1}}\right](w+1)e^{w+1} \tag{C.19}$$

or

$$\sqrt{2}\sqrt{1+we^{w+1}} + \left(\frac{2}{3}+\frac{86}{135}\left(1+we^{w+1}\right)\right)\sqrt{2}\sqrt{1+we^{w+1}}\,(w+1)\,e^{w+1} \geq \left(1+\frac{11}{12}\left(1+we^{w+1}\right)\right)(w+1)\,e^{w+1}$$

or

$$\left[1+\left(\frac{2}{3}+\frac{86}{135}\left(1+we^{w+1}\right)\right)(w+1)\,e^{w+1}\right]\sqrt{2}\sqrt{1+we^{w+1}} \geq \left(1+\frac{11}{12}\left(1+we^{w+1}\right)\right)(w+1)\,e^{w+1}$$

or

$$\left\{\left[1+\left(\frac{2}{3}+\frac{86}{135}\left(1+we^{w+1}\right)\right)(w+1)\,e^{w+1}\right]\sqrt{2}\sqrt{1+we^{w+1}}\right\}^2$$
$$\geq\left\{\left(1+\frac{11}{12}\left(1+we^{w+1}\right)\right)(w+1)\,e^{w+1}\right\}^2$$

or

$$2\left[1+\left(\frac{2}{3}+\frac{86}{135}\left(1+we^{w+1}\right)\right)(w+1)\,e^{w+1}\right]^2\left(1+we^{w+1}\right)$$
$$\geq\left(1+\frac{11}{12}\left(1+we^{w+1}\right)\right)^2\left[(w+1)\,e^{w+1}\right]^2$$

or

$$2\left[1+\frac{2}{3}\left(1+\frac{43}{45}\left(1+we^{w+1}\right)\right)(w+1)\,e^{w+1}\right]^2\left(1+we^{w+1}\right)$$
$$\geq\left(1+\frac{11}{12}\left(1+we^{w+1}\right)\right)^2\left[(w+1)\,e^{w+1}\right]^2. \qquad \text{(C.20)}$$

Consider the expression obtained by subtracting the right side of (C.20) from the left side of (C.20):

$$2\left[1+\frac{2}{3}\left(1+\frac{43}{45}\left(1+we^{w+1}\right)\right)(w+1)\,e^{w+1}\right]^2\left(1+we^{w+1}\right)-\left(1+\frac{11}{12}\left(1+we^{w+1}\right)\right)^2\left[(w+1)\,e^{w+1}\right]^2 \qquad \text{(C.21)}$$

which is a continuous function over ROI_w. To make things simpler (for now), we define the dummy function/variable $D \triangleq e^{w+1}$ and substitute D for e^{w+1} in (C.21) to get

$$2\left[1+\frac{2}{3}\left(1+\frac{43}{45}\left(1+wD\right)\right)(w+1)\,D\right]^2\left(1+wD\right)-\left(1+\frac{11}{12}(1+wD)\right)^2\left[(w+1)\,D\right]^2. \qquad \text{(C.22)}$$

Expanding (C.22) and then collecting terms, we have

$$\text{(C.22)} = term_5 + term_4 + term_3 + term_2 + term_1 + term_0, \qquad \text{(C.23)}$$

where

$$\begin{aligned}
term_5 &\triangleq \frac{14792}{18225} w^3 \left(w+1\right)^2 D^5 \\
term_4 &\triangleq \frac{320117}{97200} w^2 \left(w+1\right)^2 D^4 \\
term_3 &\triangleq \frac{1}{48600} w(w+1)(279721w{+}155881)D^3 = w(w+1)\left(\frac{279721w}{48600}+\frac{155881}{48600}\right)D^3 \\
term_2 &\triangleq \frac{1}{291600}\,(w+1)(2183687w{-}79993)D^2 = (w+1)\left(\frac{2183687w}{291600}-\frac{79993}{291600}\right)D^2 \\
term_1 &\triangleq \left(\frac{704}{135}+\frac{974}{135}w\right) D = \frac{974}{135}\left(w+1\right) D - 2D \\
term_0 &\triangleq 2.
\end{aligned}$$

If we combine $term_1$ and $term_0$, we get

$$term_{1+0} \triangleq \frac{974}{135}\left(w+1\right) D + 2\left(1-D\right).$$

Using the Taylor series expansion for e^{w+1}

$$D = \sum_{n\geq 0} \frac{1}{n!}\left(w+1\right)^n, \tag{C.24}$$

we re-express (C.23) by replacing D with its Taylor series in (C.24)

$$\begin{aligned}
\text{(C.21)} = \text{(C.22)} &= term_5 + term_4 + term_3 + term_2 + term_{1+0} \\
&= \frac{14792}{18225} w^3 \left(w+1\right)^2 D^5 + \frac{320117}{97200} w^2 \left(w+1\right)^2 D^4 + w\left(w+1\right)\left(\frac{279721w}{48600}+\right. \\
&\quad \left.\frac{155881}{48600}\right)D^3 + (w+1)\left(\frac{2183687w}{291600}-\frac{79993}{291600}\right)D^2 + \frac{974}{135}\left(w+1\right)D + 2(1-D) \\
&= \left[\frac{14792}{18225} w^3 (w+1)^2 \sum_{n\geq 0}\frac{1}{n!}\left(w+1\right)^{5n}\right] + \left[\frac{320117}{97200} w^2 (w+1)^2 \sum_{n\geq 0}\frac{1}{n!}\left(w+1\right)^{4n}\right] + \\
&\quad \left[w(w+1)\left(\frac{279721w}{48600}+\frac{155881}{48600}\right)\sum_{n\geq 0}\frac{1}{n!}\left(w+1\right)^{3n}\right] + \left[(w+1)\left(\frac{2183687w}{291600}-\right.\right. \\
&\quad \left.\left.\frac{79993}{291600}\right)\sum_{n\geq 0}\frac{1}{n!}\left(w+1\right)^{2n}\right] + \left[\frac{974}{135}\left(w+1\right)\sum_{n\geq 0}\frac{1}{n!}\left(w+1\right)^{n}\right] - \\
&\quad \left[2\sum_{n\geq 0}\frac{1}{(n+1)!}\left(w+1\right)^{n+1}\right]
\end{aligned} \tag{C.25}$$

$$\begin{aligned}
&= \sum_{n\geq 0} \frac{1}{n!}\Bigg\{ \Bigg[\frac{14792}{18225} w^3 (w+1)^2 (w+1)^{5n}\Bigg] + \Bigg[\frac{320117}{97200} w^2 (w+1)^2 (w+1)^{4n}\Bigg] + \\
&\Bigg[w(w+1)\Bigg(\frac{279721w}{48600} + \frac{155881}{48600}\Bigg)(w+1)^{3n}\Bigg] + \Bigg[(w+1)\Bigg(\frac{2183687w}{291600} - \\
&\frac{79993}{291600}\Bigg)(w+1)^{2n}\Bigg] + \Bigg[\frac{974}{135}(w+1)^{n+1}\Bigg] - \Bigg[\frac{2\cdot n!}{(n+1)!}(w+1)^{n+1}\Bigg]\Bigg\}
\end{aligned} \tag{C.26}$$

$$\begin{aligned}
&= \sum_{n\geq 0} \frac{1}{n!}\Bigg\{ \Bigg[\frac{14792}{18225} w^3 (w+1)^{5n+2}\Bigg] + \Bigg[\frac{320117}{97200} w^2 (w+1)^{4n+2}\Bigg] + \\
&\Bigg[\Bigg(\frac{279721w}{48600} + \frac{155881}{48600}\Bigg) w\,(w+1)^{3n+1}\Bigg] + \Bigg[\Bigg(\frac{2183687w}{291600} - \frac{79993}{291600}\Bigg) \times \\
&(w+1)^{2n+1}\Bigg] + \Bigg[\frac{974}{135}(w+1)^{n+1}\Bigg] - \Bigg[\frac{2}{n+1}(w+1)^{n+1}\Bigg]\Bigg\}
\end{aligned} \tag{C.27}$$

$$\begin{aligned}
&= (w+1)\sum_{n\geq 0} \frac{1}{n!}\Bigg\{ \Bigg[\frac{14792}{18225} w^3 (w+1)^{5n+1}\Bigg] + \Bigg[\frac{320117}{97200} w^2 (w+1)^{4n+1}\Bigg] + \\
&\Bigg[\Bigg(\frac{279721w}{48600} + \frac{155881}{48600}\Bigg) w\,(w+1)^{3n}\Bigg] + \Bigg[\Bigg(\frac{2183687w}{291600} - \frac{79993}{291600}\Bigg) \times \\
&(w+1)^{2n}\Bigg] + \Bigg[\frac{974}{135}(w+1)^{n}\Bigg] - \Bigg[\frac{2}{n+1}(w+1)^{n}\Bigg]\Bigg\}
\end{aligned} \tag{C.28}$$

$$\begin{aligned}
&= (w+1)\sum_{n\geq 0} \frac{1}{n!}(w+1)^n\Bigg\{ \Bigg[\frac{14792}{18225} w^3 (w+1)^{4n+1}\Bigg] + \Bigg[\frac{320117}{97200} w^2 (w+1)^{3n+1}\Bigg] + \\
&\Bigg[\Bigg(\frac{279721w}{48600} + \frac{155881}{48600}\Bigg) w\,(w+1)^{2n}\Bigg] + \Bigg[\Bigg(\frac{2183687w}{291600} - \frac{79993}{291600}\Bigg) \times \\
&(w+1)^{n}\Bigg] + \Bigg[\frac{974}{135} - \frac{2}{n+1}\Bigg]\Bigg\},
\end{aligned} \tag{C.29}$$

where:

- To get (C.25), we have used the fact that

$$1 - D = 1 - e^{w+1} = 1 - \sum_{n\geq 0}\frac{(w+1)^n}{n!} = -\sum_{n\geq 1}\frac{(w+1)^n}{n!} = -\sum_{n\geq 0}\frac{(w+1)^{n+1}}{(n+1)!}.$$

- To get (C.26), we use the fact that we can move the infinite sum to the outside and rearrange terms inside the infinite sum because we are dealing with a finite sum of absolutely convergent series on ROI_w.

- To get (C.28), we factored $(w+1)$ out of the infinite sum.
- To get (C.29), inside of the infinite sum, we factored out $(w+1)^n$ and then combined the last two terms of (C.28).

Since $w+1>0$ for $w \in ROI_w$, (C.29) is non-negative if and only if

$$\sum_{n\geq 0} \frac{1}{n!}(w+1)^n \left\{ \left[\frac{14792}{18225} w^3 (w+1)^{4n+1}\right] + \left[\frac{320117}{97200} w^2 (w+1)^{3n+1}\right] + \left[\left(\frac{279721w}{48600} + \frac{155881}{48600}\right) w (w+1)^{2n}\right] + \left[\left(\frac{2183687w}{291600} - \frac{79993}{291600}\right)(w+1)^n\right] + \left[\frac{974}{135} - \frac{2}{n+1}\right]\right\} \geq 0 \tag{C.30}$$

for all $w \in ROI_w$. If we can show that the expression on the left in (C.30) is the infinite sum of non-negative terms, then (C.30) is true for all $w \in ROI_w$ and we will have proven this lemma.

Consider the expression in (C.30) that is inside of the infinite sum (without the $\frac{1}{n!}(w+1)^n$ factor since it is always positive with the usual convention that $\frac{1}{0!}=1$):

$$A_n(w) \triangleq \left[\frac{14792}{18225} w^3 (w+1)^{4n+1}\right] + \left[\frac{320117}{97200} w^2 (w+1)^{3n+1}\right] + \left[\left(\frac{279721w}{48600} + \frac{155881}{48600}\right) \times w(w+1)^{2n}\right] + \left[\left(\frac{2183687w}{291600} - \frac{79993}{291600}\right)(w+1)^n\right] + \left[\frac{974}{135} - \frac{2}{n+1}\right], \tag{C.31}$$

where we note that as n increases, $\frac{2}{n+1}$ decreases. For each $n \geq 0$, (C.31) is a polynomial in w of order $4n+4$. Checking at $n=0$, we have

$$\begin{aligned} A_0(w) &= \left[\frac{14792}{18225} w^3 (w+1)\right] + \left[\frac{320117}{97200} w^2 (w+1)\right] + \left[\left(\frac{279721w}{48600} + \frac{155881}{48600}\right) w\right] + \\ &\quad \left[\frac{2183687w}{291600} - \frac{79993}{291600}\right] + \left[\frac{974}{135} - 2\right] \\ &= \frac{14792}{18225} w^4 + \frac{1197023}{291600} w^3 + \frac{879559}{97200} w^2 + \frac{3118973}{291600} w + \frac{1440647}{291600} \end{aligned} \tag{C.32}$$

is a 4th degree polynomial in w. To show that $A_0(w) \geq 0$ over ROI_w, first, we observe that $A_0(-1) = 0$. Next, we see that

$$\frac{d}{dw}A_0(w) = \frac{59168}{18225}w^3 + \frac{1197023}{97200}w^2 + \frac{879559}{48600}w + \frac{3118973}{291600} > 0$$

over $w \in \overline{ROI_w}$ because for $w \in ROI_w$, w is greater than the only real root of $\frac{d}{dw}A_0(w)$ which equals approximately -1.591135770. Thus, we have established that for $n = 0$, $A_n(w) \geq 0$ over ROI_w.

Now consider $A_n(w)$ when $n \geq 1$:

$$\begin{aligned} A_n(w) = {} & \underbrace{\left[\frac{14792}{18225}w^3(w+1)^{4n+1}\right]}_{a_n} + \underbrace{\left[\frac{320117}{97200}w^2(w+1)^{3n+1}\right]}_{b_n} + \\ & \underbrace{\left[\left(\frac{279721w}{48600} + \frac{155881}{48600}\right)w(w+1)^{2n}\right]}_{c_n} + \\ & \underbrace{\left[\left(\frac{2183687w}{291600} - \frac{79993}{291600}\right)(w+1)^n\right]}_{d_n} + \underbrace{\left[\frac{974}{135} - \frac{2}{n+1}\right]}_{e_n}. \end{aligned} \tag{C.33}$$

By inspection, when n is large, the term e_n *dominates* $A_n(w)$, i.e., the value of $A_n(w) \approx e_n$. This observation is supported by the fact that when n is large, the other terms in $A_n(w)$, a_n, b_n, c_n, and d_n, are scaled by the magnitude of w and $w+1$; thus for large n, $|a_n|, |b_n|, |c_n|, |d_n|$ are small since $|w|, |w+1| < 1$ for each $w \in ROI_w$. In contrast, e_n does not depend on w, and in fact, $e_n \uparrow \frac{974}{135} \approx 7.214814815$ as n increases. Then if n is large enough, since $e_n \geq \frac{704}{135} \approx 5.214814815$, there exists $n = n_0$ such that for all $n \geq n_0$, $-(a_n + b_n + c_n + d_n) < \frac{704}{135} \leq e_n$ and thus $A_n(w) \geq 0$.

Consider $n = 1$:

$$\begin{aligned} A_1(w) = {} & \underbrace{\left[\frac{14792}{18225}w^3(w+1)^5\right]}_{a_1} + \underbrace{\left[\frac{320117}{97200}w^2(w+1)^4\right]}_{b_1} + \\ & \underbrace{\left[w\left(\frac{279721w}{48600} + \frac{155881}{48600}\right)(w+1)^2\right]}_{c_1} + \\ & \underbrace{\left[\left(\frac{2183687w}{291600} - \frac{79993}{291600}\right)(w+1)\right]}_{d_1} + \underbrace{\left[\frac{974}{135} - 1\right]}_{e_1}. \end{aligned} \tag{C.34}$$

We find bounds for a_n, b_n, c_n, and d_n:

$$
\begin{aligned}
a_1 &= \frac{14792}{18225} w^3 (w+1)^5 > \frac{14792}{18225} (-1)^3 \left(LambertW\left(-\frac{2}{e^2}\right) + 1 \right)^5 \approx -0.05982980232 \\
b_1 &= \frac{320117}{97200} w^2 (w+1)^4 > 0 \\
c_1 &= w \left(\frac{279721w}{48600} + \frac{155881}{48600} \right) (w+1)^2 \\
&> -1 \cdot \underbrace{\left(\frac{279721}{48600} \cdot LambertW\left(-\frac{2}{e^2}\right) + \frac{155881}{48600} \right)}_{\text{maximium positive value in } ROI_w} \cdot \left(LambertW\left(-\frac{2}{e^2}\right) + 1 \right)^2 \\
&\approx -.3060510284 \\
d_1 &= \left(\frac{2183687w}{291600} - \frac{79993}{291600} \right)(w+1) > \left(-1 \cdot \frac{2183687}{291600} - \frac{79993}{291600} \right) \left(LambertW\left(-\frac{2}{e^2}\right) + 1 \right) \\
&\approx -4.608283146 \\
e_1 &= \left(\frac{974}{135} - 1 \right) = \frac{839}{135} \approx 6.214814815.
\end{aligned}
$$

Using these bounds, we have

$$A_1(w) > 1.240650838 \geq 0.$$

Since it is clear that for $n > 1$, e_n's domination of the expression $A_n(w)$ will increase even further, we see that for all $n \geq 0$, $A_n(w) \geq 0$. Thus, we have established that (C.30) is true for every $w \in ROI_w$ which what we needed to show to prove this lemma. ■

Proof of Lemma C.1. Using (C.1) and (C.4), we can evaluate $B(p)$ and $B_{LB}(p)$ at $p = 0$ to get

$$B(p)\Big|_{p=0} = -1 = B_{LB}(p)\Big|_{p=0}.$$

Using this fact and Lemma C.3, we conclude that $B(p) \geq B_{LB}(p)$ for $p \in \overline{ROI}_p$. ■

In Figure C.1, we have plotted the function $\frac{\frac{dB}{dp}(W)}{\frac{dB_{LB}}{dp}(W)}$ for $w \in \overline{ROI}_w$. This figure provides visual evidence to support the fact in Lemma C.3. In Figure C.2 and Figure C.3, we see that for $z \in \overline{ROI}_z$, the composition of $B_{LB}(p)$ with $p(z)$ appears to approximate $L(z)$ quite well, but that outside of $\overline{ROI}_p$, the approximation degrades considerably.

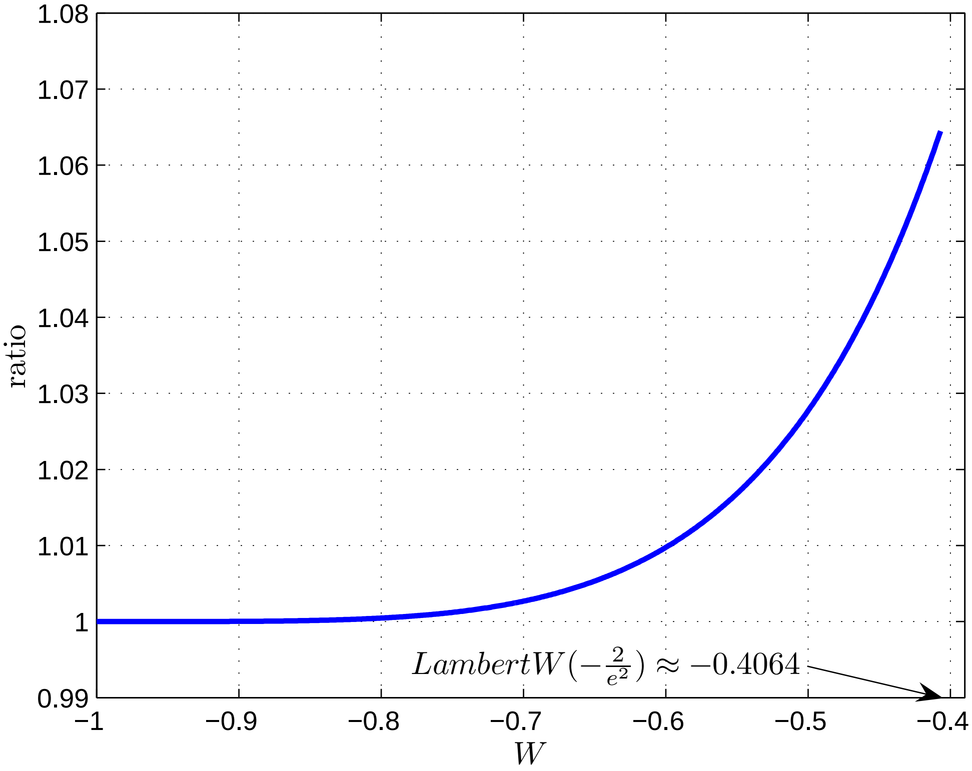


Figure C.1: Plot of $\frac{\frac{dB}{dp}(w)}{\frac{dB_{LB}}{dp}(w)}$ vs. w for $w \in \overline{ROI}_w = \left[-1, L\left(-\frac{2}{e^2}\right)\right]$.

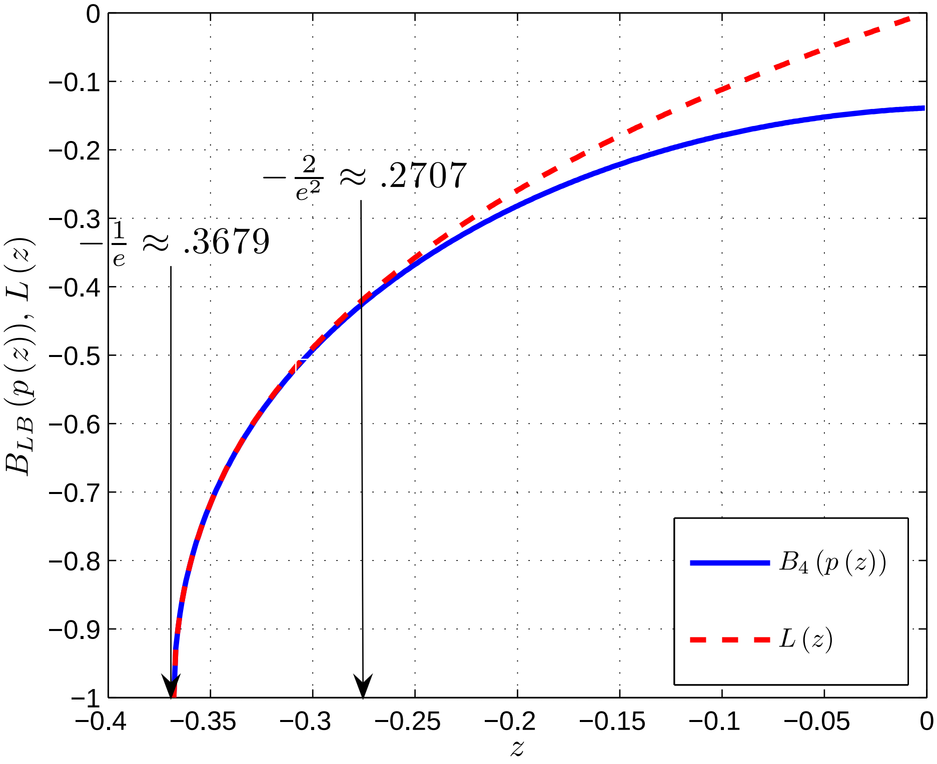


Figure C.2: Plot of $B_{LB}(p(z)) = L_{LB}(z)$ and the principal branch of the Lambert W function $L(z)$ for $z \in \left[-\frac{1}{e}, 0\right]$. Note that $ROI_z = \left(-\frac{1}{e}, \frac{2}{e^2}\right]$.

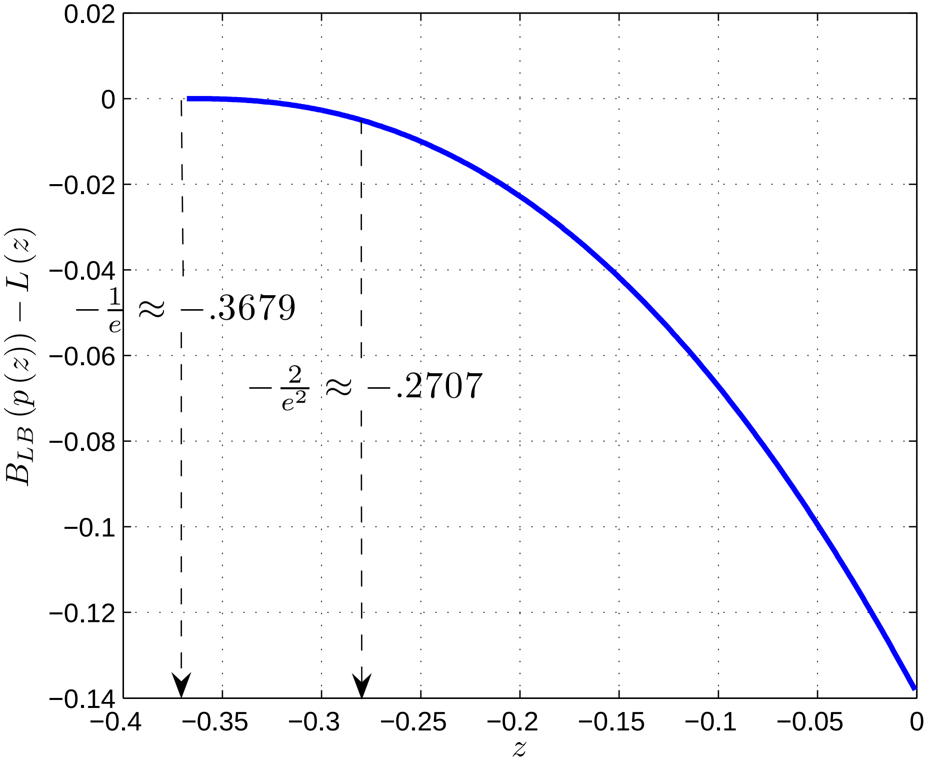


Figure C.3: Plot of $B_{LB}\left(p\left(z\right)\right) - L\left(z\right)$ for $z \in \left[-\frac{1}{e}, 0\right]$, showing that $B_{LB}\left(p\left(z\right)\right)$ is a lower bound to $L\left(z\right)$ in this region.

APPENDIX D

An Asymptotic Expansion for the Tail Function of General Exponential Sources

This appendix contains the derivation of the well-known tail function approximation used in (4.5.25) of Chapter IV. This approximation is also found in [22] for the case when $p = 2$. We remark that although the derivation to follow is for the case when f is a one-sided pdf, the derivation also holds for a two-sided pdf of the same form. Furthermore, while we have restricted ourselves to a family of pdfs for which $p \geq 1$ (which we call general exponential pdfs), the most of the formulation below only requires $p > 0$. (We will note specifically where $p \geq 1$ is required.)

Q-approximation for general exponential-type pdfs. Recall from (4.1.5), for $p \geq 1$, that a general exponential source has a pdf

$$f\left(x\right) = c_p \, e^{-\frac{x^p}{p}},$$

where $c_p > 0$ such that

$$\int_0^\infty f\left(x\right) dx = 1.$$

For such f, we observe that

$$\frac{d}{dx} f\left(x\right) = c_p \, e^{-\frac{x^p}{p}} \cdot -\frac{p x^{p-1}}{p} = f\left(x\right) \cdot -x^{p-1}.$$

Goal. We want to construct a Q-function (tail function) approximation similar to

$$Q(y) \approx \frac{f(y)}{y^{p-1}},$$

where

$$Q(y) \triangleq \int\limits_y^\infty f(u)\,du.$$

Fact.

$$\frac{d}{dx}Q(x) = -f(x).$$

For $x > 0$, we use the product rule to obtain

$$\begin{aligned}\frac{d}{dx}\frac{f(x)}{x^{p-1}} &= \frac{d}{dx}x^{-p+1}f(x) = x^{-p+1}\cdot\frac{d}{dx}f(x) + f(x)\cdot\frac{d}{dx}x^{-p+1}\\ &= x^{-p+1}\cdot\left(-x^{p-1}f(x)\right) + f(x)\cdot(-p+1)\,x^{-p}\\ &= -f(x) + f(x)\cdot\frac{-p+1}{x^p}.\end{aligned}$$

Re-arranging, we have

$$f(x) = -\left(\frac{d}{dx}\frac{f(x)}{x^{p-1}}\right) + f(x)\cdot\frac{-p+1}{x^p} = -\left(\frac{d}{dx}\frac{f(x)}{x^{p-1}}\right) + (-p+1)\,\frac{f(x)}{x^p},$$

and so for $y > 0$,

$$\begin{aligned}Q(y) &= \int\limits_y^\infty f(u)\,du = \int\limits_y^\infty -\left(\frac{d}{du}\frac{f(u)}{u^{p-1}}\right)dy + (-p+1)\int\limits_y^\infty \frac{f(u)}{u^p}dy\\ &= -\frac{f(u)}{u^{p-1}}\bigg|_y^\infty + (-p+1)\int\limits_y^\infty \frac{f(u)}{u^p}du = -0 + \frac{f(y)}{y^{p-1}} + (-p+1)\int\limits_y^\infty \frac{f(u)}{u^p}du\\ &= \frac{f(y)}{y^{p-1}} + \underbrace{(-p+1)\int\limits_y^\infty \frac{f(u)}{u^p}du}_{I_1}.\end{aligned}$$

Using IBP,

$$\begin{aligned} v &= u^{-2p+1} & dv &= \left(-2p+1\right) u^{-2p} du \\ w &= -f\left(u\right) & dw &= u^{p-1} f\left(u\right) du, \end{aligned}$$

we have

$$\begin{aligned} I_1 &= \left(-p+1\right) \int_y^\infty \frac{f\left(u\right)}{u^p} du = \left(-p+1\right) \left[-\left.\frac{f\left(u\right)}{u^{2p-1}}\right|_y^\infty + \left(-2p+1\right) \int_y^\infty \frac{f\left(u\right)}{u^{2p}} du \right] \\ &= \left(-p+1\right) \frac{f\left(y\right)}{y^{2p-1}} + \left(-p+1\right)\left(-2p+1\right) \int_y^\infty \frac{f\left(u\right)}{u^{2p}} du. \end{aligned}$$

Then for $y > 0$,

$$Q\left(y\right) = \frac{f\left(y\right)}{y^{p-1}} + \left(-p+1\right) \frac{f\left(y\right)}{y^{2p-1}} + \underbrace{\left(-p+1\right)\left(-2p+1\right) \int_y^\infty \frac{f\left(u\right)}{u^{2p}}}_{I_2}.$$

Using IBP,

$$\begin{aligned} v &= u^{-3p+1} & dv &= \left(-3p+1\right) u^{-3p} du \\ w &= -f\left(u\right) & dw &= u^{p-1} f\left(u\right) du, \end{aligned}$$

we have

$$\begin{aligned} I_2 &= \left(-p+1\right)\left(-2p+1\right) \left[-\left.\frac{f\left(u\right)}{u^{3p-1}}\right|_y^\infty + \left(-3p+1\right) \int_y^\infty \frac{f\left(u\right)}{u^{3p}} du \right] \\ &= \left(-p+1\right)\left(-2p+1\right) \frac{f\left(y\right)}{y^{3p-1}} + \left(-p+1\right)\left(-2p+1\right)\left(-3p+1\right) \int_y^\infty \frac{f\left(u\right)}{u^{3p}} du. \end{aligned}$$

Then for $y > 0$,

$$\begin{gathered} Q\left(y\right) = \frac{f\left(y\right)}{y^{p-1}} + \left(-p+1\right) \frac{f\left(y\right)}{y^{2p-1}} + \left(-p+1\right)\left(-2p+1\right) \frac{f\left(y\right)}{y^{3p-1}} + \\ \underbrace{\left(-p+1\right)\left(-2p+1\right)\left(-3p+1\right) \int_y^\infty \frac{f\left(u\right)}{u^{3p}} du}_{I_3}. \end{gathered}$$

Using IBP,

$$\begin{aligned} v &= u^{-4p+1} \quad & dv &= (-4p+1)\, u^{-4p} du \\ w &= -f(u) \quad & dw &= u^{p-1} f(u)\, du, \end{aligned}$$

we have

$$\begin{aligned} I_3 &= (-p+1)(-2p+1)(-3p+1)\left[-\frac{f(u)}{u^{4p-1}}\bigg|_y^{\infty} + (-4p+1)\int_y^{\infty} \frac{f(u)}{u^{4p}} du \right] \\ &= (-p+1)(-2p+1)(-3p+1)\frac{f(y)}{y^{4p-1}} + (-p+1)(-2p+1)(-3p+1)\cdot \\ &\quad (-4p+1)\int_y^{\infty} \frac{f(u)}{u^{4p}} du. \end{aligned}$$

Then for $y > 0$,

$$\begin{aligned} Q(y) = {} & \frac{f(y)}{y^{p-1}} + (-p+1)\frac{f(y)}{y^{2p-1}} + (-p+1)(-2p+1)\frac{f(y)}{y^{3p-1}} + (-p+1)(-2p+1)\cdot \\ & (-3p+1)\frac{f(y)}{y^{4p-1}} + \underbrace{\overbrace{(-p+1)(-2p+1)(-3p+1)(-4p+1)}^{\geq 0 \text{ for } p \geq 1}\int_y^{\infty} \frac{f(u)}{u^{4p}} du}_{I_4}. \end{aligned}$$

Using IBP,

$$\begin{aligned} v &= u^{-5p+1} \quad & dv &= (-5p+1)\, u^{-5p} du \\ w &= -f(u) \quad & dw &= u^{p-1} f(u)\, du, \end{aligned}$$

we have

$$\begin{aligned} I_4 &= (-p+1)(-2p+1)(-3p+1)(-4p+1)\left[-\frac{f(u)}{u^{5p-1}}\bigg|_y^{\infty} + (-5p+1)\int_y^{\infty} \frac{f(u)}{u^{5p}} du \right] \\ &= (-p+1)(-2p+1)(-3p+1)(-4p+1)\frac{f(y)}{y^{5p-1}} + (-p+1)(-2p+1)\cdot \\ &\quad (-3p+1)(-4p+1)(-5p+1)\int_y^{\infty} \frac{f(u)}{u^{5p}} du. \end{aligned}$$

Then for $y > 0$,

$$Q(y) = \frac{f(y)}{y^{p-1}} + (-p+1)\frac{f(y)}{y^{2p-1}} + (-p+1)(-2p+1)\frac{f(y)}{y^{3p-1}} + (-p+1)(-2p+1)\cdot (-3p+1)\frac{f(y)}{y^{4p-1}} + (-p+1)(-2p+1)(-3p+1)(-4p+1)\frac{f(y)}{y^{5p-1}} + \underbrace{\overbrace{(-p+1)(-2p+1)(-3p+1)(-4p+1)(-5p+1)}^{\leq 0 \text{ for } p\geq 1}\int_y^\infty \frac{f(u)}{u^{5p}}du}_{I_5}.$$

Summary: Expression for $\frac{Q(y)}{f(y)}$ when $y > 0$. We have

$$\frac{Q(y)}{f(y)} = \frac{1}{y^{p-1}} \times \left[1 + (-p+1)\frac{1}{y^p} + (-p+1)(-2p+1)\frac{1}{y^{2p}} + (-p+1)(-2p+1)\cdot (-3p+1)\frac{1}{y^{3p}} + (-p+1)(-2p+1)(-3p+1)(-4p+1)\frac{1}{y^{4p}} + \kappa_5 I_5\right]$$

and

$$\frac{Q(y)}{f(y)} \leq \frac{1}{y^{p-1}} \times \left[1 + (-p+1)\frac{1}{y^p} + (-p+1)(-2p+1)\frac{1}{y^{2p}} + (-p+1)(-2p+1)\cdot (-3p+1)\frac{1}{y^{3p}} + (-p+1)(-2p+1)(-3p+1)(-4p+1)\frac{1}{y^{4p}}\right]$$

since $I_5 \leq 0$ (when $p \geq 1$), and

$$\frac{Q(y)}{f(y)} \geq \frac{1}{y^{p-1}} \times \left[1 + (-p+1)\frac{1}{y^p} + (-p+1)(-2p+1)\frac{1}{y^{2p}} + (-p+1)(-2p+1)\cdot (-3p+1)\frac{1}{y^{3p}}\right]$$

since $I_4 \geq 0$ (when $p \geq 1$). Note that the upper and lower bound to $\frac{Q(y)}{f(y)}$ depend on the fact that $p \geq 1$.

Asymptotic expression for $\frac{Q(y)}{f(y)}$ using big O notation. When $y > 0$ is large,

$$\begin{aligned}\frac{Q(y)}{f(y)} &= \frac{1}{y^{p-1}} \times \left[1 + (-p+1)\frac{1}{y^p} + (-p+1)(-2p+1)\frac{1}{y^{2p}} + (-p+1)O\left(\frac{1}{y^{3p}}\right)\right]\\ &= \frac{1}{y^{p-1}} \times \left[1 + (-p+1)\left(\frac{1}{y^p} + (-2p+1)\frac{1}{y^{2p}} + O\left(\frac{1}{y^{3p}}\right)\right)\right]\end{aligned}$$

$$= \frac{1}{y^{p-1}} \times \left[1 + (-p+1)\frac{1}{y^p} + (-p+1)(-2p+1)\frac{1}{y^{2p}} + O\left(\frac{1}{y^{3p}}\right)\right],$$

or for an even cruder approximation,

$$\begin{aligned}\frac{Q(y)}{f(y)} &= \frac{1}{y^{p-1}} \times \left[1 + (-p+1)\left(\frac{1}{y^p} + O\left(\frac{1}{y^{2p}}\right)\right)\right] \\ &= \frac{1}{y^{p-1}} \times \left[1 + (-p+1)\frac{1}{y^p} + O\left(\frac{1}{y^{2p}}\right)\right].\end{aligned}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions.* Dover Publications, Inc., New York, NY, 1972.

[2] W. R. Bennett. Spectra of quantized signals. *Bell Syst. Tech. J.*, 27:446–472, 1948.

[3] J. Bucklew and Jr. N. Gallagher. A note on the computation of optimal minimum mean-square error quantizers. *IEEE Trans. Commun.*, pages 298–301, 1982.

[4] J. A. Bucklew and G. L. Wise. Multidimensional asymptotic quantization theory. *IEEE Trans. Inform. Theory*, pages 239–247, 1982.

[5] S. Cambanis and N. Gerr. A simple class of asymptotically optimal quantizers. *IEEE Trans. Inform. Thy.*, IT-29:664–676, 1983.

[6] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the lambert W function. *Adv. in Computational Math.*, 5(4):329–359, 1996.

[7] P. E. Fleischer. Sufficient conditions for achieving minimum distortion in a quantizer. *IEEE Int. Conv. Rec., Part I*, pages 104–111, 1964.

[8] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications.* John Wiley and Sons, Inc., New York, NY, 1984.

[9] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, Boston, MA, 1992.

[10] D. Hui and D. L. Neuhoff. Asymptotic analysis of optimal fixed-rate uniform scalar quantization. *IEEE Trans. Inform. Theory*, 47:957–977, 2001.

[11] A. Jeffrey. *Handbook of Mathematical Formulas and Integrals.* Academic Press, New York, NY, 1995.

[12] S. Lay. *Analysis: With an Introduction to Proof.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1990.

[13] S. P. Lloyd. Least squares quantization in PCM. *unpublished Bell Labs memorandum, July 1957; published version in IEEE Trans. Inform. Thy.*, IT-28:129–137, 1982.

[14] F. S. Lu and G. Wise. A further investigation of Max's algorithm for optimum quantization. *IEEE Trans. Commun.*, pages 746–750, 1985.

[15] J. Max. Quantizing for minimum distortion. *IEEE Trans. Inform. Thy.*, IT-6:7–12, 1960.

[16] S. Na. On the support of fixed-rate minimum mean-squared error scalar quantizers for a Laplacian source. *IEEE Trans. Inform. Thy.*, pages 937–944, May 2004.

[17] S. Na and D. Neuhoff. *Estimating the Key Parameter in Scalar Quantization.* Technical Report No 266, The University of Michigan, 1989.

[18] S. Na and D. Neuhoff. On the support of MSE-optimal, fixed-rate, scalar quantizers. *IEEE Trans. Inform. Thy.*, pages 2972–2982, 2001.

[19] K. Nitadori. Statistical analysis of ΔPCM. *Electon. Commun. in Japan*, 48:17–26, 1965.

[20] P. F. Panter and W. Dite. Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proc. IRE*, 39:44–48, 1951.

[21] H. Stark and J. W. Woods. *Probability, Random Processes, and Estimation Theory for Engineers.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1994.

[22] H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part 1.* John Wiley and Sons, Inc., New York, NY, 1989.

[23] A. V. Trushkin. Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Trans. Inform. Thy.*, IT-28:187–198, 1982.